2013-10-18

A Topics Analysis Model for Health Insurance Claims

Jared Anthony Webb
Brigham Young University - Provo

A Topics Analysis Model for Health Insurance Claims

Jared Webb

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Jeffrey Humphreys, Chair
Christopher Grant
Dennis Tolley

Department of Mathematics

Brigham Young University

October 2013

ABSTRACT

A Topics Analysis Model for Health Insurance Claims

Jared Webb
Department of Mathematics, BYU
Master of Science

Mathematical probability has a rich theory and powerful applications. Of particular note is the Markov chain Monte Carlo (MCMC) method for sampling from high dimensional distributions that may not admit a naive analysis. We develop the theory of the MCMC method from first principles and prove its relevance. We also define a Bayesian hierarchical model for generating data. By understanding how data are generated we may infer hidden structure about these models. We use a specific MCMC method called a Gibbs' sampler to discover topic distributions in a hierarchical Bayesian model called Topics Over Time. We propose an innovative use of this model to discover disease and treatment topics in a corpus of health insurance claims data. By representing individuals as mixtures of topics, we are able to consider their future costs on an individual level rather than as part of a large collective.

Keywords: Probability, Bayesian Data Analysis, Machine Learning, Markov Chains, Markov Chains, Markov Chain Monte Carlo, Bayesian Network, Latent Dirichlet Allocation, Topics Over Time

Acknowledgments

First and foremost I thank my wife for her encouragement, patience, and valuable suggestions as I worked on this project. Furthermore, I thank my committee for their suggestions, and particularly I thank Dr. Tolley for ensuring that this project was in a state that could be called "finished." I thank Dr. Humphreys for his long friendship and confidence. I acknowledge the BYU mathetmatics department for their support over the years. Finally, I wish to thank my parents and siblings for fostering a love of learning in our home and for their encouragement.

Contents

List of Tables

List of Figures

Chapter 1. Introduction

One doesn't have to listen very hard to hear the constant noise being generated in the media and culture about health care. There is hardly an issue in the public sphere that is able to stir up so much opinion and acrimony in the United States. Precisely what role government, community, actuarial fairness, and personal responsibility play is the subject of newspaper editorials, policy debates, and water cooler conversations across the country.

Despite the constant political bickering, firms with large numbers of employees participate in a relatively stable marketplace. The law of large numbers and the central limit theorem allow actuaries to effectively price insurance for these groups using standard models. However, about half of the employees in the United States are employed by small firms [4]. This market is highly volatile (see, for example, [16] for a comparison of large and small firms). Indeed, without aggregating very large numbers of employees, it is more difficult to predict the cost of health care for a population over time. Consider the definition of sample variance that we learn in an introductory statistics class:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

Here N is the size of our population and $\bar{y}$ is our sample mean. Suppose an actuary wishes to model costs for a population as a random variable, and he or she has some amount of data about the population's past consumption. Given large N, his or her random variable will have low variance — even approaching 0 as N goes to ∞. However, when N is small variance will be larger, which means higher volatility in the output of the variable, which in turn means more risk.

In this thesis we propose a new technique for modeling health care costs at an individual rather than collective level. This requires more sophisticated techniques than those used on large populations. Beginning with elementary probability theory, we will develop a rigorous framework for the Markov chain Monte Carlo method. Furthermore, we will develop a

Bayesian model for how health insurance claims are generated from some fixed number of "topic" distributions. Using a specific Markov chain Monte Carlo method called the Gibbs' sampler, we will infer what sorts of claims are produced by each topic distribution. We propose that these distributions, combined with cost data, can be used to make effective predictions about individual and small group health consumption.

1.1 History of Health Insurance in the United States

Historically, health care in the United States has been an unreliable, even dangerous, consumer good. Only in the twentieth century did it become relatively safe to consult a physician and hospitals transformed from a place to die into a place to be healed (See [14]).

Federal entanglement in health care is also a twentieth century invention, though the issue had been brought up in the previous century. In 1854 the Senate sent a bill to President Franklin Pierce calling for land grants to the states for the "indigent insane." The bill was met with veto from the President, who declared that:

"It can not be questioned that if Congress has power to make provision for the indigent insane...it has the same power to provide hospitals and other local establishments for the care and cure of every species of human infirmity...The whole field of public beneficence is thrown open to the care and culture of the Federal Government." [29]

These strong non-interventionist sentiments began to transform during the transition into the twentieth century. This new century saw existential changes to health care and the birth of the industry as we know it. In 1910 the Carnegie Foundation published the report of Robert Flexner, which harshly criticized the quantity of doctors being trained and the quality of education they were receiving. The report called for fewer, more skilled doctors rather than the current glut of ill-trained physicians [15]. In the coming years, dozens of medical schools were closed. With more resources dedicated to fewer doctors, there was a marked increase in the quality of care. This rise in quality came with a price, as the increase in the quality of care increased demand while at the same time the supply of providers

decreased [12].

The rise of a more professional cohort of physicians paralleled the development of new technologies that similarly increased the quality and cost of health care. X-rays and the development of the germ theory were beginning to transform the theory of medicine from superstition into a real science. New understanding of bacteria and sanitation now made surgery common and safe, while shifting care from homes to hospitals. All in all, a new generation of Americans were beginning to have access to a system of reliable, precise, and clean medical care [35].

These dramatic shifts led to increased demand, rising prices, and the beginnings of the health insurance industry in the United States. Prior to the spike in demand for health care, insurance companies viewed usage of health care as an uninsurable event. Insurers exposed themselves to the moral hazard of adverse risk selection, insurees that were difficult to track, and costs that were hard to quantify. Modern actuarial data, preventative care, and health diagnostics were simply unavailable. A rapidly urbanizing and suburbanizing population and improvements in communications technologies developed in parallel with the advances in care began to mitigate and quantify these risks [35].

The first health insurance policies offered in the United States were offered to cover lost wages due to illness or injury on the job similar to modern workers compensation funds. These insurance plans, however, were only offered in certain industries. This changed, however, in 1929 when Baylor University Hospital agreed to contract with several teachers in the Dallas area. The teachers agreed to pay $6.00 per year in exchange for a maximum of 21 days of hospitalization [13]. This agreement, the first of the "Blue Cross" plans, was the first modern health insurance plan offered in the United States. Fittingly, it was strictly tied to employment.

Tying health insurance to employment solved several of the problems that were facing potential insurers. By contracting with groups that existed for reasons other than buying health insurance, they could mitigate some of their adverse selection risk. Also, the pre-

payment plans offered services rather than cash rewards, thus helping to defray the moral hazard that had prevented development earlier. People were much less likely to commit fraud in order to go to the hospital than if they were offered a cash reward [13]. State governments also saw an opportunity to improve public health through these pre-payment plans, and allowed for more lax regulations and favorable tax breaks so long as the plans remained "non-profit.[35]" This was the silver bullet, and in the ensuing decades Blue Shield plans became available throughout most of the country.

Though health insurance was beginning to become more common, it was still not widely purchased nor expected employment benefit until World War II. Throughout the 1930's, the popularity of Blue Cross plans continued to grow, enough that it caught the attention of the American Medical Association. Worried about hospitals' growing influence over the market, the AMA began to actively lobby and maneuver to regulate how the new health plans could control the choices of patients. However, a government policy during World War II led to dramatically more people being insured while enshrining employment based health insurance in American culture. After the United States entered the war, strict rationing and price control policies were put into place in order to best direct resources to the war effort. These controls included wage controls on industries not deemed crucial to the war effort in order to prevent those industries from incentivizing workers away from the crucial industries [35]. In order to compete for the best workers on an unequal playing field, firms began to offer more lavish benefits. The IRS decided that health benefits did not qualify as "wages," and generous health insurance plans soon became a primary means by which firms attracted workers [22].

The last half of the twentieth century, as well as the beginning of the twenty-first, has seen the increase of federal involvement in the health care market. The Johnson administration marked the first major intervention with the passage of Medicare and Medicaid as part of the "Great Society" legislation. These programs gradually expanded in their cost and scope, but despite several attempts never guaranteed health care for all citizens. Beginning

in the 1990's, reform of these programs became an important topic, and the Clinton administration attempted to rework them from the ground up. These efforts died in legislation, however. The Bush administration greatly expanded the drug benefits enjoyed by senior citizens. Finally, the Obama administration passed the Affordable Care Act in 2009. This featured a mandate that almost all Americans purchase health insurance while at the same time greatly expanding the government's role in regulating how health insurance may be provided. Medical underwriting, except in categories determined by the federal government, will become illegal in 2014. It is estimated that tens of millions of uninsured Americans will become covered; see [9] for a detailed outline of these interventions.

It is in this volatile, uncertain market that we wish to make predictions. We begin with the basics of probability theory.

Chapter 2. Basic Probability

"If you want to make apple pie from scratch, first you must create the universe."

–Carl Sagan

2.1 Determinism vs. Random

The purpose of probability is to model uncertainty. In this chapter we will develop the tools to turn this notion into a formal theory. These tools will be used repeatedly throughout the thesis, so it is crucial that they are consistent and codified.

Mathematics provides the tools for science to model our universe. In building a model, we often make a distinction between a deterministic event and a random one. A deterministic event is one in which we can "see the end from the beginning." In other words, its outcome can be determined a priori. A random event is the complement of this idea - an event whose outcome is not completely determined by prior events. There is a subtlety worth noting here - the decision to label an event as deterministic or random often depends on the granularity

of the information available to an observer. For example, most people would call a coin flip a random event. This is because most people are not willing to look at the event with a finer eye. Certainly, the coin flip is determined by the way the flipper flips it, and so to a very careful observer this could be called deterministic. Though many things (perhaps all?) are deterministic, it is often the case that information about a process is difficult, costly, or impossible to obtain. Probability gives us the tools to still create meaning out of this often incomplete information.

2.2 Measures

In its full glory, probability theory is an application of measure theory. Though it may initially feel totally out of place with our notions of probability, we will soon see that there is no better place to start. Measure theory is a rich topic in itself (see [33]), but we will only develop it sufficient for our needs here. Let us begin by defining a σ-algebra. If A is a set, then 2^A denotes the power set of A, or the set of all subsets.

Definition 2.1 (σ-Algebra)**.** If Ω is a set and $\mathcal{A} \subset 2^\Omega$, then we say that $\mathcal{A}$ is a σ-algebra if it meets the following criteria:

(i) $\Omega \in \mathcal{A}$

(ii) $\mathcal{A}$ is closed under complements

(iii) $\mathcal{A}$ is closed under countable unions

An easy consequence of this definition is that if $\mathcal{A}$ is a σ-algebra, then $\emptyset \in \mathcal{A}$. We also clearly have that 2^Ω is a σ-algebra.

Definition 2.2 (Generated σ-algebra)**.** If A is a set or a collection of sets, then $\sigma(A)$ is the smallest σ-algebra that contains A, called the σ-algebra generated by A.

Definition 2.3 (Borel σ-algebra)**.** Let Ω be a topological space. Then we say that

$$B(\Omega) = \sigma(A \subset \Omega : A \text{ open in } \Omega)$$

is the Borel-σ algebra on Ω

We define a set function, and introduce the concepts of measure and measurable spaces:

Definition 2.4 (Set function)**.** Let $\mathcal{A} \subset 2^{\Omega}$. A set function is a map $\mu : \mathcal{A} \mapsto [0, \infty]$.

Definition 2.5 (σ-additivity)**.** A set function is σ-additive if

$$\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$$

for any countable collection of pairwise disjoint sets $A_i \in \mathcal{A}$, whose union is in $\mathcal{A}$.

Definition 2.6 (Measure)**.** A set function is a measure if $\mathcal{A}$ is a σ-algebra, $\mu(\emptyset) = 0$, and μ is σ-additive.

Definition 2.7 (Probability Measure)**.** We say a measure is a probability measure if $\mu(\Omega) = 1$.

Example 2.8. Let Ω be a non-empty finite set and let $\mathcal{A}$ be the power set of Ω. Then the set function $\mu : \mathcal{A} \mapsto \mathbb{R}$ given by

$$\mu(A) = \frac{|A|}{|\Omega|}$$

is a probability measure.

Proof. Clearly, $\mu(\Omega) = 1$ and $\mu(\emptyset) = 0$. It remains to show σ-additivity. Let $\{A_i\}$ be a countable collection of disjoint sets. Then

$$\mu(\cup_{i=1} A_i) = \frac{|\cup_{i=1} A_i|}{|\Omega|} = \sum_{i=1} \frac{|A_i|}{|\Omega|} = \sum_{i=1} \mu(A_i).$$

□

Thus μ is a probability measure.

Definition 2.9 (Measurable Space)**.** We say that a pair $(\Omega, \mathcal{A})$ where Ω is nonempty and $\mathcal{A} \subset 2^{\Omega}$ is a σ-algebra is a measurable space. The sets $A \in \mathcal{A}$ are called measurable. If in

addition we have chosen a measure, then we say that the triple $(\Omega, \mathcal{A}, \mu)$ is a measure space. Finally, if μ is a probability measure, then we re-label μ as $\mathbb{P}$ and call $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space. In this case we call $A \in \mathcal{A}$ events.

It is frequently difficult to connect rigorous definitions to applications. We will go through a simple example to help us stay grounded.

Example 2.10. Suppose that we wished to model a dice roll. Given a six sided die, we choose our events to be sets of possible outcomes of the die roll. Thus, we have

$$\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}.$$

We let $\mathcal{A} = 2^{\Omega}$. Finally, let μ be the probability measure from Example 2.8.

Consider the event $\{\boxdot\}$, or rather the event where we roll the die and we get a 1. The probability of this event is given by:

$$\mu(\{\boxdot\}) = \frac{|\{\boxdot\}|}{|\Omega|} = \frac{1}{6},$$

which is what we expect. Similarly, we could look at the probability of the event $\{\boxdot, \boxdot, \boxdot\}$, which is the probability of rolling an even number, which in this case would be $\frac{1}{2}$.

2.3 Measurable Maps and Random Variables

We now introduce the definition of a measurable map. This is a map that preserves a measure structure. An interesting analogue is that of a continuous map, where open sets in the codomain have open preimages in the domain.

Definition 2.11 (Measurable map)**.** Given Ω and Ω', and σ-algebras $\mathcal{A} \subset 2^{\Omega}$ and $\mathcal{A}' \subset 2^{\Omega'}$, we say that $X : \Omega \mapsto \Omega'$ is measurable if

$$X^{-1}(A') \in \mathcal{A}$$

for any $A' \in \mathcal{A}'$. In other words, the preimage of measurable sets are measurable.

Definition 2.12 (Image Measure)**.** Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ be measurable spaces, μ be a measure on $\mathcal{A}$, and let $X : \Omega \mapsto \Omega'$ be measurable. The image measure of μ under X is given by

$$\mu(X^{-1}) : \mathcal{A}' \mapsto [0, \infty]$$

where

$$A' \mapsto \mu(X^{-1}(A'))$$

where $A' \subset \Omega'$.

