Bayesian Statistical Methods - Study Notes

Introduction and Course Overview

Based on Hoff (2009), A First Course in Bayesian Statistical Methods

Instructor: Sudeep Regmi

Purpose of These Notes

These study notes are designed to **supplement, not replace,** the textbook and lectures. Their purpose is to serve as a guide and roadmap, helping you navigate the material more effectively. Each weekly note aims to:

- Outline the key **learning objectives** for the week's topics.
- Define and clarify crucial **concepts and terminology**.
- Provide a **conceptual overview** or "big picture" perspective.
- Summarize the **core ideas** presented in the corresponding textbook chapter(s).
- Highlight **connections** between different topics across the course.
- Point out potential **pitfalls and common points of confusion**.

The focus will be on understanding the underlying concepts, the motivation behind different techniques, and the interpretation of results, complementing the mathematical details presented in Hoff's text.

The Bayesian Approach: A Shift in Perspective

Bayesian statistics offers a different way of thinking about probability and inference compared to the classical (frequentist) approach many of you may be familiar with. Key distinctions include:

- **Probability as Degree of Belief:** Probability quantifies our uncertainty or degree of belief about unknown quantities.
- **Parameters as Random Variables:** Unknown parameters (like population means or regression coefficients) are treated as random variables having probability distributions that reflect our uncertainty about them.
- **Learning as Belief Updating:** Inference proceeds by updating our prior beliefs about parameters (**Prior Distribution**) in light of observed data (via the **Likelihood Function**) to arrive at updated beliefs (**Posterior Distribution**), using the fundamental engine of **Bayes' Theorem**.

This framework provides a coherent way to combine prior knowledge with empirical evidence and yields results (like credible intervals) that often have more direct and intuitive interpretations. Rigorous application relies heavily on the foundations of probability theory, which we will review early on.

Overview of the Notes

These 14 sets of notes will guide you through the major themes of the book, building progressively:

- **Weeks 1-2: Foundations.** We start with the core concepts of Bayesian thinking, review essential probability theory, and formally introduce the Bayesian framework components: prior, likelihood, posterior, and predictive distributions (Hoff Ch 1-3).
- **Weeks 3-6: Core Conjugate Models.** We apply the framework to fundamental statistical problems: estimating proportions (Beta-Binomial model, Hoff Ch 4), Normal means and variances (Normal-Inverse-Chi-Squared models, Hoff Ch 5), and extending to basic multivariate parameters (Dirichlet-Multinomial, Multivariate Normal, Hoff Ch 6). Understanding these "building block" models is crucial.
- **Weeks 7-10: Bayesian Computation.** We address the critical issue that analytical solutions are often intractable in Bayesian analysis. We introduce Monte Carlo simulation, the powerful Markov chain Monte Carlo (MCMC) paradigm, specific algorithms like the Gibbs Sampler and Metropolis-Hastings, and the essential practice of MCMC diagnostics (Hoff Ch 7-9). MCMC enables modern Bayesian practice.
- **Weeks 11-14: Advanced Models & Evaluation.** We explore vital topics like model checking (assessing fit using posterior predictive checks) and model comparison (Hoff Ch 10), delve into the powerful structure of Hierarchical Models for analyzing grouped data and borrowing strength (Hoff Ch 11), and apply the framework to Linear Regression (Hoff Ch 12).

How to Use These Notes Effectively

- **Read Before the Text:** Skim the relevant study note *before* diving into the corresponding chapter(s) in Hoff. This will prime you on what to look for and the key concepts to grasp.
- **Guide Your Reading:** Use the notes' summaries and objectives to focus your reading of the textbook. Pay attention to the concepts highlighted as potentially confusing.
- **Connect Ideas:** Refer back to previous notes to reinforce how topics build on each other.
- **Supplement, Don't Substitute:** The textbook contains essential mathematical details, derivations, and examples not fully replicated here. Engage deeply with Hoff's text.
- **Integrate with Lectures and Exercises:** Use the notes alongside lectures and actively work through exercises and coding examples (often in R) to solidify your understanding. Computation is a key skill in Bayesian statistics.

Mastering Bayesian methods is a rewarding journey that combines statistical theory, probability, and computational skills. These notes are intended to support you along the way. Let's get started!

Bayesian Statistical Methods - Study Notes Week 1

Foundations: Bayesian Thinking & Probability Review

Based on Hoff (2009), Chapters 1 & 2

Instructor: Sudeep Regmi

Chapters Covered

- Hoff Chapter 1: Introduction
- Hoff Chapter 2: Scientific Data and Statistical Models (Probability Review)

Learning Objectives

After engaging with this week's material, you should be able to:

- Describe the fundamental difference between Bayesian and frequentist views of probability and parameters.
- Identify and explain the three core components of Bayesian inference: prior distribution, likelihood function, and posterior distribution.
- Understand the goal of Bayesian inference as updating beliefs in light of evidence.
- State and apply basic probability rules (addition, multiplication).
- Define and calculate conditional probability.
- State and understand the Law of Total Probability.
- State, derive, and interpret Bayes' Theorem, both in its general form and in the context of parameters and data.
- Define discrete and continuous random variables, Probability Mass Functions (PMFs), and Probability Density Functions (PDFs).
- Understand the concepts of expected value and variance.
- Distinguish between joint, marginal, and conditional distributions (conceptually).

Key Concepts & Definitions

- **Parameter (θ):** A quantity governing a statistical model (e.g., population mean, regression coefficient). In Bayesian statistics, parameters are treated as *random variables* having probability distributions.
- **Data (y):** The observed evidence or measurements used to learn about parameters.

- **Prior Distribution ($p(\theta)$):** A probability distribution representing beliefs about a parameter θ *before* observing the data y. It quantifies initial uncertainty.

- **Likelihood Function ($L(\theta|y)$ or $p(y|\theta)$):** A function representing the probability (or density) of observing the data y *given* a specific value of the parameter θ. Note: As a function of θ for fixed y, it's the likelihood; as a function of y for fixed θ, it's the sampling distribution.

- **Posterior Distribution ($p(\theta|y)$):** A probability distribution representing updated beliefs about the parameter θ *after* observing the data y. It combines information from the prior and the likelihood via Bayes' Theorem.

- **Subjective Probability:** An interpretation of probability as a degree of belief or confidence in a statement being true. Central to the Bayesian viewpoint.

- **Conditional Probability ($P(A|B)$):** The probability of event A occurring given that event B has occurred. Defined as $P(A|B) = \frac{P(A \cap B)}{P(B)}$, provided $P(B) > 0$.

- **Independence:** Events A and B are independent if $P(A \cap B) = P(A)P(B)$, or equivalently, $P(A|B) = P(A)$ (if $P(B) > 0$).

- **Law of Total Probability:** If $B_1, B_2, \ldots, B_k$ form a partition of the sample space (mutually exclusive and exhaustive), then $P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$.

- **Bayes' Theorem:** A mathematical rule for updating probabilities based on new evidence. $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$.

- **Random Variable (RV):** A variable whose value is a numerical outcome of a random phenomenon.

- **Probability Mass Function (PMF):** For a discrete RV, a function $p(x) = P(X = x)$ that gives the probability of the RV taking on a specific value x. Sums to 1 over all possible values.

- **Probability Density Function (PDF):** For a continuous RV, a function $f(x)$ such that the probability of the RV falling within an interval $[a, b]$ is given by the integral $\int_a^b f(x)dx$. The total integral over the entire range is 1. Note $P(X = x) = 0$ for any specific x.

- **Expected Value (E[X] or μ):** The long-run average value of a random variable; a measure of central tendency. For discrete X, $E[X] = \sum_x xP(X = x)$. For continuous X, $E[X] = \int_{-\infty}^{\infty} xf(x)dx$.

- **Variance (Var(X) or σ^2):** A measure of the spread or dispersion of a random variable around its mean. Defined as $Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$.

Conceptual Overview / "The Big Picture"

This week introduces the core philosophy of Bayesian statistics. Unlike the frequentist approach where parameters are viewed as fixed, unknown constants and probability refers to long-run frequencies of data outcomes, the Bayesian approach treats parameters themselves as uncertain quantities described by probability distributions. The central idea is **learning as belief updating**. We start with some prior beliefs about a parameter (the prior distribution), collect data, and then use Bayes' Theorem to combine our prior beliefs with the information contained in the data (the likelihood function) to arrive at updated, more informed beliefs (the posterior distribution). All inference and conclusions are then drawn from this posterior distribution. The probability review in Chapter 2 provides the essential mathematical tools needed to perform this updating process rigorously.

Core Ideas & Summaries

Chapter 1: The Bayesian Paradigm Shift

- **Parameters as Random Variables:** This is the key conceptual leap. We express our uncertainty about unknown quantities (like the true mean μ or proportion p) using probability distributions.
- **Subjectivity vs. Objectivity:** Bayesian methods embrace subjective probability (degree of belief) in the prior. While this can seem less "objective", it makes explicit the assumptions that are often implicit in frequentist choices. Priors can range from highly informative (strong beliefs) to weakly informative or "non-informative" (letting the data speak more).
- **Inference Goal:** To obtain the posterior distribution $p(\theta|y)$, which contains all current information about the parameter θ after seeing data y.
- **The Core Components Cycle:**
  1. **Prior $p(\theta)$:** What do we believe about θ before seeing data?
  2. **Likelihood $p(y|\theta)$:** How likely is the observed data y for different possible values of θ? This connects the data to the parameter via a statistical model.
  3. **Posterior $p(\theta|y)$:** How should we update our beliefs about θ after combining the prior and the likelihood? This is achieved via Bayes' Theorem.

Chapter 2: Probability Foundations for Bayesian Inference

This chapter is largely a review but frames concepts crucial for Bayesian calculations.

- **Conditional Probability is Key:** Bayesian inference is fundamentally about conditioning on observed data. Understanding $P(A|B)$ is essential.
- **Law of Total Probability:** This is used to calculate the denominator in Bayes' Theorem, often called the *marginal likelihood* or *prior predictive distribution*, $p(y)$.

  $$p(y) = \int p(y|\theta)p(\theta)d\theta \quad \text{(for continuous } \theta\text{)}$$

  $$p(y) = \sum_{\theta} p(y|\theta)p(\theta) \quad \text{(for discrete } \theta\text{)}$$

  This represents the overall probability of observing the data y, averaged over all possible values of θ weighted by their prior probabilities.
- **Bayes' Theorem for Inference:** This is the engine driving Bayesian learning. It relates the posterior probability of a parameter value to its prior probability and the likelihood of the data given that value:

  $$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta')p(\theta')d\theta'}$$

  Interpretation of terms in the statistical context:

  - $p(\theta|y)$: **Posterior probability** (density) of parameter θ given data y.
  - $p(y|\theta)$: **Likelihood** of data y given parameter θ.
  - $p(\theta)$: **Prior probability** (density) of parameter θ.
  - $p(y)$: **Marginal likelihood** (or evidence, or prior predictive probability) of the data y. This is the normalizing constant ensuring the posterior integrates/sums to 1.

  Often, we work with proportionality, as $p(y)$ is just a constant with respect to θ:

$$\underbrace{p(\theta|y)}_{\text{Posterior}} \propto \underbrace{p(y|\theta)}_{\text{Likelihood}} \times \underbrace{p(\theta)}_{\text{Prior}}$$

  This relationship is the heart of Bayesian calculation.

- **Describing Random Variables:** We need PMFs/PDFs to define our priors and likelihoods. We use Expected Value and Variance (and other summaries like quantiles/modes) to describe the resulting posterior distributions.

Connections to Future Topics

- The concepts of prior, likelihood, and posterior introduced here will be applied repeatedly in Chapters 3, 4, 5, and beyond to specific models (Binomial, Normal, Multinomial, etc.).
- Bayes' Theorem is the fundamental equation used throughout the course to derive posterior distributions.
- The probability distributions reviewed here (and others introduced later, like Beta, Gamma, Normal) will serve as building blocks for specifying priors and likelihoods.
- The idea of calculating $p(y)$ via the Law of Total Probability becomes computationally challenging quickly, motivating the simulation-based MCMC methods introduced in Chapters 7-9.

Common Pitfalls / Points of Confusion

- **Interpreting Probability:** Confusing Bayesian degree-of-belief probability with frequentist long-run frequency. Remember, $p(\theta)$ represents belief about a single, specific (though unknown) parameter value.
- **Likelihood vs. Probability:** The likelihood $p(y|\theta)$ is a function of θ for fixed data y. It does not represent a probability distribution for θ on its own (it doesn't necessarily integrate to 1 over θ).
- **Role of the Prior:** Forgetting that the prior is a necessary component. Even "non-informative" priors represent some (often implicit) prior belief state. The influence of the prior diminishes as more data is collected.
- **The Normalizing Constant ($p(y)$):** Forgetting that $p(y)$ in Bayes' Theorem is crucial for making the posterior a valid probability distribution, even though it's often ignored when finding the *shape* of the posterior via proportionality. Its calculation can be hard!
- **Conditional Probability Direction:** Confusing $P(A|B)$ with $P(B|A)$. Bayes' Theorem is precisely the tool needed to get from one to the other (specifically, from $p(y|\theta)$ to $p(\theta|y)$).

End of Week 1 Notes.

Bayesian Statistical Methods - Study Notes Week 2

The Bayesian Framework & Conjugacy

Based on Hoff (2009), Chapter 3

Instructor: Sudeep Regmi

Chapters Covered

- Hoff Chapter 3: General Principles of Bayesian Inference

Learning Objectives

After engaging with this week's material, you should be able to:

- Formally define the prior distribution, likelihood function, and posterior distribution in the context of parameter estimation.
- Explain how the posterior distribution mathematically combines prior beliefs and information from the data.
- Use the proportionality form of Bayes' Theorem (*posterior* $\propto$ *likelihood* $\times$ *prior*) to identify the kernel of the posterior distribution.
- Understand the role of the normalizing constant (marginal likelihood) in making the posterior a valid probability distribution.
- Define, interpret, and differentiate between the prior predictive distribution and the posterior predictive distribution.
- Explain the concept of a conjugate prior distribution and why it simplifies Bayesian calculations.
- Calculate and interpret basic posterior summary statistics (mean, median, mode) as point estimates.
- Define and interpret a Bayesian credible interval.

Key Concepts & Definitions

- **Prior Distribution ($p(\theta)$):** Represents uncertainty about parameter θ *before* observing data y. Must be a valid probability distribution (integrates/sums to 1).
- **Likelihood Function ($p(y|\theta)$):** Represents the probability (or density) of the observed data y as a function of the parameter θ. It quantifies how well different values of θ explain the data. Note: Typically $p(y|\theta)$ is viewed as a function of y for fixed θ (the sampling distribution), but in inference, we fix y and examine it as a function of θ, denoted $L(\theta|y)$.

- **Posterior Distribution ($p(\theta|y)$):** Represents the updated uncertainty about parameter θ *after* observing data y. Calculated via Bayes' Theorem.
- **Bayes' Theorem (Proportionality Form):**

  $$p(\theta|y) \propto p(y|\theta)p(\theta)$$

  This states the posterior is proportional to the likelihood times the prior. This form is often used to find the "shape" or kernel of the posterior distribution.
- **Normalizing Constant / Marginal Likelihood / Evidence ($p(y)$):** The denominator in the full Bayes' Theorem:

  $$p(y) = \int p(y|\theta)p(\theta)d\theta \quad \text{(continuous } \theta\text{)}$$

  $$p(y) = \sum_{\theta} p(y|\theta)p(\theta) \quad \text{(discrete } \theta\text{)}$$

  It ensures that $\int p(\theta|y)d\theta = 1$. It represents the probability of the data averaged over all possible parameter values weighted by the prior.
- **Prior Predictive Distribution ($p(y)$):** Same as the marginal likelihood. Represents the distribution of data we would expect to see *before* any data is actually observed, based on the model and prior combined. Useful for checking if the prior specification leads to reasonable data predictions.
- **Posterior Predictive Distribution ($p(\tilde{y}|y)$):** The distribution of a potential *new* observation $\tilde{y}$ (from the same process), given the observed data y. Calculated as:

  $$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

  It averages the model's predictions $p(\tilde{y}|\theta)$ over the posterior uncertainty $p(\theta|y)$ about the parameter. Used for model checking and making predictions about future outcomes.
- **Conjugate Prior:** A family of prior distributions $\mathcal{P}$ is said to be conjugate for a likelihood function $p(y|\theta)$ if, for every prior $p(\theta) \in \mathcal{P}$, the resulting posterior distribution $p(\theta|y)$ is also in the family $\mathcal{P}$.
- **Hyperparameters:** Parameters of a prior distribution (e.g., the shape and rate parameters α, β of a Gamma prior). In conjugate families, the data updates these hyperparameters.
- **Posterior Mean, Median, Mode:** Summary statistics of the posterior distribution often used as point estimates for θ.
- **Credible Interval (or Bayesian Confidence Interval):** An interval $[a, b]$ such that the posterior probability of θ lying within the interval is a specified value (e.g., 95%).

  $$P(a \leq \theta \leq b|y) = \int_a^b p(\theta|y)d\theta = 0.95$$

  There are different ways to construct these (e.g., equal-tailed, highest posterior density). The interpretation is direct: "There is a 95% probability that the true parameter value θ lies within this interval, given the data and model."

Conceptual Overview / "The Big Picture"

This week operationalizes the Bayesian philosophy introduced last week. Chapter 3 lays out the mechanics of updating beliefs using data. We move from the abstract statement of Bayes' Theorem to its practical application in parameter estimation. The core process is identifying the mathematical forms of the prior and the likelihood, multiplying them together, and recognizing the resulting mathematical form as the (unnormalized) posterior distribution. A key practical concept introduced is **conjugacy**, a special relationship between the prior and likelihood that makes the mathematical updating process much simpler, often reducing it to simple rules for updating the prior's parameters (hyperparameters). We also explore what to *do* with the posterior distribution once we have it: summarizing it with point estimates (mean, median, mode) and intervals (credible intervals), and using it to predict future observations (posterior predictive distribution). The distinction between predicting data before observation (prior predictive) and after (posterior predictive) is also clarified.

Core Ideas & Summaries

3.1 The Bayesian Method

- Reinforces the three steps: 1. Specify a prior distribution $p(\theta)$. 2. Choose a statistical model / likelihood $p(y|\theta)$. 3. Update beliefs using Bayes' theorem to get the posterior $p(\theta|y)$.
- **Finding the Posterior:** The key computational step is often using $p(\theta|y) \propto p(y|\theta)p(\theta)$.
  - Write down the mathematical formula for the likelihood $p(y|\theta)$ (treating y as fixed constants).
  - Write down the mathematical formula for the prior $p(\theta)$.
  - Multiply them together.
  - **Crucially:** Drop any multiplicative factors that do *not* depend on θ. The remaining expression is the "kernel" of the posterior distribution.
  - Try to recognize this kernel as belonging to a known probability distribution family (e.g., Normal, Gamma, Beta).
  - The parameters of this known distribution are the parameters of the posterior. The normalizing constant is implicitly found once the distribution family is identified.

3.2 Inference and Summarization

- The posterior distribution $p(\theta|y)$ contains *all* information about θ after observing y.
- **Point Estimates:** Often we need a single number summary. Common choices:
  - **Posterior Mean:** $E[\theta|y] = \int \theta p(\theta|y) d\theta$. Optimal under squared error loss.
  - **Posterior Median:** Value m such that $P(\theta \leq m|y) = 0.5$. Optimal under absolute error loss.
  - **Posterior Mode:** Value θ that maximizes $p(\theta|y)$. Can be found by maximizing $p(y|\theta)p(\theta)$.
- **Interval Estimates (Credible Intervals):** Provide a range of plausible values for θ.
  - A $100(1-\alpha)\%$ credible interval $[a, b]$ satisfies $P(a \leq \theta \leq b|y) = 1 - \alpha$.

  - **Equal-tailed Interval:** Choose a and b such that $P(\theta < a|y) = \alpha/2$ and $P(\theta > b|y) = \alpha/2$. These are easy to compute from posterior quantiles.
  - **Highest Posterior Density (HPD) Interval:** Choose $[a, b]$ such that $p(\theta|y)$ is higher for all θ inside the interval than for any θ outside, and the total probability is $1 - \alpha$. This is the shortest possible interval, but harder to compute, especially for asymmetric or multimodal posteriors.

- **Interpretation is Key:** A 95% credible interval $[a, b]$ means there is a 95% probability (given data and model) that θ lies between a and b. This is a direct statement about θ, unlike the frequentist confidence interval interpretation.

3.3 Prediction

- **Prior Predictive Distribution $p(y)$:**
  - Formula: $p(y) = \int p(y|\theta)p(\theta)d\theta$.
  - Interpretation: The marginal distribution of the data before it's observed. What kind of data does our model+prior setup predict?
  - Use: Can be used to check if the prior makes sense. If $p(y)$ assigns very low probability to the actually observed data y_{obs}, it might suggest the prior or model is inappropriate. Also serves as the normalizing constant in Bayes' theorem.

- **Posterior Predictive Distribution $p(\tilde{y}|y)$:**
  - Formula: $p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$.
  - Interpretation: The distribution of a new, future observation $\tilde{y}$, given the data y we've already seen. It averages the predictions of the model ($p(\tilde{y}|\theta)$) weighted by our posterior beliefs about θ ($p(\theta|y)$). Accounts for parameter uncertainty.
  - Use: Making predictions about future events. Checking model fit (do predicted observations $\tilde{y}$ look similar to the observed data y?).

3.4 Conjugacy

- **Motivation:** Calculating the posterior $p(\theta|y)$ (especially the normalizing constant $p(y)$) and predictive distributions often involves difficult integrals.

- **Definition:** A prior $p(\theta)$ from family $\mathcal{P}$ is conjugate to a likelihood $p(y|\theta)$ if the posterior $p(\theta|y)$ also belongs to $\mathcal{P}$.

- **Benefit:** If we use a conjugate prior, the updating process becomes simple algebra. We just need to find the rules for updating the hyperparameters of the prior distribution based on the data. The mathematical form of the distribution remains the same.

- **Example Teaser:** The Beta distribution is conjugate for the Binomial likelihood (coming in Ch 4). The Normal distribution is conjugate for the Normal likelihood (for the mean, with known variance) (coming in Ch 5).

- **Limitation:** Conjugacy is convenient but restrictive. It might force us to use a prior that doesn't reflect our actual beliefs. For many complex models, conjugate priors don't exist. This motivates computational methods like MCMC (later chapters).

