AN ABSTRACT OF THE THESIS OF

Juliana Mbuthia for the degree of Master of Science in Electrical And Computer Engineering presented on June 4, 2014.

Title: Parameter Estimation of Gaussian Hierarchical Model Using Gibbs Sampling

Abstract approved: _____

Thinh Nguyen

Gibbs sampling method is an important tool used in parameter estimation for many probabilistic models. Specifically, for many scenarios, it is difficult to generate high-dimensional data samples from its joint distribution. The Gibbs sampling provides a way to draw high-dimensional data via the conditional distributions which are typically easier to sample. In this thesis, we study a simple generative model called Hierarchical Gaussian and an efficient method for computing its parameters using Gibbs sampling. In particular, we show that the Hierarchical Gaussian model admits closed form conditional distributions such that Gibbs sampling can be used effectively to draw the samples from the joint distribution, and perform parameter estimation.

Parameter Estimation of Gaussian Hierarchical Model Using Gibbs Sampling

by

Juliana Mbuthia

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented June 4, 2014
Commencement June 2015

Master of Science thesis of Juliana Mbuthia presented on June 4, 2014.

APPROVED:

Major Professor, representing Electrical And Computer Engineering

---

Director of the School of Electrical and Computer Engineering

---

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Juliana Mbuthia, Author

ACKNOWLEDGEMENTS

The completion of this thesis would have been impossible without the help and support from all those involved

First and foremost I thank my advisor Professor Thinh Nguyen for his unmeasurable support and technical advice. It has not been easy,but he has been very patient and flexible. I also thank my thesis Committee, Raviv Raich, Charlotte Wickham and Niess Margaret, for their comments and suggestions which helped improve this thesis.

For the last few years I have been in this University I have taken classes from so many great and wonderful professors and the experience has been great.

I would like to thank Dr, Wangai Kimenju of Nairobi University Kenya for providing us with the data. I want to thank the department chair and Laurel Scholarship awarding committee for the financial support when I need it most. I salute all my other sponsors for their financial support, AAUW, Math Department, NETL/DOE,PEO, EOP, late Mario Pastega, ISFS/NIEMI etc. Special thanks to my former host family, Shelly Murphy and my dear friend Wilma Lee Hull, for making sure we were provided for and sponsoring my daughter to come to America.

I also thank my entire family, for believing in me and being there for me all this time. I am particularly indebted to my son who has endured mum's absence in style. He has been a great source of encouragement and inspiration.

Help me God, in my next assignment.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

1.1 Overview

This thesis is about probabilistic inference or parameter estimation of higher dimensional multivariate distribution, tools and techniques of computing full conditional probability as well as running a Gibbs sampling simulation.

Before computers, such higher multidimensional problems were deemed to be too complicated to deal with and were highly avoided and the model choice was usually limited to simpler model that were often too simplified to describe the real problem. During the last decades, with incarnation of computers , modern technology, and MCMC methods, such high multivariate complex problems are highly embraced and solved without fear or favor. The model we are going to be dealing with in this thesis is an example of a well known class of Bayesian graphical model sometimes called Bayesian network or Directed acyclic Graphical model, which is higher dimensional multivariate problem, with unknown full joint distribution and posterior joint distribution but with nice hidden properties. This type of a model would have been avoided some years back because of the level of complexity or would have been simplified, not any more.

1.2 About the Bayesian Network

A Bayesian graphical model(sometimes called Bayesian Network(BNW)) is a graphical structure for representing the probabilistic relationships among a large number of random variables (attributes) and for carrying out probabilistic inference with those attributes. The graphical structure of Bayesian network provides an insightful picture of the re-

lationships among the random variables or attributes. For example, from Figure 1.1,

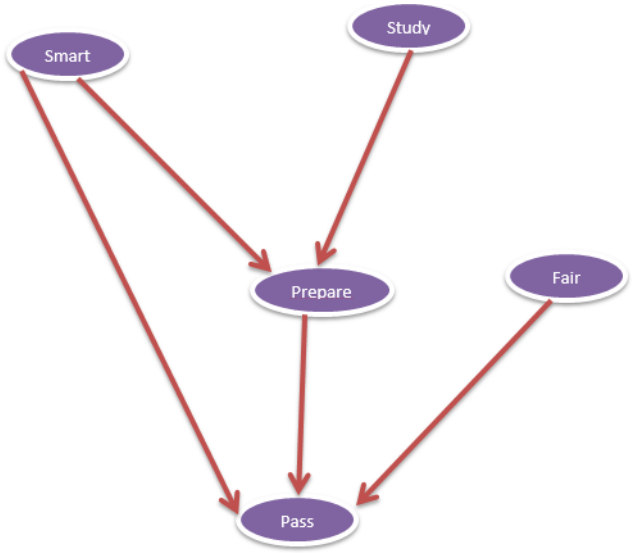

Figure 1.1: Graphical Model: Student DAG model

given a student prepared for the exam and the exam was hard, what is the probability of getting a good grade? A Bayesian graphical model is therefore a probabilistic graphical model that defines and characterizes a joint probability distribution among a given set of random variables, using a directed acyclic graph (DAG) and a conditional probability distribution for each random variable or attribute in the network given its parent set. The graphical structure and representation simply displays conditional dependencies among the variables and enables the researcher to perform the intended inference and answer queries. Bayesian networks (DAG model) thus provides a framework of incorporating diverse mixed data into a single model and existing prior knowledge with new information. By assigning probability to an event in Bayesian Network, we are simply giving an indication of how strongly we believe (uncertainty) that the event will or will not happen.

1.2.1 Why Bayesian Network?

Bayesian networks have been widely applied to a variety of fields [8] and have been used to solve so many real life problems. In almost every field of research you can think of, Bayesian network has been used. Here are some references and areas where DAG model has been extensively used and continued to be used. In the medical field, a Bayesian network has long been used for diagnosis, prognosis, and treatment of diseases [6]. In artificial intelligence area and machine learning, Bayesian networks have been used [13], vision recognition, expert system [8], [14]. Bayesian networks have been used in data mining, search engine optimization (especially by Google), criminology, agriculture, information retrieval, image processing, decision support systems, engineering, signal detection, genealogy, statistics, physics, food chain, the list continues.

Bayesian network is therefore a powerful tool for exploratory data analysis. It is such an important, incredible and powerful model in most area of research. However, applying Bayesian networks to the analysis of large-scale data, consisting of thousands of attributes, is not straightforward because of the heavy computational burden in learning, visualization and working through the attributes relationships. In this thesis, we propose a Gaussian Hierarchical model and then propose a novel method for large-scale or small data analysis based on this hierarchical Bayesian networks(HBN). There are variety of algorithms [8] and methods which can be used to deal with this type of a model. Our method and technique will help make it easier to deal with a Hierarchical Gaussian model.