We now define a random variable and the distribution of a random variable. A random variable gives us a tool to encode observations in a probability space for analysis. Frequently we have observations from an experiment without access to the event itself. For example, consider a physicist using an electron detector in an experiment. Obviously, the physicist does not observe the actual event she is studying. She does, however, have the observations from her detector. She then uses the observations in her analysis of the experiment.

Definition 2.13 (Random Variable)**.** Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ be measurable spaces, and let $X : \Omega \mapsto \Omega'$ be measurable. Then X is a random variable with values in $(\Omega', \mathcal{A}')$, and we say that X is $\mathcal{A} - \mathcal{A}'$ measurable, or just $\mathcal{A}$ measurable. If $\Omega' = \mathbb{R}$, then X is a real random variable. Finally, if X is a real random variable, say that $\{X = x\} = X^{-1}(x)$ and $\{X < x\} = X^{-1}([-\infty, x))$.

Definition 2.14 (Generated σ-Algebra)**.** Let X be a $\mathcal{A} - \mathcal{A}'$ measurable random variable. Then

$$\sigma(X) = \sigma(\{\, X^{-1}(A') : A' \in \mathcal{A}' \,\}).$$

We say that $\sigma(X)$ is the σ-algebra generated by X.

Definition 2.15 (Distribution, Density)**.** Let X be a random variable and $\mathbb{P}$ the image measure under X. This is a probability measure and called the distribution of X. We say

that the function

$$F(x) = \mathbb{P}[X \leq x] = \mathbb{P}[X^{-1}([-\infty, x))]$$

is the distribution function of X. If we can write the distribution function as

$$F(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^{x} f(t)dt,$$

then we say that f is the density function.

Definition 2.16 (Joint Distribution, Joint Density)**.** We extend the definition of a distribution and a density function to more than one variable. Let $X_1, ..., X_n$ be random variables. Then

$$F(x_1, ..., x_n) = \mathbb{P}[X_1 < x_1, ..., X_n < x_n] = \mathbb{P}[\cap_{i=1}^{n} X_i^{-1}([-\infty, x_i))]$$

is called the joint distribution. If we can write the joint distribution as

$$F(x_1, ..., x_n) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f(t_1, ..., t_n)dt_1...dt_n,$$

then we say that f is the joint density function.

Definition 2.17 (Marginal Density Function)**.** If $X_1, ..., X_n$ are continuous random variables with joint density function $f(x_1, ..., x_n)$, then we say that

$$f_{X_i}(x_i) = \int f(x_1, ..., x_n)dx_1...dx_{i-1}\ dx_{i+1}...dx_n$$

is the marginal density function for X_i.

The concept of distribution, and specific distributions, will factor largely in developments in later chapters. We now describe some of the distributions of random variables that will be used in this thesis. Often in working on applications of probability theory we avoid explicitly constructing probability spaces and focus instead on random variables and their distributions.

Example 2.18 (Bernoulli Distribution)**.** Let $p \in [0,1]$, $\mathbb{P}[X=1] = p$, and $\mathbb{P}[X=0] = 1-p$. Then $\mathbb{P}_X$ is called the Bernoulli Distribution with parameter p

Example 2.19 (Categorical Distribution)**.** We can generalize the idea of the Bernoulli distribution by allowing for a countable number of points in Ω', rather than just 0 and 1. Let $k > 0$ and $p_1, ..., p_k \in [0,1]$ such that $\sum_i p_i = 1$. Then $p = [p_1, ..., p_k]$ is called a probability vector. Let $\mathbb{P}[X = i] = p_i$. Then we say that $\mathbb{P}_X$ is the categorical distribution with parameter p.

Example 2.20 (Normal Distribution)**.** Let $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, and let X be a real random variable with

$$\mathbb{P}[X \leq x] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt.$$

We call $\mathbb{P}_X$ a Gaussian or normal distribution with the parameters μ and σ^2. If $\mu = 0$ and $\sigma^2 = 1$, then we say that $\mathbb{P}_X$ is the standard normal distribution.

Example 2.21 (Dirichlet Distribution)**.** Finally, we discuss a more challenging distribution. Let α be a vector in $\mathbb{R}^k$ with $\alpha_i > 0$ for all i. Then for any vector $x \in \mathbb{R}^k$ such that $x_i > 0$ and $\sum_i x_i = 1$, we have

$$\mathbb{P}[X = x] = \frac{1}{B(\alpha) \prod_{i=1}^{k} x_i^{\alpha_i - 1}}$$

where

$$B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)}.$$

What does a graph of this probability density function look like? We provide some examples where $k = 3$ in Figure 2.1.

Note that Ω' for Dirichlet distributed random variables is the space of categorical distributions with k categories. Certain distributions have higher probabilities of occurring given our parameter vector α.

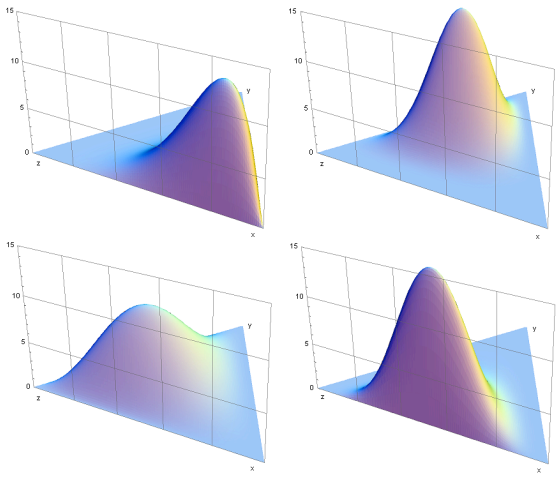


Figure 2.1: Above are four Dirichlet distributions on three dimensions, each with different parameter vector α. Note how changing this vector changes the graph. The graph on the top left has $\alpha = (6, 2, 2)$, bottom left has $\alpha = (2, 3, 4)$, bottom right has $\alpha = (6, 2, 6)$ and finally the top right has $\alpha = (3, 7, 5)$. Source: Wikipedia [3]

2.4 Independence

It is often useful to consider events and random variables that are independent from each other. In fact, Shiryaev states that independence of events is "precisely the concept that distinguishes probability from the general theory of measure spaces." [34]

We say that two events $A, B \in \mathcal{A}$ are independent if:

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B].$$

Example 2.22. Let's build an example of independence to again connect our intuition to our formal theory. We would assume that a coin flip and a die roll would be events that have little do with each other, and thus are independent. Let's check to make sure that our theory matches with our intuition. First, we construct our probability space. To model a coin flip followed by a die roll, we have the following:

$$\Omega = \{H, T\} \times \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$$

and

$$\mathcal{A} = 2^{\Omega}.$$

We will assume that the coin and die are fair, and so our probability distribution will be uniform (see Example 2.10). Now consider two events — getting a heads on our coin flip followed by rolling an even number with our die.

$$A = \{H\} \times \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$$

$$B = \{H, T\} \times \{\boxdot, \boxdot, \boxdot\}$$

A is the event of the coin flip followed by any outcome in our die roll. B is an even roll preceded by any outcome from the coin flip. Note that $|\Omega| = 12$. Thus, we have

$$\begin{aligned}
\mathbb{P}[A \cap B] &= \mathbb{P}[\{H\} \times \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\} \cap \{H, T\} \times \{\boxdot, \boxdot, \boxdot\}] \\
&= \mathbb{P}[\{H\} \times \{\boxdot, \boxdot, \boxdot\}] \\
&= \frac{|\{H\} \times \{\boxdot, \boxdot, \boxdot\}|}{|\Omega|} = \frac{3}{12} \\
&= \frac{1}{4}.
\end{aligned}$$

Which is what we expect. Now note

$$\begin{aligned}
\mathbb{P}[A]\mathbb{P}[B] &= \mathbb{P}[\{H\} \times \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}]\ \mathbb{P}[\{H, T\} \times \{\boxdot, \boxdot, \boxdot\}] \\
&= \frac{|\{H\} \times \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}|}{12} \cdot \frac{|\{H, T\} \times \{\boxdot, \boxdot, \boxdot\}|}{12} \\
&= \frac{6}{12} \cdot \frac{6}{12} = \frac{1}{2} \cdot \frac{1}{2} \\
&= \frac{1}{4}.
\end{aligned}$$

Thus we see that $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$, and the events are independent.

We can generalize this definition of independence to an arbitrary collection of events.

Definition 2.23 (Independence)**.** Let I be an index set and let $(A_i)_{i\in I}$ be a family of events indexed by I. This family is called independent if given any finite $J \subset I$ we have:

$$\mathbb{P}[\cap_{j\in J} A_j] = \Pi_{j\in J}\mathbb{P}[A_j].$$

This generalization gives us considerable power to make statements about the nature of probability. An example of an independent family of events is an indefinitely repeated experiment. See [24] Chapter 2 for a treatment of this. In addition to concrete examples, independence is crucial in the general analysis of probability. The Borel-Cantelli Lemma and the Kolmogorov 0-1 Law, for example, are important theorems that rely exclusively on the notion of independence; see [23].

The definition of independence can be further extended to σ-algebras and random variables:

Definition 2.24 (Independent σ-Algebras)**.** Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let $\mathcal{F}, \mathcal{F}'$ be sub σ-algebras of $\mathcal{A}$. We say that $\mathcal{F}$ and $\mathcal{F}'$ are independent if given arbitrary $F \in \mathcal{F}$ and $F' \in \mathcal{F}'$, we have

$$\mathbb{P}[F \cap F'] = \mathbb{P}[F]\mathbb{P}[F'].$$

Definition 2.25 (Independent Random Variables)**.** Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. Let $(X_i)_{i\in I}$ be a family of $\mathcal{A}$-measurable random variables. We say that the family is independent if the family $(\sigma(X_i))_{i\in I}$ of generated σ-algebras is independent.

We finish our introduction on probability with a brief description of expectations. A thorough treatment of the subject can be found in [33]. We only use them briefly for a few definitions and results in this thesis, and as such we simply state the definition, without rigorously defining the integral operator.

Definition 2.26 (Expectation)**.** The expectation of a random variable, denoted as $E[X]$, is

$$E[X] = \int_\Omega X d\mathbb{P}.$$

Here $d\mathbb{P}$ indicates that we are integrating with respect to the probability measure.

Chapter 3. Conditional Probabilities and Bayes Theorem

3.1 Introductory Intuition

To begin our Bayesian adventure, let us consider a very simple example. Suppose that we want to flip a coin. We hypothesize that the coin is a fair coin, i.e., that after repeated flippings the ratio of heads to tails will be around 1:1. We now test our hypothesis by flipping the coin. Our data will be the outcome of the flips, or experiments.

Let us suppose that we have flipped the coin twenty times, and each time the result has been heads. Would a reasonable person conclude that this coin is fair? While it is still possible that the coin is fair, with each flip we are becoming less and less convinced. We feel as though it is very unlikely that the coin is a fair coin after so many heads in a row. If we were playing a game of chance, we would begin to suspect that foul play was involved to skew the outcome.

Bayes' Theorem takes this "feeling" and turns it into something more formal, and rigorously analyzable. To develop these ideas further, we make a definition.

3.2 Formalizing a Feeling

Let us define the conditional probability of an event.

Definition 3.1 (Conditional Probability)**.** Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and $A \in \mathcal{A}$. We define the conditional probability given A for any $B \in \mathcal{A}$ by

$$\mathbb{P}[B|A] = \begin{cases} \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]} & \text{if } \mathbb{P}[A] > 0 \\ 0 & \text{if } \mathbb{P}[A] = 0. \end{cases}$$

We have a similar definition for density functions.

Definition 3.2 (Conditional Probability For Density Functions)**.** If X and Y are random variables, then we have

$$f\big(y|X=x\big) = \frac{f_{X,Y}(x,y)}{f_X(x)},$$

when $f_X(x) > 0$.

There are a few obvious implications that follow from our definition that are worth noting, if only because they jive well with our intuition about probability and give us confidence that our theory of probability is still good for something. First:

Theorem 3.3. *If $\mathbb{P}[A] > 0$, then $\mathbb{P}[\cdot|A]$ is a probability measure.*

Proof. First, note that

$$\frac{\mathbb{P}[\Omega \cap A]}{\mathbb{P}[A]} = \frac{\mathbb{P}[A]}{\mathbb{P}[A]} = 1.$$

Similarly, $\mathbb{P}[\emptyset \cap A] = 0$. Now we show σ-additivity. Given disjoint $\{B_i\}$, we have

$$\mathbb{P}[\cup B_i|A] = \frac{\mathbb{P}[\cup(B_i \cap A)]}{\mathbb{P}[A]} = \sum \mathbb{P}[B_i|A].$$

□

Our second result concerns independence.

Theorem 3.4. *If $A,\ B \in \mathcal{A}$ and $\mathbb{P}[A],\ \mathbb{P}[B] > 0$, then the following are equivalent:*

(i) A, B *independent*

(ii) $\mathbb{P}[B|A] = \mathbb{P}[B]$

(iii) $\mathbb{P}[A|B] = \mathbb{P}[A]$.

Proof. This follows directly from the definition of independence and conditional probability.

□

One very powerful idea that, here, seems obvious and simple, concerns independence and conditional probability. The consequences of the following provide an important analytic technique that we will make use of later.

Theorem 3.5. *If A and B are independent, then given some event C we have*

$$\mathbb{P}[A \cap B|C] = \mathbb{P}[A|B]\mathbb{P}[B|C].$$

Proof. This trivially follows from the definition of independence with the probability measure $\mathbb{P}[\cdot|C]$ □

One more direct consequence of the definition of conditional probability is the summation formula.

Theorem 3.6 (Summation Formula)**.** *Let I be a countable set and let $(B_i)_{i\in I}$ be pairwise disjoint sets such that $\mathbb{P}[\cup_{i\in I} B_i] = 1$. Then, given any $A \in \mathcal{A}$ we have*

$$\mathbb{P}[A] = \sum_{i\in I} \mathbb{P}[A|B_i]\mathbb{P}[B_i].$$

Proof. This again follows from the definition of conditional probability and the σ-additivity of measures:

$$\mathbb{P}[A] = \mathbb{P}[\cup_{i\in I}(A \cap B_i)] = \sum_{i\in I} \mathbb{P}[A \cap B_i] = \sum_{i\in I} \mathbb{P}[A|B_i]\,\mathbb{P}[B_i].$$

□

3.3 Bayes' Theorem

We saw in the last section that when we condition on data we get a new probability distribution. Also, given two independent events A and B, conditioning on B does not change the probability of A and vice versa. Thus, if we wish to further understand the probability of

an event, we must seek condition on meaningful events, or in other words, seek meaningful data. This sets the stage for Bayes' Theorem:

Theorem 3.7 (Bayes' Theorem—Discrete Version)**.** *Let I be countable and $(B_i)_{i\in i}$ be pairwise disjoint subsets of Ω such that $\mathbb{P}[\cup B_i] = 1$. Then for any $A \in \mathcal{A}$ where $\mathbb{P}[A] > 0$ and $k \in I$ we have*

$$\mathbb{P}[B_k|A] = \frac{\mathbb{P}[A|B_k]\,\mathbb{P}[B_k]}{\sum_{i\in I}\mathbb{P}[A|B_i]\,\mathbb{P}[B_i]}.$$

Proof. This follows directly from the definition of conditional probability and the summation formula. From the definition of conditional probability, we have

$$\begin{aligned}\mathbb{P}[B_k|A] &= \frac{\mathbb{P}[B_k \cap A]}{\mathbb{P}[A]} \\ &= \frac{\mathbb{P}[A|B_k]\,\mathbb{P}[B_k]}{\mathbb{P}[A]}.\end{aligned}$$

From our previous theorem, we have $\mathbb{P}[A] = \sum_{i\in I}\mathbb{P}[A|B_i]\,\mathbb{P}[B_i]$, and so we have

$$\mathbb{P}[B_k|A] = \frac{\mathbb{P}[A|B_k]\,\mathbb{P}[B_k]}{\sum_{i\in I}\mathbb{P}[A|B_i]\,\mathbb{P}[B_i]}.$$

□

Despite the simplicity of the mathematics of this theorem, the consequences are profound and far reaching. Suppose we have some hypothesis about the probability of some event B_i happening. Bayes' theorem allows us to update our belief about the probability of B_i given some data A. We are in essence building a new probability measure on Ω that is normalized by A. In other words, we are re-aligning our probability space to reflect some new information we may have discovered. Thus A becomes, in a sense, the new universal set; any event that does not allow for the event A now has probability 0.