Connections to Previous/Future Topics

- **Week 1:** This week applies the core concepts (Bayes' theorem, probability distributions) introduced last week to the specific task of parameter inference.
- **Chapter 4 (Next Week):** Will provide the first detailed example of this entire framework using the Beta prior, Binomial likelihood, and resulting Beta posterior (a conjugate family).
- **Chapter 5:** Will apply the framework to Normal data, introducing Normal-Normal and Normal-Inverse-Gamma conjugate families.
- **Chapters 7-9 (Computation):** The potential difficulty in calculating integrals for non-conjugate posteriors or for predictive distributions motivates the need for simulation methods (MCMC).
- **Chapter 10 (Model Checking):** Posterior predictive distributions are a primary tool for Bayesian model checking.

Common Pitfalls / Points of Confusion

- **Confusing $p(y|\theta)$ and $p(\theta|y)$:** Remembering which is the likelihood (data probability given parameter) and which is the posterior (parameter probability given data).
- **Ignoring the Normalizing Constant:** Treating $p(y|\theta)p(\theta)$ as the actual posterior density, rather than just being proportional to it. This matters when calculating probabilities or comparing posterior heights.
- **Prior vs. Posterior Predictive:** Mixing up $p(y)$ (prediction before data, average over prior) and $p(\tilde{y}|y)$ (prediction after data, average over posterior).
- **Interpretation of Credible Intervals:** Stating the frequentist interpretation (about the interval capturing the true value in repeated experiments) instead of the direct Bayesian probability statement about the parameter itself.
- **Algebra Errors:** Simple mistakes when multiplying likelihood and prior functions and trying to identify the kernel of the resulting posterior distribution. Keep track of terms involving θ versus constants.
- **Over-reliance on Conjugacy:** Assuming conjugacy is always possible or necessary. It's a useful tool for simple models but not the entirety of Bayesian practice.

End of Week 2 Notes.

Bayesian Statistical Methods - Study Notes Week 3

Inference for Proportions: The Beta-Binomial Model

Based on Hoff (2009), Chapter 4

Instructor: Sudeep Regmi

Chapters Covered

- Hoff Chapter 4: Inference for a Proportion

Learning Objectives

After engaging with this week's material, you should be able to:

- Identify scenarios where the Binomial distribution is an appropriate model for the data-generating process.
- Write down the likelihood function for a Binomial experiment based on the number of successes (y) in a fixed number (n) of trials, parameterized by the unknown success probability θ.
- Describe the properties of the Beta distribution ($Beta(a, b)$) and explain why it serves as a flexible prior for a parameter θ constrained between 0 and 1.
- Interpret the shape parameters (a, b) of the Beta distribution, sometimes viewed as 'pseudo-counts' of prior successes and failures.
- Demonstrate and explain the conjugacy between the Beta prior and the Binomial likelihood.
- Derive the posterior distribution for θ as a $Beta(a + y, b + n - y)$ distribution.
- Calculate and interpret the posterior mean, mode, and variance for θ.
- Construct and interpret equal-tailed credible intervals for θ using the quantiles of the Beta posterior distribution.
- Perform Bayesian hypothesis tests by calculating the posterior probability that θ lies in a specific range (e.g., $P(\theta > c|y)$).
- Understand, calculate (conceptually), and interpret the prior predictive distribution (the Beta-Binomial distribution).
- Calculate and interpret the posterior predictive probability for a future Bernoulli trial.

Key Concepts & Definitions

- **Bernoulli Trial:** A single random experiment with exactly two possible outcomes, "success" and "failure", where the probability of success is θ.
- **Binomial Distribution ($Y \sim$ Binomial(n, θ)):** Describes the probability of obtaining y successes in n independent Bernoulli trials, each with success probability θ.
  - **PMF:** $p(y|n, \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$ for $y = 0, 1, \ldots, n$.
  - **Likelihood Function $L(\theta|y, n)$:** The PMF viewed as a function of θ for fixed data y, n. Proportional to $\theta^y(1-\theta)^{n-y}$.
- **Proportion Parameter (θ):** The unknown probability of success in a Bernoulli trial, $0 \leq \theta \leq 1$.
- **Beta Distribution ($\theta \sim$ Beta(a, b)):** A continuous probability distribution defined on the interval $[0, 1]$, parameterized by two positive shape parameters, a and b.
  - **PDF:** $p(\theta|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$ for $0 \leq \theta \leq 1$.
  - **Shape Parameters (a, b):** Control the shape of the distribution. Often interpreted as representing a prior 'successes' and b prior 'failures' (or more accurately, $a-1$ and $b-1$ pseudo-counts contributing to the exponents).
  - **Mean:** $\mathbb{E}[\theta] = \frac{a}{a+b}$.
  - **Variance:** $\text{Var}(\theta) = \frac{ab}{(a+b)^2(a+b+1)}$.
  - **Mode:** $\frac{a-1}{a+b-2}$ (for $a, b > 1$).
  - **Special Cases:** $Beta(1, 1)$ is the Uniform(0,1) distribution. $Beta(0.5, 0.5)$ is Jeffreys' prior for the Binomial parameter.
- **Beta-Binomial Conjugacy:** The property that if the prior distribution for θ is Beta and the likelihood function is Binomial, then the posterior distribution for θ is also Beta.
- **Posterior Distribution in Beta-Binomial Model:** If prior is Beta(a, b) and data is y successes in n trials, the posterior is $\theta|y \sim \text{Beta}(a + y, b + n - y)$.
- **Posterior Mean:** $\mathbb{E}[\theta|y] = \frac{a+y}{a+b+n}$. This is a weighted average of the prior mean $\frac{a}{a+b}$ and the sample proportion $\frac{y}{n}$, with weights determined by the prior 'sample size' $a+b$ and the data sample size n.
- **Credible Interval for θ:** An interval $[L, U]$ such that $P(L \leq \theta \leq U|y) = 1-\alpha$. Typically found using quantiles of the Beta$(a + y, b + n - y)$ distribution.
- **Bayesian Hypothesis Testing:** Evaluating hypotheses by calculating their posterior probability. E.g., testing $H_1 : \theta > c$ by calculating $P(\theta > c|y) = \int_c^1 p(\theta|y)d\theta$.
- **Prior Predictive Distribution ($p(y|a, b, n)$):** The marginal distribution of the data y before it is observed, obtained by integrating the likelihood over the prior. For this model, it follows the **Beta-Binomial distribution**:

  $$p(y|a, b, n) = \binom{n}{y}\frac{B(a + y, b + n - y)}{B(a, b)}$$

  where $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the Beta function.
- **Posterior Predictive Probability (for one trial):** The probability of success ($\tilde{y} = 1$) in a single future Bernoulli trial, given the observed data y:

  $$P(\tilde{y} = 1|y, n, a, b) = \mathbb{E}[\theta|y] = \frac{a + y}{a + b + n}$$

Conceptual Overview / ”The Big Picture”

This week provides the first concrete application of the Bayesian inference framework developed in Weeks 1 and 2. We focus on the common problem of estimating an unknown proportion or probability of success, θ. The Binomial distribution naturally models the number of successes in a fixed number of trials. We introduce the Beta distribution as a mathematically convenient and flexible prior for θ, as it is defined on $[0, 1]$ and can take various shapes depending on its hyperparameters a and b. The central result is the **conjugacy** of the Beta prior with the Binomial likelihood. This means the mathematical form of our belief distribution remains Beta after observing data, simplifying calculations significantly. The update rule is intuitive: the posterior parameters are simply the prior parameters plus the observed counts of successes and failures. We then explore how to extract meaningful summaries (point estimates, intervals) and make predictions using this Beta posterior distribution.

Core Ideas & Summaries

4.1 The Binomial Likelihood

- Assumes n independent trials, constant success probability θ, and we observe y successes.
- The likelihood function, crucial for updating beliefs about θ, is:

$$L(\theta|y, n) = p(y|n, \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

- As a function of θ, the term $\binom{n}{y}$ is a constant. Therefore, the kernel of the likelihood is:

$$L(\theta|y, n) \propto \theta^y(1-\theta)^{n-y}$$

  This kernel is what interacts with the prior in Bayes’ theorem.

4.2 The Beta Prior Distribution

- Needed: A distribution on $[0, 1]$ to represent prior beliefs about θ.
- The Beta distribution, $p(\theta|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$, fits perfectly.
- Interpretation of a, b:
  - Larger $a + b$ implies stronger prior beliefs (distribution is more peaked).
  - Ratio a/b determines the location of the peak (prior mean is $a/(a + b)$).
  - $a = 1, b = 1$ gives $Beta(1, 1) = Uniform(0, 1)$, a common ”non-informative” prior, suggesting all values of θ are equally likely beforehand.
  - $a = 0.5, b = 0.5$ gives Jeffreys’ prior, another non-informative choice derived from information theory principles.
- The kernel of the prior is $\theta^{a-1}(1-\theta)^{b-1}$.

4.3 The Beta Posterior Distribution

- Apply Bayes’ Theorem: $p(\theta|y, n, a, b) \propto p(y|n, \theta) \times p(\theta|a, b)$.

- Multiply the kernels:

$$p(\theta|y) \propto [\theta^y(1-\theta)^{n-y}] \times [\theta^{a-1}(1-\theta)^{b-1}]$$

$$p(\theta|y) \propto \theta^{y+a-1}(1-\theta)^{(n-y)+b-1}$$

- Recognize the Result: This is the kernel of a Beta distribution with updated parameters.
- Posterior Distribution: $\theta|y, n, a, b \sim \text{Beta}(a', b')$, where:
  - $a' = a + y$ (Prior 'successes' + Observed successes)
  - $b' = b + n - y$ (Prior 'failures' + Observed failures)
- The effect of the data is to add the observed counts y and $n - y$ to the prior parameters a and b. The prior acts like having observed $a + b$ 'pseudo-observations' consisting of a successes and b failures.

4.4 Posterior Summarization and Inference

- **Point Estimates:** Use properties of the $\text{Beta}(a + y, b + n - y)$ distribution.
  - Posterior Mean: $\mathbb{E}[\theta|y] = \frac{a+y}{a+b+n}$
  - Posterior Mode: $\frac{a+y-1}{a+b+n-2}$ (if $a + y > 1, b + n - y > 1$)
  - Posterior Median: Requires finding m such that $P(\theta \leq m|y) = 0.5$ (usually via numerical quantile function).
- **Credible Intervals:** A $100(1 - \alpha)\%$ equal-tailed credible interval is $[\theta_L, \theta_U]$, where θ_L is the $\alpha/2$ quantile and θ_U is the $1 - \alpha/2$ quantile of the $\text{Beta}(a + y, b + n - y)$ distribution. Interpretation: "Given the data and prior, there is a $(1 - \alpha)$ probability that the true value of θ lies between θ_L and θ_U."
- **Hypothesis Testing:** Calculate the probability of the hypothesis directly from the posterior. Example: To test $H_1 : \theta > 0.5$ vs $H_0 : \theta \leq 0.5$, calculate $P(\theta > 0.5|y) = \int_{0.5}^{1} p(\theta|y, a', b')d\theta$. This is found using the CDF of the $\text{Beta}(a', b')$ distribution. The result is the posterior probability that H_1 is true.

4.5 Prediction

- **Prior Predictive** $p(y|n, a, b)$**:** Predicts the number of successes y in n trials *before* observing data. Averages Binomial probabilities over the $\text{Beta}(a, b)$ prior. Results in the Beta-Binomial distribution. Useful for checking if prior assumptions lead to reasonable predictions.
- **Posterior Predictive** $p(\tilde{y}|y, n, a, b)$**:** Predicts the outcome $\tilde{y}$ of *future* trials after observing data y. Averages the Binomial probability $p(\tilde{y}|\theta)$ over the $\text{Beta}(a + y, b + n - y)$ posterior.
  - For a single next trial ($\tilde{n} = 1$), the probability of success is $P(\tilde{y} = 1|y) = \mathbb{E}[\theta|y] = \frac{a+y}{a+b+n}$.
  - For $\tilde{k}$ successes in $\tilde{n}$ future trials, the distribution is Beta-Binomial with parameters $\tilde{n}$, $a' = a + y$, $b' = b + n - y$.

Connections to Previous/Future Topics

- **Week 2:** Directly applies the general principles (prior, likelihood, posterior, conjugacy, prediction) discussed last week.
- **Chapter 5 (Next Weeks):** Will follow a very similar structure (Likelihood + Conjugate Prior -¿ Posterior) but for normally distributed data and estimating means/variances. The concept of updating parameters based on data statistics will reappear.
- **Later Chapters (MCMC):** If we chose a non-Beta prior for θ, the posterior would not be a standard distribution, and we would need computational methods (Ch 7-9) to approximate it.

Common Pitfalls / Points of Confusion

- **Mixing up a, b with $y, n-y$:** Remember a, b are prior parameters, while $y, n-y$ are data counts. The posterior parameters combine them: $a_{post} = a_{prior} + y$, $b_{post} = b_{prior} + (n-y)$.
- **Interpretation of a, b:** While the "pseudo-count" idea is helpful, a and b don't have to be integers. They are shape parameters defining the prior density. $a = 1, b = 1$ means a prior 'sample size' of 2.
- **Likelihood vs. Prior Kernels:** Forgetting to drop constants ($\binom{n}{y}$, Gamma functions) when identifying the kernel for proportionality arguments. The kernel only includes terms involving θ.
- **Credible Interval Interpretation:** Stating the frequentist confidence interval interpretation. Reiterate: The Bayesian interval makes a direct probability statement about the parameter θ.
- **Hypothesis Testing:** Comparing the Bayesian posterior probability $P(H_1|y)$ directly to a frequentist p-value. They are conceptually different. The Bayesian approach gives the probability of the hypothesis being true, given the data.
- **Predictive Distributions:** Confusing the prior predictive (what data was expected before observation?) with the posterior predictive (what data is expected next, given what we saw?).

End of Week 3 Notes.

Bayesian Statistical Methods - Study Notes Week 4

Inference for Normal Means (Known Variance)

Based on Hoff (2009), Chapter 5 (Sections 5.1-5.2)

Instructor: Sudeep Regmi

Chapters Covered

- Hoff Chapter 5: Inference for Normal Mean and Variance
  - Section 5.1: The Normal Model
  - Section 5.2: Inference for the Mean (Known Variance)

Learning Objectives

After engaging with this week's material, you should be able to:

- Write down the likelihood function for the mean μ of a Normal distribution, assuming the variance σ^2 is known, based on n independent observations $y_1, \ldots, y_n$.
- Identify the kernel of the Normal likelihood function with respect to the mean μ.
- Describe the properties of the Normal distribution when used as a prior for the mean μ.
- Interpret the hyperparameters of the Normal prior for μ: the prior mean m_0 and prior variance s_0^2.
- Demonstrate and explain the conjugacy between the Normal prior (for μ) and the Normal likelihood (with known σ^2).
- Derive the posterior distribution for μ as a Normal distribution, including its mean m_n and variance s_n^2.
- Understand and interpret the posterior mean m_n as a precision-weighted average of the prior mean m_0 and the sample mean $\bar{y}$.
- Understand and interpret the posterior precision $(1/s_n^2)$ as the sum of the prior precision $(1/s_0^2)$ and the data precision (n/σ^2).
- Construct and interpret credible intervals for μ based on the Normal posterior distribution.
- Compare and contrast the Bayesian posterior distribution and credible interval for μ with the frequentist sampling distribution of $\bar{y}$ and the corresponding confidence interval.

Key Concepts & Definitions

- **Normal Distribution ($X \sim \mathcal{N}(\mu, \sigma^2)$):** A continuous probability distribution characterized by its mean μ and variance σ^2.
  - **PDF:** $p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$.
- **Likelihood Function for μ (Known σ^2):** Given i.i.d. data $y_1, \dots, y_n \sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 known, the likelihood function for μ is:

$$L(\mu|\mathbf{y}, \sigma^2) = p(\mathbf{y}|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu)^2\right\}$$

$$L(\mu|\mathbf{y}, \sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}$$

  Further simplification using $\sum(y_i-\mu)^2 = \sum(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2$ shows the kernel depends on μ only through the $n(\bar{y}-\mu)^2$ term:

$$L(\mu|\mathbf{y}, \sigma^2) \propto \exp\left\{-\frac{n}{2\sigma^2}(\mu-\bar{y})^2\right\}$$

  This is the kernel of a $\mathcal{N}(\bar{y}, \sigma^2/n)$ distribution (as a function of μ).
- **Sample Mean ($\bar{y}$):** $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. It is the sufficient statistic for μ when σ^2 is known.
- **Normal Prior for μ:** We assume a prior belief about μ that follows a Normal distribution: $\mu \sim \mathcal{N}(m_0, s_0^2)$.
  - m_0: Prior mean (our best guess for μ before seeing data).
  - s_0^2: Prior variance (our uncertainty about μ before seeing data). Larger s_0^2 means less certainty.
- **Precision:** The reciprocal of variance ($1/\sigma^2$). Higher precision means lower variance/uncertainty. Often used in Bayesian Normal models for algebraic convenience.
  - Prior Precision: $1/s_0^2$.
  - Data Precision (per observation): $1/\sigma^2$.
  - Precision of Sample Mean: n/σ^2.
- **Normal-Normal Conjugacy (Known σ^2):** If the prior for μ is Normal and the likelihood for μ (from Normal data with known σ^2) is Normal, then the posterior for μ is also Normal.
- **Posterior Distribution for μ:** If prior is $\mu \sim \mathcal{N}(m_0, s_0^2)$ and data are $y_1, \dots, y_n \sim \mathcal{N}(\mu, \sigma^2)$ (known σ^2), then the posterior is $\mu|\mathbf{y} \sim \mathcal{N}(m_n, s_n^2)$, where:
  - **Posterior Mean:** $m_n = \frac{\frac{1}{s_0^2}m_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{s_0^2} + \frac{n}{\sigma^2}}$
  - **Posterior Variance:** $s_n^2 = \left(\frac{1}{s_0^2} + \frac{n}{\sigma^2}\right)^{-1}$
  - **Posterior Precision:** $\frac{1}{s_n^2} = \frac{1}{s_0^2} + \frac{n}{\sigma^2}$ (Posterior Precision = Prior Precision + Data Precision)
- **Credible Interval for μ:** A $100(1-\alpha)\%$ credible interval for μ is given by $[m_n - z_{\alpha/2}s_n, m_n + z_{\alpha/2}s_n]$, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard Normal distribution $\mathcal{N}(0, 1)$.

Conceptual Overview / "The Big Picture"

This week shifts focus from discrete outcomes (counts, proportions via Binomial) to continuous outcomes modeled by the Normal distribution. We address the fundamental problem of estimating the population mean μ when data $y_1, \ldots, y_n$ are assumed to come from a $\mathcal{N}(\mu, \sigma^2)$ distribution. To simplify things initially, we assume the population variance σ^2 is known. Following the pattern from the Beta-Binomial model, we specify a prior distribution for the unknown parameter μ. Since μ can take any real value, the Normal distribution itself is a natural and mathematically convenient choice for the prior, $\mu \sim \mathcal{N}(m_0, s_0^2)$. The core result is demonstrating that this Normal prior is conjugate to the Normal likelihood (when σ^2 is known). This means the posterior distribution for μ is also Normal. We derive the parameters (mean and variance) of this posterior distribution and see how they intuitively combine information from the prior (via m_0, s_0^2) and the data (via $\bar{y}, n, \sigma^2$).

Core Ideas & Summaries

5.1 The Normal Model Likelihood (Known σ^2)

- Data: $y_1, \ldots, y_n$ are i.i.d from $\mathcal{N}(\mu, \sigma^2)$, with σ^2 known.
- Likelihood Function: $L(\mu|\mathbf{y}, \sigma^2) = p(\mathbf{y}|\mu, \sigma^2) = \prod p(y_i|\mu, \sigma^2)$.
- Simplifying the exponent: The sum $\sum(y_i - \mu)^2$ is key. By adding and subtracting $\bar{y}$ inside the square and expanding, we get $\sum(y_i - \mu)^2 = \sum(y_i - \bar{y} + \bar{y} - \mu)^2 = \sum(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2$.
- Kernel of the Likelihood: Since $\sum(y_i - \bar{y})^2$, n, and σ^2 do not depend on μ, the likelihood is proportional to:

  $$L(\mu|\mathbf{y}, \sigma^2) \propto \exp\left\{-\frac{n}{2\sigma^2}(\mu - \bar{y})^2\right\}$$

  This is recognized as the kernel of a Normal PDF for μ, centered at $\bar{y}$ with variance σ^2/n. The sample mean $\bar{y}$ encapsulates all the information in the data relevant to μ.

5.2 Normal Prior, Normal Posterior (Known σ^2)

- Prior Specification: Assume prior belief $\mu \sim \mathcal{N}(m_0, s_0^2)$. The PDF is:

  $$p(\mu|m_0, s_0^2) \propto \exp\left\{-\frac{1}{2s_0^2}(\mu - m_0)^2\right\}$$

- Apply Bayes' Theorem: $p(\mu|\mathbf{y}) \propto L(\mu|\mathbf{y}) \times p(\mu)$.

  $$p(\mu|\mathbf{y}) \propto \exp\left\{-\frac{n}{2\sigma^2}(\mu - \bar{y})^2\right\} \times \exp\left\{-\frac{1}{2s_0^2}(\mu - m_0)^2\right\}$$

  $$p(\mu|\mathbf{y}) \propto \exp\left\{-\frac{1}{2}\left[\frac{n}{\sigma^2}(\mu - \bar{y})^2 + \frac{1}{s_0^2}(\mu - m_0)^2\right]\right\}$$

- Completing the Square: The term inside the square brackets is a quadratic in μ. Expanding the squares and collecting terms involving μ^2 and μ:

  $$\cdots = \left(\frac{n}{\sigma^2} + \frac{1}{s_0^2}\right)\mu^2 - 2\left(\frac{n\bar{y}}{\sigma^2} + \frac{m_0}{s_0^2}\right)\mu + \text{(terms not involving } \mu)$$

  This expression has the form $A\mu^2 - 2B\mu + C$. To match the Normal kernel $\exp\{-\frac{1}{2s_n^2}(\mu - m_n)^2\} \propto \exp\{-\frac{1}{2}(\frac{1}{s_n^2}\mu^2 - \frac{2m_n}{s_n^2}\mu + \ldots)\}$, we equate coefficients:

  - $\frac{1}{s_n^2} = A = \frac{n}{\sigma^2} + \frac{1}{s_0^2}$ (Posterior Precision = Data Precision + Prior Precision)
  - $\frac{m_n}{s_n^2} = B = \frac{n\bar{y}}{\sigma^2} + \frac{m_0}{s_0^2}$

- Solving for Posterior Parameters:
  - Posterior Variance: $s_n^2 = \left(\frac{1}{s_0^2} + \frac{n}{\sigma^2}\right)^{-1}$
  - Posterior Mean: $m_n = s_n^2 \times B = \left(\frac{1}{s_0^2} + \frac{n}{\sigma^2}\right)^{-1} \left(\frac{m_0}{s_0^2} + \frac{n\bar{y}}{\sigma^2}\right) = \frac{\frac{1}{s_0^2}m_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{s_0^2} + \frac{n}{\sigma^2}}$

- Interpretation of Posterior Mean: m_n is a weighted average of the prior mean m_0 and the data mean $\bar{y}$, weighted by their respective precisions. As $n \to \infty$ (or as prior variance $s_0^2 \to \infty$, a "non-informative" prior), the posterior mean $m_n \to \bar{y}$.