1.3 Our Contribution

The purpose of this thesis is to propose and study a simple generative model called Hierarchical Gaussian model and show efficient method of estimating parameters of interest using Gibbs sampling. To achieve our desired result we will start by proposing a hierar-

chical Gaussian model which is higher dimensional with unknown multivariate posterior distribution and marginal distributions which are not so easy to compute and may be even intractable but the full conditional distributions are nice close-formed uni variate standard distributions. Our objective is to estimate parameters by drawing samples from the multivariate joint posterior. Since the joint posterior of our proposed model is an unknown distribution it will not be easy to sample from the joint posterior. What do we do? Try marginal distributions? But the marginals are also not easy to compute. We are left with the option of using the full conditional distributions. Hammersley Clifford Theorem tells us that we can decompose a joint distribution into full conditional distributions (distribution of a random variable conditioned on or given everything else). Our other contribution is showing that the full conditional distributions admit closed form (uni variate standard distribution). We will do this by literally computing the full conditional distributions of our Hierarchical Gaussian model. With right tools, model and techniques, the full conditionals are are not hard to compute. This brings us to our other contribution, that of developing the required tools and technique. We will prove some probability results, specifically the Normal distribution equalities. We will then show how the proofs and techniques used in proving them are very handy and effective tools for computing the full conditional distributions of our model. With all the necessary tool and ingredient we will run a simulation given a real data using Gibbs sampling.

1.4 Outline of The Thesis

This thesis is divided into 6 chapters. In Chapter 2 we provide brief background and review of Bayesian analysis graphs, graphical models and Bayesian Graphical model which is necessary for everyone who is not familiar with these concepts. Chapter 3 provides an insight about probabilistic graphical model and graphical inferences. In this chapter we will look at various probabilistic inference technique and graphical inference

algorithms. The list of algorithms we cover is nowhere near being exhaustive. In Chapter 4 we look at the techniques of parameter estimation for Hierarchical Gaussian model using Gibbs Sampling. In this chapter we will introduce and justify the use of Gibbs Sampling and and also prove some probability results (Normal equalities) . In Chapter 5 we will propose a simple and multiple populations hierarchical Gaussian models and compute the full conditional distributions of the two models. We will conclude Chapter 5 by doing Gibbs sampling parameter estimation for a multiple groups hierarchical Gaussian model given real data by running a matlab simulation. Chapter 6 discusses the summary of the thesis and proposes some possible future research.

Chapter 2: Background and Bayesian Analysis Preliminaries

In this chapter we provide a brief review of the concepts necessary for this thesis. It is neither comprehensive nor is it detailed. The reader or any interested person is assumed to have some knowledge of probability, Bayesian analysis, graph theory and inference algorithms [10]. Will explain some of the details if need be and give necessary references.

2.1 Bayesian Review and Genesis

Bayes rule: Is named after its founder Reverend Thomas Bayes, Figure 2.1. Bayes, a British mathematician and theologian, created the basic law of probability referred to as Bayes rule which allows updating of probabilities given new evidence. He is therefore the father of Bayesian analysis. The Bayesian method involves updating our existing knowledge as new evidence is obtained. The Bayesian was once highly disputed but in the recent past, it being highly accepted and used in the scientific world of research and it is becoming extremely popular with researchers. However, despite its fame and acceptance the frequentist Statistician still have trouble accepting the the Bayesian results especially when an informative or subjective prior is involved. The frequentist have no problem with Bayesian method if non informative or objective prior is used. The Bayesian analysis is a simple concept by itself, but it is difficult to implement and it is especially difficult when it comes to selecting a good prior as there is no standard rule of choosing a prior. The computation involved is also sometimes tasking and even intractable. The normalizing constant involves computing integrals and we all know that integrals are hard and most often, no closed form solution.

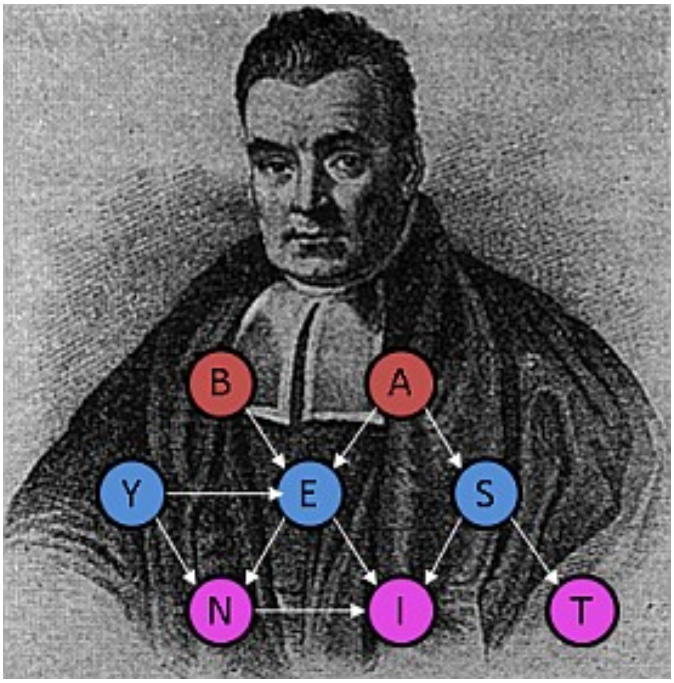

Figure 2.1: Thomas Bayes: Father of Bayesian Analysis

2.1.1 Subjectivity in Bayesian Inference and Remedy

A Bayesian analysis is subjective in that two different researchers/research groups (e.g., EPA and DOE/NETL or an OSU and a PSU researcher) may observe the same data X and yet arrive at different Bayesian conclusions about the unknown parameter of interest, call it μ. This usually happens when the two groups or two researchers have different opinions or different existing knowledge about μ. This subjectivity of Bayesian paradigm is a great source of controversy, a genuine case and legitimate concern about the application of Bayesian inference approach in science. This is because, it will seem "unscientific" for a researcher's personal opinion or prior knowledge to influence and affect important research inputs and therefore affecting the conclusions of a deemed scientific study [6]. The problem is on choosing prior distribution as there is no explicit way or method of choosing a prior. However, when a large amount of data are available, the posterior mean and distribution tend to lean towards the sampling distribution since

the prior mean does not change with the increase in the number of observations. The prior thus becomes subdued as we increase our observations and hence has little effect on the posterior, unless the prior was extremely certain or accurate where in that case the prior has great influence on posterior. In the same token it also means that a sharp likelihood neutralizes the prior effect and the study is thus more scientific. We can also dilute the effect of researcher's opinion by using a more rigorous and objective approach through non-informative prior, which expresses a researcher's ignorance about the unknown parameters. A non informative or vague prior like using a constant for our prior has little or no effect at all on Bayesian result. On the same note, today's posterior can be used as prior in future.