In Bayesian data analysis, the probability measure prior to conditioning in Bayes' theorem is called the prior distribution. When building a model, a prior distribution may be chosen for the model based on prior understanding and expert opinion, before any data is collected. The

distribution that results from applying Bayes' theorem is called the posterior distribution.

Unfortunately, this formulation of Bayes' theorem is insufficient as it stands for some important kinds of analysis. We illustrate this with another example. Suppose we wish to use Bayes' Theorem to update our belief about a person's gender given their hair length. While gender is a discrete event, hair length is usually modeled using a continuous random variable. Herein lies our dilemma—the probability that a persons hair length is exactly 12.0341 inches long is 0. Yet, if we know that the length of a person's hair is 12.0341 inches long, we would be inclined to say that it is more likely that this person is a woman.

Rigorously developing the mathematics to handle this case, unfortunately, requires a good deal of formalism. Rather than explore the technical details that do little to advance our intuition or understanding, we refer the reader to [34] Chapter 2.7 for a discussion on regular conditional distribution, which is the tool needed to rigorously develop the continuous case of Bayes' theorem.

Finally, we also present Bayes' theorem for density functions.

Theorem 3.8 (Bayes' Theorem for Density Functions)**.** *If X and Y are random variables, then we have*

$$f_X\big(x \mid Y = y\big) = \frac{f_Y\big(y \mid X = x\big)\ f_X(x)}{f_Y(y)}.$$

Chapter 4. Markov Chains

4.1 Introduction

The Markov Chain Monte Carlo method is widely considered to be one of the most important algorithms of the 20th century. Its discovery, along with the rise of computation, helped transform Bayesian statistics from a fringe theory to an important and practical methodology for data analysis. The algorithm itself depends on a beautiful mathematical theory that we

will discuss in the following chapters.

In this chapter we will build and justify the essential tools that will make understanding the model possible. We will first develop a bare-bones theory of stochastic processes, and use it to define a special process called a Markov chain. In the following chapter we will develop the Markov chain Monte Carlo method.

4.2 Stochastic Process

A stochastic process gives us a way to model the evolution of random events over time. For example, one may wish to model a stock or consumer good price, growth of human males over their life span, or the parameters of a distribution given some data.

Definition 4.1 (Stochastic Process)**.** Let $I \subset \mathbb{R}$. Then a family of random variables $X = (X_t : t \in I)$ with values in $(\mathbb{R}, B(\mathbb{R}))$ is called a stochastic process.

Thus, in a stochastic process we have a real random variable associated with every time $t \in I$.

4.3 Filtrations

We now define a filtration and what it means for a stochastic process to be adapted to a filtration.

Definition 4.2 (Filtration)**.** Let $\mathbb{F} = (\mathcal{F}_t, t \in I)$ be a family of σ-algebras with $\mathcal{F}_t \subset \mathcal{F}$ for all $t \in I$. $\mathbb{F}$ is called a filtration if $\mathcal{F}_s \subset \mathcal{F}_t$ for all $s \leq t$

Definition 4.3. A stochastic process $X = (X_t : t \in I)$ is adapted to the filtration $\mathbb{F}$ if X_t is $\mathcal{F}_t$-measurable for all $t \in I$. If $\mathcal{F}_t = \sigma(X_s, s \leq t)$ for all $t \in I$, then we denote $\mathbb{F} = \sigma(X)$, the filtration generated by X.

Intuitively, we understand the filtration to signify what "can happen" as our stochastic process evolves. As this may be difficult to connect to the practical, we provide a simple example to illustrate.

Example 4.4. Consider a random walk of 3 steps. Our stochastic process is $X = (X_0, X_1, X_2, X_3)$, where $\mathbb{P}[X_i = 1] = \mathbb{P}[X_i = 0] = \frac{1}{2}$. We can explicitly construct a filtration that is adapted to this process. We define our event space as the eight possible ordered outputs from the process:

$$\Omega = \{(1,1,1),(1,1,0),(1,0,0),(1,0,1),(0,1,1),(0,0,1),(0,1,0),(0,0,0)\}.$$

At the zeroth step of our process, before anything has happened, any sequence of ones and zeros is possible. To reflect this, the corresponding σ-algebra $\mathcal{F}_1$ contains only two elements:

$$\mathcal{F}_0 = \{\emptyset, \Omega\}.$$

After the first step, we have seen either a zero or one. Our next σ-algebra reflects this by adding events—one for the possible outcomes after seeing a one first, and the other for the possible outcomes after seeing a zero first:

$$\mathcal{F}_1 = \{\emptyset, \Omega, \{(1,1,1),(1,0,0),(1,0,1),(1,1,0)\}, \{(0,0,0),(0,1,1),(0,1,0),(0,0,1)\}\}.$$

At $t = 2$, we have our second step in our random walk and we drill down even further and our σ-algebra becomes

$$\begin{aligned}\mathcal{F}_2 =&\{\emptyset, \Omega, \{(1,1,1),(1,0,0),(1,0,1),(1,1,0)\}, \{(0,0,0),(0,1,1),(0,1,0),(0,0,1)\},\\ &\{(1,1,0),(1,1,1)\}, \{(1,0,1),(1,0,0)\}, \{(0,1,1),(0,1,0)\}, \{(0,0,1),(0,0,0)\}\}.\end{aligned}$$

We omit $\mathcal{F}_3$ for the sake of not being tedious. The interested reader may construct it easily enough.

To see that X is $\mathcal{F}_i$ measurable for each $i \in I$, note that at $t = 1$ we have

$$X_1^{-1}(1, *, *) = \{(1, 1, 1), (1, 1, 0), (1, 0, 1), (1, 0, 0)\},$$
$$X_1^{-1}(0, *, *) = \{(0, 0, 0), (0, 1, 0), (0, 0, 1), (0, 0, 0)\}.$$

At $t = 2$ we have

$$X_2^{-1}(1, 1, *) = \{(1, 1, 1), (1, 1, 0)\},$$
$$X_2^{-1}(1, 0, *) = \{(1, 0, 1), (1, 0, 0)\},$$
$$X_2^{-1}(0, 0, *) = \{(0, 0, 1), (0, 0, 0)\},$$
$$X_2^{-1}(0, 1, *) = \{(0, 1, 0), (0, 1, 1)\}.$$

The $t = 3$ case is similar.

4.4 The Markov Property and Markov Chains

The statement of the Markov property is easily made; however, its applications are far reaching. Applications of Markov chains are found in social and hard sciences; see [17] for a short list. The literature that studies them as mathematical objects is also extensive. In this section we will begin with the statement of the Markov property, define Markov chains, and then investigate several interesting properties that have important consequences in applications. This will set the stage for the next chapter where we will discuss Markov chain Monte-Carlo methods.

In this chapter, $X = (X_t)_{t \in I}$ is a stochastic process on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ that is adapted to the filtration generated by the process, or $\mathbb{F} = (\mathcal{F}_t)_{t \in I} = \sigma(X)$. In full generality, we require only that the process takes values in a Polish space (see [24]). In this thesis, however, we will only concern ourselves with real valued processes. In the following E represents values that our process can take, which we call the state space. Now, we begin

with a definition.

Definition 4.5 (Stochastic Kernel)**.** Let $(\Omega_1, \mathcal{A}_1), (\Omega_2, \mathcal{A}_2)$ be measurable spaces. Let $\kappa : \Omega_1 \times \mathcal{A}_2 \to [0, \infty]$. We call κ a stochastic kernel if

(i) $\omega \mapsto \kappa(\omega, A)$ is $\mathcal{A}_1$ measurable for all $A \in \mathcal{A}_2$,

(ii) $A \mapsto \kappa(\omega, A)$ is a σ finite measure on $(\Omega_2, \mathcal{A}_2)$ for any $\omega \in \Omega_1$,

(iii) $\kappa(\omega, \Omega_2) = 1$ for all $\omega \in \Omega_1$. In other words, $\kappa(\omega, \cdot)$ is a probability measure on $(\Omega_2, \mathcal{A}_2)$ for any choice of $\omega \in \Omega_1$.

We now define the Markov property.

Definition 4.6 (Markov Property)**.** X has the Markov Property if for every $A \in B(E)$ and all $s, t \in I$ with $s \leq t$,

$$\mathbb{P}[X_t \in A | \mathcal{F}_s] = \mathbb{P}[X_t \in A | X_s].$$

The notation $X_t \in A$ means $X_t^{-1}(A)$. Also, note that $\mathbb{P}[\cdot | X_s] = \mathbb{P}[\cdot | \sigma(X_s)]$.

Intuitively, the Markov property states that our future outcomes only depend on our current situation. Indeed, if we take our $I = \mathbb{N}_0$, we have

$$\mathbb{P}[X_t = i_t | X_0 = i_0, X_1 = i_1, ..., X_{t-1} = i_{t-1}] = \mathbb{P}[X_t = i | X_{t-1} = i_{t-1}]$$

In other words, the conditional probability measure at time t depends on the value the process takes at time $t - 1$; see [24] Chapter 17 for a more general treatment.

We now define a Markov chain.

Definition 4.7 (Markov Chain)**.** Let $I = \mathbb{N}_0$. A stochastic process $X = (X_t)_{t \in I}$ is called a Markov chain with distributions $(\mathbb{P}_x)_{x \in E}$ on the space $(\Omega, \mathcal{A})$ if

(i) For every $x \in E$, X is a stochastic process on $(\Omega, \mathcal{A}, \mathbb{P}_x)$ with $\mathbb{P}_x[X_0 = x] = 1$,

(ii) The map $\kappa : E \times \mathcal{B}(E)^{\times I} \to [0, 1]$, where $(x, B) \mapsto \mathbb{P}_x[X \in B]$ is a stochastic kernel,

(iii) For every $A \in \mathcal{B}(E)$, $x \in E$, and $s, t \in I$, we have

$$\mathbb{P}_x[X_{t+s} \in A | \mathcal{F}_s] = \kappa_t(X_s, A),$$

where for every $t \in I$, $\kappa_t : E \times \mathcal{B}(E) \to [0, 1]$ is the stochastic kernel defined for $x \in E$ and $A \in \mathcal{B}(E)$, defined as:

$$\kappa_t(x, A) = \mathbb{P}_x[X_t \in A].$$

What does all this mean? We comment on each item of the definition.

(i) Given any state $x \in E$, we have an associated probability measure $\mathbb{P}_x$ such that X is a stochastic process on $(\Omega, \mathcal{A}, \mathbb{P}_x)$. Intuitively, this measure is the probability of any given event in the state space given that we are in state x.

(ii) This condition says we have a stochastic kernel associated with each state x. This is essential bookkeeping in the case that we don't have a countable state space or we wish to condition on sets of measure zero.

(iii) At any time, and given any specific state, we have a transition kernel that allows us to calculate the probability of an event A given that we are in state x.

Markov chains with a countable state space can be expressed succinctly with a matrix of transition probabilities.

Definition 4.8 (Transition Matrix)**.** A matrix p is called a transition matrix if

$$p_{(x,y)} = \mathbb{P}[X_n = y | X_{n-1} = x].$$

In other words, the matrix is the transition kernel for the Markov chain. Since the state space is countable, we can express the probability of moving from a fixed state x to any other state as an ordered categorical distribution. We call each entry (x, y) the transition probability of moving from state x to state y.

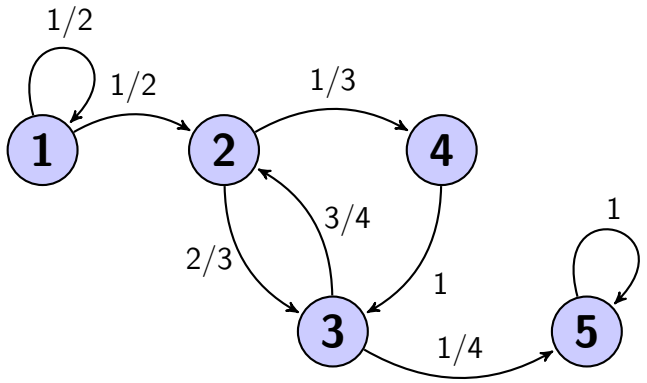


Figure 4.1: A Markov chain with 5 states. Each node represents a state and the edges represent the probability of moving from one state to another.

Theorem 4.9. *The transition matrix for a Markov chain is a stochastic kernel.*

Proof. We need to show two properties hold, namely

$$p(x, \cdot) = \kappa(x, \cdot)$$

is a probability measure and

$$p(\cdot, A) = \kappa(\cdot, A)$$

is measurable.

To show the first property, we note that

$$p(x, \cdot) = \mathbb{P}[\cdot | X_{n-1} = x] = \kappa(x, \cdot)$$

which is a probability measure. Since $p(\cdot, y) = \kappa(\cdot, y)$ is a countable function, it is clearly measurable. Thus the transition matrix for a Markov chain is a stochastic kernel.

□

These definitions are perhaps further illuminated when we describe Markov chains graphically. We do this by representing states as nodes and edges to represent transition probabilities from one state to another. For example, see Figure 4.1.

The transition matrix for the Markov chain in Figure 4.1 is

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} & 0 \\ 0 & \frac{3}{4} & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

4.5 Properties of Markov Chains

There are properties of Markov chains that are desirable in applications. In this section, we will define conditions on a Markov chain that will eventually allow us to describe its long term behavior. In the remainder of this thesis, we will consider Markov chains with countable state space E and transition matrix p.

Definition 4.10 (Entrance Time)**.** For any $x \in E$, let $\tau_x := \inf\{n > 0 : X_n = x\}$. Then the random variable τ_x is the entrance time of X for x, or in other words, the time when the Markov chain X enters state x.

We now define a function $F : E \times E \to [0, 1]$ by

$$F(x, y) = \mathbb{P}_x[\tau_y < \infty] = \mathbb{P}_x[X_n = y \text{ for some } n < \infty].$$

In other words, F is the probability that we will ever arrive at state y from state x.