- Interpretation of Posterior Precision/Variance: The posterior precision is simply the sum of the prior precision and the precision contributed by the data. This means the posterior variance s_n^2 is always smaller than both the prior variance s_0^2 and the variance of the sample mean σ^2/n. Uncertainty decreases by combining information.

Comparison with Frequentist Inference

- **Frequentist:** Treats μ as fixed and unknown. Inference based on the sampling distribution of the estimator $\bar{y}$, which is $\bar{y} \sim \mathcal{N}(\mu, \sigma^2/n)$.
  - Point Estimate: $\hat{\mu} = \bar{y}$.
  - Confidence Interval: $\bar{y} \pm z_{\alpha/2}\sqrt{\sigma^2/n}$. Interpreted as: if we repeated the experiment many times, $(1-\alpha)\%$ of the *intervals* constructed this way would contain the true fixed μ.

- **Bayesian:** Treats μ as random, described by the posterior $\mu|\mathbf{y} \sim \mathcal{N}(m_n, s_n^2)$.
  - Point Estimate: Could be posterior mean m_n, median (also m_n here), or mode (also m_n). Note $m_n \neq \bar{y}$ unless the prior is non-informative ($1/s_0^2 = 0$) or $m_0 = \bar{y}$.
  - Credible Interval: $m_n \pm z_{\alpha/2}\sqrt{s_n^2}$. Interpreted as: there is a $(1-\alpha)\%$ probability that the true value of μ lies within this specific interval, given the data and prior.

- **Key Difference:** Interpretation. Bayesian inference yields direct probability statements about the parameter μ. Frequentist inference makes probability statements about the procedure and data, conditional on the unknown fixed parameter.

Connections to Previous/Future Topics

- **Week 2 3:** Continues the pattern of Likelihood × Conjugate Prior → Posterior, now with continuous data and parameters. The idea of posterior mean being a weighted average was also seen implicitly in the Beta-Binomial model.

- **Week 5 (Next Week):** Will address the more realistic case where σ^2 is also unknown. This requires a joint prior for (μ, σ^2) and leads to a different posterior distribution (related to the Student's t-distribution). The concepts of likelihood and prior specification remain central.

- **Limit of large n or $s_0^2 \to \infty$:** Shows how Bayesian results can sometimes converge to frequentist results under non-informative priors or large datasets, though interpretations remain distinct.

Common Pitfalls / Points of Confusion

- **Known vs. Unknown Variance:** Forgetting that σ^2 is assumed known throughout this specific derivation. The results change when σ^2 is unknown.
- **Algebra Errors:** Mistakes in the "completing the square" step when deriving the posterior parameters. Carefully track terms involving μ^2 and μ.
- **Confusing Variances:** Mixing up the (known) data variance σ^2, the prior variance s_0^2, the variance of the sample mean σ^2/n, and the posterior variance s_n^2. Keep track of what quantity each variance refers to.
- **Notation:** Confusing prior parameters (m_0, s_0^2) with posterior parameters (m_n, s_n^2).
- **Interpretation of Intervals:** Incorrectly stating the frequentist interpretation for a Bayesian credible interval, or vice versa.
- **Precision vs. Variance:** While equivalent, switching between precision $(1/\sigma^2)$ and variance (σ^2) can sometimes cause confusion if not done carefully. The "sum of precisions" rule is often simpler algebraically.

End of Week 4 Notes.

Bayesian Statistical Methods - Study Notes Week 5

Inference for Normal Mean and Variance (Unknown Variance)

Based on Hoff (2009), Chapter 5 (Sections 5.3-5.5)

Instructor: Sudeep Regmi

Chapters Covered

- Hoff Chapter 5: Inference for Normal Mean and Variance
  - Section 5.3: Inference for the Variance (Known Mean - Conceptual Basis)
  - Section 5.4: Joint Inference for Mean and Variance
  - Section 5.5: Marginal Inference

Note: Section 5.3 introduces priors for variance assuming a known mean, which helps build intuition, but the main focus is the joint inference in 5.4 and the resulting marginals in 5.5.

Learning Objectives

After engaging with this week's material, you should be able to:

- Understand the need for a joint prior distribution $p(\mu, \sigma^2)$ when both parameters are unknown.
- Recognize the Scaled Inverse Chi-Squared distribution (Inv-$\chi^2(\nu, S^2)$) as a suitable prior for the variance σ^2.
- Understand the structure of the standard conjugate prior for (μ, σ^2): $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$ and $\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0)$.
- Interpret the hyperparameters of the conjugate prior: $\mu_0, \kappa_0, \nu_0, \sigma_0^2$.
- Write down the likelihood function for (μ, σ^2) based on i.i.d. Normal data $\mathbf{y} = (y_1, \ldots, y_n)$.
- Identify the sufficient statistics for Normal data: sample size n, sample mean $\bar{y}$, and sample variance s^2 (or sum of squares $(n-1)s^2$).
- Understand that the joint posterior distribution $p(\mu, \sigma^2|\mathbf{y})$ under the conjugate prior is also Normal-Inverse-Chi-Squared, and know the rules for updating the hyperparameters to $\mu_n, \kappa_n, \nu_n, \sigma_n^2$.
- Derive or state the marginal posterior distribution for the variance $\sigma^2|\mathbf{y}$, which is Inv-$\chi^2(\nu_n, \sigma_n^2)$.
- Derive or state the marginal posterior distribution for the mean $\mu|\mathbf{y}$, which follows a Student's t-distribution. Specifically, $(\mu - \mu_n)/\sqrt{\sigma_n^2/\kappa_n} \sim t_{\nu_n}$.
- Explain why the marginal posterior for μ is a t-distribution rather than a Normal distribution.

- Construct and interpret credible intervals for μ using quantiles of the appropriate Student's t-distribution.
- Construct and interpret credible intervals for σ^2 using quantiles of the appropriate Inverse Chi-Squared distribution.

Key Concepts & Definitions

- **Likelihood Function $L(\mu, \sigma^2|\mathbf{y})$:** For i.i.d $y_1, \dots, y_n \sim \mathcal{N}(\mu, \sigma^2)$:

$$L(\mu, \sigma^2|\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\}$$

$$L(\mu, \sigma^2|\mathbf{y}) \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}$$

  Using $\sum(y_i - \mu)^2 = (n-1)s^2 + n(\bar{y} - \mu)^2$, where $s^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2$:

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

- **Sufficient Statistics:** For (μ, σ^2) in the Normal model: the sample size n, sample mean $\bar{y}$, and sample variance s^2 (or equivalently, $(n-1)s^2 = \sum(y_i - \bar{y})^2$).
- **Scaled Inverse Chi-Squared Distribution ($\sigma^2 \sim$ Inv-$\chi^2(\nu, S^2)$):** A distribution for positive continuous random variables, often used as a prior for variances. Parameterized by degrees of freedom ν and scale S^2. (Related to Inverse Gamma: Inv-$\chi^2(\nu, S^2) \equiv$ Inv-Gamma$(\nu/2, \nu S^2/2)$). Hoff uses σ_0^2 for the scale parameter in the prior.
  - PDF: $p(\sigma^2|\nu, \sigma_0^2) \propto (\sigma^2)^{-(\nu/2+1)} \exp(-\nu\sigma_0^2/(2\sigma^2))$.
- **Conjugate Prior (Normal-Inverse-Chi-Squared):** A joint prior structure for (μ, σ^2):
  1. $\sigma^2 \sim$ Inv-$\chi^2(\nu_0, \sigma_0^2)$
  2. $\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0)$

  Hyperparameters:

  - μ_0: Prior mean for μ.
  - κ_0: Prior effective sample size for μ. Controls how strongly prior mean μ_0 influences the posterior mean relative to the data mean $\bar{y}$. Larger κ_0 means stronger prior belief about μ.
  - ν_0: Prior degrees of freedom for σ^2. Controls prior certainty about σ^2. Larger ν_0 means stronger prior belief.
  - σ_0^2: Prior scale parameter for σ^2. Related to the prior estimate of the variance.
- **Reference Prior / Non-informative Prior:** A commonly used prior attempting to let the data dominate: $p(\mu, \sigma^2) \propto 1/\sigma^2$. This corresponds to the limit of the conjugate prior as $\kappa_0 \to 0, \nu_0 \to -1$ (or sometimes $\nu_0 \to 0$).
- **Joint Posterior Distribution:** Under the conjugate prior and Normal likelihood, the joint posterior is also Normal-Inverse-Chi-Squared:
  1. $\sigma^2|\mathbf{y} \sim$ Inv-$\chi^2(\nu_n, \sigma_n^2)$

2. $\mu|\sigma^2, \mathbf{y} \sim \mathcal{N}(\mu_n, \sigma^2/\kappa_n)$

Updated Hyperparameters:

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

- **Marginal Posterior for σ^2:** $\sigma^2|\mathbf{y} \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2)$. Found by integrating the joint posterior $p(\mu, \sigma^2|\mathbf{y})$ over μ.

- **Marginal Posterior for μ (Student's t-distribution):** Found by integrating the joint posterior $p(\mu, \sigma^2|\mathbf{y})$ over σ^2. The result is that the standardized quantity follows a t-distribution:
$$\frac{\mu - \mu_n}{\sqrt{\sigma_n^2/\kappa_n}} \sim t_{\nu_n}$$
This means μ follows a *location-scale t-distribution* with location μ_n, scale $\sqrt{\sigma_n^2/\kappa_n}$, and ν_n degrees of freedom.

Conceptual Overview / "The Big Picture"

This week addresses the standard, realistic scenario for normally distributed data where both the mean μ and the variance σ^2 are unknown and need to be estimated. Since both parameters are unknown, we need a joint prior distribution $p(\mu, \sigma^2)$ to represent our initial beliefs. We introduce the standard conjugate prior structure, the Normal-Inverse-Chi-Squared distribution, which allows for tractable updating. The key idea is that our beliefs about μ might depend on σ^2 (hence $\mu|\sigma^2 \sim Normal$), and we have separate beliefs about σ^2 itself ($\sigma^2 \sim \text{Inv-}\chi^2$). The joint posterior distribution combines the prior information with the likelihood (summarized by $n, \bar{y}, s^2$). While the joint posterior contains all information, we are often interested in one parameter at a time. This requires finding the *marginal posterior distributions* by integrating out the other parameter. The marginal posterior for σ^2 remains an Inverse Chi-Squared distribution. Crucially, the marginal posterior for μ becomes a Student's t-distribution. The t-distribution has heavier tails than the Normal distribution derived in Week 4 (when σ^2 was known), reflecting the additional uncertainty introduced by not knowing the true variance σ^2.

Core Ideas & Summaries

5.3 Inference for Variance (Known Mean)

- Sets the stage by considering how to put a prior on σ^2.
- If μ were known, the likelihood for σ^2 involves $\sum(y_i - \mu)^2$.
- The $\text{Inv-}\chi^2(\nu_0, \sigma_0^2)$ distribution is conjugate for σ^2 in this simpler (known μ) case. It's defined for $\sigma^2 > 0$.

5.4 Joint Inference for Mean and Variance

- **Likelihood:** Depends on both μ and σ^2, involving the sufficient statistics $n, \bar{y}, (n-1)s^2$.

- **Prior Choice - Conjugate:** The Normal-Inv-$\chi^2(\mu_0, \sigma^2/\kappa_0; \nu_0, \sigma_0^2)$ structure is chosen for mathematical convenience. It implies prior dependence between μ and σ^2 (variance of μ depends on σ^2).
- **Joint Posterior:** Multiplying the likelihood and the joint prior results in a kernel that corresponds to another Normal-Inverse-Chi-Squared distribution. The update rules (given in Key Concepts) show how the prior parameters are modified by the data statistics $n, \bar{y}, s^2$.
  - μ_n is a weighted average of μ_0 and $\bar{y}$, weighted by κ_0 and n.
  - κ_n and ν_n are simply sums of prior and data counts/degrees of freedom.
  - $\nu_n\sigma_n^2$ (the posterior sum of squares parameter) combines the prior sum of squares ($\nu_0\sigma_0^2$), the data sum of squares ($(n-1)s^2$), and a term accounting for the difference between the prior mean μ_0 and data mean $\bar{y}$.
- **Reference Prior $p(\mu, \sigma^2) \propto 1/\sigma^2$:** Leads to a proper posterior (also Normal-Inv-ChiSq) provided $n \geq 2$. The posterior parameters correspond to using $\kappa_0 = 0, \nu_0 = -1, \nu_0\sigma_0^2 = 0$ (or similar limiting values, e.g., $\nu_0 = 0$) in the update formulas. Often $\mu_n = \bar{y}$, $\kappa_n = n$, $\nu_n = n-1$, $\nu_n\sigma_n^2 = (n-1)s^2$.

5.5 Marginal Inference

- **Goal:** Get $p(\mu|\mathbf{y})$ and $p(\sigma^2|\mathbf{y})$ from $p(\mu, \sigma^2|\mathbf{y})$.
- **Marginal for σ^2:** Integrating $p(\mu, \sigma^2|\mathbf{y}) = p(\mu|\sigma^2, \mathbf{y})p(\sigma^2|\mathbf{y})$ over μ. Since $p(\mu|\sigma^2, \mathbf{y})$ is a Normal PDF in μ, it integrates to 1 (w.r.t μ). The result leaves $p(\sigma^2|\mathbf{y})$ which is Inv-$\chi^2(\nu_n, \sigma_n^2)$. Credible intervals for σ^2 can be found using quantiles of this distribution.
- **Marginal for μ:** Integrating $p(\mu, \sigma^2|\mathbf{y})$ over σ^2. This involves integrating the product of a Normal kernel for μ (conditional on σ^2) and an Inverse Chi-Squared kernel for σ^2. This specific integral mathematically results in the kernel of a Student's t-distribution PDF for μ.
  - Distribution: $\mu|\mathbf{y} \sim t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n)$, meaning $(\mu - \mu_n)/\sqrt{\sigma_n^2/\kappa_n} \sim t_{\nu_n}$.
  - Parameters: Location μ_n, Scale $\sqrt{\sigma_n^2/\kappa_n}$, Degrees of freedom ν_n.
  - **Why t-distribution?** It arises from averaging Normal distributions (for $\mu|\sigma^2$) over the uncertainty in σ^2 (described by its Inv-ChiSq posterior). The heavier tails account for this extra uncertainty about the scale.
  - Credible Interval for μ: $[\mu_n - t_{\nu_n,\alpha/2}\sqrt{\sigma_n^2/\kappa_n}, \mu_n + t_{\nu_n,\alpha/2}\sqrt{\sigma_n^2/\kappa_n}]$, where $t_{\nu_n,\alpha/2}$ is the $1-\alpha/2$ quantile of the standard t_{ν_n} distribution.

Connections to Previous/Future Topics

- **Week 4:** Contrasts directly with the known-variance case. The posterior for μ was Normal(m_n, s_n^2) there; here it's Student's t, reflecting the added uncertainty about σ^2. As $n \to \infty$ (thus $\nu_n \to \infty$), the t_{ν_n} distribution approaches a Normal distribution, and the Bayesian results often converge towards frequentist results based on the t-test.
- **Bayesian Linear Regression (Chapter 12):** The Normal-Inverse-Chi-Squared prior structure and the resulting t-distribution marginal posteriors for coefficients form the basis of standard Bayesian linear regression models.
- **Hierarchical Models (Chapter 11):** Understanding priors on variance parameters (σ^2) is crucial for hierarchical models where variances themselves are modeled across groups.

Common Pitfalls / Points of Confusion

- **Joint vs. Marginal vs. Conditional:** Confusing $p(\mu, \sigma^2|\mathbf{y})$ (joint), $p(\mu|\mathbf{y})$ (marginal), $p(\sigma^2|\mathbf{y})$ (marginal), and $p(\mu|\sigma^2, \mathbf{y})$ (conditional). Keep track of what is being held fixed and what is integrated out.

- **Parameters of the t-distribution:** Remembering the correct location (μ_n), scale ($\sqrt{\sigma_n^2/\kappa_n}$), and degrees of freedom (ν_n) for the marginal posterior of μ. Note the scale involves the posterior *scale* parameter σ_n^2, not just the sample variance s^2.

- **Hyperparameter Roles:** Misinterpreting κ_0 (prior strength for mean) or ν_0 (prior strength for variance).

- **Algebra:** The formula for $\nu_n \sigma_n^2$ can look intimidating. Focus on understanding its components: prior sum of squares, data sum of squares, and the adjustment term for difference between prior/data means.

- **Inverse Gamma vs. Inverse Chi-Squared:** Different texts use slightly different parameterizations for the variance prior/posterior (Inv-Gamma vs. Scaled-Inv-ChiSq). Be mindful of the specific definitions and parameters used (Hoff uses Scaled Inv-ChiSq).

- **Why t, not Normal?** Forgetting the conceptual reason for the t-distribution: it explicitly accounts for the uncertainty in σ^2 when making inferences about μ.

End of Week 5 Notes.

Bayesian Statistical Methods - Study Notes Week 6

Multiparameter Inference: Proportions and Means

Based on Hoff (2009), Chapter 6

Instructor: Sudeep Regmi

Chapters Covered

- Hoff Chapter 6: Inference for Multivariate Parameters
  - Multinomial Model for Categorical Data
  - Dirichlet Prior and Posterior Distributions
  - Multivariate Normal Model (focus on known covariance)

Learning Objectives

After engaging with this week's material, you should be able to:

- Identify scenarios where the Multinomial distribution is an appropriate model for count data across multiple categories.
- Write down the Multinomial likelihood function for the category probabilities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ given observed counts $\mathbf{y} = (y_1, \ldots, y_k)$.
- Describe the properties of the Dirichlet distribution (Dirichlet$(\alpha_1, \ldots, \alpha_k)$) and explain why it serves as a flexible prior for a probability vector $\boldsymbol{\theta}$ (where components are positive and sum to 1).
- Interpret the concentration parameters $(\alpha_1, \ldots, \alpha_k)$ of the Dirichlet distribution, often viewed as prior pseudo-counts for each category.
- Demonstrate and explain the conjugacy between the Dirichlet prior and the Multinomial likelihood.
- Derive the posterior distribution for $\boldsymbol{\theta}$ as a Dirichlet$(\alpha_1 + y_1, \ldots, \alpha_k + y_k)$ distribution.
- Calculate posterior means for individual probabilities θ_j.
- Understand that the marginal posterior distribution for a single probability θ_j is a Beta$(\alpha_j + y_j, \sum_{i \neq j}(\alpha_i + y_i))$ distribution.
- Define the parameters (mean vector $\boldsymbol{\mu}$, covariance matrix Σ) of a Multivariate Normal (MVN) distribution.
- Write down the MVN likelihood function for the mean vector $\boldsymbol{\mu}$ assuming the covariance matrix Σ is known.
- Specify a conjugate MVN prior for the mean vector $\boldsymbol{\mu}$.
- Derive the posterior distribution for $\boldsymbol{\mu}$ (when Σ is known) as another MVN distribution, understanding the role of precision matrices in the update formulas.

Key Concepts & Definitions

- **Multinomial Distribution ($\mathbf{Y} \sim$ Multinomial$(n, \boldsymbol{\theta})$):** Generalization of the Binomial for $k \geq 2$ categories. Describes the probability of obtaining counts $\mathbf{y} = (y_1, \ldots, y_k)$ in k categories after n independent trials, where $\sum y_i = n$. The vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ contains the probability of falling into each category, with $\theta_i \geq 0$ and $\sum \theta_i = 1$.
  - **PMF:** $p(\mathbf{y}|n, \boldsymbol{\theta}) = \frac{n!}{y_1! y_2! \cdots y_k!} \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k}$.
  - **Likelihood Function $L(\boldsymbol{\theta}|\mathbf{y}, n)$:** The PMF viewed as a function of $\boldsymbol{\theta}$. Proportional to $\prod_{i=1}^k \theta_i^{y_i}$.

- **Probability Vector ($\boldsymbol{\theta}$):** A vector $(\theta_1, \ldots, \theta_k)$ such that $\theta_i \geq 0$ for all i and $\sum_{i=1}^k \theta_i = 1$. Lives on the standard $(k-1)$-simplex.

- **Dirichlet Distribution ($\boldsymbol{\theta} \sim$ Dirichlet$(\boldsymbol{\alpha})$):** A multivariate continuous distribution defined on the simplex, generalizing the Beta distribution. Parameterized by a vector of positive concentration parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$.
  - **PDF:** $p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k \theta_i^{\alpha_i - 1}$ for $\boldsymbol{\theta}$ on the simplex.
  - **Concentration Parameters (α_i):** Control the shape and concentration. Often interpreted as prior pseudo-counts for category i. Larger $\sum \alpha_i = \alpha_0$ means the distribution is more concentrated.
  - **Mean:** $\mathbb{E}[\theta_j] = \frac{\alpha_j}{\sum_{i=1}^k \alpha_i} = \frac{\alpha_j}{\alpha_0}$.
  - **Marginal Distribution:** If $\boldsymbol{\theta} \sim$ Dirichlet$(\alpha_1, \ldots, \alpha_k)$, then the marginal distribution of a single component θ_j is $\theta_j \sim$ Beta$(\alpha_j, \sum_{i \neq j} \alpha_i)$.
  - **Special Case:** Dirichlet$(1, 1, \ldots, 1)$ is the uniform distribution over the simplex.

- **Dirichlet-Multinomial Conjugacy:** The property that if the prior for $\boldsymbol{\theta}$ is Dirichlet and the likelihood is Multinomial, the posterior for $\boldsymbol{\theta}$ is also Dirichlet.