2.2 Bayesian Analysis

In a nutshell we can summarize Bayesian analysis by just looking at the Bayesian theorem.

$$f(\theta|X) = \frac{p(\theta, X)}{m(X)} = \frac{f(x|\theta)p(\theta)}{\int f(x|\theta)p(\theta)d\theta}$$

- We want to make inference about our unknown parameter of interest θ.
- We update our uncertainty about θ after observing new evidence or dataset X.
- $p(\theta)$ reflects our prior knowledge of parameter.
- $p(X|\theta)$ is the likelihood : Distribution of data given the unknown parameter θ. It is actually a function of θ once we observe X.
- $f(\theta|X)$ is the posterior distribution which gives the Bayesian result. If θ is the unknown mean of a distribution we can obtain and estimate it from the posterior by finding the expectation of the posterior distribution.

- m(X) is the marginal likelihood of the observed data also known as the normalizing constant, usually the hardest to compute as it involves integrals.

2.2.1 The Bayesian Approach to Distribution and Parameter Estimation.

The Bayesian approach to density estimation is to form a posterior probability distribution function $p(\mu|x)$ using the Bayes theorem in Section 2.2. The Bayes theorem estimates the density of μ given the new observed dataset. We then estimate our parameter of interest from the posterior distribution. Consider for instance a problem where the population variance σ^2 is a known parameter, but mean μ is unknown and thus a random variable and our parameter of interests. We thus to seek obtain a posterior distribution $p(\mu|x))$ and then estimate mean μ from the posterior. The first task is to decide on the density of our prior $p(\mu)$. For computational purposes we take $p(\mu)$ to be a Gaussian distribution, sampling distribution to be Gaussian as well, by conjugate priors, the posterior is also Gaussian. The situation is modeled and graphically represented in Figure **??**, it is also called the Naive Bayes

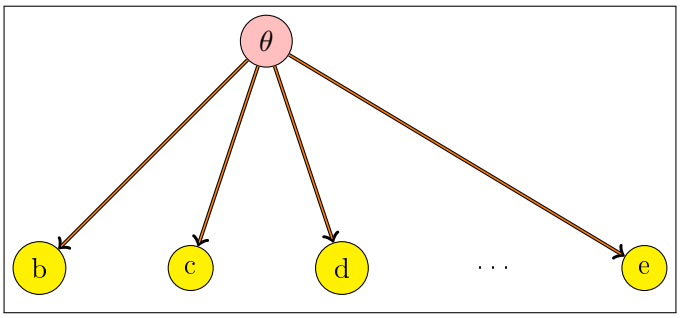

Figure 2.2: Naive Bayesian.

2.3 Graphs and Graphical Models

What is a graph? A graph is a representation consisting of nodes which are connected by of edges and is usually denoted as G(V,E). Edges also known as links may be directed or undirected. Edges can also have associated weights [10]. A graph with all edges directed is called a directed graph else it is an undirected graph, see Figure 2.3. A directed graph may be cyclic or acyclic as shown in Figure 2.4 . If we let the nodes to represent random variables and the links or edges to represent conditional representation among random variables the graph becomes a graphical model (Probabilistic model to be exact). Probabilistic graphical model integrates both graph and probability. A graphical model may be directed or undirected graphical model

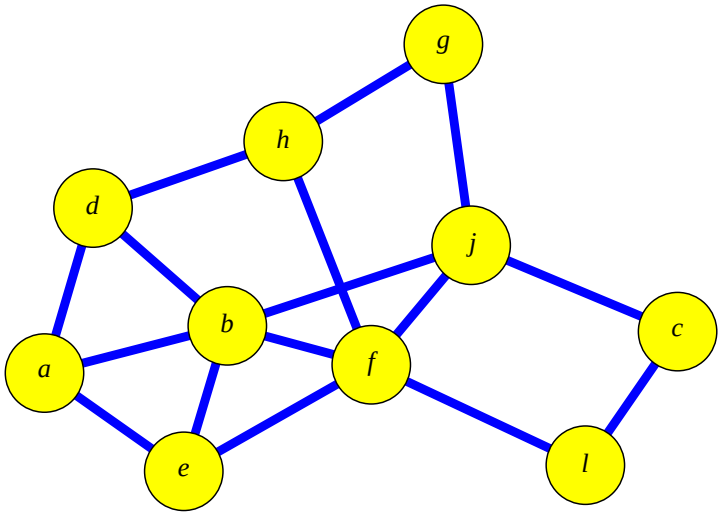

Figure 2.3: Undirected Graph

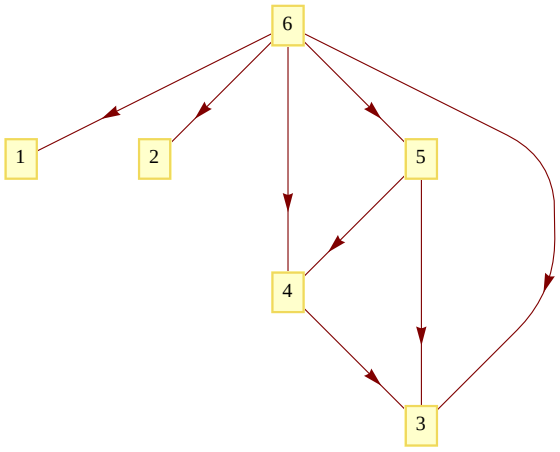

Figure 2.4: Directed Acyclic graph

2.3.1 Directed Acyclic Graphical Model (DAG)

A DAG is a graph with edges between the nodes which are as shown in Figure 2.5. A DAG model is acyclic which means that there is no complete cycle and once we leave a node there is no going back and that node cannot be revisited. In a DAG a node has either parents and ancestors or children and descendants. Parents of a node are those nodes that have a direct path to the node. In Figure 2.5 node d is a parent of node b. Ancestors of a node are those nodes who have a directed path ending at node. In Figure 2.5 the ancestors of node c are nodes b,d,e. Conversely, the descendants of a node are those nodes who have a directed path starting at the node. In Figure 2.5, nodes a,b,c,d are descendant of node e. Nodes a and c are children of node b.

If the graph is not directed it is then undirected Figure 2.6. An undirected graph is sometimes called an undirected network or Markov Random Fields. In an undirected graphical model every node is an ancestor or a descendant of every other node.