We now build define properties that states of a Markov chain may have. These will be used in subsequent sections for proofs about the long term behavior of Markov chains.

Definition 4.11 (Properties of states)**.** A state $x \in E$ is

(i) recurrent if $F(x, x) = 1$,

(ii) positive recurrent if $E[\tau_x] < \infty$,

(iii) transient if $F(x,x) < 1$,

(iv) absorbing if $p(x,x) = 1$.

A Markov chain where every state is recurrent is called recurrent. If every recurrent state is absorbing, then we say that the Markov chain is transient.

These properties figure prominently in Section 4.6 where we analyze the long term behavior of Markov chains.

Definition 4.12 (Irreducible)**.** We say a Markov chain is irreducible if $F(x,y) > 0$ for all $x, y \in E$.

Example 4.13 (Recurrent Markov chain)**.** See Figure 4.2.

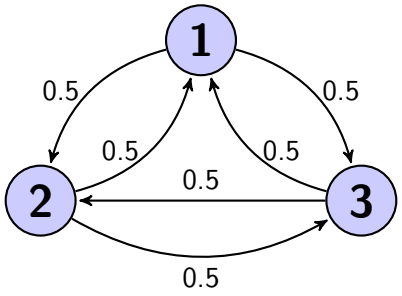


Figure 4.2: A recurrent Markov chain. With probability one, each state will be visited an infinite number of times.

Example 4.14 (Transient Markov chain)**.** See Figure 4.3.

Example 4.15 (A reducible Markov Chain)**.** See Figure 4.4.

We now define a function that will be of great use to us in analyzing some of these properties.

Definition 4.16 (Green function)**.** We let $\mathbf{E}_x$ be the expectation with respect to the probability measure at state x. Let $\mathbf{1}_{(\cdot)}$ be the characteristic function. Let $N(y) = \sum_{n=0}^{\infty} \mathbf{1}_{X_n = y}$, or the total number of times the chain visits state y. Thus,

$$G(x,y) = \mathbf{E}_x[N(y)] = \sum_{n=0}^{\infty} p^n(x,y).$$

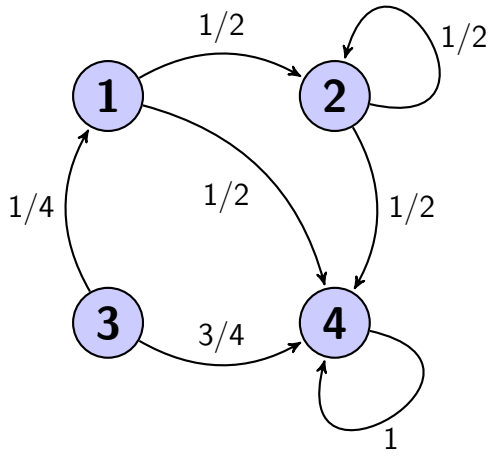


Figure 4.3: A transient Markov chain. When the process enters state 4, it will never visit another state. State 4 is the only recurrent state, and so the process is transient.

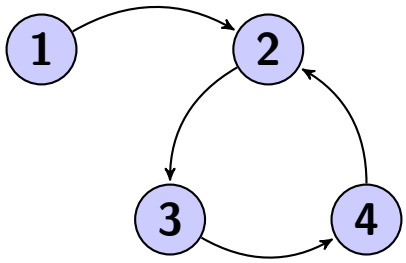


Figure 4.4: A reducible Markov chain. If we remove state 1 then the Markov chain is irreducible.

We call $G(x, y)$ the Green function of X.

Clearly, a state x is recurrent if $G(x, x) = \infty$ and vice versa.

The Green function gives us the expected number of times we will arrive at state y from state x. We will use this function as a tool to rigorously prove some intuitive results about Markov chains. This function is actually the discrete case of a more general continuous case that can be used to study Markov processes with uncountable state space; see [23]. It is also interesting to note that this function is left inverse of the Laplace operator; see [10]. In other words, this function is far from a contrived tool to prove the results that follow.

Theorem 4.17. *If x is recurrent and $F(x,y) > 0$, then y is also a recurrent state and $F(x,y) = F(y,x) = 1$.*

Proof. Let $F(x,y) > 0$. Then there is a chain of states from x to y with positive probability, i.e. $x, x_1, ..., x_k = y$ such that

$$\mathbb{P}_x[X_i = x_i \text{ for all } i = 1, .., k] > 0.$$

Note that we have $p^k(x,y) > 0$ (recall that p is the transition matrix). Now, from the definition of F we have

$$1 - F(x,x) = \mathbb{P}_x[\tau_x = \infty].$$

Or in other words, $1 - F(x,x)$ is the probability that we will never return to x. Recall that x is recurrent, and so $F(x,x) = 1$. Now we have

$$\begin{aligned}
&\mathbb{P}_x[\tau_x = \infty] \\
&\geq \mathbb{P}_x[X_1 = x_1, X_2 = x_2, ...X_k = x_k, \tau_x = \infty] && \text{(Monotonicity of Probability)} \\
&= \mathbb{P}_x[X_1 = x_1, ..., X_k = x_k]\mathbb{P}_x[\tau_x = \infty | X_1 = x_1, ..., X_k = x_k] && \text{(Conditional Probability)} \\
&= \mathbb{P}_x[X_1 = x_1, ..., X_k = x_k]\mathbb{P}_x[\tau_x = \infty | X_k = x_k] && \text{(Markov Property)} \\
&= \mathbb{P}_x[X_1 = x_1, ..., X_k = x_k]\mathbb{P}_y[\tau_x = \infty] && \text{(Definition of } \mathbb{P}_y\text{)} \\
&= \mathbb{P}_x[X_1 = x_1, ..., X_k = x_k](1 - F(y,x)).
\end{aligned}$$

Now, since $\mathbb{P}_x[X_1 = x_1, ..., X_k = x_k] > 0$, we have that $1 - F(y,x) = 0$ and so $F(y,x) = 1$. Thus, there is a $l \in \mathbb{N}$ with $p^l(y,x) > 0$ and

$$p^{l+n+k}(y,y) \geq p^l(y,x)p^n(x,x)p^k(x,y).$$

This implies

$$G(y,y) \geq \sum_{n=0}^{\infty} p^{l+n+k}(y,y) \geq p^l(y,x)p^k(x,x)G(x,x) = \infty,$$

since $G(x,x) = \infty$. Therefore, $G(y,y) = \infty$ and y is recurrent.

□

We have the following useful corollary of this theorem:

Theorem 4.18. *An irreducible discrete Markov chain is either recurrent or transient. The only irreducible Markov chain with an absorbing state is the trivial one.*

Proof. This follows immediately from the previous theorem. □

We end this section by defining one more special property a Markov chain may have, namely reversibility.

Definition 4.19 (Reversibility)**.** Let π be a measure. We say that $X = (X_i)_{i \in I}$ is reversible with respect to π if

$$\pi(\{x\})p(x,y) = \pi(\{y\})p(y,x)$$

for all $x, y \in E$. We say that X is reversible if it is reversible with respect to some measure π.

Reversibility, combined with some of the properties in Definition 4.11, is a very strong property that allows us to guarantee certain kinds of long term behavior. These properties are the building blocks of the Markov Chain Monte-Carlo method.

4.6 Invariant Distributions

A common question we ask about Markov chains is given a starting distribution or a starting state, what is the long term behavior of the chain? Does starting at one state lead to different behavior than starting at another? Under what conditions can we guarantee that the long term behavior of the Markov chain is the same regardless of its initial state?

Definition 4.20. If μ is a measure, then we write:

$$\mu p(\{x\}) = \sum_{y \in E} \mu(\{y\})p(y,x)$$

if the sum converges.

Definition 4.21 (Invariant measure)**.** A σ-finite measure μ on E is called an invariant measure if

$$\mu p = \mu.$$

A probability measure that is an invariant measure is called an invariant distribution.

Theorem 4.22. *If every state of X is transient, then it has no invariant distribution.*

Proof. Since $G(x, y) < \infty$ for all $x, y \in E$, we have $p^n(x, y) \to 0$ as $n \to \infty$. Thus, for any probability measure on E we have $\mu p^n(\{x\}) \to 0$. So, μp does not converge to μ and X has no invariant distribution. □

We note that Theorem 4.22 only makes a statement regarding invariant distributions, not invariant measures in general.

Theorem 4.23. *If X is irreducible, then X has at most one invariant distribution.*

Proof. This proof is not difficult, but requires either a significant amount of space to develop some mathematics that would serve no other purpose, or even more space using definitions that we have already developed. Instead of taking up the space, we refer the reader to [8, Chapter 3.2]. □

Theorem 4.24. *If X is reversible with respect to π, then π is an invariant measure for X.*

Proof.

$$\begin{aligned}\pi p(\{x\}) &= \sum_{y \in E} \pi(\{y\}) p(y, x) \\ &= \sum_{y \in E} \pi(\{x\}) p(x, y) \\ &= \pi(\{x\}).\end{aligned}$$

□

If a Markov chain meets certain criteria, then given a starting distribution the Markov chain will converge to a unique invariant distribution eventually. Convergence of Markov chains requires more theory than we have developed here as well as a careful implementation. It is not a trivial exercise. We refer the interested reader to [24, Chapter 18.1–2].

Invariant distributions will play a critical role in the coming chapter. Suppose we wish to sample from a distribution that does not have a closed form expression. Given certain conditions, we will show that we are able to construct a Markov chain that has the same invariant distribution as the distribution from which we want to sample. While it may be impossible or difficult to sample from the original distribution, we will be able to sample from the invariant distribution of the Markov chain.

Chapter 5. Markov Chain Monte Carlo

5.1 Introduction

One of the most important applications of Markov chains are Markov chain Monte Carlo (MCMC) methods. These techniques are widely used in applications ranging from the social sciences to theoretical physics. The technique was first discovered by Metropolis et. al. [27] in 1953, followed by improvements by Hastings [21] in 1970. It has been listed as one of the ten most important algorithms of the twentieth century [11] and the papers by Metropolis and Hastings have both been cited several thousands of times. In this chapter, we will discuss an MCMC method called the Gibbs' sampler, explicitly build its corresponding Markov chain, and then briefly connect it to its MCMC algorithm for computer simulation.

As a brief historical note, it is perhaps unfortunate that we use Metropolis's name so frequently in probability. Oral histories indicate that he had little or nothing to do with the development of the algorithm except providing computer time (see [5], [19]).

5.2 Motivation

The Gibbs' sampler has been called "...the workhorse of the MCMC world" [32]. A common problem in applications of probability is determining the set of parameters that best fit some model. For example, we may have some model that depends on parameters θ_1 and θ_2. Though we know that the model follows the behavior of certain distributions that depend on θ_1 and θ_2, we do not necessarily know what values of θ_1 and θ_2 produce the optimal model given some data. The Gibbs' sampler will provide a strategy for solving this problem.

5.3 The Gibbs Sampler

We now present the Gibbs' sampler in terms of our rigorous probability theory. Suppose that we have chosen a model for some data, and we wish to choose the optimal parameters to match the model to the data. In short, we wish to know the Bayesian posterior distribution, $\mathbb{P}[\theta|data]$, where θ is the vector of parameters. In order to do this we are going to carefully construct a Markov chain that has the same invariant distribution as this posterior distribution. We will do this by way of the Metropolis algorithm, of which the Gibbs' sampler is a special case.

Let q be the transition matrix of an arbitrary irreducible Markov chain with some state space E that has mostly zeros or low probabilities. This speeds up convergence, but it is not actually necessary; see Hastings' paper [21]. Using this matrix, we will define a new stochastic matrix called the Metropolis matrix.

Definition 5.1. Let π be the posterior distribution that we wish to sample from. Define a matrix p on E by

$$p(x,y) = \begin{cases} q(x,y)\min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right), & \text{if } x \neq y, q(x,y) > 0 \\ 0, & \text{if } x \neq y, q(x,y) = 0 \\ 1 - \sum_{z \neq x} p(x,z), & \text{if } x = y. \end{cases}$$

We call p the Metropolis matrix and q the proposal matrix. The entry (i, j) of p is called the acceptance probability for moving from state i to state j.

Theorem 5.2. *The Metropolis matrix is the transition matrix for a reversible Markov chain.*

Proof. This is a simple proof that follows directly from the definition. In the case where $x \neq y$ and $q(x, y) > 0$, we have

$$\begin{aligned}\pi(x)p(x, y) &= \pi(x)q(x, y)\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} \\ &= \pi(y)q(y, x).\end{aligned}$$

In the case when $x = y$,

$$\begin{aligned}\pi(x)p(x, y) = \pi(x)p(x, x) &= \pi(x)(1 - \sum_{z \neq x} p(x, z)) \\ &= \pi(y)(1 - \sum_{z \neq y} p(y, z)) \\ &= \pi(y)p(y, y) = \pi(y)p(y, x).\end{aligned}$$

□

So long as the irreducibility condition holds, we may choose our matrix q arbitrarily. One typical method is to use a multivariate normal distribution centered at x that gives a probability of moving to any other state. Since $p(x, y)$ is irreducible and is reversible with respect to π it follows that π is the unique invariant distribution of p by Theorem 4.24

Once we have chosen the matrix q, we may proceed with the description of the Metropolis algorithm.

Algorithm 5.3 (Metropolis Algorithm)**.**

```
input: arbitrary starting point, proposal matrix q
output: samples from posterior distribution

current_state = starting point

for each iteration
    propose new_state from q(current_state, *)
    generate acceptance probability a = p(current_state, new_state)
    current_state = new_state with probability a
    store current_state
```

This algorithm simulates the Markov chain in Definition 5.1, and it converges to the posterior distribution π. While we used π in the construction of the Metropolis matrix p, we only calculated ratios of π evaluated at two states. In practice, it is often possible to derive these ratios without calculating the posterior distribution directly.

5.4 The Gibbs' Sampler Algorithm

We will now describe the algorithm for the Gibbs' sampler and provide an example with two parameters. Following this description, we will explain how this algorithm is connected to a special case of the Metropolis algorithm.

Algorithm 5.4 (Gibbs' Sampler)**.**

```
input: prior distributions p_i, list of parameters x_i, data
output: samples from posterior distribution

for each parameter x_i:
   draw x_i0 from p_i
```

```
for j = 1 to N:
   for each parameter x_i:
      draw x_ij from p(x_i | x_1, x_2, ..., x_i-1, x_i+1,...x_k)
  \\These are draws from the conditional distribution
   store (x_1j, x_2j, ..., x_kj) as X_j
return X
```

Since our first state depends solely on our prior distributions, and since these are not the correct distribution in general, we have to wait several iterations for the algorithm to wander towards areas of high probability. This period is colloquially known as "burn-in."

It is curious that the Gibbs' sampler allows us to learn about a joint distribution by only examining its conditionals. However, work by Hammersley and Clifford [20] and Besag [6] has shown that any joint distribution is uniquely determined by its conditionals. This theorem, appropriately called the Hammersley-Clifford theorem, proved important in the development of MCMC methods, but was never published. See [31] for an excellent historic sketch of the development of MCMC methods, including the influence of this theorem.