- **Posterior Distribution (Dirichlet-Multinomial):** If prior is $\boldsymbol{\theta} \sim$ Dirichlet$(\boldsymbol{\alpha})$ and data counts are $\mathbf{y} = (y_1, \ldots, y_k)$ from n trials, the posterior is $\boldsymbol{\theta}|\mathbf{y} \sim$ Dirichlet$(\boldsymbol{\alpha} + \mathbf{y}) =$ Dirichlet$(\alpha_1 + y_1, \ldots, \alpha_k + y_k)$.

- **Multivariate Normal Distribution ($\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$):** Generalization of the Normal distribution to p dimensions. Characterized by a $p \times 1$ mean vector $\boldsymbol{\mu}$ and a $p \times p$ symmetric positive semi-definite covariance matrix Σ.
  - **PDF:** $p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}$.
  - $\boldsymbol{\mu}$: Vector of means, $\mathbb{E}[X_i] = \mu_i$.
  - Σ: Covariance matrix, $\Sigma_{ij} = \text{Cov}(X_i, X_j)$. Diagonal elements $\Sigma_{ii} = \text{Var}(X_i)$.

- **Precision Matrix (Λ):** The inverse of the covariance matrix, $\Lambda = \Sigma^{-1}$. Often used in Bayesian calculations.

- **MVN Likelihood $L(\boldsymbol{\mu}|\mathbf{y}_1, \ldots, \mathbf{y}_n, \Sigma)$ (Known Σ):** For i.i.d $\mathbf{y}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$:

$$L(\boldsymbol{\mu}| \ldots) \propto \exp\left\{-\frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{y}_i - \boldsymbol{\mu})\right\}$$

$$L(\boldsymbol{\mu}| \ldots) \propto \exp\left\{-\frac{n}{2} (\boldsymbol{\mu} - \bar{\mathbf{y}})^T \Sigma^{-1} (\boldsymbol{\mu} - \bar{\mathbf{y}})\right\}$$

where $\bar{\mathbf{y}} = \frac{1}{n} \sum \mathbf{y}_i$ is the sample mean vector. This is the kernel of a $\mathcal{N}_p(\bar{\mathbf{y}}, \Sigma/n)$ distribution for $\boldsymbol{\mu}$.

- **MVN Prior for $\boldsymbol{\mu}$:** $\boldsymbol{\mu} \sim \mathcal{N}_p(\mathbf{m}_0, S_0)$.
- **Posterior for $\boldsymbol{\mu}$ (Known Σ):** If prior is $\boldsymbol{\mu} \sim \mathcal{N}_p(\mathbf{m}_0, S_0)$ and likelihood is from $\mathbf{y}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$, the posterior is $\boldsymbol{\mu}|\mathbf{y}_1, \ldots, \mathbf{y}_n \sim \mathcal{N}_p(\mathbf{m}_n, S_n)$, where:
  - Posterior Precision: $S_n^{-1} = S_0^{-1} + n\Sigma^{-1}$.
  - Posterior Mean: $\mathbf{m}_n = S_n(S_0^{-1}\mathbf{m}_0 + n\Sigma^{-1}\bar{\mathbf{y}})$.

Conceptual Overview / "The Big Picture"

This week extends the Bayesian framework from single parameters (θ, μ, σ^2) to multiple parameters considered simultaneously. We cover two fundamental multiparameter models: 1. **Multinomial Model:** Used for categorical data where each observation falls into one of k distinct categories (e.g., survey responses, types of outcomes). We estimate the vector of probabilities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ for these categories. The Dirichlet distribution serves as the conjugate prior, generalizing the Beta distribution. The inference process closely parallels the Beta-Binomial model, with prior pseudo-counts updated by observed counts. 2. **Multivariate Normal Model:** Used for continuous multivariate data where observations are vectors $\mathbf{y} = (y_1, \ldots, y_p)$ assumed to follow a Multivariate Normal distribution. This model accounts for correlations between the components of $\mathbf{y}$. We focus on the case of estimating the mean vector $\boldsymbol{\mu}$ when the covariance structure Σ is known. The MVN distribution itself serves as the conjugate prior for $\boldsymbol{\mu}$, and the update rules generalize the univariate Normal-Normal case using matrix algebra, particularly precision matrices.

Core Ideas & Summaries

Multinomial-Dirichlet Model

- **Likelihood:** Data $\mathbf{y} = (y_1, \ldots, y_k)$ with $\sum y_i = n$. Parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ with $\sum \theta_i = 1$.
$$L(\boldsymbol{\theta}|\mathbf{y}) \propto \prod_{i=1}^{k} \theta_i^{y_i}$$
- **Prior:** $\boldsymbol{\theta} \sim$ Dirichlet$(\alpha_1, \ldots, \alpha_k)$. Kernel: $\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$.
- **Posterior:** Apply Bayes' Theorem $p(\boldsymbol{\theta}|\mathbf{y}) \propto L(\boldsymbol{\theta}|\mathbf{y}) \times p(\boldsymbol{\theta})$.
$$p(\boldsymbol{\theta}|\mathbf{y}) \propto \left(\prod_{i=1}^{k} \theta_i^{y_i}\right) \times \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}\right) = \prod_{i=1}^{k} \theta_i^{(\alpha_i + y_i) - 1}$$
This is the kernel of a Dirichlet$(\alpha_1 + y_1, \ldots, \alpha_k + y_k)$ distribution.
- **Interpretation:** The posterior parameters are simply the prior pseudo-counts plus the observed counts for each category. $\alpha_0 = \sum \alpha_i$ acts as the prior effective sample size.
- **Summarization:**
  - Posterior Mean: $\mathbb{E}[\theta_j|\mathbf{y}] = \frac{\alpha_j + y_j}{\sum_{i=1}^{k}(\alpha_i + y_i)} = \frac{\alpha_j + y_j}{\alpha_0 + n}$. This is a weighted average of the prior mean α_j/α_0 and the sample proportion y_j/n.
  - Posterior Mode: Mode$(\theta_j) = \frac{\alpha_j + y_j - 1}{\alpha_0 + n - k}$ (if all $\alpha_i + y_i > 1$).

- **Marginal Inference:** Often interested in a specific θ_j. The marginal posterior distribution is Beta:

$$\theta_j|\mathbf{y} \sim \text{Beta}(\alpha_j + y_j, \sum_{i\neq j}(\alpha_i + y_i)) = \text{Beta}(\alpha_j + y_j, (\alpha_0 + n) - (\alpha_j + y_j))$$

  We can then compute credible intervals for individual θ_j using this Beta distribution, just like in Chapter 4.

Multivariate Normal Model (Known Covariance Σ)

- **Likelihood:** Data $\mathbf{y}_1, \ldots, \mathbf{y}_n$ i.i.d from $\mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$, Σ known. Summarized by sample mean vector $\bar{\mathbf{y}}$.

$$L(\boldsymbol{\mu}|\ldots) \propto \exp\left\{-\frac{n}{2}(\boldsymbol{\mu} - \bar{\mathbf{y}})^T\Sigma^{-1}(\boldsymbol{\mu} - \bar{\mathbf{y}})\right\}$$

  Kernel of $\mathcal{N}_p(\bar{\mathbf{y}}, \Sigma/n)$.

- **Prior:** $\boldsymbol{\mu} \sim \mathcal{N}_p(\mathbf{m}_0, S_0)$. Kernel $\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{m}_0)^T S_0^{-1}(\boldsymbol{\mu} - \mathbf{m}_0)\right\}$.

- **Posterior:** Apply Bayes' Theorem. The exponent involves summing two quadratic forms:

$$-\frac{1}{2}\left[(\boldsymbol{\mu} - \mathbf{m}_0)^T S_0^{-1}(\boldsymbol{\mu} - \mathbf{m}_0) + n(\boldsymbol{\mu} - \bar{\mathbf{y}})^T\Sigma^{-1}(\boldsymbol{\mu} - \bar{\mathbf{y}})\right]$$

  Completing the square in vector/matrix form (analogous to the univariate case) shows this is a quadratic form corresponding to a MVN distribution. The resulting posterior is $\boldsymbol{\mu}|\text{data} \sim \mathcal{N}_p(\mathbf{m}_n, S_n)$.

- **Posterior Parameters (using Precision Matrices $\Lambda_0 = S_0^{-1}$, $\Lambda = \Sigma^{-1}$):**
  - Posterior Precision: $\Lambda_n = S_n^{-1} = \Lambda_0 + n\Lambda$. (Posterior Precision = Prior Precision + Data Precision).
  - Posterior Mean: $\mathbf{m}_n = \Lambda_n^{-1}(\Lambda_0\mathbf{m}_0 + n\Lambda\bar{\mathbf{y}}) = S_n(S_0^{-1}\mathbf{m}_0 + n\Sigma^{-1}\bar{\mathbf{y}})$.

- **Interpretation:** The posterior mean $\mathbf{m}_n$ is a matrix-weighted average of the prior mean $\mathbf{m}_0$ and the sample mean $\bar{\mathbf{y}}$, weighted by their respective precision matrices. The posterior precision is the sum of the prior and data precisions.

- **Unknown Σ (Brief Mention):** If Σ is unknown, we need a prior for it, typically the Inverse Wishart distribution (multivariate analogue of Inv-ChiSq/Inv-Gamma). The joint posterior for $(\boldsymbol{\mu}, \Sigma)$ is Normal-Inverse-Wishart, and the marginal posterior for $\boldsymbol{\mu}$ becomes a Multivariate Student's t-distribution. (This is more advanced, analogous to Ch 5.4-5.5).

Connections to Previous/Future Topics

- **Beta-Binomial (Ch 4):** The Dirichlet-Multinomial model is a direct generalization. If $k = 2$, the Multinomial becomes Binomial, and the Dirichlet(α_1, α_2) becomes Beta(α_1, α_2).

- **Normal-Normal (Ch 5.2):** The MVN inference with known Σ is a direct generalization of the univariate Normal case with known σ^2. The update rules for mean and precision have the same structure, just using matrix operations.

- **Normal-InvChiSq (Ch 5.4-5.5):** The (uncovered) case of unknown Σ using the Normal-Inverse-Wishart posterior parallels the Normal-InvChiSq structure and leads to marginal t-distributions for mean components.

- **Linear Regression (Ch 12):** Estimating multiple regression coefficients $\boldsymbol{\beta}$ often involves Multivariate Normal priors and posteriors.
- **Hierarchical Models (Ch 11):** Often involve vectors of parameters (e.g., means for different groups) which might be modeled using MVN distributions.

Common Pitfalls / Points of Confusion

- **Notation Overload:** Keeping track of vectors $(\boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{y}, \boldsymbol{\mu}, \mathbf{m}_0, \mathbf{m}_n, \bar{\mathbf{y}})$ and matrices $(\Sigma, S_0, S_n, \Lambda, \Lambda_0, \Lambda_n)$. Bold symbols help!
- **Dirichlet Parameters:** Understanding that $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ determines the shape. Larger α_i relative to others pulls the distribution towards the i-th corner of the simplex. $\alpha_0 = \sum \alpha_i$ controls overall concentration.
- **Dirichlet Marginals:** Remembering that the marginal distribution of θ_j is Beta, and knowing how to find its parameters $(\alpha_j + y_j, \alpha_0 + n - (\alpha_j + y_j))$.
- **MVN Covariance vs. Precision:** Understanding the role of both Σ and its inverse $\Lambda = \Sigma^{-1}$. Bayesian updates are often cleaner using precision matrices.
- **Matrix Algebra:** Following the derivation of the MVN posterior requires comfort with basic matrix operations (transpose, inverse, multiplication) and the definition of quadratic forms. The structure mirrors the univariate case, but the mechanics are matrix-based.
- **Known vs. Unknown Σ:** Being clear that the simpler MVN results derived here assume Σ is known. The unknown case leads to multivariate t-distributions (analogous to Ch 5.5).

End of Week 6 Notes.

Bayesian Statistical Methods - Study Notes Week 7

Introduction to Bayesian Computation: Monte Carlo Methods

Based on Hoff (2009), Chapter 7

Instructor: Sudeep Regmi

Chapters Covered

- Hoff Chapter 7: Bayesian Computation

Learning Objectives

After engaging with this week's material, you should be able to:

- Explain why computational methods are essential for practical Bayesian inference, particularly in non-conjugate models or complex scenarios.
- Define Monte Carlo integration and explain how it uses random samples to approximate integrals (especially expectations).
- State the Law of Large Numbers (LLN) and explain its relevance as the theoretical foundation for Monte Carlo integration.
- Describe how to use a sample $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ drawn from the posterior distribution $p(\theta|y)$ to estimate posterior summaries like means, variances, probabilities, and quantiles (credible intervals).
- Explain the concept of Monte Carlo error and how to estimate the standard error of a Monte Carlo estimate (particularly the posterior mean).
- Describe methods for generating random samples from target distributions:
  - Direct sampling (using built-in functions for standard distributions).
  - Inverse CDF method (conceptually).
  - Rejection Sampling algorithm: Understand its requirements, steps, and efficiency considerations.
- Define target density, proposal density, and acceptance probability in the context of rejection sampling.
- Recognize the limitations of basic Monte Carlo methods, motivating the need for Markov chain Monte Carlo (MCMC).

Key Concepts & Definitions

- **Intractability Problem:** Many Bayesian analyses require calculating integrals that lack closed-form analytical solutions. Examples include the normalizing constant $p(y)$, posterior means $\mathbb{E}[g(\theta)|y]$, posterior probabilities $P(\theta \in A|y)$, and predictive distributions $p(\tilde{y}|y)$.
- **Monte Carlo Integration:** A numerical method using random sampling to approximate the value of an integral. Specifically, to approximate $\mathbb{E}[g(\theta)|y] = \int g(\theta)p(\theta|y)d\theta$, we draw S samples $\theta^{(1)}, \dots, \theta^{(S)}$ from the posterior distribution $p(\theta|y)$ and compute the sample average:
$$\widehat{E[g(\theta)|y]} = \bar{g}_S = \frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)})$$
- **Law of Large Numbers (LLN):** States that, under general conditions, the sample average $\bar{g}_S$ converges to the true expectation $\mathbb{E}[g(\theta)|y]$ as the sample size $S \to \infty$. This guarantees that Monte Carlo integration works for large enough S.
- **Posterior Sample:** A collection of random draws $\{\theta^{(1)}, \dots, \theta^{(S)}\}$ from the posterior distribution $p(\theta|y)$. This sample is the basis for Monte Carlo inference.
- **Monte Carlo Error:** The difference between the Monte Carlo estimate (e.g., $\bar{g}_S$) and the true value it approximates (e.g., $\mathbb{E}[g(\theta)|y]$). It arises because we use a finite sample size S.
- **Standard Error of Monte Carlo Mean:** A measure of the typical size of the Monte Carlo error for the posterior mean estimate. Estimated as $\text{SE}(\bar{\theta}_S) = \sqrt{\widehat{\text{Var}}(\theta)/S}$, where $\widehat{\text{Var}}(\theta)$ is the sample variance of the posterior draws $\{\theta^{(1)}, \dots, \theta^{(S)}\}$.
- **Direct Sampling:** Generating random variables directly from a target distribution using built-in computer functions (e.g., 'rnorm', 'rbeta' in R) when the target is a standard distribution.
- **Inverse CDF Method:** A general method to sample from a distribution F by sampling $u \sim \text{Uniform}(0, 1)$ and calculating $x = F^{-1}(u)$. Requires knowing and being able to invert the cumulative distribution function F. Often difficult in practice.
- **Rejection Sampling:** An algorithm to generate samples from a target density $f(\theta)$ (which we may only know up to a normalizing constant, e.g., $f(\theta) = p(y|\theta)p(\theta)$) using samples from an easier proposal density $g(\theta)$.
  - **Target Density ($f(\theta)$):** The density we want to sample from (e.g., the posterior density $p(\theta|y)$ or its unnormalized kernel $p(y|\theta)p(\theta)$).
  - **Proposal Density ($g(\theta)$):** A density we know how to sample from that loosely resembles $f(\theta)$.
  - **Majorizing Constant (M):** A constant such that $f(\theta) \le Mg(\theta)$ for all θ. Must exist for the method to work.
  - **Acceptance Probability (for a given θ^*):** The probability of accepting a proposed value $\theta^* \sim g(\theta)$, given by $\alpha(\theta^*) = \frac{f(\theta^*)}{Mg(\theta^*)}$. Note $0 \le \alpha(\theta^*) \le 1$.
  - **Overall Acceptance Rate:** The proportion of proposals that are accepted. Approximately $1/M$ if f and g are normalized densities.

- **Importance Sampling (Mentioned briefly):** Another Monte Carlo technique, often used for approximating expectations rather than generating an unweighted sample from the target. Weights are assigned to samples from the proposal distribution. Less focus in Hoff Ch 7 compared to Rejection Sampling.

Conceptual Overview / "The Big Picture"

Chapters 3-6 focused on models where Bayesian inference was mathematically tractable, largely due to conjugate priors yielding posterior distributions of known forms (Beta, Normal, Dirichlet, Inv-ChiSq, t). However, in many realistic and complex models, we encounter:

- Non-conjugate priors (chosen for better belief representation).
- Complex likelihoods (e.g., from hierarchical models, non-linear models).
- High-dimensional parameter spaces.

In these situations, deriving the posterior distribution $p(\theta|y)$ analytically, or calculating integrals involving it, becomes impossible or impractical. Chapter 7 introduces the fundamental computational solution: **Monte Carlo methods**. The core idea is to replace analytical integration with numerical approximation based on random sampling. If we can generate a large sample of parameter values $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ *directly from the posterior distribution* $p(\theta|y)$, we can approximate almost any feature of that posterior distribution using the properties of the sample (mean, variance, quantiles, histogram). This chapter explains *why* this works (Law of Large Numbers) and explores basic methods (like Rejection Sampling) for generating the required posterior samples when direct sampling isn't feasible.

Core Ideas & Summaries

7.1 The Need for Computation

- Reviews the components of Bayesian inference: prior $p(\theta)$, likelihood $p(y|\theta)$, posterior $p(\theta|y) \propto p(y|\theta)p(\theta)$.
- Highlights common computational challenges:
  - Finding the normalizing constant $p(y) = \int p(y|\theta)p(\theta)d\theta$.
  - Calculating posterior expectations $\mathbb{E}[g(\theta)|y] = \int g(\theta)p(\theta|y)d\theta$.
  - Finding posterior quantiles for credible intervals.
  - Calculating posterior probabilities $P(\theta \in A|y) = \int_A p(\theta|y)d\theta$.
  - Obtaining predictive distributions $p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$.
- Concludes that for many models, these require numerical approximation.

7.2 Monte Carlo Integration

- Introduces the main idea: approximate $\mathbb{E}[g(\theta)|y]$ with $\bar{g}_S = \frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)})$, where $\theta^{(s)} \sim p(\theta|y)$.
- **Justification:** The Law of Large Numbers ensures $\bar{g}_S \to \mathbb{E}[g(\theta)|y]$ as $S \to \infty$.
- **Applications:**
  - Posterior Mean: Use $g(\theta) = \theta$, estimate with $\bar{\theta}_S = \frac{1}{S}\sum \theta^{(s)}$.

  - Posterior Variance: Use $g(\theta) = (\theta - \mathbb{E}[\theta|y])^2$. Estimate with $\frac{1}{S}\sum(\theta^{(s)} - \bar{\theta}_S)^2$ (or use $(S-1)$ denominator for sample variance).
  - Posterior Probability $P(\theta \in A|y)$: Use $g(\theta) = I(\theta \in A)$ (indicator function). Estimate with $\frac{1}{S}\sum I(\theta^{(s)} \in A)$ = proportion of samples in A.
  - Posterior Quantiles: Use the sample quantiles of $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$. E.g., the 2.5% and 97.5% sample quantiles approximate a 95% credible interval.

- **Visualizing the Posterior:** Histograms or density estimates of the sample $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ provide an approximation of the posterior PDF $p(\theta|y)$.

7.3 Generating Posterior Samples

- The crucial step: How to get the sample $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ from $p(\theta|y)$?

- **Direct Sampling:** Possible if $p(\theta|y)$ is a known distribution family for which random number generators exist (e.g., Beta, Normal, Gamma). This applied to the conjugate models in Ch 3-6.

- **Rejection Sampling:** Useful when $p(\theta|y)$ is not standard, but we can evaluate its kernel $f(\theta) = p(y|\theta)p(\theta)$ and find an appropriate proposal density $g(\theta)$ and constant M such that $f(\theta) \leq Mg(\theta)$.
  - **Algorithm:** 1. Sample a candidate $\theta^* \sim g(\theta)$. 2. Generate $u \sim \text{Uniform}(0, 1)$. 3. Calculate the acceptance ratio $r = f(\theta^*)/(Mg(\theta^*))$. (Note: if f is unnormalized, this ratio is correct). 4. If $u \leq r$, accept $\theta^{(s)} = \theta^*$. 5. If $u > r$, reject θ^* and return to step 1.
  - **Intuition:** It samples uniformly under the curve $Mg(\theta)$ and keeps only points that are also under the (possibly unnormalized) target curve $f(\theta)$. The accepted points correctly represent the distribution proportional to $f(\theta)$.
  - **Efficiency:** The probability of accepting a proposal is $1/M$ (if f, g normalized). Requires $M = \sup_\theta[f(\theta)/g(\theta)]$. Finding a $g(\theta)$ that is easy to sample from, similar in shape to $f(\theta)$, and leads to a small M is key for efficiency. Can be very inefficient if f and g are dissimilar or in high dimensions.

- **Inverse CDF Other Methods:** Mentioned but rejection sampling is the main new general-purpose algorithm introduced here.

7.4 Monte Carlo Error

- Since S is finite, the estimate $\bar{g}_S$ is not exactly $\mathbb{E}[g(\theta)|y]$.

- The Central Limit Theorem (CLT) often applies to $\bar{g}_S$, suggesting $\bar{g}_S \dot{\sim} \mathcal{N}(\mathbb{E}[g(\theta)|y], \text{Var}(g(\theta))/S)$.

- The **Monte Carlo Standard Error (MCSE)** for the mean estimate $\bar{\theta}_S$ is $\text{SE}(\bar{\theta}_S) = \sqrt{\text{Var}(\theta)/S}$.

- We estimate this using the sample variance $\widehat{\text{Var}}(\theta) = \frac{1}{S-1}\sum(\theta^{(s)} - \bar{\theta}_S)^2$:

$$\widehat{\text{SE}}(\bar{\theta}_S) = \sqrt{\frac{\widehat{\text{Var}}(\theta)}{S}} = \frac{\text{sample standard deviation}}{\sqrt{S}}$$

- MCSE tells us about the precision of our Monte Carlo approximation due to finite sample size S. It should be small relative to the posterior standard deviation $\sqrt{\widehat{\text{Var}}(\theta)}$ for reliable inference. Decreases as S increases.