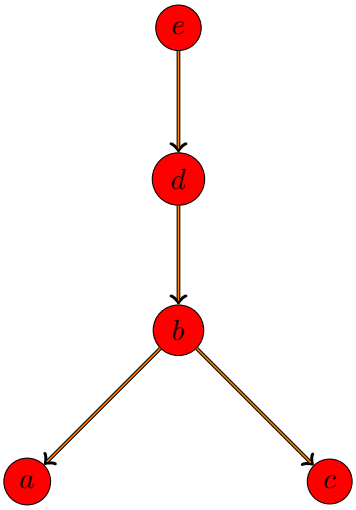

Figure 2.5: DAG: Ancestors, Parents and Chidren

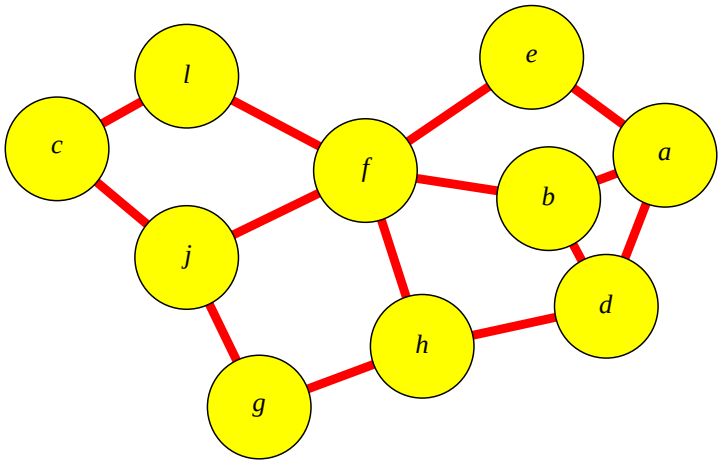

Figure 2.6: Undirected Graph Model

2.3.2 Plates

Sometimes the graph involves so many attributes rendering it hard to visualize the structure. In such a situation we use plate notation on the repeated structure to represent complex graphical models. We use plates or rectangular boxes to enclose a repeated

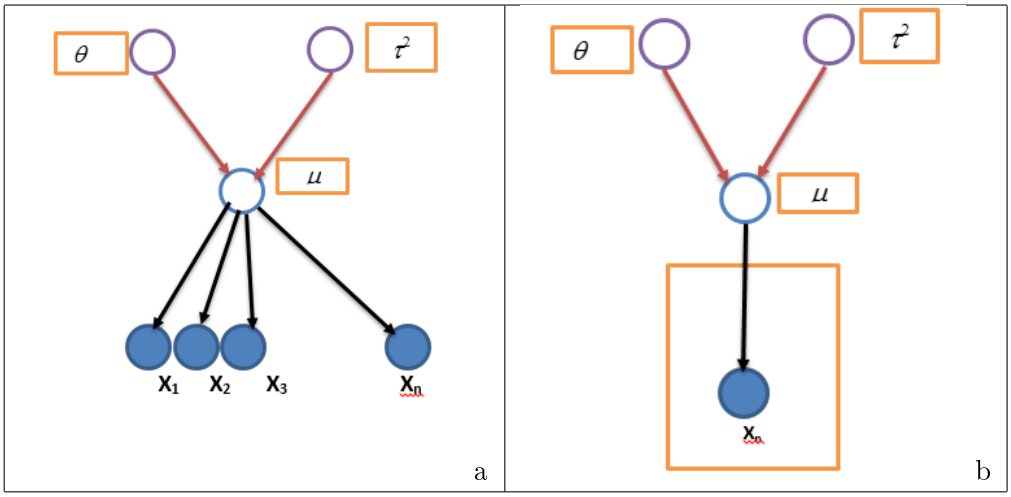

Figure 2.7: Graphical Model and Plate representation

structure. We do this by pasting the structure inside the plate N times, where N is the Number of repeats and N is indicated in the lower right corner of the plate. The Figure 2.7 gives a graphical model and its associated plate.

2.4 Probabilistic Graphical Models (PGM)

Given a set of discrete random variable $X = X_1, X_2, ..., X_N$, the full joint distribution for directed acyclic graphical model is given as

$$p(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} p(X_i|paX_i)$$

where paX represent the parents of X. From the joint distribution we can obtain the marginal probabilities of respective variables, the conditional distributions of variables given other variables. A probabilistic graphical model is therefore just a nice representation of a joint distribution, from which we can obtain marginals and conditional probabilities. There are several advantages of the PGM over other representation.

- It is much more compact hence occupying smaller space (especially if we use plates).

- It is much more time efficient.
- It is easier to understand and communicate ("A picture is worth a thousand words").
- It is easier to build and learn.

We now show how to compute the full joint probability distribution using Figure 2.8. Since the full joint is a product of local conditionals where each variable is conditioned on it parents, we just follow the definition to the letter and use the the fundamental probability chain rule.

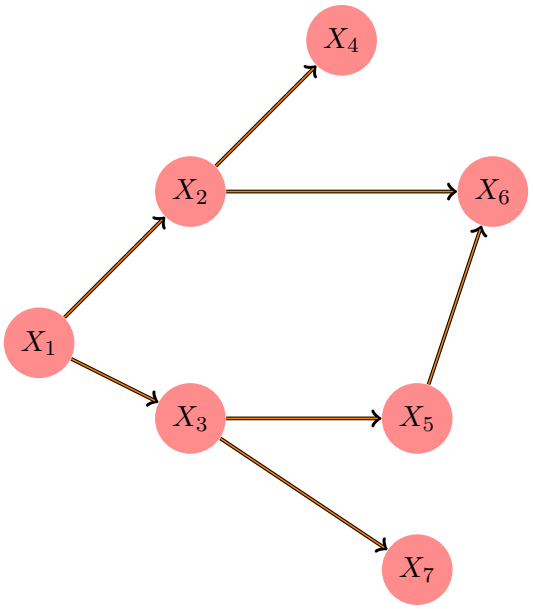

Figure 2.8: DAG for Full Joint Probability distribution .