Example 5.5 (Gibbs Sampler for two Parameters)**.** Suppose that we have some data drawn from a normal distribution. We wish to estimate the mean and standard variation of the distribution given the data. We are able to solve this problem analytically in this case, but in general this is impossible. We will use this example in order to compare the MCMC solution to the analytic solution.

A normal distribution has two parameters μ and σ^2. To make our analysis simpler, we will use $\tau = \frac{1}{\sigma}$. We assume that our parameters μ and τ have prior distributions, namely μ distributed as $\mathcal{N}(\alpha, \beta)$ and τ distributed as $Gamma(a, b)$.

We will draw μ_0, τ_0 from our prior distributions. This draw corresponds to the initial state of our Markov chain. Now, we calculate the conditional distribution of μ given τ and

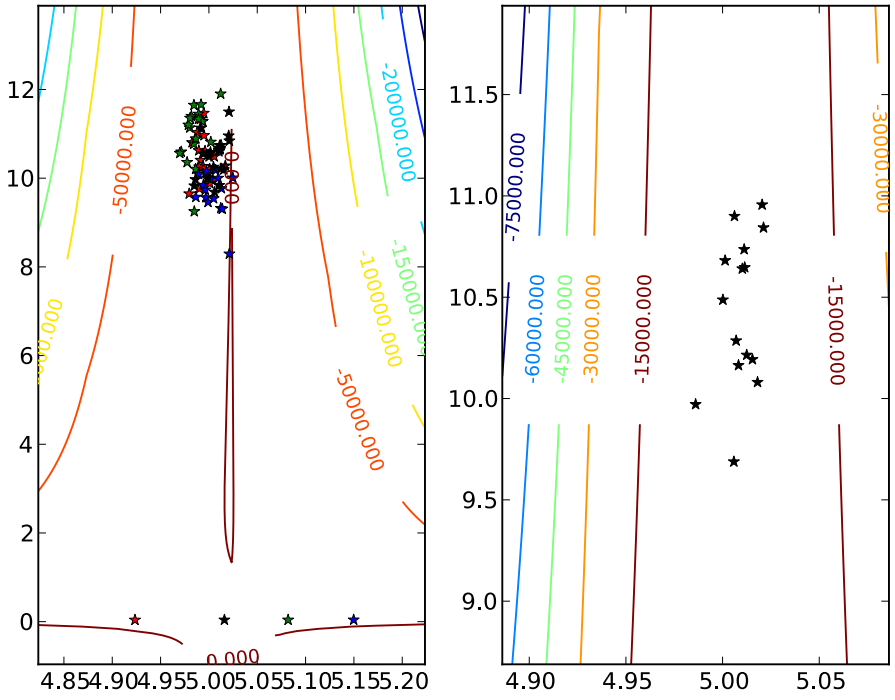


Figure 5.1: The left plot has choices for μ on the x-axis and choices for τ on the y-axis. The contour lines show the log probability of the analytic solution for the posterior probability, i.e. $\mathbb{P}[\mu, \tau | data]$. The different colors represent different runs of the sampler starting from different initial states. On the right hand side we see the last points drawn from the Gibbs' sampler for one chain. The contour lines give us level sets of the function $\mathbb{P}[\mu, \tau | data]$. The Gibbs' sampler samples area of high probability most since the Markov chain it simulates has the same the posterior distribution for its invariant distribution.

τ given μ for the Gibbs' sampler. We have

$$\mathbb{P}[\tau|\mu] \propto \Gamma\left[\alpha + \frac{n}{2},\ \beta + \frac{1}{2}\sum_{x \in data}(x-\mu)^2\right]$$

and

$$\mathbb{P}[\mu|\tau] \propto \mathcal{N}\left(\frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0},\ \sqrt{\frac{1}{n\tau + \tau_0}}\right),$$

where n is the number of data points that we have and $\bar{x}$ is the mean of the data. Figure 5.1 shows several points saved from a Gibbs' sampler using these conditional distributions.

How is this algorithm a special case of the Metropolis algorithm? How are we proposing

transitions? Which do we reject? In fact, the Gibbs' sampler leverages a carefully chosen proposal matrix so that we accept every proposition. First we choose an arbitrary ordering of the parameters from the distribution that we wish to sample. Then for each choice of parameter in order, let x^- be the other parameters. Now we choose the matrix q such that

$$q(x, y) \propto \pi(y|x^-, data))$$

where x and y represent states, or some choice of parameters. The state y will be the same as the state x except for the new parameter being chosen. Thus, when we calculate the Metropolis matrix p using Definition 5.1, we have

$$p(x, y) = q(x, y)\frac{\pi(y)\pi(x|x^-, data)}{\pi(x)\pi(y|x^-, data)}.$$

The entry $q(x, y)$ represents a proposed transition from state x to state y, and so $\pi(\cdot)$ is conditioned on x^- as well. This gives us an acceptance probability of

$$\frac{\pi(y|x^-, data)\pi(x|x^-, data)}{\pi(x|x^-, data)\pi(y|x^-, data)},$$

which evaluates to 1. Thus the Gibbs' sampler is a Metropolis algorithm where every proposed transition is accepted. The Markov chain we have built must converge for the same reasons it converges for the general Metropolis case (see above discussion about reversibility). We have not delved deeply into some of the details regarding the Hammersley-Clifford theorem and the Gibbs'-Markov equivalence. The theory necessary to be fully rigorous is admittedly difficult. The best treatment we have seen so far is in Chapter 9 and 10 of Casella and Roberts [32], but even they skip several details.

Chapter 6. Bayesian Networks

6.1 Introduction and Motivation

A Bayesian network is a directed acyclic graph that connects parameters by their probabilistic dependencies. For example, if some parameter x has a prior distribution with parameters y and z, then y and z are dependencies of x. Specifically, nodes represent random variables and edges represent probabilistic dependencies. These objects do have a rigorous definition (see [25, Chapter 2]), but for our purposes they are best understood intuitively. We begin with a simple example of a Bayesian network. Then we will continue with an explanation of a minor tweak on Bayesian networks called Bayesian plate notation. We will also provide an example using Bayesian plate notation.

6.2 Definition

Though these objects are best understood intuitively for our purposes, we offer here a partial definition to help our intuition.

Definition 6.1 (Bayesian Network)**.** We say a directed, acyclic graph is a Bayesian network if the following hold:

(i) Each node has an associated random variable X_i,

(ii) Each node with no parents has an associated probability distribution that is the distribution of the random variable, $\mathbb{P}[X_i]$,

(iii) Each node with a non-empty parent set $Pa_{X_i} = \{X_1, .., X_d\}$ has an associated probability distribution $\mathbb{P}[X_i|X_1, ..., X_d]$.

Thus, the joint distribution of the network can be written as:

$$\prod_i \mathbb{P}[X_i|Pa_{X_i}],$$

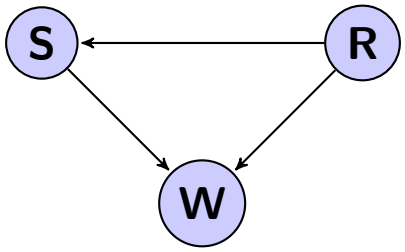


Figure 6.1: A simple Bayesian network that can be used to determine the probability that the grass is wet given information about sprinklers and rain. Node R is the random variable telling us if it has rained or not, S tells us if the sprinklers have been turned on, and W tells us if the grass is wet.

where $\mathbb{P}[X_i|Pa_{X_i}] = \mathbb{P}[X_i]$ when Pa_{X_i} is empty.

Why do we build Bayesian networks? A network carries within it all the information required to construct a joint probability distribution. However, if we were to examine just the joint distribution, we would not immediately see the dependency structure that is immediately communicated via the graph structure. It is easier for a human to understand nodes and edges than a high dimensional function.

6.3 Some Examples

We begin with a trivial example to clarify the intuitive notions we developed in the introduction.

Example 6.2 (Is the grass wet?)**.** Let us consider a classic example from probability—is the grass wet or isn't it? Suppose for a moment that the only way that grass could get wet is from a sprinkler or from rain. Further suppose that there is a probability of rain, a probability of using sprinklers given that it has rained, and a probability of using sprinklers given that it hasn't rained. We may encode all of this information in a Bayesian network; see Figure 6.1.

We may use this network to answer questions about the random variables associated with the nodes. For example, what is the probability that it has rained given that the grass is wet?

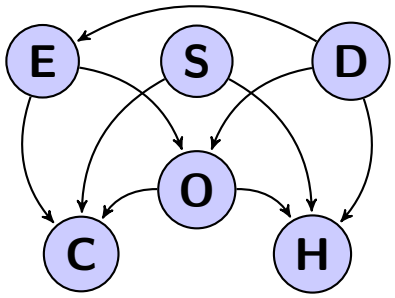


Figure 6.2: A more complicated example of a Bayesian network. This gives a model for some of the most common chronic conditions in the United States based on lifestyle choices. We see that the dependencies are more complicated than in the previous example, but we are still able to succinctly communicate some of the behavior of the model. Should we choose distributions for the random variables, we would also be able to answer questions about the probability of a person getting sick given their lifestyle choices.

$$\mathbb{P}[R = T|W = T] = \frac{\mathbb{P}[R = T, W = T]}{\mathbb{P}[W = T]} = \frac{\sum_{S=\{T,F\}} \mathbb{P}[R = T, S, W = T]}{\sum_{S,R=\{T,F\}} \mathbb{P}[S, W = T, R]}.$$

If we were to choose distributions for each random variable, we could calculate a number for this probability; see [2] for a complete example with this same network.

Example 6.3 (Modelling Sickness)**.** Bayesian networks allow for much more complicated systems than our previous example. We will consider a more complicated model for disease occurrence. Suppose we wished to connect lifestyle decisions like diet (D), exercise (E), and smoking (S) to major long term health problems—heart disease (H), cancer (C), and obesity (O). Lifestyle decisions like smoking, diet, and exercise affect the likelihood of each of these problems. Also, obesity occurring affects the probability of cancer and heart disease; see Figure 6.2.

6.4 Bayesian Plate Notation

Bayesian plate notation is a Bayesian network that includes some shorthand for easier expression of more complicated models. Anything written with Bayesian plate notation can be written as a Bayesian network. However, the new notation allows for much more compact

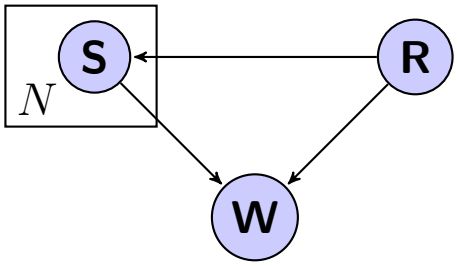


Figure 6.3: Bayesian plate notation. Instead of drawing N distinct sprinkler nodes, we draw a box around them and label it with an N.

expression without sacrificing information. The notation is especially useful for describing models that generate some data.

Example 6.4. Bayesian plate notation uses boxes, or plates, to consolidate many random variables into one. Let us expand on our previous grass example. Suppose that we need to water grass over some extremely wide area, so wide that some parts of the lawn might get wet in a rain storm and others not. Say that we need N sprinklers to cover this whole lawn. When N is large, expressing the joint distribution on grass wetness becomes a tedious exercise, as we would need to draw N nodes, each affected by R in different ways. It also defeats the purpose of easily and succinctly describing a model. Using Bayesian plate notation, this same model would be written as in Figure 6.3.

Chapter 7. Latent Dirichlet Allocation and Topics Over Time

7.1 Introduction

In this chapter, we develop models for generating corpora of documents. The models we discuss admit several parameters that allow for different kinds of data to be generated. A common task then is to discover the best choice of parameters for a model given data. These parameters may then give us insight into the data.

Parameter	Meaning
N	Number of documents
M	Number of topics
V	Number of unique words in corpus
N_i	Number of words in document i
α	Dirichlet prior on document topic distribution
β	Dirichlet prior on topic word distribution
θ_i	Topic distribution for document i
ϕ_k	Word distribution for topic k
z_{ij}	Topic for word j in document i
w_{ij}	Word j in document i

Table 7.1: Parameters for Latent Dirichlet Allocation generative model. See Chapter 1 for a refresher on how Dirichlet distributions work.

7.2 Latent Dirichlet Allocation

The Bayesian networks in the previous chapter were simple and illustrative. We now consider a model called Latent Dirichlet Allocation (LDA) that has enjoyed a great deal of success since its original publication [7]. LDA is a simplified explanation of how documents are generated from a mixture of topics. When Blei et. al. first published the technique, it was used to model the generation of corpora of documents based on topics that are determined by word co-occurrence. This understanding of how data is being generated allows us to use Gibbs' sampling to discover hidden information. In the following example we will describe how this may be done.

Example 7.1 (Latent Dirichlet Allocation)**.** First we will describe the parameters of the model, then give its plate notation, and finally describe the generative process. A corpus is a collection of documents. Given that we wish to generate N documents from K topics, we have the parameters in Table 7.1.

Using these parameters, we can express the generative model using a Bayesian plate notation; see Figure 7.1.

How, then, are data generated by this model? The step by step process by which documents are generated is as follows:

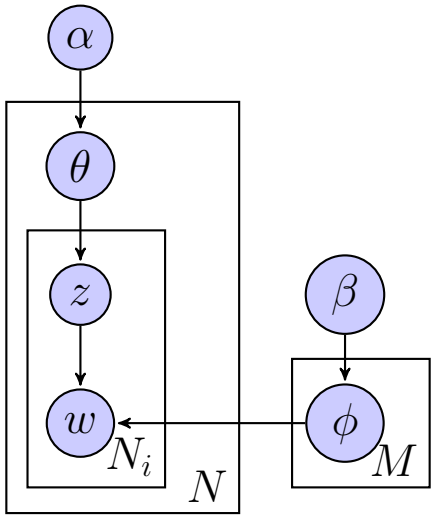


Figure 7.1: Bayesian network describing how LDA generates a corpus of data with dirichlet priors α and β. See Chapter 1 for a description of Dirichlet distributions. See Algorithm 7.2.

Algorithm 7.2.

```
For each document i, choose a topic distribution theta_i from Dir(alpha)

For each topic k, choose a word distribution \phi_k from Dir(beta)

For each of the N_i words in document i:

    Choose topic z_ij from theta_i

    Choose word w_ij from phi_(z_ij)
```

Of particular interest to us is using the information available to us from the generative process and our choices of prior distributions to determine the word distributions for each topic. In other words, given a topic i, what are the words that are most likely to be sampled? This is inferable from data and the model. The derivations are quite lengthy and technical, and we will forgo the calculation here. See [7] for a long treatment and [18] for a much more approachable, but wordier and lighter, derivation. We may use these derivations to infer the posterior distributions for our topics, i.e.

$$\mathbb{P}[z|\theta, \phi, \alpha, \beta, data]$$

parameter	meaning
m_{dz_i}	Number of words in document assigned to topic z_i
$n_{z_i w}$	Number of times word w is assigned topic z_i
n_{z_i}	Total number of words assigned topic z_i

Table 7.2: Parameters specific to LDA posterior calculation

where our data are the words in a corpus of documents. We are able to write this distribution down in a general sense, but we are incapable of calculating it. (See again [18]). Fortunately, we can use Bayes' rule and Gibbs' sampling to sample from the distribution. Though we don't know what the probability of a topic given a word is, we do know the probability of a topic and what the topic of a word given a topic is (see Figure 7.1). We are thus able to find conditional distributions:

$$\mathbb{P}[z_i = j | z_{-i}, \theta, \phi, \alpha, \beta, data],$$

where z_{-i} represents the topic assignments to the other words in the data. The result of the calculation of this conditional distribution gives us

$$\mathbb{P}[z_i = j | z_{-i}, data] \propto (m_{dz_i} + \alpha) \frac{n_{z_i w} + \beta}{n_{z_i} + V\beta}.$$

See Table 7.1 for an explanation of parameters.