Connections to Previous/Future Topics

- **Chapters 3-6:** Provided examples where direct sampling from the posterior was often possible due to conjugacy. This chapter addresses what to do when it's not.

- **Chapters 8-9 (MCMC):** Rejection sampling can be difficult or impossible to apply efficiently, especially in high dimensions (finding good g, M is hard). Markov chain Monte Carlo (MCMC) methods, like Gibbs sampling (Ch 8) and Metropolis-Hastings (Ch 9), provide more powerful and generally applicable algorithms for generating samples that approximate the posterior distribution, even without knowing its normalizing constant. MCMC is the workhorse of modern Bayesian computation.

Common Pitfalls / Points of Confusion

- **Goal of MC:** Thinking the goal is just to calculate one number. The real goal is often to get a whole sample representing the posterior, from which *any* summary can be calculated.

- **Rejection Sampling Requirements:** Forgetting the need for M such that $f(\theta) \leq Mg(\theta)$, or difficulty in finding such an M or a suitable g.

- **Using Unnormalized Target:** In the acceptance ratio $f(\theta^*)/(Mg(\theta^*))$, it's crucial that $f(\theta)$ can be the unnormalized posterior kernel $p(y|\theta)p(\theta)$. The normalizing constants cancel.

- **Efficiency of Rejection Sampling:** Underestimating how quickly rejection sampling can become inefficient if the proposal g poorly matches the target f, or in multiple dimensions. Acceptance rates can become astronomically low.

- **MC Error vs. Posterior Variance:** Confusing the uncertainty due to the simulation (MCSE, decreases with S) with the inherent uncertainty about the parameter captured by the posterior variance (doesn't decrease with simulation effort S).

- **MC vs. MCMC:** Monte Carlo is a general term for using random numbers to solve problems. Rejection sampling is one type of MC algorithm. MCMC (next chapters) is a more advanced class of MC algorithms specifically designed to sample from complex probability distributions by constructing a Markov chain.

End of Week 7 Notes.

Bayesian Statistical Methods - Study Notes Week 8

Markov Chain Monte Carlo I: The Gibbs Sampler

Based on Hoff (2009), Chapter 8

Instructor: Sudeep Regmi

Chapters Covered

- Hoff Chapter 8: Markov Chain Monte Carlo
  - Section 8.1: Markov Chains
  - Section 8.2: The Gibbs Sampler
  - Section 8.3: Why Gibbs Sampling Works (Invariance)
  - Section 8.4: Example: Bivariate Normal Distribution

Learning Objectives

After engaging with this week's material, you should be able to:

- Define a Markov chain and the Markov property.
- Define the state space and transition kernel of a Markov chain.
- Define a stationary (or invariant) distribution for a Markov chain.
- Understand the basic principle of Markov chain Monte Carlo (MCMC): simulate a Markov chain whose stationary distribution is the target posterior distribution $p(\boldsymbol{\theta}|y)$.
- Explain why MCMC methods are particularly useful for sampling from complex, high-dimensional posterior distributions where direct sampling or rejection sampling are difficult.
- Describe the Gibbs sampling algorithm for drawing samples from a multivariate posterior distribution $p(\theta_1, \ldots, \theta_p|y)$.
- Define the full conditional distribution $p(\theta_j|\boldsymbol{\theta}_{-j}, y)$ and explain its role in Gibbs sampling.
- Derive the full conditional distributions for parameters in simple models (e.g., bivariate Normal).
- Understand that Gibbs sampling requires the ability to draw random samples from each full conditional distribution.
- Explain the concept of "burn-in" and why initial samples from an MCMC run are often discarded.
- Recognize that MCMC generates a dependent sequence of samples, unlike i.i.d. samples from direct or rejection sampling.

Key Concepts & Definitions

- **Markov Chain:** A sequence of random variables $\{\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots\}$ where the distribution of the next state $\boldsymbol{\theta}^{(s+1)}$ depends only on the current state $\boldsymbol{\theta}^{(s)}$, not on the past states $\{\boldsymbol{\theta}^{(0)}, \dots, \boldsymbol{\theta}^{(s-1)}\}$. This is the **Markov Property**.

- **State Space:** The set of all possible values that the states $\boldsymbol{\theta}^{(s)}$ can take (e.g., $\mathbb{R}^p$ for p continuous parameters).

- **Transition Kernel ($K(\boldsymbol{\theta}'|\boldsymbol{\theta})$):** Defines the probability (or density) of moving from state $\boldsymbol{\theta}$ to state $\boldsymbol{\theta}'$ in one step. $P(\boldsymbol{\theta}^{(s+1)} \in A | \boldsymbol{\theta}^{(s)} = \boldsymbol{\theta}) = \int_A K(\boldsymbol{\theta}'|\boldsymbol{\theta}) d\boldsymbol{\theta}'$.

- **Stationary Distribution ($\pi(\boldsymbol{\theta})$):** A probability distribution π such that if the current state $\boldsymbol{\theta}^{(s)}$ is drawn from π, then the next state $\boldsymbol{\theta}^{(s+1)}$ will also be drawn from π. Mathematically, $\pi(\boldsymbol{\theta}') = \int K(\boldsymbol{\theta}'|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$. Also called an **invariant distribution**.

- **Ergodicity:** A property of Markov chains (implying irreducibility and aperiodicity) which ensures that the distribution of $\boldsymbol{\theta}^{(s)}$ converges to the unique stationary distribution $\pi(\boldsymbol{\theta})$ as $s \to \infty$, regardless of the starting state $\boldsymbol{\theta}^{(0)}$. (Sufficient conditions often hold for MCMC algorithms).

- **Markov chain Monte Carlo (MCMC):** A class of algorithms for sampling from a target probability distribution $\pi(\boldsymbol{\theta})$ (typically a posterior $p(\boldsymbol{\theta}|y)$) by constructing an ergodic Markov chain whose stationary distribution is π. We simulate the chain and use the states visited (after convergence) as approximate samples from π.

- **Gibbs Sampling:** A specific MCMC algorithm for sampling from a multivariate distribution $\pi(\boldsymbol{\theta}) = p(\theta_1, \dots, \theta_p|y)$ by iteratively sampling each component θ_j from its full conditional distribution.

- **Full Conditional Distribution ($p(\theta_j|\boldsymbol{\theta}_{-j}, y)$):** The conditional distribution of a single parameter (or block of parameters) θ_j given all the other parameters $\boldsymbol{\theta}_{-j} = (\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_p)$ and the data y.

  - **Key Property:** $p(\theta_j|\boldsymbol{\theta}_{-j}, y) \propto p(\boldsymbol{\theta}|y)$. To find it, take the joint posterior $p(\boldsymbol{\theta}|y) \propto p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})$ and view it only as a function of θ_j, treating all other θ_i ($i \neq j$) and y as constants.

- **Burn-in Period:** The initial portion of an MCMC sequence $\{\boldsymbol{\theta}^{(0)}, \dots, \boldsymbol{\theta}^{(B)}\}$ that is discarded because the chain has not yet converged to its stationary distribution. The remaining samples $\{\boldsymbol{\theta}^{(B+1)}, \dots, \boldsymbol{\theta}^{(S)}\}$ are used for inference.

- **Dependent Samples:** Unlike direct or rejection sampling which produce independent draws, MCMC algorithms produce samples where $\boldsymbol{\theta}^{(s+1)}$ is correlated with $\boldsymbol{\theta}^{(s)}$. This dependence affects the efficiency of Monte Carlo estimates (discussed more in Ch 9/10).

Conceptual Overview / "The Big Picture"

Week 7 established the power of Monte Carlo methods for approximating posterior distributions but showed limitations of basic techniques like rejection sampling, especially in high dimensions. MCMC provides a more robust and general framework. The core idea is ingenious: instead of trying to sample *independently* from the complex target posterior $p(\boldsymbol{\theta}|y)$, we construct a Markov chain designed to *explore* the parameter space in such a way that the regions it visits most often correspond precisely to regions of high posterior probability. If the chain is run long enough (ergodicity), the distribution of states visited converges to the target posterior. Gibbs sampling is the first major MCMC algorithm we study. It's applicable when the multivariate parameter vector $\boldsymbol{\theta}$ can be broken down, and we can easily sample from the conditional distribution of each component (or block) θ_j given the current values of all other components $\boldsymbol{\theta}_{-j}$. By cyclically sampling from these simpler "full conditional" distributions, Gibbs constructs a Markov chain that converges to the desired joint posterior $p(\boldsymbol{\theta}|y)$. It cleverly turns a hard high-dimensional sampling problem into a sequence of potentially easier low-dimensional sampling problems.

Core Ideas & Summaries

8.1 Markov Chains

- Formalizes the idea of state dependence: $p(\boldsymbol{\theta}^{(s+1)}|\boldsymbol{\theta}^{(s)}, \dots, \boldsymbol{\theta}^{(0)}) = p(\boldsymbol{\theta}^{(s+1)}|\boldsymbol{\theta}^{(s)})$.
- Transition kernel $K(\boldsymbol{\theta}'|\boldsymbol{\theta})$ governs the one-step dynamics.
- Stationary distribution π is the equilibrium distribution; if $\boldsymbol{\theta}^{(s)} \sim \pi$, then $\boldsymbol{\theta}^{(s+1)} \sim \pi$.
- Ergodicity is crucial: guarantees convergence $\boldsymbol{\theta}^{(s)} \xrightarrow{d} \pi$ as $s \to \infty$. MCMC algorithms are designed to be ergodic with the target posterior as π.

8.2 The Gibbs Sampler

- **Goal:** Sample from $p(\boldsymbol{\theta}|y)$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$.
- **Requirement:** Must be able to sample from all full conditional distributions $p(\theta_j|\boldsymbol{\theta}_{-j}, y)$ for $j = 1, \dots, p$.
- **Algorithm:**
  1. Choose a starting value $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \dots, \theta_p^{(0)})$.
  2. For iteration $s = 1, \dots, S$: a. Sample $\theta_1^{(s)} \sim p(\theta_1|\theta_2^{(s-1)}, \theta_3^{(s-1)}, \dots, \theta_p^{(s-1)}, y)$. b. Sample $\theta_2^{(s)} \sim p(\theta_2|\theta_1^{(s)}, \theta_3^{(s-1)}, \dots, \theta_p^{(s-1)}, y)$. c. Sample $\theta_3^{(s)} \sim p(\theta_3|\theta_1^{(s)}, \theta_2^{(s)}, \dots, \theta_p^{(s-1)}, y)$. d. ... e. Sample $\theta_p^{(s)} \sim p(\theta_p|\theta_1^{(s)}, \theta_2^{(s)}, \dots, \theta_{p-1}^{(s)}, y)$.
  3. The sequence $\{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(S)}\}$ contains the samples.
- **Key Aspects:**
  - Uses the most recently sampled values when conditioning.
  - Order of sampling components usually doesn't matter for convergence (though can affect efficiency).
  - Components θ_j can be blocks of parameters instead of singletons.
- **Finding Full Conditionals:** Use $p(\theta_j|\boldsymbol{\theta}_{-j}, y) \propto p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Treat everything except θ_j as fixed constants and identify the resulting distribution for θ_j. Conjugacy often helps here (e.g., if prior is Normal and likelihood involves θ_j normally, conditional might be Normal).

8.3 Why Gibbs Sampling Works

- The Gibbs update steps implicitly define a transition kernel $K(\boldsymbol{\theta}'|\boldsymbol{\theta})$.
- This kernel can be shown to leave the target joint posterior distribution $p(\boldsymbol{\theta}|y)$ invariant. If $\boldsymbol{\theta}^{(s-1)} \sim p(\boldsymbol{\theta}|y)$, then after one full Gibbs cycle, $\boldsymbol{\theta}^{(s)} \sim p(\boldsymbol{\theta}|y)$.
- Under mild conditions (positivity of the posterior density), the chain is ergodic, guaranteeing convergence to $p(\boldsymbol{\theta}|y)$.

8.4 Example: Bivariate Normal Distribution

- Suppose the target is $p(\theta_1, \theta_2|y)$, which is a Bivariate Normal distribution with mean (μ_1, μ_2), variances (σ_1^2, σ_2^2), and correlation ρ.
- From properties of the MVN, the full conditionals are univariate Normal:
  - $\theta_1|\theta_2, y \sim \mathcal{N}(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(\theta_2 - \mu_2), \sigma_1^2(1 - \rho^2))$
  - $\theta_2|\theta_1, y \sim \mathcal{N}(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(\theta_1 - \mu_1), \sigma_2^2(1 - \rho^2))$
- **Gibbs Sampler:** 1. Start with $(\theta_1^{(0)}, \theta_2^{(0)})$. 2. For $s = 1, \ldots, S$: a. Sample $\theta_1^{(s)}$ from the Normal distribution for $\theta_1|\theta_2^{(s-1)}, y$. b. Sample $\theta_2^{(s)}$ from the Normal distribution for $\theta_2|\theta_1^{(s)}, y$.
- The sequence $(\theta_1^{(s)}, \theta_2^{(s)})$ converges in distribution to the target BVN. Visualizing the path shows steps parallel to the axes exploring the target density contours.

Using MCMC Output

- Discard initial burn-in samples (e.g., first B iterations).
- Use the remaining samples $\{\boldsymbol{\theta}^{(B+1)}, \ldots, \boldsymbol{\theta}^{(S)}\}$ for Monte Carlo integration (Chapter 7):
  - Estimate posterior means, variances, probabilities, quantiles (CIs).
  - Create histograms/density plots of marginal posteriors (e.g., plot just the $\theta_1^{(s)}$ values).
- Remember samples are correlated, which affects MCSE (addressed in Ch 9/10).

Connections to Previous/Future Topics

- **Ch 7 (Monte Carlo):** Gibbs sampling is a method to generate the posterior samples needed for Monte Carlo integration when simpler methods fail.
- **Ch 5, 6 (Normal, MVN, Multinomial models):** Deriving full conditionals often relies on the structure of these models and conjugate relationships (e.g., Normal conditional from Normal-InvChiSq joint posterior).
- **Ch 9 (Metropolis-Hastings):** Provides a more general MCMC algorithm needed when sampling directly from full conditionals is not possible. Gibbs can be seen as a special case of Metropolis-Hastings.
- **Ch 10 (originally Ch 9 in text - Diagnostics):** Deals with crucial practical issues: How long should burn-in be? Has the chain converged? How many samples S are needed? How does correlation affect estimates?

Common Pitfalls / Points of Confusion

- **Deriving Full Conditionals:** Making errors by not correctly isolating terms involving only the target parameter θ_j or by incorrectly identifying the resulting distribution family. Remember $p(\theta_j|\boldsymbol{\theta}_{-j}, y) \propto p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})$.
- **Confusing Conditionals and Marginals:** The full conditional $p(\theta_j|\boldsymbol{\theta}_{-j}, y)$ is NOT the same as the marginal posterior $p(\theta_j|y)$. Gibbs uses conditionals to eventually sample from the joint, from which marginals can be estimated.

- **Need for Sampling Conditionals:** Forgetting that you must be able to *draw random samples* from each $p(\theta_j|\boldsymbol{\theta}_{-j}, y)$, not just write down its formula.
- **Burn-in:** Not understanding its purpose (letting the chain forget its starting point and reach equilibrium) or how to choose its length (requires diagnostics, Ch 9/10).
- **Dependence:** Treating MCMC output as i.i.d. samples. The correlation means that S dependent samples usually contain less information than S independent samples.
- **Gibbs vs. MCMC:** Thinking Gibbs is the only MCMC method. It's powerful when applicable, but Metropolis-Hastings is more general.

End of Week 8 Notes.

Bayesian Statistical Methods - Study Notes Week 9

Markov Chain Monte Carlo II: The Metropolis-Hastings Algorithm

Based on Hoff (2009), Chapter 9 (Sections 9.1-9.3)

Instructor: Sudeep Regmi

Chapters Covered

- Hoff Chapter 9: The Metropolis-Hastings Algorithm
  - Section 9.1: Why Gibbs Sampling Isn't Always Enough
  - Section 9.2: The Metropolis-Hastings Algorithm
  - Section 9.3: Different Types of MH Proposals (Random Walk, Independence)

Note: Section 9.4 on Diagnostics might be covered next week.

Learning Objectives

After engaging with this week's material, you should be able to:

- Explain situations where the Gibbs sampler cannot be easily applied (intractable full conditionals).
- Describe the general logic of the Metropolis-Hastings (MH) algorithm: propose a move, then accept or reject it based on a calculated probability.
- Define the target distribution $\pi(\theta)$ (usually the posterior $p(\theta|y)$ or its kernel) and the proposal distribution $q(\theta'|\theta^{(s)})$.
- Write down and interpret the Metropolis-Hastings acceptance probability $\alpha(\theta'|\theta^{(s)})$.
- Explain the role of the detailed balance (reversibility) condition in ensuring that the MH algorithm has the target distribution as its stationary distribution.
- Implement the MH algorithm steps.
- Describe and differentiate between common types of proposal distributions:
  - Random Walk Metropolis-Hastings (symmetric proposal).
  - Independence Sampler (proposal independent of current state).
- Understand the concept of tuning a proposal distribution (e.g., the variance of a random walk) to achieve reasonable acceptance rates and efficient exploration of the state space.
- Recognize that Gibbs sampling is a special case of the Metropolis-Hastings algorithm.

Key Concepts & Definitions

- **Target Distribution ($\pi(\theta)$):** The distribution we want to sample from, typically the posterior $p(\theta|y)$. We often only need to know it up to a normalizing constant, i.e., we work with $f(\theta) \propto \pi(\theta)$, where $f(\theta) = p(y|\theta)p(\theta)$.
- **Proposal Distribution ($q(\theta'|\theta)$):** A distribution used to generate candidate values θ' based on the current state θ. The user chooses this distribution.
- **Metropolis-Hastings Algorithm:** An MCMC algorithm that generates a sequence of samples $\{\theta^{(0)}, \theta^{(1)}, \dots\}$ which converges to the target distribution $\pi(\theta)$.
- **Acceptance Probability ($\alpha(\theta'|\theta)$):** The probability of accepting a proposed move from state θ to state θ'. Defined as:

  $$\alpha(\theta'|\theta) = \min\left(1, \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}\right)$$

  Since we usually work with $f(\theta) \propto \pi(\theta)$, the ratio $\pi(\theta')/\pi(\theta)$ can be replaced by $f(\theta')/f(\theta)$:

  $$\alpha(\theta'|\theta) = \min\left(1, \frac{f(\theta')q(\theta|\theta')}{f(\theta)q(\theta'|\theta)}\right)$$

  This ratio depends on the target density at the proposed and current points, and the proposal densities of moving forward ($\theta \to \theta'$) and backward ($\theta' \to \theta$).
- **Detailed Balance (Reversibility):** A sufficient condition for a Markov chain kernel K to have π as its stationary distribution. It requires the rate of flow from θ to θ' to equal the rate of flow from θ' to θ when the chain is in equilibrium:

  $$\pi(\theta)K(\theta'|\theta) = \pi(\theta')K(\theta|\theta')$$

  The MH acceptance probability is constructed precisely to satisfy detailed balance.
- **Metropolis Algorithm:** A special case of MH where the proposal distribution is symmetric, i.e., $q(\theta'|\theta) = q(\theta|\theta')$. The acceptance probability simplifies to:

  $$\alpha(\theta'|\theta) = \min\left(1, \frac{\pi(\theta')}{\pi(\theta)}\right) = \min\left(1, \frac{f(\theta')}{f(\theta)}\right)$$
- **Random Walk Metropolis-Hastings:** Uses a symmetric proposal centered around the current state. E.g., propose $\theta' = \theta^{(s)} + \epsilon$, where ϵ is drawn from a symmetric distribution like $\mathcal{N}(0, \sigma^2_{prop})$. Uses the simpler Metropolis acceptance ratio. Requires tuning the proposal step size (e.g., σ^2_{prop}).
- **Independence Sampler:** Uses a proposal distribution $q(\theta'|\theta^{(s)}) = q(\theta')$ that does *not* depend on the current state $\theta^{(s)}$. The acceptance probability becomes:

  $$\alpha(\theta'|\theta^{(s)}) = \min\left(1, \frac{f(\theta')q(\theta^{(s)})}{f(\theta^{(s)})q(\theta')}\right)$$

  Requires the proposal $q(\theta')$ to have heavier tails than the target $\pi(\theta)$. Can be efficient if q is a good approximation to π, but poor if not. Related conceptually to rejection sampling.
- **Tuning:** The process of adjusting parameters of the proposal distribution q (e.g., variance of random walk, parameters of independence proposal) to optimize the MCMC sampler's performance (balancing acceptance rate and exploration). Often aims for acceptance rates around 20-50%.

Conceptual Overview / "The Big Picture"

Last week, we learned about Gibbs sampling, a powerful MCMC technique. However, its application hinges on our ability to sample from all full conditional distributions $p(\theta_j|\boldsymbol{\theta}_{-j}, y)$. What if one or more of these conditionals are non-standard distributions from which we don't know how to easily draw samples? The Metropolis-Hastings (MH) algorithm provides a more general solution. It allows us to construct a Markov chain converging to the target posterior $p(\theta|y)$ even without being able to sample directly from conditionals. The MH strategy is based on a "propose and verify" idea: 1. From the current state $\theta^{(s)}$, propose a candidate move to a new state θ' using a proposal distribution $q(\theta'|\theta^{(s)})$. 2. Evaluate whether this proposed move is "good" by comparing the target density at the proposed point $\pi(\theta')$ versus the current point $\pi(\theta^{(s)})$, while also accounting for how likely the proposal mechanism was to suggest the move ($\theta^{(s)} \to \theta'$) versus the reverse move ($\theta' \to \theta^{(s)}$). 3. Accept the move to $\theta^{(s+1)} = \theta'$ with probability $\alpha(\theta'|\theta^{(s)})$. If rejected, stay put: $\theta^{(s+1)} = \theta^{(s)}$. The clever construction of the acceptance probability α ensures the resulting Markov chain satisfies detailed balance with respect to $\pi(\theta)$, guaranteeing that $\pi(\theta)$ is the stationary distribution. MH is incredibly flexible because the user has freedom in choosing the proposal distribution q.

Core Ideas & Summaries

9.1 Motivation: Beyond Gibbs

- Gibbs sampling requires sampling from $p(\theta_j|\boldsymbol{\theta}_{-j}, y)$.
- These conditionals might not be recognizable standard distributions, especially with non-conjugate priors or complex likelihoods.
- MH provides an alternative MCMC approach that only requires evaluating the posterior density (up to a constant), not sampling from complex conditionals.