$$
\begin{aligned}
p(x_1, x_2, x_3, x_4, x_5, x_6, x_7) &= p(x_7, x_6, x_5, x_4, x_3, x_2, x_1) \\
&= p(x_7|x_6, x_5, x_4, x_3, x_2, x_1)p(x_6, x_5, x_4, x_3, x_2, x_1) \\
&= p(x_7|x_3)p(x_6, x_5, x_4, x_3, x_2, x_1) \\
&= p(x_7|x_3)p(x_6|x_5, x_4, x_3, x_2, x_1)p(x_5, x_4, x_3, x_2, x_1) \\
&= p(x_7|x_3)p(x_6|x_5, x_2)p(x_5, x_4, x_3, x_2, x_1) \\
&= p(x_7|x_3)p(x_6|x_5, x_2)p(x_5|x_4, x_3, x_2, x_1)p(x_4, x_3, x_2, x_1) \\
&= p(x_7|x_3)p(x_6|x_5, x_2)p(x_5|x_3)p(x_4, x_3, x_2, x_1) \\
&= p(x_7|x_3)p(x_6|x_5, x_2)p(x_5|x_3)p(x_4|x_3, x_2, x_1)p(x_3, x_2, x_1) \\
&= p(x_7|x_3)p(x_6|x_5, x_2)p(x_5|x_3)p(x_4|x_2)p(x_3, x_2, x_1) \\
&= p(x_7|x_3)p(x_6|x_5, x_2)p(x_5|x_3)p(x_4|x_2)p(x_3|x_2, x_1)p(x_2, x_1) \\
&= p(x_7|x_3)p(x_6|x_5, x_2)p(x_5|x_3)p(x_4|x_2)p(x_3|x_1)p(x_2, x_1) \\
&= p(x_7|x_3)p(x_6|x_5, x_2)p(x_5|x_3)p(x_4|x_2)p(x_3|x_1)p(x_2|x_1)p(x_1) \\
&= p(x_1)\prod_{i=2}^{7} p(x_i|pa(x_i))
\end{aligned}
$$

This gives us the joint distribution as a product of local conditional distributions.

2.4.1 Bayesian Networks and Hierarchical Bayesian Models

A Bayesian network is a Directed Acyclic Graphical (DAG) model where the nodes represent random variables and edges represent probabilistic dependencies. Hierarchical Bayes models is a Bayesian Network model written in a hierarchical form. It is called hierarchical because the model is nested and the graphical representation is structured

in such a way that they sit on top of each other forming a hierarchical structure.

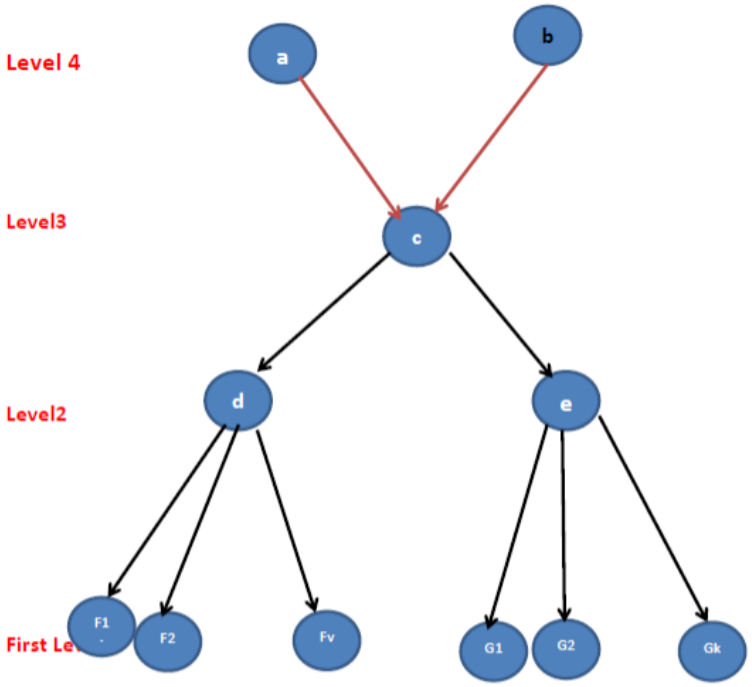

Figure 2.9: Hierarchical DAG model

The Bayesian framework treats all model attributes as random variables. These attributes include observed data, hidden variables, parameters, nuisance parameters as random variables. Although Bayesian models can be represented using either directed or undirected graphical model, it is the directed model that is commonly used in many real life problems. In particular, in hierarchical Bayesian models, the naive Bayesian (simple Bayesian) is modified where prior distributions involves additional parameters known as hyper parameters,and the hyper parameters when treated as random variables will have their own distributions and the overall hierarchical model is thus a set of conditional probabilities linking hyper parameters, parameters and observed data.

.

Chapter 3: Probabilistic Inference in Graphical Models

In this chapter we are going to discuss the various algorithms and techniques applied in probabilistic and statistical inferences for our graphical models. We will in particular discuss at length Gibbs sampling which is the heart of this thesis.

3.1 Modeling and Inference

Once we identify the attributes (variables) to include in our model, our goal or objective is to describe how these variables can interact and associate with each other. This is achieved using graphical representations and forming joint distribution.

Once we are done constructing the probabilistic graphical model, we can embark on answering all the questions of interest by performing inference on the distribution. The questions that can be answered from the model are called queries. For example, common queries are to infer the value of unobserved data points or missing data points. We might be interested in estimating the maximum likelihood estimate (MLE). the maximum a posteriori (MAP) estimate. The Bayesian rule (Posterior distribution happens to be our mantra), is commonly used to answer queries.

Statistical inference is concerned with drawing conclusions, from numerical data, about quantities that are not observed i.e we try to estimate the population parameters from using sample statistics. For example, a clinical trial of a new throat cancer drug might be designed to compare the survival rate in a population given the new drug with population on another treatment or no treatment at all. These survival rate refer to a large population of patients, and it is neither viable, feasible nor ethically acceptable to experiment the entire population. The estimates about the survival rate are thus based on

a sample of patients. That said, graphical models are usually complex if a large number of attributes are involved, and to thus make inferences we need some handy tools.

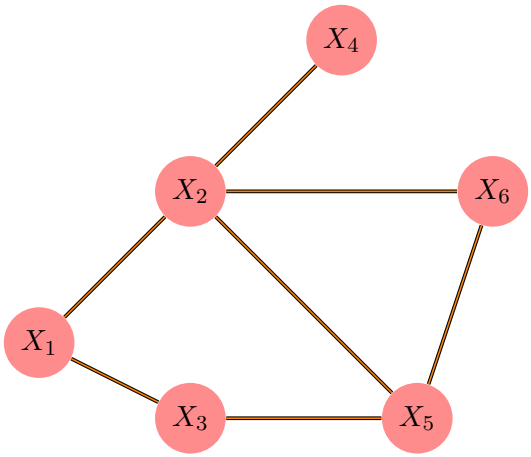

Figure 3.1: Directed acyclic Graphical Model.