Thus, we may follow the following procedure to calculate the topic distributions:

Algorithm 7.3 (Gibbs' Sampler to Discover Topic Distributions)**.**

```
For each word in each document, assign a random topic
For each iteration
    For each document
        For each word
            draw new topic from conditional distribution
            update parameters
```

After burn in, we can iterate until we have sufficient samples to construct the word distribution for each topic.

7.3 Topics Over Time

We now introduce a modification to the LDA model that incorporates time. This model was first implemented by Wang and McCallum in 2006 [37]. This model is non-Markovian, but still seeks to incorporate some sort temporal information to produce better topic distributions. If, for example, we were to run LDA on a corpus of American State of the Union speeches, we would expect to see topics relating to war, economic crises, and social policy. Is there a way to further distinguish these topics so that wars in the nineteenth century are put into different topics than twenty-first century conflicts? Wang and McCallum's Topic Over Time (TOT) model seeks to do just that.

We alter the LDA model slightly and attempt to infer topics based not just on which words occur together, but also when they occur together. The TOT algorithm perhaps could be improved by a modification that makes it Markovian. This could possibly improve the description of the evolution of topics over time.

Again, we begin by naming parameters. Given that we wish to generate N documents from K topics, we have the parameters in Table 7.3.

The plate notation for the model is seen in Figure 7.2.

The data is generated by TOT similarly to the way LDA generates data:

Algorithm 7.4 (Topics Over Time)**.**

```
For each document i, choose a topic distribution theta_i from Dir(alpha)
For each topic k, choose a word distribution phi_k from Dir(beta)
For each of the N_i words in document i:
    Choose topic z_ij from theta_i
    Choose word w_ij from phi_i
    Choose timestamp t_ij from psi_(z_ij)
```

Parameter	Meaning
N	Number of documents
M	Number of topics
V	Number of unique words in corpus
α	Dirichlet prior on document topic distribution
β	Dirichlet prior on topic word distribution
θ_i	Topic distribution for document i
ϕ_k	Word distribution for topic k
z_{ij}	Topic for word j in document i
w_{ij}	Word j in document i
ψ_z	Beta time distribution for topic z
t_{ij}	Timestamp for word j in document i

Table 7.3: Parameters for the Topics Over Time generative model. These parameters are exactly the same as those that we have in table 7.1, except with the addition of the temporal parameters found in the last two entries.

Fortunately, adding temporal data does not dramatically alter the conditional probability of a topic against other topics. We have

$$\mathbb{P}[z_i = j | z_{-i}, data] \propto (m_{dz_i} + \alpha - 1)\ \frac{n_{z_i w} + \beta - 1}{n_z + V\beta - 1}\ \frac{(1 - t_{w_i})^{\psi_{z_i 1} - 1}\ t_{w_i}^{\psi_{z_i 2} - 1}}{B(\psi_{z_i 1}, \psi_{z_i 1})}.$$

Parameter	Meaning
m_{dz_i}	Number of words in document assigned to topic z_i
$n_{z_i w}$	Number of times word w is assigned topic z_i
n_{z_i}	Total number of words assigned topic z_i
t_w	Timestamp of word w
$\psi_{z_i 1}$	First parameter for beta time distribution associated with topic z_i
$\psi_{z_i 2}$	Second parameter for beta time distribution associated with topic z_i

Table 7.4: Parameters for the Topics Over Time conditional probability calculation.

We reiterate that the derivation of this conditional distribution is quite tedious and space consuming. Full derivation can be found in Wang and McCallum's paper. See Table 7.4 for an explanation of the parameters in the TOT conditional probability calculation. The Gibbs' sampling algorithm for TOT then becomes:

Algorithm 7.5 (Gibbs' Sampler for TOT)**.**

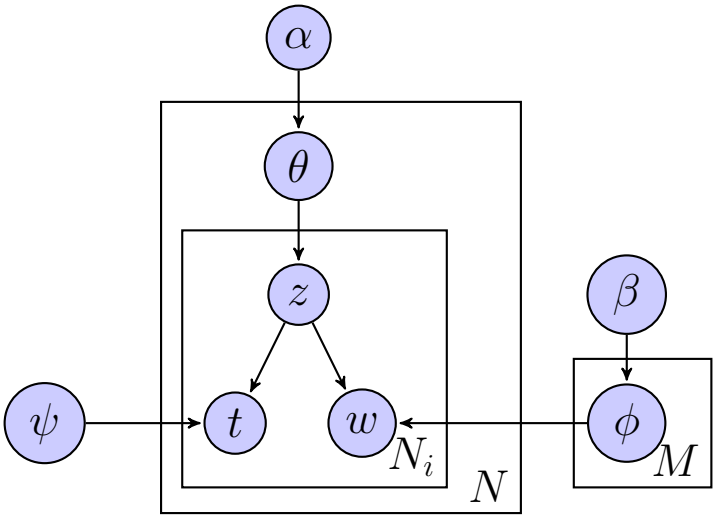


Figure 7.2: Bayesian network describing how TOT generates a corpus of data; see Algorithm 7.4.

```
For each word in each document, assign a random topic
For each iteration
    For each document
        For each word
            draw new topic from conditional distribution
            update parameters
For each topic:
    update topic time distribution
```

We have not yet determined how we will update the topic time distribution. Wang and McCallum suggest using a method of moments calculation to re-estimate the parameters after each sweep of the sampler. This update will then alter how the conditional probabilities are calculated in the next sweep, which will in turn alter the topic time distributions. While this indeed appears to be a coupled Markov chain, there has been no work done thus far to try to provide rigor for this heuristic. The method of moments technique updates the time parameters as follows:

$$\psi_{z1} = \bar{t}_z \left(\frac{\bar{t}_z(1 - \bar{t}_z)}{s_z^2} - 1 \right),$$

$$\psi_{z2} = (1 - \bar{t}_z) \left(\frac{\bar{t}_z(1 - \bar{t}_z)}{s_z^2} - 1 \right).$$

7.4 Interpreting topics

Though the only guide for determining which terms are assigned which topics are given by the data at hand, a human being must determine how to interpret a given topic distribution. For example, after we have run a Gibbs' sampler on the data from a corpus of political speeches, we will have several topic distributions with different words being most likely in each one. If there were a topic where words like 'stimulus', 'confidence', and 'rates' were very likely, we would interpret that topic to be about the economy. Another topic might have words like 'terrorism', 'weapons', and 'threat'. We would interpret this topic to be relating to terrorism and modern conflict.

Chapter 8. Methodology and Results

8.1 Overview

As mentioned in our introduction, health insurance companies have an interest in measuring the riskiness of a small population. These markets represent the majority of workers with employment based insurance. It is also more volatile than other health insurance markets. Consider Figure 8.1, which shows the distribution of the monthly log-costs of health insurance claims for individuals in a population of 60000 adult males in a metropolitan area of the United States. While the mode of the data is on the order of 100 dollars, there is a fat-tail of thousands of claims that cost more than 1000 dollars. Some claims reach into the 100000 dollar range.

If we wished to model the costs for a large population, we can expect to have enough healthy individuals to subsidize care for the unhealthy who incur expensive treatment. These populations are easier to price in part because the variance of a population tends to decrease

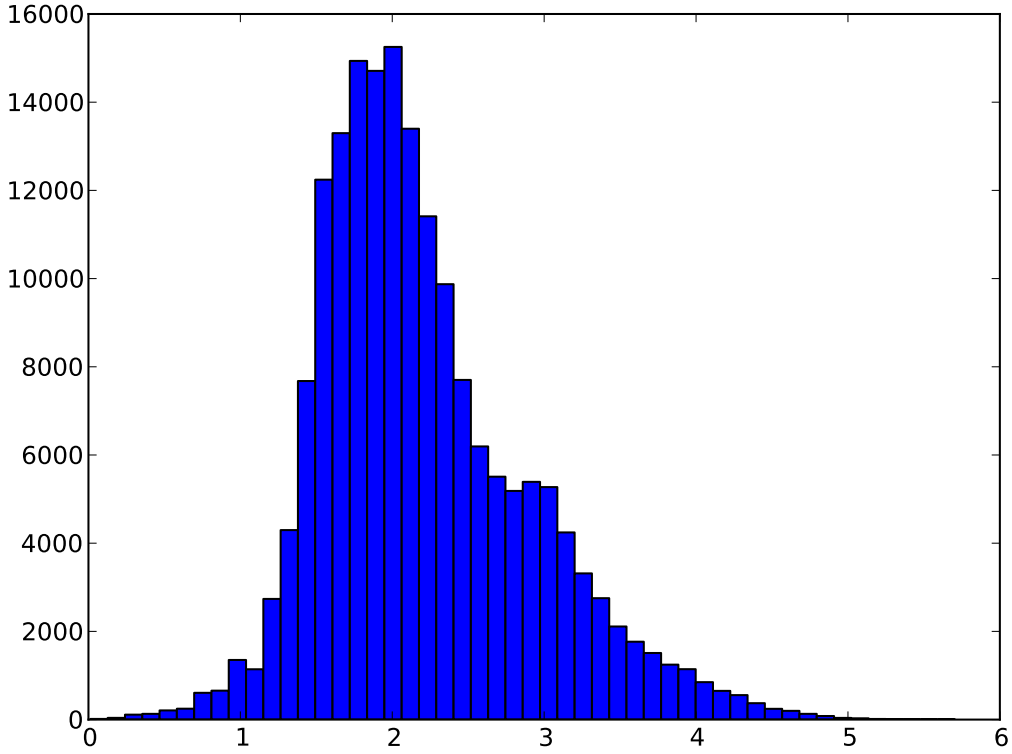


Figure 8.1: This Figure shows the non-zero monthly costs of health care claims from individuals in a sample population of 60000 adult males. The x-axis is the log-costs of claims, and the y-axis is the number of claims at cost x. These are also called Per-Member Per-Month (PMPM) costs. Note the fat tail and skew to the right. This distribution makes it particularly difficult to predict future claims of small groups because of the relative high frequency of very expensive individuals.

as the size of the population increases (see Introduction). Conversely, it is difficult to predict costs for a small population due to its higher variance. It is also more difficult for the healthy to subsidize the unhealthy in a small population.

In this chapter we will discuss implementing the mathematics developed in this thesis to create a model for the financial risk of a small population of which we know relatively little.

8.2 The Data

We begin tackling our problem of predicting risk by describing the data at our disposal. In partnership with a firm in the health care industry, we have obtained the anonymized

health records of several hundred thousand people from a major metropolitan area of the United States. We will use these data to hypothesize the underlying rules of how people consume healthcare. Combining this with known outcomes we build a model for assessing the riskiness of an individual or group.

Health records are complex. A simple doctors visit may entail pages of forms documenting procedures that occurred, prescriptions prescribed, assessments of a patients health, etc. Insurance companies require documentation from doctors with whom they do business. The data available to us are the coded information on these forms. There are several types of data that can be extracted from these forms. Of interest for our purposes will be the diagnosis and procedure codes attached to billable medical events. Insurance companies negotiate with providers what they will pay for a given service, and they use diagnosis and procedure codes as the key by which they calculate their payments. In other words, the insurance companies agree to pay a certain amount for each code, and they will only pay for diagnoses that are recorded by health care providers. We thus have reason to hope that these data are a good way to learn about how people consume health care.

The diagnosis and procedure codes come from the International Classification of Diseases (ICD). Since 1948, the World Health Organization has maintained this index of medical conditions and procedures. Originally created to measure causes of death across nations, this classification system is also used by hospitals and insurance companies to contract how much care will cost. These codes are quite extensive and can be used to precisely describe an injury, sickness, medical procedure, etc. In the version of the ICD that we will be using, called the ICD-9, these codes are five digits. The first three digits are a prefix designating a specific etiology, or causation. The remaining digits, separated by a decimal point, are used to specify the specific part of the anatomy being affected. For example, all codes with prefixes 800-999 are used to code injuries and poisoning. 820-829 describe fractures of lower limbs, and 820 describes the fracture of the neck or femur. Finally, 820.22 codes a closed fracture of subtrochanteric section of neck or femur. There are around 13000 dignosis codes

total in the ICD-9. Procedure codes are similarly specific, but follow a different scheme.

We are not overly concerned with the exact mechanics of how these codes are produced. For our purposes, we treat the process that generates them as a measurable map, or rather a random variable, mapping medical events to codes. We will add another layer of encoding by creating a bijection from the codes to a subset of the positive integers for convenience. In summary, we are defining a random variable X to model a medical event:

$$X : \text{ Medical Events } \rightarrow \text{ Diagnosis Codes.}$$

Given the information available, we will define a member of a health insurance plan as a collection of random variables, one for each health event. Our task is to use our past experience, or in other words, our data, to predict which members will have particularly costly health event. Given the large volume of treatment codes, however, a model attempting to explain costs based on the occurrence of any single diagnosis code would be at the same time noisy, sparse, and computationally prohibitive to implement. To mitigate these daunting obstacles, we first discover structure within the treatment codes and exploit this to build a simplified, yet useful model.

8.3 The Model

We wish to reduce the complexity of determining which individuals in a group are the most costly. We will do this by implementing the topics over time model described in chapter 7. Rather than describing an individual as a collection of various diagnosis codes, we will assign a topic to each diagnosis code that is assigned to the individual and thus describe an individual as a collection of topics. Also associated with each individual claim is the age of the patient when the claim was made. This is not so easy as going through the data and assigning a single topic to each diagnosis code. While the ICD-9 is quite extensive and precise, there are factors that we wish to take into account that are orthogonal to its

descriptive power. For example, a young person being diagnosed with pneumonia presents a different risk to an insurer than a senior citizen who has a weakened immune system.

Topic analysis models such as TOT are a relatively recent technique for data analysis. Since LDA was first published in 2003 [37] as a technique for determining topics in text based data, several novel applications have been proposed for taming large amounts of data. See for example [38] and [36]. We propose using TOT as the generative model for health insurance claims. However, instead of considering when the claim was made, we consider how old the patient is when the claim was made. We call this adjustment Topics Over Age (TOA) rather than Topics Over Time.

Recall from Chapter 7 the topics over time model. Rather than using the process to generate documents, we generate health insurance claims data.