9.2 The Metropolis-Hastings Algorithm

- **Goal:** Sample from target $\pi(\theta)$ (or $f(\theta) \propto \pi(\theta)$).
- **Ingredients:** Target kernel $f(\theta)$, Proposal distribution $q(\theta'|\theta)$.
- **Algorithm:** 1. Choose starting value $\theta^{(0)}$. 2. For $s = 0, \ldots, S-1$: a. Propose $\theta' \sim q(\theta'|\theta^{(s)})$. b. Calculate acceptance probability:

$$\alpha(\theta'|\theta^{(s)}) = \min\left(1, \frac{f(\theta')q(\theta^{(s)}|\theta')}{f(\theta^{(s)})q(\theta'|\theta^{(s)})}\right)$$

  c. Generate $u \sim \text{Uniform}(0, 1)$. d. If $u \leq \alpha(\theta'|\theta^{(s)})$, accept: $\theta^{(s+1)} = \theta'$. e. Else (if $u > \alpha$), reject: $\theta^{(s+1)} = \theta^{(s)}$.

- **Why it works (Detailed Balance):** The construction ensures $\pi(\theta)K(\theta'|\theta) = \pi(\theta')K(\theta|\theta')$, where K is the MH transition kernel. If the proposal density q allows the chain to reach any part of the state space (irreducibility) and avoid getting stuck in cycles (aperiodicity), the chain will converge to π.
- **Handling the Normalizing Constant:** Note that the ratio $f(\theta')/f(\theta)$ is used, so the unknown normalizing constant of the posterior cancels out. This is a major advantage!

9.3 Proposal Distributions

- The choice of q is crucial for efficiency.
- **Random Walk MH:**
  - Proposal: $\theta' = \theta^{(s)} + \epsilon$, where $\epsilon \sim g(\epsilon)$ and g is symmetric ($g(\epsilon) = g(-\epsilon)$), e.g., $\epsilon \sim \mathcal{N}(0, \sigma^2_{prop})$.
  - Symmetry means $q(\theta'|\theta^{(s)}) = g(\theta' - \theta^{(s)}) = g(\theta^{(s)} - \theta') = q(\theta^{(s)}|\theta')$.
  - Acceptance probability simplifies (Metropolis): $\alpha(\theta'|\theta^{(s)}) = \min(1, f(\theta')/f(\theta^{(s)}))$.
  - **Tuning σ^2_{prop}:**
    - * If too small: High acceptance rate, but chain explores slowly (high autocorrelation).
    - * If too large: Proposes moves to low-density regions, high rejection rate, chain gets stuck.
    - * Rule of thumb: Aim for acceptance rates around 25-45% for moderate dimensions.
- **Independence Sampler:**
  - Proposal: $\theta' \sim q(\theta')$, independent of $\theta^{(s)}$.
  - Acceptance probability: $\alpha(\theta'|\theta^{(s)}) = \min\left(1, \frac{f(\theta')q(\theta^{(s)})}{f(\theta^{(s)})q(\theta')}\right)$. Sometimes written using weights $w(\theta) = f(\theta)/q(\theta)$ as $\min(1, w(\theta')/w(\theta^{(s)}))$.
  - Requires $q(\theta)$ to have heavier tails than $\pi(\theta)$ (i.e., support$(f) \subseteq$ support(q)).
  - Works well if q is a good approximation of π. Can be very poor otherwise.
- **Other proposals:** Many sophisticated proposals exist (e.g., Metropolis-Adjusted Langevin Algorithm - MALA, Hamiltonian Monte Carlo - HMC).

Relationship to Gibbs

- Gibbs sampling can be viewed as a sequence of MH steps. To update θ_j, the proposal distribution is the full conditional itself: $q(\theta'_j|\boldsymbol{\theta}^{(s)}_{-j}) = p(\theta'_j|\boldsymbol{\theta}^{(s)}_{-j}, y)$.
- Plugging this into the MH acceptance ratio for updating only θ_j (keeping $\boldsymbol{\theta}_{-j}$ fixed), the ratio becomes exactly 1.
- So, Gibbs always accepts the proposal drawn from the full conditional. MH generalizes this by allowing proposals from other distributions and using the acceptance probability to correct for the discrepancy.

Connections to Previous/Future Topics

- **Ch 8 (Gibbs):** MH provides a necessary tool when Gibbs is infeasible. MH can be used *within* a Gibbs scan (Metropolis-within-Gibbs) if one specific full conditional is hard to sample from. Gibbs is a special case of MH.
- **Ch 7 (Rejection Sampling):** Both MH and rejection sampling use proposals, but the mechanisms differ. Rejection sampling produces i.i.d samples but requires knowing M. MH produces dependent samples but avoids needing M and works directly with density ratios.

- **Ch 10 (originally Ch 9.4 - Diagnostics):** Practical use of MH requires assessing convergence and efficiency. How to choose burn-in, evaluate mixing, estimate MCSE for dependent samples? These are crucial next steps.

Common Pitfalls / Points of Confusion

- **Calculating Acceptance Ratio:** Forgetting the proposal ratio term $q(\theta^{(s)}|\theta')/q(\theta'|\theta^{(s)})$ when the proposal is not symmetric. Getting the forward and backward proposal densities mixed up.

- **Target vs. Proposal:** Confusing the target density $f(\theta)$ (which determines the desired equilibrium) with the proposal density $q(\theta'|\theta)$ (which determines how moves are suggested).

- **Tuning:** Spending too little or too much effort tuning. Realizing tuning is often problem-specific and may require experimentation. Setting proposal variance too high/low in Random Walk MH.

- **Interpretation of Rejection:** When a proposal is rejected, the chain *stays* at the current state $\theta^{(s+1)} = \theta^{(s)}$. This is necessary for the chain to spend the correct amount of time in high-density regions. Don't mistakenly discard rejected steps entirely.

- **Independence Sampler Requirements:** Forgetting the need for the proposal q to have support covering the target π (or f), ideally with heavier tails.

- **Log Scale:** Calculations involving products of densities are often performed on the log scale for numerical stability: $\log(\alpha) = \min(0, \log f(\theta') - \log f(\theta^{(s)}) + \log q(\theta^{(s)}|\theta') - \log q(\theta'|\theta^{(s)}))$. Then compare $\log(u)$ to $\log(\alpha)$.

End of Week 9 Notes.

Bayesian Statistical Methods - Study Notes Week 10

MCMC Diagnostics & Output Analysis

Based on Hoff (2009), Chapter 9 (Section 9.4) & Related Concepts

Instructor: Sudeep Regmi

Chapters Covered

- Hoff Chapter 9: The Metropolis-Hastings Algorithm
  - Section 9.4: Assessing Sampler Performance (and related practical issues)
- Concepts often bundled with MCMC practicalities (may draw slightly from ideas leading into Ch 10).

Learning Objectives

After engaging with this week's material, you should be able to:

- Explain why assessing MCMC convergence and efficiency is crucial before using the output for inference.
- Use trace plots to visually assess whether a Markov chain appears to have reached stationarity (convergence).
- Understand the utility of running multiple MCMC chains from dispersed starting points.
- Describe the concept of burn-in and apply visual methods (trace plots) to determine an appropriate burn-in length.
- Explain autocorrelation in MCMC samples and its implications for sampler efficiency.
- Interpret Autocorrelation Function (ACF) plots for MCMC output.
- Define thinning and understand its purpose and limitations (often not necessary for parameter estimation).
- Define Effective Sample Size (ESS) and explain its interpretation as the number of equivalent independent samples.
- Understand how ESS relates to autocorrelation.
- Calculate the Monte Carlo Standard Error (MCSE) for posterior mean estimates using ESS.
- Explain why reporting MCSE is important for evaluating the precision of Monte Carlo estimates.
- Summarize MCMC output appropriately (posterior means, medians, credible intervals) using post-burn-in samples, accounting for MCSE.

Key Concepts & Definitions

- **Convergence (Practical):** The state where an MCMC chain has "forgotten" its starting value and the distribution of its states is close to the target stationary distribution (the posterior $p(\theta|y)$). We never know *exactly* when convergence occurs, but diagnostics help assess if it likely has.
- **Stationarity:** The theoretical property where the distribution of $\theta^{(s)}$ remains the same for all subsequent steps s. MCMC algorithms are designed to have the posterior as their stationary distribution.
- **Trace Plot:** A plot of the sampled parameter values ($\theta^{(s)}$) against the iteration number (s). A primary tool for visual convergence assessment.
- **Multiple Chains:** Running two or more MCMC chains simultaneously, starting from widely dispersed initial values $\theta^{(0)}$. Convergence is more confidently assessed if all chains appear to converge to the same distribution.
- **Burn-in (B):** The initial number of iterations discarded from an MCMC run to mitigate the influence of the starting value and allow the chain time to reach approximate stationarity.
- **Autocorrelation:** The correlation between samples in the MCMC sequence at different lags. $\text{Corr}(\theta^{(s)}, \theta^{(s+k)})$ for lag k. High autocorrelation means the chain mixes slowly and explores the parameter space inefficiently.
- **Autocorrelation Function (ACF) Plot:** A plot of the autocorrelation against the lag k. Used to visualize how quickly the correlation between samples decreases as the distance between them increases. Ideally, ACF drops quickly towards zero.
- **Thinning:** Subsampling the MCMC chain by keeping only every k-th sample (e.g., keep samples $B + k, B + 2k, B + 3k, \dots$). Reduces autocorrelation in the stored output and saves memory, but discards information and is often unnecessary for estimating posterior means or standard deviations if using ESS correctly.
- **Effective Sample Size (ESS):** An estimate of the number of independent samples that would provide the same amount of information about the posterior mean as the $S - B$ autocorrelated samples actually obtained. Calculated as:

  $$\text{ESS} = \frac{S - B}{1 + 2\sum_{k=1}^{\infty} \rho_k}$$

  where ρ_k is the autocorrelation at lag k. In practice, the sum is truncated. Packages like ‘coda‘ estimate ESS automatically. ESS $\leq S - B$. Low ESS indicates high autocorrelation and inefficient sampling.
- **Monte Carlo Standard Error (MCSE):** The standard deviation of a Monte Carlo estimate (e.g., posterior mean $\bar{\theta}$) considered as a random variable depending on the specific MCMC run. For the posterior mean estimated from $N = S - B$ post-burn-in samples:

  $$\text{MCSE}(\bar{\theta}) = \sqrt{\frac{\widehat{\text{Var}}(\theta)}{\text{ESS}}}$$

  where $\widehat{\text{Var}}(\theta)$ is the sample variance of the N post-burn-in samples. This correctly accounts for autocorrelation via ESS. A small MCSE (relative to the posterior standard deviation $\sqrt{\widehat{\text{Var}}(\theta)}$) indicates a reliable Monte Carlo estimate.

Conceptual Overview / "The Big Picture"

We've learned how to build MCMC samplers (Gibbs, MH) that theoretically converge to the target posterior distribution. However, MCMC provides only an *approximation* based on a finite simulation run. Before trusting the results (posterior means, credible intervals, etc.) derived from the MCMC output, we **must** assess the quality of that approximation. This involves two key questions: 1. **Convergence:** Has the simulation run long enough for the chain to plausibly reach its stationary distribution (the posterior) and forget its arbitrary starting point? We need to determine a suitable burn-in period. 2. **Efficiency/Accuracy:** How well do the post-burn-in samples represent the posterior? Due to autocorrelation, N dependent MCMC samples typically contain less information than N independent samples. We need to quantify this efficiency (using ESS) to understand the precision of our Monte Carlo estimates (via MCSE). Diagnostics are therefore not optional extras but an essential part of any MCMC-based Bayesian analysis. They provide evidence (though not absolute proof) that the MCMC output is reliable enough for inference.

Core Ideas & Summaries

Assessing Convergence

- **Goal:** Determine if the chain has reached a stable state representative of the stationary distribution.
- **Trace Plots (Visual):** Plot $\theta^{(s)}$ vs s for each parameter (or key parameters).
  - Look for: Rapid up-and-down movement around a stable mean, covering the parameter space quickly (good "mixing"). The plot should look like "white noise" without long-term trends or drifts.
  - Bad signs: Obvious upward/downward trends, slow drifts, getting stuck in one region for long periods.
- **Multiple Chains:** Run ≥ 2 chains starting from different, overdispersed points.
  - Plot trace plots for the same parameter from all chains overlaid.
  - Look for: All chains converging to overlap in the same distribution/region. If chains remain separate, convergence is highly suspect.
- **Burn-in Determination:** Based on trace plots (especially from multiple chains), identify the point B after which all chains appear to be stable and mixing around the same distribution. Discard iterations $1, \ldots, B$. Be conservative (choose a larger B if unsure).
- **Formal Tests (e.g., Gelman-Rubin $\hat{R}$ statistic):** Quantitative methods often compare within-chain variance to between-chain variance. Values close to 1 suggest convergence. (Hoff mentions briefly; often covered in more detail elsewhere or via software like ‘coda‘).

Assessing Efficiency and Accuracy

- **Goal:** Understand how much information is in the post-burn-in samples and quantify the precision of estimates.
- **Autocorrelation (ACF Plots):** Plot $\text{Corr}(\theta^{(s)}, \theta^{(s+k)})$ vs lag k.
  - Look for: ACF dropping rapidly to near zero.

  - Bad signs: ACF remains high for many lags, indicating strong dependency and slow mixing. The chain needs to run much longer to get useful information.

- **Thinning:** Keeping every k-th sample reduces ACF in the *stored* samples and saves disk space. However:
  - Information is discarded. Estimating the mean from N/k thinned samples usually gives a less precise estimate (higher true MCSE) than using all N samples.
  - Often unnecessary if using software that calculates ESS correctly from unthinned chains.
  - Can be useful for algorithms that rely on approximate independence (rare) or simply to make trace/ACF plots clearer if autocorrelation is extremely high.

- **Effective Sample Size (ESS):** Measures the information content, accounting for autocorrelation.
  - High ESS (approaching $N = S - B$) $\implies$ low autocorrelation, efficient sampler.
  - Low ESS ($\ll N$) $\implies$ high autocorrelation, inefficient sampler. Need a longer run (S) to achieve desired precision.
  - Rule of thumb: Need ESS of at least several hundred, preferably thousands, for stable estimates of means and quantiles.

- **Monte Carlo Standard Error (MCSE):** Quantifies accuracy of posterior mean estimate.
  - $\text{MCSE} = \sqrt{\widehat{\text{Var}}(\theta)/\text{ESS}}$. Directly uses ESS.
  - Should be small relative to the posterior standard deviation $\sqrt{\widehat{\text{Var}}(\theta)}$. If MCSE is large, the Monte Carlo estimate $\bar{\theta}$ is unreliable. Solution: Run the MCMC longer to increase ESS.

Summarizing MCMC Output

- Use samples $\{\theta^{(B+1)}, \ldots, \theta^{(S)}\}$ (post-burn-in, possibly thinned if desired, but generally use unthinned). Let $N = S - B$.
- **Posterior Mean:** $\bar{\theta} = \frac{1}{N} \sum_{s=B+1}^{S} \theta^{(s)}$. Report with $\text{MCSE}(\bar{\theta})$.
- **Posterior Median:** Sample median of $\{\theta^{(B+1)}, \ldots, \theta^{(S)}\}$.
- **Posterior Standard Deviation:** Sample standard deviation of $\{\theta^{(B+1)}, \ldots, \theta^{(S)}\}$.
- **Credible Intervals:** Sample quantiles (e.g., 2.5% and 97.5% quantiles for a 95% CI) of $\{\theta^{(B+1)}, \ldots, \theta^{(S)}\}$.
- **Posterior Densities:** Histograms or kernel density estimates of the samples.

Connections to Previous/Future Topics

- **Ch 7-9 (MC, Gibbs, MH):** Diagnostics are essential for validating the output from the simulation methods developed in these chapters.
- **Ch 10+ (Model Checking, Comparison, Specific Models):** All subsequent Bayesian inference using MCMC relies on having reliable posterior samples. Poor MCMC performance invalidates subsequent modeling conclusions. MCSE informs the precision of quantities like posterior predictive checks or DIC values.

Common Pitfalls / Points of Confusion

- **Assuming Convergence Too Soon:** Relying on short runs or single chains. Failing to use multiple chains from dispersed starting values.
- **Ignoring Burn-in:** Using initial samples that are heavily influenced by the starting value.
- **Misinterpreting Trace Plots:** Thinking a flat trace plot means low variance (it means the chain is stuck), or thinking slow drift is acceptable.
- **Misinterpreting ACF:** Thinking high ACF is okay. It signals inefficiency.
- **Unnecessary/Harmful Thinning:** Thinning excessively when ESS could be calculated on the full chain, thereby reducing precision. Thinking thinning *removes* the need to worry about autocorrelation (it only hides it in the stored output).
- **Ignoring ESS:** Calculating standard errors as $\sqrt{\widehat{\text{Var}}(\theta)/N}$ instead of $\sqrt{\widehat{\text{Var}}(\theta)/\text{ESS}}$, leading to underestimated uncertainty.
- **Not Reporting MCSE:** Presenting posterior means without indicating the simulation uncertainty (MCSE).
- **Confusing MCSE and Posterior SD:** MCSE relates to the simulation accuracy; posterior SD relates to the actual uncertainty about the parameter θ given the data. MCSE should be much smaller than posterior SD.

End of Week 10 Notes.

Bayesian Statistical Methods - Study Notes Week 11

Model Checking and Comparison

Based on Hoff (2009), Chapter 10

Instructor: Sudeep Regmi

Chapters Covered

- Hoff Chapter 10: Model Checking and Selection

Learning Objectives

After engaging with this week's material, you should be able to:

- Explain the importance of model checking in the Bayesian workflow.
- Define the posterior predictive distribution and explain how to simulate replicated datasets (y^{rep}) from it using MCMC output.
- Implement and interpret graphical posterior predictive checks (e.g., comparing histograms, densities, or ECDFs of y_{obs} and y^{rep}).
- Define a test statistic $T(y)$ and explain its role as a discrepancy measure in posterior predictive checks.
- Implement and interpret numerical posterior predictive checks by comparing $T(y_{\text{obs}})$ to the distribution of $T(y^{\text{rep}})$.
- Calculate and interpret a Bayesian p-value (p_B) derived from a test statistic.
- Understand the concept of sensitivity analysis and how to assess the impact of prior distributions or likelihood choices on posterior inferences.
- Understand the basic idea behind model comparison criteria like the Deviance Information Criterion (DIC).
- Interpret the components of DIC: posterior mean deviance ($\bar{D}$) as a measure of fit, and effective number of parameters (p_D) as a measure of complexity.
- Recognize the limitations of model checking techniques (e.g., dependence on chosen test statistics, interpretation of p-values).

Key Concepts & Definitions

- **Model Checking:** The process of evaluating whether a fitted statistical model adequately represents the observed data and captures its key features. It assesses the model's descriptive accuracy.

- **Posterior Predictive Distribution ($p(\tilde{y}|y_{\text{obs}})$):** The distribution of future (or replicated) data $\tilde{y}$ given the observed data y_{obs}. It averages the likelihood $p(\tilde{y}|\theta)$ over the posterior distribution $p(\theta|y_{\text{obs}})$:

$$p(\tilde{y}|y_{\text{obs}}) = \int p(\tilde{y}|\theta)p(\theta|y_{\text{obs}})d\theta$$

- **Replicated Data (y^{rep}):** A dataset simulated from the posterior predictive distribution. Represents data that *could* have been generated under the fitted model. Simulation process: 1. Draw $\theta^{(s)} \sim p(\theta|y_{\text{obs}})$ (from MCMC output). 2. Draw $y^{\text{rep}(s)} \sim p(y|\theta^{(s)})$ (from the likelihood/sampling model). The collection $\{y^{\text{rep}(s)} : s = 1, \ldots, S\}$ approximates the posterior predictive distribution.

- **Posterior Predictive Check (PPC):** A method for model checking that involves comparing the observed data y_{obs} to replicated datasets y^{rep} drawn from the posterior predictive distribution. Any systematic differences suggest model misfit.

- **Test Statistic / Discrepancy Measure ($T(y)$):** A function of the data (observed or replicated) chosen to probe a specific aspect of model fit (e.g., mean, variance, minimum, maximum, proportion of zeros).

- **Bayesian p-value (p_B):** The posterior predictive probability that a chosen test statistic $T(y^{\text{rep}})$ calculated on replicated data is more extreme (usually greater than or equal to) than the statistic calculated on the observed data $T(y_{\text{obs}})$:

$$p_B = P(T(y^{\text{rep}}) \geq T(y_{\text{obs}})|y_{\text{obs}})$$

  Estimated from MCMC output as the proportion of replicated datasets where $T(y^{\text{rep}(s)}) \geq T(y_{\text{obs}})$. Values close to 0 or 1 indicate potential model misfit *with respect to the chosen statistic T*.

- **Sensitivity Analysis:** Assessing how posterior inferences (e.g., means, credible intervals) change when assumptions of the analysis are varied, particularly the choice of prior distribution or likelihood specification.

- **Deviance:** A measure of model misfit, often defined as $D(\theta) = -2 \log p(y_{\text{obs}}|\theta)$. Lower deviance generally indicates better fit.

- **Deviance Information Criterion (DIC):** A criterion for Bayesian model comparison that balances model fit and complexity, analogous to AIC.

$$\text{DIC} = \bar{D} + p_D$$

  where

  - $\bar{D} = \mathbb{E}_{\theta|y_{\text{obs}}}[D(\theta)]$: Posterior mean deviance (measures fit). Estimated by averaging $D(\theta^{(s)})$ over MCMC samples.
  - $p_D = \bar{D} - D(\bar{\theta})$: Effective number of parameters (measures complexity). Calculated using the posterior mean $\bar{\theta}$ of θ. (Other definitions of p_D exist, e.g., $\frac{1}{2}\text{Var}_{\theta|y_{\text{obs}}}[D(\theta)]$).

  Lower DIC values suggest a better trade-off between fit and complexity. Useful for comparing models fit to the same data.