3.2 Exact Algorithm

3.2.1 Variable Elimination Algorithm

We will use an example to explain the variable elimination algorithm which is just one example of the Exact Algorithms. Consider the model in Figure 3.1 and suppose we wish to compute the marginal probability $p(x_1)$. We will obtain the marginal by summing (assuming a discrete case) over the remaining variables (unobserved variables)[17]

$$p(x_1) = \sum_{x_2}\sum_{x_3}\sum_{x_4}\sum_{x_5}\sum_{x_6}\frac{1}{Z}\phi(x_1,x_2)\phi(x_1,x_3)\phi(x_2,x_4)\phi(x_3,x_5)\phi(x_2,x_5,x_6)$$

Naively, each of these sums is applied to a summand involving six variables and thus the computational complexity scales as k^6 assuming each variable has k possible outcomes.

We can simplify this complexity by exploiting the distributive law.

$$
\begin{aligned}
p(x_1) &= \sum_{x_2}\sum_{x_3}\sum_{x_4}\sum_{x_5}\sum_{x_6}\frac{1}{Z}\phi(x_1,x_2)\phi(x_1,x_3)\phi(x_2,x_4)\phi(x_3,x_5)\phi(x_2,x_5,x_6) \\
&= \frac{1}{Z}\sum_{x_2}\phi(x_1,x_2)\sum_{x_3}\phi(x_1,x_3)\sum_{x_4}\phi(x_2,x_4)\sum_{x_5}\phi(x_3,x_5)\sum_{x_6}\phi(x_2,x_5,x_6) \\
&= \frac{1}{Z}\sum_{x_2}\phi(x_1,x_2)\sum_{x_3}\phi(x_1,x_3)\sum_{x_4}\phi(x_2,x_4)\sum_{x_5}\phi(x_3,x_5)f_6(x_2,x_5) \\
&= \frac{1}{Z}\sum_{x_2}\phi(x_1,x_2)\sum_{x_3}\phi(x_1,x_3)\sum_{x_4}\phi(x_2,x_4)f_5(x_2,x_3) \\
&= \frac{1}{Z}\sum_{x_2}\phi(x_1,x_2)\sum_{x_3}\phi(x_1,x_3)f_5(x_2,x_3)\sum_{x_4}\phi(x_2,x_4) \\
&= \frac{1}{Z}\sum_{x_2}\phi(x_1,x_2)\sum_{x_3}\phi(x_1,x_3)f_5(x_2,x_3)f_4(x_2) \\
&= \frac{1}{Z}\sum_{x_2}\phi(x_1,x_2)f_4(x_2)\sum_{x_3}\phi(x_1,x_3)f_5(x_2,x_3) \\
&= \frac{1}{Z}\sum_{x_2}\phi(x_1,x_2)f_4(x_2)f_3(x_2,x_3) \\
&= \frac{1}{Z}f_2(x_1)
\end{aligned}
$$

where we have defined the intermediate factors or functions as f_i, we obtain the value of Z, and hence the marginal, by summing the final expression with respect to x_1. If on the other hand we assume the variables in Figure 3.1 are continuous the same will hold. We will only need to interchange the summation with an integral as follows

$$
\begin{aligned}
p(x_1) &= \int_{x_2}\int_{x_3}\int_{x_4}\int_{x_5}\int_{x_6} \frac{1}{Z}\phi(x_1,x_2)\phi(x_1,x_3)\phi(x_2,x_4)\phi(x_3,x_5)\phi(x_2,x_5,x_6)d(x_6)d(x_5)d(x_4)d(x_3)d(x_2) \\
&= \frac{1}{Z}\int_{x_2}\phi(x_1,x_2)\int_{x_3}\phi(x_1,x_3)\int_{x_4}\phi(x_2,x_4)\int_{x_5}\phi(x_3,x_5)\int_{x_6}\phi(x_2,x_5,x_6)d(x_6)d(x_5)d(x_4)d(x_3)d(x_2) \\
&= \frac{1}{Z}\int_{x_2}\phi(x_1,x_2)\int_{x_3}\phi(x_1,x_3)\int_{x_4}\phi(x_2,x_4)\int_{x_5}\phi(x_3,x_5)f_6(x_2,x_5)d(x_5)d(x_4)d(x_3)d(x_2) \\
&= \frac{1}{Z}\int_{x_2}\phi(x_1,x_2)\int_{x_3}\phi(x_1,x_3)\int_{x_4}\phi(x_2,x_4)f_5(x_2,x_3)d(x_4)d(x_3)d(x_2) \\
&= \frac{1}{Z}\int_{x_2}\phi(x_1,x_2)\int_{x_3}\phi(x_1,x_3)f_5(x_2,x_3)\int_{x_4}\phi(x_2,x_4)d(x_4)d(x_3)d(x_2) \\
&= \frac{1}{Z}\int_{x_2}\phi(x_1,x_2)\int_{x_3}\phi(x_1,x_3)f_5(x_2,x_3)f_4(x_2)d(x_3)d(x_2) \\
&= \frac{1}{Z}\int_{x_2}\phi(x_1,x_2)f_4(x_2)\int_{x_3}\phi(x_1,x_3)f_5(x_2,x_3)d(x_3)d(x_2) \\
&= \frac{1}{Z}\int_{x_2}\phi(x_1,x_2)f_4(x_2)f_3(x_2,x_3)d(x_2) \\
&= \frac{1}{Z}f_2(x_1)
\end{aligned}
$$

3.3 Approximate Algorithms

The above exact algorithms focused on the algebraic and graphical structure of probabilistic graphical model inference. As the number of random variable increase exact algorithm are no longer feasible. It is not an easy task to do carry out elimination algorithm if for instance we have 1000 attributes. Thus enters the approximate algorithms The approximate Algorithms on the other hand use the law of large numbers and the Central Limit theorem to exploit the probability theory of the graphical model. One popular class of approximate algorithms is based on the Markov Chain Monte Carlo method which is discussed next.

3.3.1 Markov Chain Monte Carlo(MCMC= $(MC)^2$)

MCMC methods are techniques to approximate integrals and especially intractable integrals using simulated values. From our directed graphical model the joint distribution is higher dimension multivariate distribution which is usually unknown. Computing marginals from the joint distribution is also not easy and in most cases the integrals are intractable with no known closed form solution. We thus use MCMC sampling methods to sample from the joint distribution. For the purpose of this thesis we will discuss Gibbs sampling as our example of MCMC method.