(i) For each Person i, choose a health topic distribution θ_i from $Dir(\alpha)$

(ii) For each health topic k, choose a diagnosis code distribution ϕ_k from $Dir(\beta)$

(iii) For each of the N_i claims for person i:

  (a) Choose health topic z_{ij} from θ_i

  (b) Choose claim code w_{ij} from ϕ_i

  (c) Choose age of person t_{ij} from $\psi_{z_{ij}}$

In other words, rather than having a distribution for each topic in a corpus of documents, we have a distribution for each health topic over a population. Where TOA for a corpus of political speeches would have topics like war and economics with their associated words, a corpus of health insurance claims data will have distributions for topics like routine care, diabetes, cancer, etc. with their associated claims codes. We are able to infer the parameters of these distributions using a Gibbs' sampler via the same routine we described in Chapter 7. Thus, we may find out what insurance claims codes are likely to be produced given that we are sampling from a specific topic.

After using a Gibb's sampler to estimate the parameters in the model, we have for each topic in our model a distribution of diagnosis codes. For each topic we have a probability of each code showing up. For example, in a chronic kidney disease topic we will have codes for dialysis, kidney failure, and muscle weakness with high probability while a code for acne will have quite a low probability. On the other hand, a topic dealing with adolescent health will have a high probability of an acne code and a low probability of a kidney failure code.

8.4 Topic Results

As mentioned in Chapter 7, though our learning is unsupervised, we require a human to examine the resulting topics and recognize the topics for what they are. See for example Figure 8.2, where we have three plots of topic distributions related to some of the most common health issues in the United States. These particular distributions were generated on a training set of 75000 adult males from our data. In each graph the x-axis represents age and the y-axis represents the number of times that a term on our training set was assigned that topic after burn in. Below each graph we have the most likely ICD-9 codes in the topic distribution as well as their description. They are arranged in order of probability.

We see that routine, common treatments are detected and grouped appropriately in this model. Along with these common topics, there are also topics that are less prevalent in the population whose distributions are much less smooth. It is possible, for example, for a small handful of people of varying ages to have a rare disease with a rigorous treatment regime. This would produce strong co-occurrences between diagnosis codes that are then detected by the Gibb's sampler. For example, from the same population as the topic distributions from Figure 8.2, we have the topic distribution in Figure 8.3.

The most likely terms in a topic distribution also give us results about the co-morbidity, or co-occurrence, of diagnoses in specific topics. For some topics the most likely diagnosis codes seemed to be unrelated. However, consulting the medical literature revealed that these seemingly unrelated diagnoses are actually connected. See Figure 8.4 for two examples of

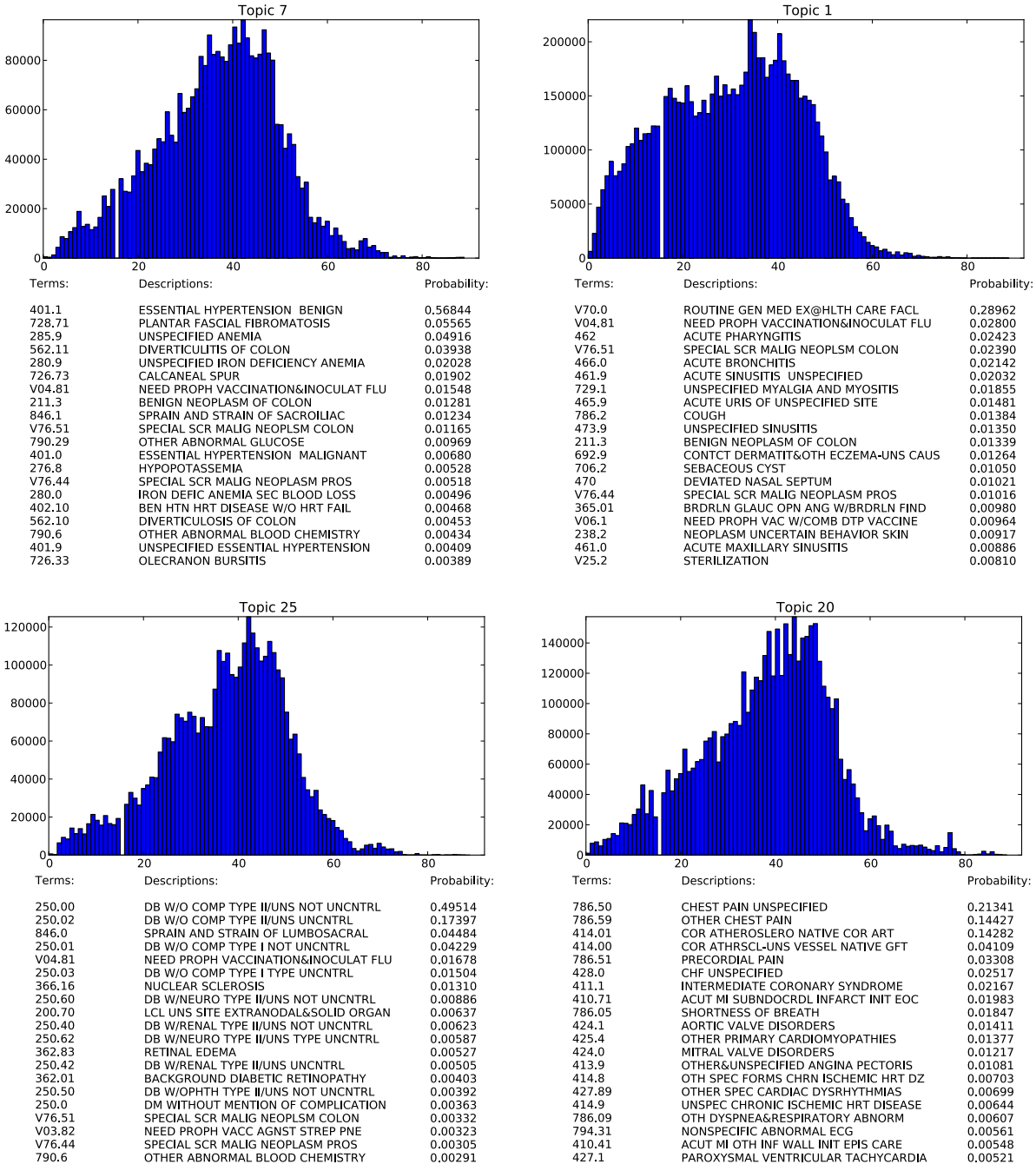
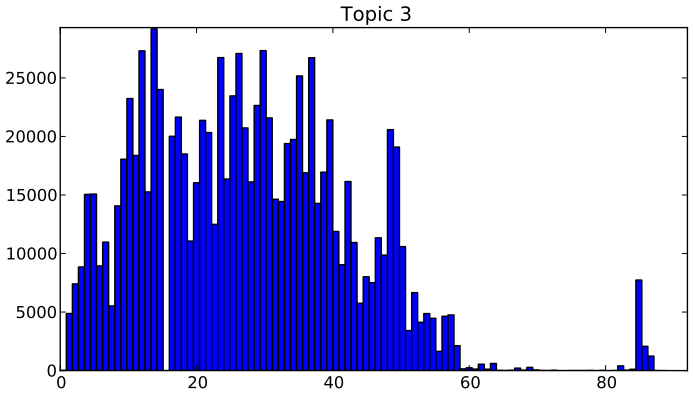
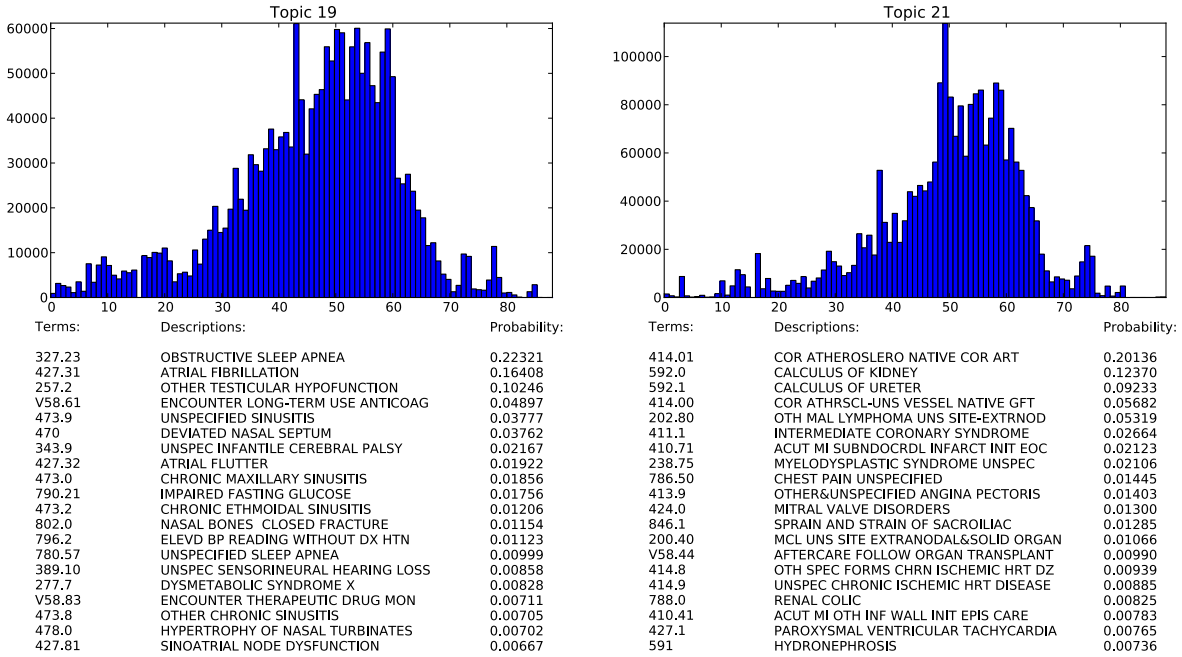


Figure 8.2: Term distributions for 4 topics. By observing these distributions, we see that the first topic relates primarily to hypertension, or high blood pressure, as well as other maladies associated with aging. The second topic, on the other hand, deals primarily with codes relating to routine medical care. Note that the y-axis for the second topic goes significantly higher than in the first topic, indicating its more common occurrence. The third topic deals with diabetes, and the fourth deals with heart disease.



Terms:	Descriptions:	Probability:
724.8	OTHER SYMPTOMS REFERABLE TO BACK	0.19238
723.3	CERVICOBRACHIAL SYNDROME	0.17735
844.2	SPRAIN&STRAIN CRUCIATE LIG KNEE	0.12293
724.1	PAIN IN THORACIC SPINE	0.09748
553.1	UMB HERN W/O MENTION OBST/GANGREN	0.05268
717.83	OLD DISRUPTION OF ACL	0.01725
552.1	UMBILICAL HERNIA WITH OBSTRUCTION	0.01663
826.0	CLOS FRACTURE 1/MORE PHALANGES FOOT	0.01513
850.5	CONCUSSION W/LOC UNSPEC DURATION	0.01395
200.83	OTH VARNTS LYMPHOSARCOMA INTRA-ABD	0.01393
823.00	CLOSED FRACTURE UPPER END TIBIA	0.01390
729.1	UNSPECIFIED MYALGIA AND MYOSITIS	0.01340
805.4	CLOS FX LUMB VERT W/O SP CORD INJR	0.01298
812.00	CLOS FX UNSPEC PART UPPER END HUM	0.01199
733.13	PATHOLOGIC FRACTURE OF VERTEBRAE	0.00951
805.2	CLOS FX DORS VERT W/O SP CORD INJR	0.00942
736.32	COXA VARA	0.00928
739.0	NONALLOPATHIC LES HEAD REGION NEC	0.00925
996.78	OTH COMP OTH INT ORTH DEV IMPL&GFT	0.00922
359.9	UNSPECIFIED MYOPATHY	0.00829

Figure 8.3: This term distribution represents more uncommon ICD-9 codes. Note the spike on the right of the graph. This represents a small number of older men producing codes that are assigned this topic. This topic treats unspecified back pain and cervicobrachial syndrome. This syndrome is a vague diagnosis that has fallen out of use and is only rarely employed by doctors, who now prefer diagnosis codes dealing with shoulder and neck pain; see [1].



Terms:	Descriptions:	Probability:
327.23	OBSTRUCTIVE SLEEP APNEA	0.22321
427.31	ATRIAL FIBRILLATION	0.16408
257.2	OTHER TESTICULAR HYPOFUNCTION	0.10246
V58.61	ENCOUNTER LONG-TERM USE ANTICOAG	0.04897
473.9	UNSPECIFIED SINUSITIS	0.03777
470	DEVIATED NASAL SEPTUM	0.03762
343.9	UNSPEC INFANTILE CEREBRAL PALSY	0.02167
427.32	ATRIAL FLUTTER	0.01922
473.0	CHRONIC MAXILLARY SINUSITIS	0.01856
790.21	IMPAIRED FASTING GLUCOSE	0.01756
473.2	CHRONIC ETHMOIDAL SINUSITIS	0.01206
802.0	NASAL BONES CLOSED FRACTURE	0.01154
796.2	ELEVD BP READING WITHOUT DX HTN	0.01123
780.57	UNSPECIFIED SLEEP APNEA	0.00999
389.10	UNSPEC SENSORINEURAL HEARING LOSS	0.00858
277.7	DYSMETABOLIC SYNDROME X	0.00828
V58.83	ENCOUNTER THERAPEUTIC DRUG MON	0.00711
473.8	OTHER CHRONIC SINUSITIS	0.00705
478.0	HYPERTROPHY OF NASAL TURBINATES	0.00702
427.81	SINOATRIAL NODE DYSFUNCTION	0.00667

Terms:	Descriptions:	Probability:
414.01	COR ATHEROSLERO NATIVE COR ART	0.20136
592.0	CALCULUS OF KIDNEY	0.12370
592.1	CALCULUS OF URETER	0.09233
414.00	COR ATHRSCL-UNS VESSEL NATIVE GFT	0.05682
202.80	OTH MAL LYMPHOMA UNS SITE-EXTRNOD	0.05319
411.1	INTERMEDIATE CORONARY SYNDROME	0.02664
410.71	ACUT MI SUBNDOCRDL INFARCT INIT EOC	0.02123
238.75	MYELODYSPLASTIC SYNDROME UNSPEC	0.02106
786.50	CHEST PAIN UNSPECIFIED	0.01445
413.9	OTHER&UNSPECIFIED ANGINA PECTORIS	0.01403
424.0	MITRAL VALVE DISORDERS	0.01300
846.1	SPRAIN AND STRAIN OF SACROILIAC	0.01285
200.40	MCL UNS SITE EXTRANODAL&SOLID ORGAN	0.01066
V58.44	AFTERCARE FOLLOW ORGAN TRANSPLANT	0.00990
414.8	OTH SPEC FORMS CHRN ISCHEMIC HRT DZ	0.00939
414.9	UNSPEC CHRONIC ISCHEMIC HRT DISEASE	0.00885
788.0	RENAL COLIC	0.00825
410.41	ACUT MI OTH INF WALL INIT EPIS CARE	0.00783
427.1	PAROXYSMAL VENTRICULAR TACHYCARDIA	0.00765
591	HYDRONEPHROSIS	0.00736

Figure 8.4: Two topic distributions from the same population. On the left, note that obstructive sleep apnea is seemingly connected to several diagnoses related to heart problems, notably atrial fibrillation. Our first inclination was to dismiss this connection. However, a review of the medical literature revealed [30], an article published in 2012 calling for more research to investigate the connection between the two diagnoses. Similarly, on the right, we see a connection between atherosclerosis of the heart, a heart condition, and calculus of the kidney, or buildup of calcium deposits in the kidney. This connection also seemed unlikely, but [26] recently detected a possible connection.

co-morbidities discovered by the Gibbs' sampler that have only recently been discovered by medical researchers. These results were unexpected and lent even more credibility to the topic distributions found by the Gibbs' sampler.