Conceptual Overview / ”The Big Picture”

After fitting a Bayesian model and obtaining posterior distributions (often via MCMC), a crucial step is to assess how well the model actually fits the data. A model might converge nicely but still be a poor representation of reality. Model checking addresses this ”goodness-of-fit” question. The central Bayesian approach is the **Posterior Predictive Check (PPC)**. The logic is intuitive: if the model is a good fit, then data generated *from the fitted model* should look similar to the data we actually observed. We operationalize this by simulating many ”replicated” datasets (y^{rep}) from the posterior predictive distribution (which incorporates uncertainty about the parameters θ). We then compare these y^{rep} datasets to our real dataset y_{obs}. Comparisons can be graphical (e.g., histograms) or numerical using test statistics ($T(y)$) sensitive to particular features (like variance, skewness, range). Extreme Bayesian p-values suggest the model fails to capture the aspect of the data measured by $T(y)$. Beyond checking fit, we also need to assess the robustness of our conclusions using **Sensitivity Analysis** – how much do results depend on subjective choices like the prior? Finally, if we have multiple candidate models, we need methods like **DIC** to compare them based on both how well they fit the data and how complex they are.

Core Ideas & Summaries

10.1 Philosophy of Model Checking

- ”All models are wrong, but some are useful.” (George Box)
- Goal is not to find the ”true” model, but a model that is adequate for the purpose at hand and captures important features of the data.
- Checking involves examining model assumptions and the congruence between model predictions and observations.

10.2 Posterior Predictive Checks

- **Simulating y^{rep}:** Requires posterior samples $\theta^{(s)}$ (from MCMC). For each $\theta^{(s)}$, draw $y^{\text{rep}(s)} \sim p(y|\theta^{(s)})$. This generates a sample from $p(y^{\text{rep}}|y_{\text{obs}})$.
- **Graphical Checks:** Visually compare y_{obs} to several (e.g., 10-20) y^{rep} datasets, or compare y_{obs} to the distribution of all y^{rep}’s combined.
  - Examples: Histograms, density plots, ECDFs, scatterplots (for multivariate data).
  - Look for systematic discrepancies: Does the location, spread, shape, presence of outliers, etc., of y_{obs} look markedly different from the typical y^{rep}?
- **Numerical Checks (Test Statistics):**
  - Choose $T(y)$ to measure a feature of interest (e.g., $T(y) = \text{variance}(y)$, $T(y) = \min(y)$, $T(y) = \max(y)$, $T(y) = \text{count}(y = 0)$).
  - Calculate $T(y_{\text{obs}})$.
  - Calculate $T(y^{\text{rep}(s)})$ for each replicated dataset $s = 1, \ldots, S$.
  - Compare $T(y_{\text{obs}})$ to the histogram/distribution of $\{T(y^{\text{rep}(s)})\}$. Is $T(y_{\text{obs}})$ in the main body or tails of the distribution?
  - Calculate Bayesian p-value: $p_B = \frac{1}{S}\sum_{s=1}^{S} I(T(y^{\text{rep}(s)}) \geq T(y_{\text{obs}}))$.
  - Interpretation: $p_B \approx 0$ or $p_B \approx 1$ suggests the model does not capture the feature measured by $T(y)$. $p_B \approx 0.5$ suggests good fit *for that specific statistic*.

  - Caution: p_B depends heavily on the choice of $T(y)$. Check multiple relevant statistics. p_B is not a measure of $P(\text{Model is True})$.

10.3 Sensitivity Analysis

- **Purpose:** Assess robustness of conclusions to modeling choices.
- **Prior Sensitivity:** Re-run the analysis with different plausible priors (e.g., more/less informative, different families). Do key posterior summaries (means, CIs) change substantially? If so, the data provides limited information relative to the prior for those quantities.
- **Likelihood Sensitivity:** If unsure about the sampling distribution (e.g., Normal vs. t-distribution for potential outliers), fit models with different likelihoods and compare results.

10.4 Model Comparison (Brief Overview)

- Sometimes need to choose between competing models $(M_1, M_2, \dots)$.
- **DIC:** Balances fit ($\bar{D}$) and complexity (p_D). Lower DIC preferred.
  - $\bar{D}$ = average deviance. Estimated by averaging $D(\theta^{(s)}) = -2 \log p(y_{\text{obs}}|\theta^{(s)})$ over MCMC samples.
  - p_D = effective number of parameters. Measures how much the deviance is expected to decrease from the prior to the posterior. $p_D = \bar{D} - D(\bar{\theta})$.
  - DIC $= \bar{D} + p_D$. Can be viewed as estimating predictive error.
  - Calculated easily from MCMC output if likelihood $p(y_{\text{obs}}|\theta)$ can be evaluated.
  - Limitations: Can be unstable, different definitions of p_D exist, interpretation requires care, less theoretically grounded than Bayes Factors for model *selection*. More useful for comparing similar models.
- **Bayes Factors (Mentioned):** Compares the marginal likelihood $p(y_{\text{obs}}|M)$ across models. $BF_{12} = p(y_{\text{obs}}|M_1)/p(y_{\text{obs}}|M_2)$. Can be difficult to compute reliably. Not the focus of Hoff Ch 10.

Connections to Previous/Future Topics

- **Ch 7-9 (MCMC):** Posterior predictive checks rely heavily on having reliable posterior samples $\theta^{(s)}$ from MCMC. Poor MCMC invalidates PPCs. DIC calculation also needs MCMC output.
- **Ch 3 (Prediction):** Formalizes the use of the posterior predictive distribution $p(\tilde{y}|y)$, introduced earlier conceptually.
- **Specific Models (Ch 11, 12+):** Model checking is an essential step when applying hierarchical models, regression models, etc. Did the chosen model structure capture the data well?

Common Pitfalls / Points of Confusion

- **Confusing y^{rep} and y_{obs}:** y_{obs} is fixed; y^{rep} is random, simulated from the fitted model.
- **Confusing Posterior Predictive and Prior Predictive:** Prior predictive uses $p(\tilde{y}) = \int p(\tilde{y}|\theta)p(\theta)d\theta$ (averages over prior); posterior predictive averages over the posterior $p(\theta|y_{\text{obs}})$.
- **Misinterpreting Bayesian p-values:** Treating p_B like a frequentist p-value for H_0: Model Fits. Low p_B only indicates misfit *with respect to T(y)*. A value near 0.5 doesn't "prove" the model is correct.
- **Over-reliance on one Test Statistic:** A model might fit well for $T(y) = \text{mean}(y)$ but poorly for $T(y) = \text{variance}(y)$. Need to check aspects relevant to the research question.
- **Ignoring Graphical Checks:** Relying solely on p-values can miss obvious visual misfits.
- **Calculation Errors:** Mistakes computing $T(y^{\text{rep}(s)})$ or $D(\theta^{(s)})$.
- **DIC Interpretation:** Treating small differences in DIC as definitive evidence for one model over another. Rules of thumb exist (e.g., difference ¿ 5-10 considered substantial), but context matters. Forgetting DIC relies on the chosen likelihood form.
- **Not Doing Sensitivity Analysis:** Assuming the chosen prior/likelihood is perfect without checking robustness.

End of Week 11 Notes.

Bayesian Statistical Methods - Study Notes Week 12

Hierarchical Models I: Concepts and Structure

Based on Hoff (2009), Chapter 11

Instructor: Sudeep Regmi

Chapters Covered

- Hoff Chapter 11: Hierarchical Modeling
  - Sections introducing motivation, exchangeability, Beta-Binomial hierarchy, Normal hierarchy (structure and basic concepts). Computation details may span into next week.

Learning Objectives

After engaging with this week's material, you should be able to:

- Explain the motivation for using hierarchical models, including "borrowing strength" across related units and modeling structured populations.
- Define exchangeability and explain its role as a justification for hierarchical structures.
- Describe the multi-level structure typical of hierarchical models: data level, parameter (process) level, and hyperparameter level.
- Formulate a simple hierarchical model, such as the Beta-Binomial model for multiple proportions or the Normal-Normal model for multiple means.
- Understand the role of population parameters (hyperparameters) and hyperpriors in hierarchical models.
- Interpret the concept of shrinkage in hierarchical models: how individual parameter estimates are pulled towards a common population mean.
- Explain intuitively how hierarchical models pool information across groups or units.
- Recognize that posterior inference in hierarchical models often requires MCMC (typically Gibbs sampling).

Key Concepts & Definitions

- **Hierarchical Model (HM) / Multilevel Model:** A statistical model specified in multiple levels, allowing for the modeling of dependencies and structured populations. Typically involves parameters whose distributions depend on other parameters (hyperparameters).

- **Borrowing Strength:** The phenomenon in HMs where estimates for one unit (e.g., group mean θ_j) are improved by using information from other related units, effectively increasing the sample size for each estimate.

- **Shrinkage:** The effect in HMs where individual parameter estimates (e.g., θ_j) are "shrunk" from their individual maximum likelihood estimates (e.g., $\hat{\theta}_j = y_j/n_j$) towards a common population mean (e.g., estimated from the hyperparameter posterior). The amount of shrinkage depends on the precision of the individual estimate and the variability across units.

- **Exchangeability:** A set of random variables $(\theta_1, \ldots, \theta_J)$ is exchangeable if their joint distribution $p(\theta_1, \ldots, \theta_J)$ is invariant under permutation of the indices. I.e., the labels $(1, \ldots, J)$ contain no information. If we believe units are exchangeable, it justifies modeling their parameters θ_j as i.i.d. draws from some common population distribution G.

- **De Finetti's Theorem:** Provides a formal link: If $(\theta_1, \theta_2, \ldots)$ is an infinitely exchangeable sequence, then their joint distribution can be represented as if they were i.i.d draws from some distribution G, averaged over uncertainty about G. $p(\theta_1, \ldots, \theta_J) = \int \left[\prod_{j=1}^{J} p(\theta_j|\phi)\right] p(\phi)d\phi$. This justifies the hierarchical structure where $p(\theta_j|\phi)$ defines the population distribution parameterized by hyperparameters ϕ.

- **Model Structure (Typical 3 Levels):**

  1. **Level 1 (Data):** $y_j \sim p(y_j|\theta_j)$ for each unit $j = 1, \ldots, J$. (Likelihood for individual unit's data given its parameter).
  2. **Level 2 (Process/Parameters):** $\theta_j \sim p(\theta_j|\phi)$. (Parameters for units are drawn from a common population distribution governed by hyperparameters ϕ). This level links the units.
  3. **Level 3 (Hyperparameters):** $\phi \sim p(\phi)$. (Prior distribution, or "hyperprior", for the population parameters).

- **Hyperparameters (ϕ):** Parameters governing the distribution of the unit-specific parameters θ_j. E.g., the mean and variance of the population distribution from which group means are drawn.

- **Hyperprior ($p(\phi)$):** The prior distribution placed on the hyperparameters. Often chosen to be weakly informative.

- **Pooling:**

  - **Complete Pooling:** Assumes all θ_j are identical ($\theta_1 = \cdots = \theta_J = \theta$). Ignores variation between units. Equivalent to combining all data.
  - **No Pooling:** Assumes all θ_j are independent and unrelated. Estimates each θ_j using only data from unit j. Ignores potential similarities.
  - **Partial Pooling (Hierarchical Model):** Assumes θ_j are related via a common population distribution $p(\theta_j|\phi)$. Allows borrowing strength and provides a compromise between complete and no pooling, guided by the data.

Conceptual Overview / "The Big Picture"

Often, we have data structured in groups or related units (e.g., students within schools, patients within hospitals, experiments repeated in different labs, measurements on multiple related species). We might want to estimate a parameter θ_j for each unit j.

- Should we analyze each unit completely separately (”no pooling”)? This ignores potential similarities and can yield poor estimates if some units have little data.

- Should we assume all units are identical and just combine all the data (”complete pooling”)? This ignores real differences between units.

Hierarchical models provide a principled compromise. The core idea is to assume that the parameters θ_j for the different units, while potentially different, are related because they are drawn from a common *population distribution.* This population distribution is governed by hyperparameters ϕ (e.g., population mean, population variance). We specify a model reflecting this structure: data depends on unit parameters (θ_j), unit parameters depend on population parameters (ϕ), and we put a prior (hyperprior) on ϕ. The concept of **exchangeability** provides a justification: if we believe the units are similar in the sense that their labels don’t provide any information beyond the data itself, then modeling their parameters as draws from a common (unknown) distribution is appropriate. By fitting this multi-level structure (often using MCMC), we achieve **partial pooling**. The estimate for each θ_j is informed by both its own data y_j and by the estimates from other units (via the estimated population distribution). This leads to **shrinkage**, where individual estimates are pulled towards the estimated population mean, especially for units with noisy or limited data. HMs allow units to ”borrow strength” from each other, leading to more stable and often more realistic estimates.

Core Ideas & Summaries

11.1-11.2 Motivation and Exchangeability

- HMs address situations with structured data where parameters might be related.

- Examples: estimating disease rates in multiple cities, student test scores in multiple schools, treatment effects across different studies (meta-analysis).

- Exchangeability assumption: If $p(\theta_1, \dots, \theta_J)$ is symmetric in its arguments, we can model θ_j’s as draws from a common distribution G. We then model G using hyperparameters ϕ, leading to the structure $\theta_j|\phi \sim p(\theta_j|\phi)$.

11.3 Example: Beta-Binomial Hierarchical Model

- **Scenario:** Estimating proportions θ_j (e.g., success rates) for J different units, based on binomial data $y_j \sim \text{Binomial}(n_j, \theta_j)$ for each unit.

- **HM Structure:**

  1. Data: $y_j|\theta_j \sim \text{Binomial}(n_j, \theta_j)$ for $j = 1, \dots, J$.
  2. Process: $\theta_j|a, b \sim \text{Beta}(a, b)$ (assuming proportions come from a common Beta population distribution). Here $\phi = (a, b)$.
  3. Hyperparameters: Need hyperpriors $p(a, b)$. Choosing these well can be tricky. Hoff often reparameterizes in terms of mean $\mu = a/(a + b)$ and concentration $\kappa = a + b$. $p(\mu, \kappa)$.

- **Joint Posterior:** $p(\boldsymbol{\theta}, a, b|\mathbf{y}) \propto p(a, b) \prod_{j=1}^{J} [p(\theta_j|a, b)p(y_j|\theta_j)]$. Usually intractable analytically.

- **Computation (Gibbs Preview):** Sample from full conditionals:

  - $p(\theta_j|\text{rest}, \mathbf{y}) \propto p(y_j|\theta_j)p(\theta_j|a, b) \implies \theta_j|\cdots \sim \text{Beta}(a + y_j, b + n_j - y_j)$. (Easy to sample!)

  - $p(a, b|\text{rest}, \mathbf{y}) \propto p(a, b) \prod_{j=1}^{J} p(\theta_j|a, b)$. This depends on the choice of hyperprior $p(a, b)$ and often requires a Metropolis step if not conjugate.

- **Shrinkage Interpretation:** The posterior mean $\mathbb{E}[\theta_j|\mathbf{y}]$ will be between the unit-specific MLE y_j/n_j and the estimated population mean $\mathbb{E}[a/(a+b)|\mathbf{y}]$. Units with small n_j shrink more towards the population mean.

11.4 Example: Normal Hierarchical Model

- **Scenario:** Estimating means μ_j for J different units, based on Normal data $y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)$ (assuming common known variance σ^2 for simplicity first, or estimate unit variances σ_j^2). Often summarized by $\bar{y}_j \sim \mathcal{N}(\mu_j, \sigma^2/n_j)$.

- **HM Structure (common σ^2 known):**

  1. Data (Summarized): $\bar{y}_j|\mu_j \sim \mathcal{N}(\mu_j, \sigma^2/n_j)$ for $j = 1, \ldots, J$.
  2. Process: $\mu_j|\mu_\theta, \tau^2 \sim \mathcal{N}(\mu_\theta, \tau^2)$ (unit means come from a Normal population with mean μ_θ and variance τ^2). $\phi = (\mu_\theta, \tau^2)$.
  3. Hyperparameters: Need hyperpriors $p(\mu_\theta)$ (e.g., Normal or flat) and $p(\tau^2)$ (e.g., Inverse Gamma/Inv-ChiSq).

- **Computation (Gibbs Preview):** Sample from full conditionals:

  - $p(\mu_j|\text{rest}, \mathbf{y}) \propto p(\bar{y}_j|\mu_j)p(\mu_j|\mu_\theta, \tau^2) \implies \mu_j|\cdots \sim \mathcal{N}(\tilde{\mu}_j, \tilde{V}_j)$ where $\tilde{\mu}_j$ is precision-weighted average of $\bar{y}_j$ and μ_θ, $\tilde{V}_j^{-1} = n_j/\sigma^2 + 1/\tau^2$. (Easy!)
  - $p(\mu_\theta|\text{rest}, \mathbf{y}) \propto p(\mu_\theta) \prod_{j=1}^{J} p(\mu_j|\mu_\theta, \tau^2)$. Often Normal if $p(\mu_\theta)$ is Normal/flat. (Easy!)
  - $p(\tau^2|\text{rest}, \mathbf{y}) \propto p(\tau^2) \prod_{j=1}^{J} p(\mu_j|\mu_\theta, \tau^2)$. Often Inv-Gamma/Inv-ChiSq if $p(\tau^2)$ is conjugate. (Easy!)

- **Shrinkage Interpretation:** Posterior mean $\mathbb{E}[\mu_j|\mathbf{y}]$ is shrunk from $\bar{y}_j$ towards the estimated population mean $\mathbb{E}[\mu_\theta|\mathbf{y}]$. Shrinkage is stronger when n_j is small (high σ^2/n_j) or when population variance τ^2 is small.

Connections to Previous/Future Topics

- **Ch 3-6 (Basic Models):** HMs build upon these basic likelihoods (Binomial, Normal) and priors (Beta, Normal, InvGamma).

- **Ch 7-10 (Computation Diagnostics):** HMs almost always require MCMC (esp. Gibbs) due to the complexity of the joint posterior. All the diagnostic tools are essential here.

- **Ch 12 (Linear Models):** Random effects models are a type of hierarchical linear model, where intercepts or slopes are allowed to vary across groups according to a population distribution.

Common Pitfalls / Points of Confusion

- **Mixing up Levels:** Confusing data y_j, parameters θ_j, and hyperparameters ϕ. Keep the dependencies straight.

- **Exchangeability Assumption:** Forgetting that the standard HM relies on the assumption that units are exchangeable a priori. If units have known covariate differences, these should be included in the model.
- **Hyperprior Choice:** Underestimating the difficulty and importance of choosing appropriate hyperpriors $p(\phi)$, especially for variance parameters like τ^2. Standard "non-informative" choices can sometimes be problematic.
- **Interpretation of Shrinkage:** Thinking shrinkage is an artificial modification. It's a natural consequence of the model structure and Bayesian updating when parameters are assumed to come from a common population.
- **Computation:** Assuming the posterior is simple. Realizing that MCMC is usually needed and deriving the full conditionals correctly (esp. for hyperparameters) can be challenging.
- **Pooling:** Confusing partial pooling (HM) with complete or no pooling.

End of Week 12 Notes.

Bayesian Statistical Methods - Study Notes Week 13

Hierarchical Models II: Computation and Interpretation

Based on Hoff (2009), Chapter 11

Instructor: Sudeep Regmi

Chapters Covered

- Hoff Chapter 11: Hierarchical Modeling (focus on computation, full conditional derivations, interpretation of parameters and shrinkage).

Learning Objectives

After engaging with this week's material, you should be able to:

- Derive the full conditional distributions necessary for Gibbs sampling in common hierarchical models (e.g., Normal-Normal model with unknown population variance τ^2).
- Implement a Gibbs sampler for a hierarchical model.
- Use MCMC output to obtain posterior estimates and credible intervals for both unit-specific parameters (θ_j or μ_j) and population hyperparameters (ϕ).
- Interpret the posterior distribution of hyperparameters (e.g., population mean, population variance τ^2).
- Quantify and interpret the amount of shrinkage observed in the posterior estimates of unit-specific parameters.
- Explain how the data informs both the unit-level parameters and the population-level hyperparameters simultaneously.
- Apply MCMC diagnostics (from Week 10) to assess the convergence and efficiency of the Gibbs sampler for a hierarchical model.
- Understand conceptually how posterior predictive checks can be performed for hierarchical models.

Key Concepts & Definitions

- **Full Conditional Derivation (Recap):** To find $p(\text{param}|\text{rest}, \text{data})$, use proportionality: $p(\text{param}|\dots) \propto p(\text{data}|\text{params}) \times p(\text{params}|\text{hyperparams}) \times p(\text{hyperparams})$. Identify the terms involving 'param' and recognize the resulting distribution kernel.
- **Gibbs Sampling for HMs:** An iterative MCMC algorithm cycling through updates for:
    1. All unit-specific parameters (θ_j or μ_j for $j = 1, \dots, J$).

2. All hyperparameters (ϕ, e.g., μ_θ, τ^2).

Each parameter (or block) is sampled from its full conditional distribution, given the most recent values of all other parameters and the data.

- **Shrinkage Factor:** In the Normal-Normal model (with known $\sigma_j^2 = \sigma^2/n_j$), the posterior mean for μ_j is $\tilde{\mu}_j = w_j \bar{y}_j + (1 - w_j)\mu_\theta$, where the weight $w_j = (\sigma_j^{-2})/(\sigma_j^{-2} + \tau^{-2})$ comes from the precision-weighted average formula. The amount of shrinkage towards μ_θ is $1 - w_j = (\tau^{-2})/(\sigma_j^{-2} + \tau^{-2})$. Shrinkage is high if unit precision σ_j^{-2} is low, or if population variance τ^2 is low.

- **Posterior Inference for Hyperparameters:** The MCMC output provides samples from the marginal posteriors $p(\mu_\theta|\mathbf{y})$ and $p(\tau^2|\mathbf{y})$. Summaries (mean, median, CI) of these samples tell us about the estimated population center and spread.

- **Posterior Predictive Checks for HMs:** Simulating involves simulating the entire hierarchy: 1. Draw $(\phi^{(s)}, \boldsymbol{\theta}^{(s)})$ from the joint posterior MCMC sample. 2. Draw ${}_j^{(s)} \sim p(y_j|\theta_j^{(s)})$ for each unit j. Alternatively, to check the population distribution assumption: 1. Draw $\phi^{(s)}$ from its marginal posterior sample. 2. Draw $\theta_{j,\text{rep}}^{(s)} \sim p(\theta_j|\phi^{(s)})$ for $j = 1, \ldots, J$. 3. Compare the distribution of $\{\theta_{j,\text{rep}}^{(s)}\}$ to the posterior sample $\{\theta_j^{(s)}\}$.