3.3.2 Gibbs Sampling

Gibbs sampling or a Gibbs sampler is Markov chain Monte Carlo (MCMC) algorithm for sampling a non independent sequence of random values from full conditionals of a given joint distribution. This sampling is done when direct sampling is not possible i.e when not possible to sample from the multivariate distribution. The sample observed can be used to approximate the joint distribution, bi variate distribution, marginal distributions and even conditional distributions

Gibbs sampling is commonly used as a means of statistical inference and parameter estimation, especially Bayesian inference. Since Gibbs sampling is the heart of this thesis we will devote the next chapter to the discussion of Gibbs method.

Chapter 4: Parameter Estimation For Hierarchical Gaussian Model Via Gibbs Sampling

4.1 Gibbs Sampling

Gibbs sampling is a technique to draw samples from a joint distribution based on the full conditional distributions of all the associated random variables. Gibbs sampling can be traced back to the work of [11], whereas the concept of Gibbs sampler introduced by [2] to the field of image processing. Given a Bayesian graphical model, the hyper prior and prior distributions and the likelihood which generates the data,we can easily construct joint posterior distribution of all parameters and random variables involved. From the joint distribution, the marginal distribution of any random variable is obtained by integrating out all other parameters or variables in the model. With a marginal distribution, parameters such as the mean, mode, median and variance of each parameter can then be obtained from marginal posterior distribution. However, this integration is usually computationally tedious and in most cases not analytically tractable and has no closed form solution, hence we use numerical methods. Approximations to the marginal distributions were proposed by [6].

Gibbs sampling [2] is thus a numerical integration method using MCMC. The method generates random values from the marginal posterior distributions by repetitively sampling from full conditional distributions.The properties of Gibbs and its power as an MCMC numerical integration tool can be found in [5].

4.1.1 Why Gibbs Sampling?

Most real life problem high dimensional. Such high dimensional problems involve full joint distribution or joint posterior distribution that are high dimensional multivariate distribution. For instance if a graphical model has 100 attributes and each attribute can take 10 values, then we are talking of a sample space of 10^{100} configurations. To simulate a random configuration from such a sample space is next to impossible.
Gibbs sampling simulate and generate Markov Chain sequence from univariate distributions called full conditional distributions. Gibbs sampling theory shows the generated n^{th} value is a random value from a distribution that is close to the target or true distribution. We might never hit the target but we will get pretty close 'for any chemical reaction to take place'. Can we trust that the full conditionals will get us close to the target distribution? The law of large numbers and the central limit theorem results cemented by the Hammersley- Clifford Theorem are the main reason why Gibbs sampling works.

Definition 4.1.1 *Let $p(y_1, ..., y_k)$ be the joint density of a random vector $X_1, ..., X_p$ and let $p^i(x_i)$ denote the marginal density of X_i. If $p^i(x_i > 0)$ for every $i = 1, .., p$ implies $p(x_1, ..., x_p) > 0$, then the joint density is said to satisfy the positivity condition [19].*

4.1.2 The Hammersley- Clifford Theorem

A very important property of full conditionals, and which the Gibbs sampler is based on is that full conditionals fully specify and describe the joint distribution. The Theorem proves that full conditionals, which the Gibbs sampler is based on, fully specify the joint distribution [19, 1, 8]. That is the joint density can be decomposed and easily be derived from the conditional densities.

Theorem 1 *Let $(X_1, ..., X_p)$ satisfy the postivity and joint condition and have joint den-*

sity $f(x_1, ..., x_p)$. *Then for all* $\xi_1, .., \xi_p \in supp(f)$

$$f(x_1, ..., x_p) \propto \prod_{i=1}^{p} \frac{f_{X_i|X_{-i}}(x_{i|x_1,...x_{1-1},\xi_{i+1},...,\xi_p})}{f_{X_i|X_{-i}}(\xi_{i|x_1,...x_{1-1},\xi_{i+1},...,\xi_p})}$$

Proof 1 *The proof and explanation can be found in [1, 8, 19].*

4.2 Normal Distribution Equalities

For higher dimension problems the joint posterior is a distribution whose kernel is unknown and it is not easy to associate it with any known standard distribution. If we want to sample from such a joint posterior it will not be directly possible. All is not lost, as there are other techniques of sampling and inferences. In this part we will specifically do Gibbs Sampling.

What do we need? To use Gibbs Sampler we need closed form (standard) full conditional distributions. Since full conditions distributions are uni variate we can easily sample if the distribution known. In this section we will proof some normal distribution equalities from the scratch. These equalities which will resurface frequently from now on, happen to be very a powerful tool when computing full conditionals. With a cleverly chosen data model, parameter and hyper parameter models we were able to derive closed form full conditional distributions. We now state and prove the theorems, in which case we will find some have trivial proof but some are intense.

Theorem 2 *These inequalities hold for Normal density function [2]*

1 $N(x; \mu, \sigma^2) = N(\mu; x, \sigma^2)$

2 $N(ax; \mu, \sigma^2) = \frac{1}{a} N(x; \frac{\mu}{a}, \frac{\sigma^2}{a^2})$

3 $N(x; \mu_1, \sigma_1^2) N(x; \mu_2, \sigma_2^2) = KN\left(x, \frac{\sigma_2^2\mu_1 + \sigma_1^2\mu_2}{\sigma_1^2 + \sigma_2^2}, \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)$ *where* K *is constant.*

We define

$$N(ax;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2\sigma^2}(ax-\mu)^2) \tag{4.1}$$

We are now going to prove these e-qualities one at a time, some proofs are trivial others will require some techniques and algebra manipulation

Proof 2 *The proof of the first equality is trivial via definition. To thus prove that* $N(x;\mu,\sigma^2) = N(\mu;x,\sigma^2)$ *we use the definition. We want to Prove* $N(x;\mu,\sigma^2) = N(\mu;x,\sigma^2)$

$$\begin{aligned} N(x;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2\sigma^2}(x-\mu)^2) \\ &= \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2\sigma^2}(\mu-x)^2) \\ &= N(\mu;x,\sigma^2)\bullet \end{aligned}$$

Proof 3 *The proof of the second equality uses the definition and some algebra. Thus the proof of* $N(ax;\mu,\sigma^2) = \frac{1}{a}N(x;\frac{\mu}{a},\frac{\sigma^2}{a^2})$ *is as follows:*

$$\begin{aligned} N(ax;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2\sigma^2}(ax-\mu)^2) \\ &= \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2\sigma^2}(a(x-\mu/a)^2) \\ &= \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{a^2}{2\sigma^2}((x-\mu/a)^2) \\ &= \frac{1}{a\sqrt{2\pi}\sigma/a} exp(-\frac{1}{2\sigma^2/a^2}(x-\mu/a)^2) \\ &= \frac{1}{a}N(x;\mu/a,\sigma^2/a^2)\bullet \end{aligned}$$