After we have used a Gibbs' sampler to determine the parameters on our topic distributions, we say that we have trained our model. We can now confidently use a new random variable in our model, mapping health events into a handful of topics rather than hundreds of thousands of ICD-9 codes:

$$X : \text{ Medical Events } \rightarrow \text{ Topic Number.}$$

Each member of the health insurance plan may be thought of as one such random variable

with a categorical distribution with unique parameters determined by their past claims and the claims of others in their sample population. Because we are interested in assessing the riskiness of individuals or small groups of individuals, we also seek to create a measure of risk over the topics. The next step will be to do this by connecting a topic to a price distribution.

8.5 Pricing The Topics

Given that an individual's medical events are being assigned a certain topic, what can we understand about their riskiness? We use the monetary cost to the insurance company as our measure of risk and establish a price distribution for each topic. In other words, we have a random variable P_t for each topic t such that

$$P_t: \text{ Medical Event } \rightarrow \text{ Prices.}$$

We assume that prices are distributed log-normally for each topic. We now create distributions for each topic based on our data. When we discussed the nature and structure of the data, we noted that each medical event is given a diagnosis code and an allowable amount representing what the insurance company is willing to pay for that medical event/diagnosis code. Since each diagnosis in our training set is assigned a topic number, and since each diagnosis code has an allowable amount linked to it, for each topic we have data points $p_1, p_2, p_3, ..., p_{n_t}$ where n_t is the number of medical events assigned topic t after the model has been trained. We use the maximum likelihood estimator for a log-normal distribution to determine the best parameters for modeling the price of a given topic:

$$\mu_t = \frac{\sum_k \ln x_k}{n_t}$$

$$\sigma_t^2 = \frac{\sum_k (\ln x_k - \mu_t)^2}{n_t}.$$

Thus for each topic t we have parameters μ_t, σ_t^2 that optimally fit the distribution to the

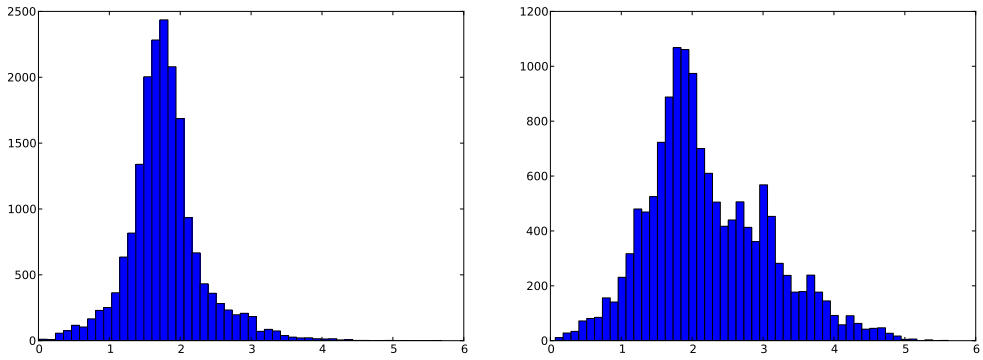


Figure 8.5: Price histograms of terms assigned a specific topic. The topic associated with the histogram on the left has a high probability of diagnosis codes dealing with minor respiratory problems and knee sprains and pains. Medical events that are assigned this topic generally cost the insurance company less than 100 dollars. On the other hand, the topic on the right has a high probability of diagnosis codes dealing with hyperlipidemia, or very high fat content in the blood. Medical events coded with this topic typically cost more than 100 dollars.

data; see Figure 8.5.

Chapter 9. Prediction and Classification

9.1 Classification

In the previous chapter we discussed discovering topic distributions for health insurance claims codes from a large population. We will call this process training and we call the data we used the training set. Armed with these topic distributions, we may classify smaller populations with a small amount of claims data by 'picking up where we left off.' For example, previously we learned topic distributions for the insurance claims codes in a population of 75000 men. Now suppose that we have a small population of men of which we have 3 months of claims data. Taking those claims, we can use the same procedure we used to learn the topic distributions to assign topics to these new claims:

Algorithm 9.1 (Classification of Claims)**.**

`For each code associated with each person, assign a random topic`

```
For each iteration
    For each person
        For each code
            draw new topic from conditional distribution
            update parameters
For each topic:
update topic time distribution
```

Assuming that these new people are from a similar population, we expect the topics to have the same parameters that we learned from the training set. Thus, after a quick burn-in, we will have assigned topics to each of the diagnosis codes for each person.

9.2 A Cost Model For Future Claims

We wish to predict how expensive an individual may be given topics present in their claims history. We present a simple model to demonstrate one way that this may be done. The model will have strict limitations that relegate this example to an illustrative case. We will discuss some of these limitations when we conclude this section.

Given a large number of people we may train a topics model to determine what sorts of patterns diseases follow in the population. Suppose then that we wish to predict future costs for an individual that matches the demographics of the group that we have trained on. As this individual makes claims, we assign them topics using the Algorithm 9.1.

We now wish to represent the individual as a mixture of topics. Recall that along with the term distributions for each topic we are also able to find a price distribution (See Figure 8.5). We model the individuals future cost by sampling from these topic price distributions proportionally to the frequency that each topic appears in his or her history. Thus we build an individuals price distribution with the following algorithm:

Algorithm 9.2.

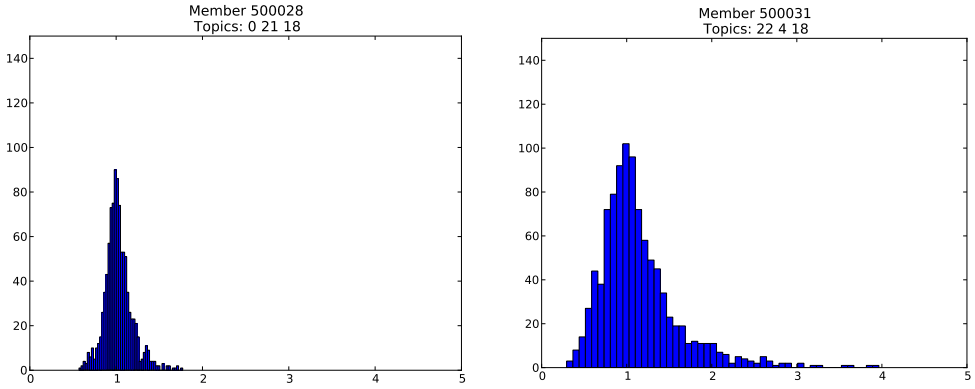


Figure 9.1: Price distributions created for two individuals based on their past claims history using Algorithm 9.2 with 1000 samples. The x-axis represents log costs and the y-axis represents the number of samples. The topics most present in the individual on the left are Type I Diabetes and Routine/General Health Examination. Both of these topics have manageable, predictable costs. On the other hand, the individual to the right typically has claims in topics dealing with Kidney Transplantation and Cystic Fibrosis. These topics are more expensive and we see that reflected in the individual price distribution which skews towards higher costs. We would expect the next claims from the individual on the right to be more expensive than the individual on the left.

```
w_0 = 0

For each topic i:

    Divide number of occurences of topic in claims history by total number of claims

    Save result as w_i


For 1 to number_of_samples:

    Draw x from a uniform distribution on [0,1]

    For i = 0 to number of topics:

        If w_i < x < w_i+1

            Draw a sample from price distribution of topic i+1
```

This process will sample from the price distribution of the individuals next claim according to the assumption of our model. See Figure 9.1 for some examples for individual price distributions using this model.

There are several limitations to this model. One important consideration that is not

modelled is how frequently individuals seek care. For example, an individual may have a relatively inexpensive condition that requires frequent visits to the doctor. This individual may become, over time, more expensive than someone that has fewer and more serious claims. In other words, in addition to modeling what sorts of claims will appear next in an individual's history, we must also model how topics affect the frequency with which individuals make claims. Modelling this correctly will be crucial to building a more robust model. We also fail to consider the way that the topics being assigned to an individual's claims may be evolving over time.

Chapter 10. Conclusion

In this thesis we proposed the Topics Over Age model for describing the process that generates health insurance claims data. This describes each person's claims history as a mixture of topic distributions. Using the Gibbs' sampler, we were able to infer distributions that predict which topic produces which diagnoses over a population.

Future work in this direction must decide how robust this model is for predicting future costs and assessing the risk of an individual or a small group of employees. There are a few directions that can be taken to do this that we are currently investigating. First, we propose describing individuals as a mixture of topics. In other words, given the topics discovered by the Gibbs' sampler, we examine an individual's initial claim history and try to predict their future costs based on topic price distributions.

Another direction we are investigating is using machine learning algorithms to train on a population's topic assignments against its health care consumption. Once trained, we attempt to predict a new population's future consumption based on a small initial history, say three to six months. We are currently using the `scikit-learn`[28] open source library for `python` to investigate this avenue, with some promising initial results. There are hurdles that need to be overcome still—insurance costs have been climbing drastically in the years

we have available in our data, and so we tend to under-predict future costs.

Both of these techniques must also consider the frequency with which individuals make claims. Understanding which topics may be chronic and requiring long term care is necessary to correctly model costs. An individual that makes a 50 dollar claim every month is more expensive in the long run than an individual that only makes one 200-dollar claim per year. Since we do not consider these usage characteristics in our model, we would predict that the individual that is making monthly claims is less of a risk than the person making the yearly claim.

Finally, though we incorporated temporal information into our Topics Over Age model, any analysis we have done on the topic distributions of individuals does not consider the temporal evolution of topics. For example, we may look at the topics of all their claims in the past year, but we do not attempt to analyze how the topics assigned to their claims evolve over time. Perhaps trying to examine a Markov structure on the claims history could provide insight into individual cost distributions.

Bibliography

[1] Cervicobrachial syndrome. http://www.mdguidelines.com/cervicobrachial-syndrome, 2012.

[2] Bayesian network. http://en.wikipedia.org/wiki/Bayesian_network, 2013.

[3] Dirichlet distribution. http://en.wikipedia.org/wiki/Dirichlet_distribution, 2013.

[4] U.S. Small Business Administration. Frequently asked questions. http://www.sba.gov/sites/default/files/FAQ_Sept_2012.pdf, 2013.

[5] Kai-Henrik Barth. Oral history transcript - dr. marshall rosenbluth. http://www.aip.org/history/ohilist/28636_1.html, 2003.

[6] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

[7] David M. Blei, Andrew Y. Yng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[8] Pierre Bremaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, 1999.

[9] Thomas C Buchmueller and Alan C Monheit. Employer-sponsored health insurance and the promise of health insurance reform. *Inquiry*, 46(2):187–202, 2009.

[10] Kai Lai Chung and John B. Walsh. *Markov processes, Brownian motion, and time symmetry*, volume 249 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, New York, second edition, 2005.

[11] Barry A Cipra. The best of the 20th century: Editors name top 10 algorithms. *SIAM news*, 33(4):1–2, 2000.

[12] Malcolm Cox, David M Irby, Molly Cooke, David M Irby, William Sullivan, and Kenneth M Ludmerer. American medical education 100 years after the Flexner report. *New England journal of medicine*, 355(13):1339–1344, 2006.

[13] David M Cutler and Sarah J Reber. Paying for health insurance: the trade-off between competition and adverse selection. *The Quarterly Journal of Economics*, 113(2):433–466, 1998.

[14] John Duffy. *From Humors to Medical Science: A History of American Medicine*. University of Illinois Press, 1993.

[15] Abraham Flexner. Medical education in the United States and Canada bulletin number four (the flexner report). *New York (NY): The Carnegie Foundation for the Advancement of Teaching*, 1910.

[16] Kaiser Family Foundation, Health Research, and Educational Trust. Employer health benefits. http://kff.org/health-costs/report/employer-health-benefits-annual-survey-archives/, 2013.

[17] Walter R Gilks, Sylvia Richardson, and David J Spiegelhalter. *Markov chain Monte Carlo in practice*, volume 2. Chapman & Hall/CRC, 1996.

[18] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 2004.

[19] JE Gubernatis. Marshall rosenbluth and the metropolis algorithm. *Physics of plasmas*, 12:057303, 2005.

[20] J Hammersley and P Clifford. Markov fields on finite graphs and lattices. *Unpublished manuscript*, 1971.

[21] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 1970.

[22] Robert B Helms. Tax policy and the history of the health insurance industry. *Using taxes to reform health insurance. Washington (DC): Brookings Institution*, pages 13–35, 2008.

[23] Olav Kallenberg. *Foundations of Modern Probability*. Springer, 2002.

[24] Achim Klenke. *Probability Theory: A Comprehensive Course*. Springer, 2008.

[25] Timo Koski and John M. Noble. *Bayesian Networks An Introduction*. Wiley, 2009.

[26] Kuanrong Li, Rudolf Kaaks, Jakob Linseisen, and Sabine Rohrmann. Associations of dietary calcium intake and calcium supplementation with myocardial infarction and stroke risk and overall cardiovascular mortality in the Heidelberg cohort of the european prospective investigation into cancer and nutrition study (epic-heidelberg). *Heart*, 98(12):920–925, 2012.

[27] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[29] Franklin Pierce. Veto message. http://www.presidency.ucsb.edu/ws/?pid=67850, 1854.

[30] Susan Redline and Stuart F Quan. Sleep apnea: A common mechanism for the deadly triadcardiovascular disease, diabetes, and cancer? *American journal of respiratory and critical care medicine*, 186(2):123–124, 2012.

[31] Christian Robert and George Casella. A short history of markov chain monte carlo: subjective recollections from incomplete data. *Statistical Science*, 26(1):102–115, 2011.

[32] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

[33] Walter Rudin. *Real and Complex Analysis*. Tata McGraw-Hill Education, 2006.

[34] A.N. Shiryaev. *Probability*. Springer, 1989.

[35] Melissa A Thomasson. From sickness to health: the twentieth-century development of us health insurance. *Explorations in Economic History*, 39(3):233–253, 2002.

[36] Xiaogang Wang and Eric Grimson. Spatial latent dirichlet allocation. *Advances in neural information processing systems*, 20:1577–1584, 2007.

[37] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM, 2006.

[38] Yi Wang, Hongjie Bai, Matt Stanton, Wen-Yen Chen, and Edward Y Chang. Plda: Parallel latent dirichlet allocation for large-scale applications. In *Algorithmic Aspects in Information and Management*, pages 301–314. Springer, 2009.