Conceptual Overview / "The Big Picture"

Last week, we introduced the structure and philosophy of hierarchical models as a way to borrow strength across related units by modeling parameters (θ_j) as draws from a population distribution governed by hyperparameters (ϕ). This week, we focus on how to actually fit these models using MCMC, specifically Gibbs sampling, and how to interpret the rich output they provide. The key computational step is deriving and sampling from the full conditional distribution for each parameter and hyperparameter in the model. This often involves combining information from the data level (likelihood), the process level (link between θ_j and ϕ), and the hyperprior level. Once we have the MCMC samples, we can simultaneously learn about individual units (via the posteriors for θ_j, which exhibit shrinkage) and about the population they come from (via the posteriors for ϕ). Interpreting the posterior for population variance (τ^2 in the Normal model) is particularly important, as it tells us how much heterogeneity exists across units and influences the degree of shrinkage.

Core Ideas & Summaries

Computation: Gibbs Sampling for HMs

- **General Strategy:** Identify all parameters (θ_j's) and hyperparameters (ϕ). Find the full conditional distribution for each by treating all other variables as fixed and examining the joint posterior $p(\text{data}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\phi)p(\phi)$ as a function of the parameter of interest.

- **Normal-Normal Example (Unknown τ^2, known σ^2):**

  - Model:
    - $\bar{y}_j|\mu_j \sim \mathcal{N}(\mu_j, \sigma^2/n_j)$ (Data, σ^2 known)
    - $\mu_j|\mu_\theta, \tau^2 \sim \mathcal{N}(\mu_\theta, \tau^2)$ (Process)
    - $p(\mu_\theta) \propto 1$ (Hyperprior for population mean, flat)
    - $\tau^2 \sim \text{Inv-Gamma}(a_0, b_0)$ (Hyperprior for population variance)

  - Full Conditionals for Gibbs:
    1. **For each μ_j ($j = 1, \dots, J$):** $p(\mu_j | \dots) \propto p(\bar{y}_j|\mu_j)p(\mu_j|\mu_\theta, \tau^2)$
    $\propto \exp\left(-\frac{n_j}{2\sigma^2}(\mu_j - \bar{y}_j)^2\right) \exp\left(-\frac{1}{2\tau^2}(\mu_j - \mu_\theta)^2\right)$ This is a Normal kernel for μ_j. Completing the square yields: $\mu_j|\dots \sim \mathcal{N}(\tilde{\mu}_j, \tilde{V}_j)$, where $\tilde{V}_j^{-1} = n_j/\sigma^2 + 1/\tau^2$ $\tilde{\mu}_j = \tilde{V}_j((n_j/\sigma^2)\bar{y}_j + (1/\tau^2)\mu_\theta)$
    2. **For μ_θ:** $p(\mu_\theta| \dots) \propto p(\mu_\theta) \prod_{j=1}^J p(\mu_j|\mu_\theta, \tau^2) \propto 1 \times \prod_{j=1}^J \exp\left(-\frac{1}{2\tau^2}(\mu_j - \mu_\theta)^2\right)$
    $\propto \exp\left(-\frac{1}{2\tau^2}\sum(\mu_j - \mu_\theta)^2\right) \propto \exp\left(-\frac{J}{2\tau^2}(\mu_\theta - \bar{\mu})^2\right)$, where $\bar{\mu} = \frac{1}{J}\sum \mu_j$.
    This is a Normal kernel for μ_θ. $\mu_\theta|\dots \sim \mathcal{N}(\bar{\mu}, \tau^2/J)$
    3. **For τ^2:** $p(\tau^2| \dots) \propto p(\tau^2) \prod_{j=1}^J p(\mu_j|\mu_\theta, \tau^2)$
    $\propto (\tau^2)^{-(a_0+1)} e^{-b_0/\tau^2} \times \prod_{j=1}^J (\tau^2)^{-1/2} \exp\left(-\frac{1}{2\tau^2}(\mu_j - \mu_\theta)^2\right)$
    $\propto (\tau^2)^{-(a_0+J/2+1)} \exp\left(-\frac{1}{\tau^2}\left[b_0 + \frac{1}{2}\sum(\mu_j - \mu_\theta)^2\right]\right)$
    This is an Inverse-Gamma kernel for τ^2. $\tau^2|\dots \sim \text{Inv-Gamma}(a_0+J/2, b_0+\frac{1}{2}\sum(\mu_j - \mu_\theta)^2)$
  - **Gibbs Implementation:** Initialize $(\mu_1^{(0)}, \dots, \mu_J^{(0)}, \mu_\theta^{(0)}, (\tau^2)^{(0)})$. Cycle through sampling from the derived Normal and Inv-Gamma full conditional distributions, using the most recent values for conditioning.

- **Other Models (e.g., Beta-Binomial):** Similar process. Full conditional for θ_j is often Beta. Full conditionals for hyperparameters (a, b or μ, κ) depend on hyperpriors and may require Metropolis-Hastings steps if not conjugate.

- **MCMC Diagnostics:** Absolutely essential. Check trace plots for all θ_j's (or a subset) and all ϕ's. Ensure convergence, assess mixing via ACF/ESS, calculate MCSE. HMs can sometimes have high correlation between parameters (e.g., μ_θ and μ_j's, or τ^2 and μ_j's), requiring longer runs.

Interpretation of HM Output

- **Unit-Level Parameters (θ_j, μ_j):**
  - The MCMC output gives samples from $p(\theta_j|\mathbf{y})$. Summarize with posterior mean, median, credible interval.
  - Compare $\mathbb{E}[\theta_j|\mathbf{y}]$ to the non-hierarchical estimate (e.g., $\hat{\theta}_j = y_j/n_j$). Observe the shrinkage towards the population mean $\mathbb{E}[\phi|\mathbf{y}]$.
  - Shrinkage is adaptive: units with less precise individual information (small n_j or large σ_j^2) are shrunk more strongly.
  - Posterior credible intervals for θ_j are often narrower than non-pooled CIs due to borrowing strength, especially for small units.

- **Hyperparameters (ϕ, e.g., μ_θ, τ^2):**
  - MCMC output gives samples from $p(\phi|\mathbf{y})$. Summarize with posterior mean, median, credible interval.
  - $\mathbb{E}[\mu_\theta|\mathbf{y}]$ estimates the overall population mean.
  - The posterior for τ^2 (population variance) is critical.
    - Small $\mathbb{E}[\tau^2|\mathbf{y}]$ (and CI close to 0) suggests units are very similar $\implies$ strong shrinkage (closer to complete pooling).
    - Large $\mathbb{E}[\tau^2|\mathbf{y}]$ (and CI far from 0) suggests substantial heterogeneity between units $\implies$ weak shrinkage (closer to no pooling).
  - The data informs τ^2: If unit estimates $\bar{y}_j$ are very spread out, the posterior for τ^2 will favor larger values.

- **Simultaneous Learning:** The model learns about θ_j's and ϕ together. Information flows both ways: estimates of θ_j's inform the population parameters ϕ, and the estimated population distribution ϕ informs (shrinks) the estimates of θ_j.

Model Checking for HMs

- Can perform PPCs similar to non-hierarchical models.
- Simulate incorporating all levels of uncertainty.
- Check overall fit (e.g., distribution of all y_{ij}) and group-level fit (e.g., distribution within specific group j, or distribution of group means $\bar{y}_j$).
- Test statistics can focus on individual units, or population features (e.g., variance of the $\bar{y}_j$'s).
- Check assumptions: e.g., normality of residuals within groups, normality assumption for $\theta_j|\phi$.

Connections to Previous/Future Topics

- **Ch 8-10 (MCMC & Diagnostics):** Direct application of Gibbs sampling techniques and essential use of diagnostic tools to ensure reliable inference from the complex HM posterior.
- **Ch 12 (Linear Models):** Random effects models are a direct application of the Normal hierarchical model structure within a regression context (e.g., random intercepts model $\beta_{0j} \sim \mathcal{N}(\mu_{\beta_0}, \tau^2_{\beta_0})$).

Common Pitfalls / Points of Confusion

- **Errors in Full Conditionals:** Algebraic mistakes when deriving the conditional distributions, especially for hyperparameters involving products over j.
- **Slow Mixing/Convergence:** HMs can sometimes exhibit slow mixing, particularly between hyperparameters (like τ^2) and unit parameters (μ_j). Requires careful diagnostics and potentially long MCMC runs or reparameterizations.
- **Hyperprior Sensitivity:** Posterior inference, especially for hyperparameters like τ^2 when J is small, can be sensitive to the choice of hyperprior $p(\tau^2)$.
- **Overinterpreting Shrinkage:** Forgetting that shrinkage magnitude is learned from the data (via τ^2) and not arbitrarily imposed.
- **Ignoring Diagnostics:** Failing to check convergence and mixing for *all* relevant parameters, including hyperparameters. Poor mixing for τ^2, for example, can affect all μ_j estimates.
- **Coding Errors:** Bugs in implementing the Gibbs sampler steps.

End of Week 13 Notes.

Bayesian Statistical Methods - Study Notes Week 14

Bayesian Linear Models

Based on Hoff (2009), Chapter 12

Instructor: Sudeep Regmi

Chapters Covered

- Hoff Chapter 12: Linear Regression

Note: Depending on course focus, this might also include special topics or serve as a review/wrap-up week.

Learning Objectives

After engaging with this week's material, you should be able to:

- Formulate the standard linear regression model within the Bayesian framework.
- Write down the likelihood function for the regression model parameters $(\boldsymbol{\beta}, \sigma^2)$ based on observed data $(\mathbf{y}, \mathbf{X})$.
- Specify common prior distributions for regression coefficients $\boldsymbol{\beta}$ and error variance σ^2.
- Describe the standard conjugate prior for $(\boldsymbol{\beta}, \sigma^2)$, the Multivariate Normal-Inverse-Chi-Squared distribution.
- Interpret the hyperparameters of the conjugate prior.
- Understand the structure of the joint posterior distribution $p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X})$ under the conjugate prior.
- State the marginal posterior distribution for the error variance σ^2 (Inverse-Chi-Squared).
- State the marginal posterior distribution for the regression coefficients $\boldsymbol{\beta}$ (Multivariate Student's t-distribution).
- Explain how to perform posterior inference (point estimates, credible intervals) for individual regression coefficients β_j.
- Outline how Gibbs sampling can be used to obtain posterior samples when using the conjugate prior.
- Understand how posterior predictive distributions can be used for prediction and model checking in the regression context.

Key Concepts & Definitions

- **Linear Regression Model:** Models a continuous outcome variable y_i as a linear function of p predictor variables $x_{i1}, \dots, x_{ip}$, plus random error ϵ_i.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

  where $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$ is the predictor vector for observation i (including intercept), and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ is the vector of regression coefficients. The errors are typically assumed i.i.d $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

- **Matrix Notation:** $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is $n \times 1$, $\mathbf{X}$ is $n \times (p+1)$, $\boldsymbol{\beta}$ is $(p+1) \times 1$, $\boldsymbol{\epsilon}$ is $n \times 1$, and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

- **Likelihood Function $L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X})$:** Based on the $\mathcal{N}$ error assumption:

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2\right)$$

$$p(\mathbf{y}|\dots) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right)$$

  Let $\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ (OLS estimate) and $SSR = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$ (Sum of Squared Residuals). The exponent can be rewritten involving $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{OLS})$.

- **Sufficient Statistics:** $(\mathbf{X}^T\mathbf{X})$ and $(\mathbf{X}^T\mathbf{y})$, or equivalently $\hat{\boldsymbol{\beta}}_{OLS}$ and SSR.

- **Conjugate Prior (MVN-Inv-ChiSq):** A joint prior structure for $(\boldsymbol{\beta}, \sigma^2)$:

  1. $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$
  2. $\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}_{p+1}(\boldsymbol{\beta}_0, \sigma^2\Sigma_0)$

  Hyperparameters:

  - $\boldsymbol{\beta}_0$: Prior mean vector for $\boldsymbol{\beta}$.
  - Σ_0: Prior covariance matrix structure for $\boldsymbol{\beta}$ (scaled by σ^2). Often diagonal, e.g., $\Sigma_0 = c(\mathbf{X}^T\mathbf{X})^{-1}$ (Zellner's g-prior) or $\Sigma_0 = \tau_0^2\mathbf{I}$. Σ_0^{-1} is the prior precision matrix structure.
  - ν_0: Prior degrees of freedom for σ^2.
  - σ_0^2: Prior scale for σ^2.

- **Reference Prior / Non-informative Prior:** Often $p(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$. Corresponds to limits of the conjugate prior (e.g., $\nu_0 \to -1$, $\Sigma_0^{-1} \to \mathbf{0}$).

- **Joint Posterior Distribution:** Under the conjugate prior, the posterior is also MVN-Inv-ChiSq:

  1. $\sigma^2|\mathbf{y}, \mathbf{X} \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2)$
  2. $\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X} \sim \mathcal{N}_{p+1}(\boldsymbol{\beta}_n, \sigma^2\Sigma_n)$

  Updated Hyperparameters (Conceptual):

  - $\boldsymbol{\beta}_n$ is a precision-weighted average of prior mean $\boldsymbol{\beta}_0$ and OLS estimate $\hat{\boldsymbol{\beta}}_{OLS}$.
  - Σ_n^{-1} involves sum of prior precision Σ_0^{-1} and data precision related to $\mathbf{X}^T\mathbf{X}$.

  - $\nu_n = \nu_0 + n$.
  - $\nu_n \sigma_n^2$ combines prior sum of squares ($\nu_0 \sigma_0^2$), data residual sum of squares (SSR), and a term for discrepancy between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}_{OLS}$.

  (Exact formulas analogous to Hoff Eq 12.5-12.8, similar structure to Ch 5 Normal model).

- **Marginal Posterior for σ^2:** $\sigma^2 | \mathbf{y}, \mathbf{X} \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2)$.
- **Marginal Posterior for $\boldsymbol{\beta}$ (Multivariate Student's t):**

$$(\boldsymbol{\beta} - \boldsymbol{\beta}_n)^T (\sigma_n^2 \Sigma_n)^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_n)/(p+1) \sim F_{p+1,\nu_n}$$

  Equivalently, the distribution is a Multivariate Student's t-distribution with location $\boldsymbol{\beta}_n$, scale matrix $\sigma_n^2 \Sigma_n$, and degrees of freedom ν_n.

- **Marginal Posterior for β_j (Univariate Student's t):** The marginal posterior for a single coefficient β_j follows a univariate Student's t-distribution:

$$\frac{\beta_j - \beta_{nj}}{\text{se}(\beta_j)} \sim t_{\nu_n}$$

  where β_{nj} is the j-th element of $\boldsymbol{\beta}_n$ and $\text{se}(\beta_j) = \sqrt{(\sigma_n^2 \Sigma_n)_{jj}}$ is the posterior standard error (square root of the j-th diagonal element of the posterior scale matrix).

Conceptual Overview / "The Big Picture"

Linear regression is a workhorse of statistical modeling. This week, we frame it within the Bayesian paradigm. Instead of finding point estimates and standard errors based on sampling distributions (frequentist), we aim to find the joint posterior distribution of the unknown parameters: the regression coefficients $\boldsymbol{\beta}$ and the error variance σ^2. Assuming Normal errors leads to a likelihood function similar in structure to the multivariate Normal case. We introduce the standard conjugate prior, the Multivariate Normal-Inverse-Chi-Squared distribution, which allows for analytical derivation of the posterior (also MVN-Inv-ChiSq). This prior structure specifies beliefs about the coefficients $\boldsymbol{\beta}$ (centered at $\boldsymbol{\beta}_0$) conditional on the variance σ^2, and separate beliefs about σ^2. The resulting marginal posterior for individual coefficients β_j is a Student's t-distribution, analogous to the univariate Normal case with unknown variance (Week 5), reflecting uncertainty in both $\boldsymbol{\beta}$ and σ^2. While analytical results exist for the conjugate prior, MCMC (often Gibbs sampling) is commonly used, especially if non-conjugate priors are preferred or if the model becomes more complex (e.g., variable selection, robust errors). Bayesian regression allows for direct probabilistic statements about coefficients (e.g., $P(\beta_1 > 0|\text{data})$) and straightforward prediction incorporating parameter uncertainty.

Core Ideas & Summaries

12.1 The Linear Regression Model

- Setup: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.
- Likelihood function $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)$ derived from the Normal PDF. Proportionality depends on $(\sigma^2)^{-n/2}$ and the sum of squares term $\exp(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))$.

12.2 Prior Distributions

- Need a joint prior $p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2)$.
- **Conjugate Prior:** $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$ and $\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}(\boldsymbol{\beta}_0, \sigma^2\Sigma_0)$.
  - Choice of $\boldsymbol{\beta}_0$: Often $\mathbf{0}$ if variables are centered/standardized.
  - Choice of Σ_0: Controls prior variance and correlation of $\boldsymbol{\beta}$.
    - $\Sigma_0 = \tau_0^2\mathbf{I}$: Independent priors on coefficients (after scaling by σ^2). Requires choosing large τ_0^2 for weak information.
    - Zellner's g-prior: $\Sigma_0 = g(\mathbf{X}^T\mathbf{X})^{-1}$. Links prior covariance structure to data structure. Requires choosing g.
  - Choice of ν_0, σ_0^2: Often chosen for weak information on σ^2 (e.g., small ν_0, σ_0^2 near sample variance).
- **Reference Prior:** $p(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$. Yields posterior results often numerically similar to frequentist OLS results (but interpretation differs).

12.3 Posterior Distributions (Conjugate Case)

- Combining likelihood and conjugate prior yields a MVN-Inv-ChiSq posterior $p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X})$.
- **Marginal for σ^2:** $\sigma^2|\mathbf{y}, \mathbf{X} \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2)$ where $\nu_n = \nu_0 + n$ and $\nu_n\sigma_n^2$ involves SSR. Allows CIs for σ^2.
- **Marginal for $\boldsymbol{\beta}$:** $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$ is Multivariate Student's t with ν_n degrees of freedom, location $\boldsymbol{\beta}_n$ (posterior mean), and scale matrix $\sigma_n^2\Sigma_n$.
- **Marginal for β_j:** $p(\beta_j|\mathbf{y}, \mathbf{X})$ is univariate Student's t. Allows calculation of posterior mean β_{nj}, standard deviation $\text{se}(\beta_j)$, and credible intervals $[\beta_{nj} \pm t_{\nu_n,\alpha/2}\text{se}(\beta_j)]$.
- Can directly compute posterior probabilities like $P(\beta_j > c|\mathbf{y}, \mathbf{X})$ using the t-distribution CDF.

12.4 Computation (Gibbs Sampling)

- Even with conjugate priors, Gibbs sampling is often convenient, especially as models grow.
- **Full Conditionals (Conjugate Prior):** 1. $p(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X})$ is $\mathcal{N}_{p+1}(\boldsymbol{\beta}_n, \sigma^2\Sigma_n)$. Easy to sample from. 2. $p(\sigma^2|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$ is $\text{Inv-}\chi^2(\nu_0 + n, \frac{\nu_0\sigma_0^2 + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\nu_0+n})$. Also easy to sample from.
- **Gibbs Algorithm:** Initialize $\boldsymbol{\beta}^{(0)}, (\sigma^2)^{(0)}$. Iterate sampling σ^2 given current $\boldsymbol{\beta}$, then sample $\boldsymbol{\beta}$ given newly sampled σ^2.
- Diagnostics (Week 10) are essential.

Inference and Prediction

- **Coefficient Inference:** Use marginal posteriors $p(\beta_j|\mathbf{y}, \mathbf{X})$ (often t-distributions) for CIs and hypothesis tests ($P(\beta_j > 0|\text{data})$).
- **Prediction:** For a new vector of predictors $\mathbf{x}_{new}$:
  - Predicted mean response: $\eta_{new} = \mathbf{x}_{new}^T\boldsymbol{\beta}$. Can find posterior distribution $p(\eta_{new}|\mathbf{y}, \mathbf{X})$ by passing MCMC samples $\boldsymbol{\beta}^{(s)}$ through the linear predictor.

  - Predictive distribution for a new observation y_{new}: Incorporates both uncertainty in $\boldsymbol{\beta}$ and the error variance σ^2. $p(y_{new}|\mathbf{y}, \mathbf{X}) = \iint p(y_{new}|\mathbf{x}_{new}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) d\boldsymbol{\beta} d\sigma^2$. Can simulate draws $y_{new}^{(s)} \sim \mathcal{N}(\mathbf{x}_{new}^T \boldsymbol{\beta}^{(s)}, (\sigma^2)^{(s)})$.

- **Model Checking:** Use posterior predictive checks (Week 11). Simulate replicated $\mathbf{y}_{rep}$ from $p(\mathbf{y}_{rep}|\mathbf{y}, \mathbf{X})$ and compare to $\mathbf{y}$. Useful statistics $T(y)$ include residual summaries, R-squared, etc. Check residual plots against predictors.

- **Model Comparison:** Use DIC or Bayes Factors (with caution) to compare different sets of predictors or different error models.

Connections to Previous/Future Topics

- **Ch 5 (Normal Model):** Bayesian linear regression builds directly on the Normal model inference framework (MVN-Inv-ChiSq posterior, marginal t-distribution for mean/coefficients). Regression is like estimating the mean where the mean depends linearly on predictors.

- **Ch 6 (MVN):** Explicitly uses MVN distributions for priors and conditional posteriors of $\boldsymbol{\beta}$.

- **Ch 11 (Hierarchical Models):** Linear regression can be extended hierarchically. For example, random effects models allow intercepts (β_{0j}) or slopes (β_{1j}) to vary across groups j according to a population distribution: $\beta_{0j} \sim \mathcal{N}(\mu_0, \tau_0^2)$.

- **Further Topics:** Bayesian variable selection, robust regression (using t-distributed errors), generalized linear models (logistic, Poisson regression), etc.

Common Pitfalls / Points of Confusion

- **Matrix Algebra:** Keeping track of dimensions and operations for $\boldsymbol{\beta}, \mathbf{X}, \Sigma_0, \Sigma_n$.

- **Prior Specification:** Choosing appropriate priors $\boldsymbol{\beta}_0, \Sigma_0, \nu_0, \sigma_0^2$, especially achieving weak information without causing computational issues. Understanding the implications of choices like g-priors.

- **Marginal vs. Conditional:** Distinguishing the marginal posterior for $\boldsymbol{\beta}$ (Multivariate t) from the conditional posterior $p(\boldsymbol{\beta}|\sigma^2, \text{data})$ (MVN).

- **Interpretation:** Correctly interpreting credible intervals for β_j and posterior probabilities like $P(\beta_j > 0|\text{data})$.

- **Computation:** Errors implementing Gibbs sampler, particularly sampling from the MVN for $\boldsymbol{\beta}$. Need for diagnostics.

- **Collinearity:** High correlation between predictors in $\mathbf{X}$ can lead to high posterior correlation between corresponding β_j's and slow MCMC mixing, similar to frequentist issues.

End of Week 14 Notes.