Proof 4 *Here we use the definition, algebra and one simple trick of completing square on the exponent.Thus the proof of* $N(x;\mu_1,\sigma_1^2)N(x;\mu_2,\sigma_2^2) = KN\left(x,\frac{\sigma_2^2\mu_1+\sigma_1^2\mu_2}{\sigma_1^2+\sigma_2^2},\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)$

where K does not depend on X is as follows:

$$\begin{aligned}
N(x;\mu_1,\sigma_1^2)N(x;\mu_2,\sigma_2^2) &= \frac{1}{\sqrt{2\pi}\sigma_1}\left(exp(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right)\frac{1}{\sqrt{2\pi}\sigma_2}exp\left(-\frac{1}{2\sigma_2^2}(x-\mu_2)^2\right)\\
&= \frac{1}{2\pi\sigma_1\sigma_2}exp\left(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2-\frac{1}{2\sigma_2^2}(x-\mu_2)^2\right)\\
&= \frac{1}{2\pi\sigma_1\sigma_2}exp\left(-\frac{1}{2\sigma_1^2}(x^2-2x\mu_1+\mu_1^2\right)exp\left(-\frac{1}{2\sigma_2^2}(x^2-2x\mu_2+\mu_2^2)\right)\\
&= \frac{1}{2\pi\sigma_1\sigma_2}exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)x^2-2\left(\frac{\mu_1}{\sigma_1^2}+\frac{\mu_2}{\sigma_2^2}\right)x\right)exp\left(-\frac{\mu^2}{2\sigma_1^2}-\frac{\mu_2^2}{2\sigma_2^2}\right)\\
&= \frac{C}{2\pi\sigma_1\sigma_2}exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)x^2-2\left(\frac{\mu_1}{\sigma_1^2}+\frac{\mu_2}{\sigma_2^2}\right)x\right)\\
&= \frac{C^1}{2\pi\sigma_1\sigma_2}exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)\left(x-\frac{\left(\frac{\mu_1}{\sigma_1^2}+\frac{\mu_2}{\sigma_2^2}\right)}{\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)}\right)^2\right)\\
&= \frac{C^1}{2\pi\sigma_1\sigma_2}exp\left(-\frac{1}{2}\left(\frac{\sigma_2^2+\sigma_1^2}{\sigma_1^2\sigma_2^2}\right)\left(x-\frac{\left(\frac{\sigma_2^2\mu_1+\sigma_1^2\mu_2}{\sigma_1^2\sigma_2^2}\right)}{\left(\frac{\sigma_2^2+\sigma_1^2}{\sigma_1^2\sigma_2^2}\right)}\right)^2\right)\\
&= \frac{C^1}{2\pi\sigma_1\sigma_2}exp\left(-\frac{1}{2}\left(\frac{\sigma_2^2+\sigma_1^2}{\sigma_1^2\sigma_2^2}\right)\left(x-\frac{\sigma_2^2\mu_1+\sigma_1^2\mu_2}{\sigma_1^2+\sigma_2^2}\right)^2\right)\\
&= KN\left(x,\frac{\sigma_2^2\mu_1+\sigma_1^2\mu_2}{\sigma_1^2+\sigma_2^2},\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)
\end{aligned}$$

Thus

$$N(x;\mu_1,\sigma_1^2)N(x;\mu_2,\sigma_2^2) = KN\left(x,\frac{\sigma_2^2\mu_1+\sigma_1^2\mu_2}{\sigma_1^2+\sigma_2^2},\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)$$

Chapter 5: Hierarchical Bayesian Models, Joint Posterior and Full Conditional Distributions

5.1 Simple Hierarchical Normal DAG Model with Known Variance

In this section we are going to compute joint distributions and full conditional distributions given a simple hierarchical model.

Consider the following directed a cyclic graph model

$$x_i|\mu \sim N(\mu, \sigma^2)$$

$$\mu|\theta, \tau^2 \sim N(\theta, \tau^2)$$

$$\theta \backsim N(a, b^2)$$

$$\tau^2 \sim IG(c, d)$$

where a, b, c, d are given or specified and σ^2 is assumed to be known usually estimated to be the sample variance.

The simple model (data from a single population) can be represented graphically as in Figure 5.1

We could also represent the simple hierarchical graphical model using graph and plates as shown in Figure 5.2

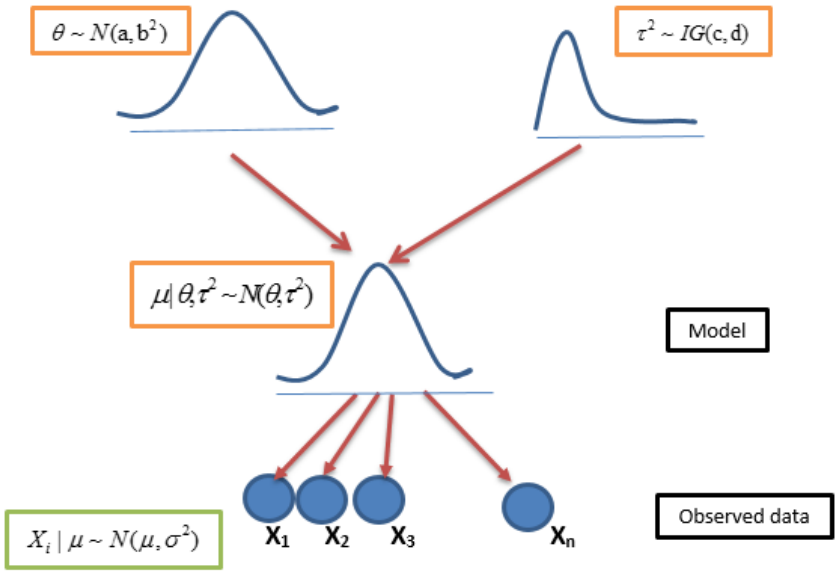

Figure 5.1: Data from several Normal Distributions

5.1.1 Joint Posterior Distribution

We start by deriving the joint posterior distribution.

$$\begin{aligned}
p(\theta, \tau^2, \mu|x) &= \frac{p(\theta, \tau^2, \mu, x)}{p(x)} \\
&= \frac{p(\theta, \tau)p(\mu|\theta, \tau)p(x|\mu)}{p(x)} \\
&\propto p(\theta, \tau)p(\mu|\theta, \tau)p(x|\mu) \\
&\propto p(\theta)p(\tau^2)N(\mu; \theta, \tau^2)\prod_{i=1}^{n} N(x_i; \mu, \sigma^2) \\
&\propto p(\theta)p(\tau^2)N(\mu; \theta, \tau^2)N(\bar{x}; \mu, \sigma^2/n) \\
&= p(\theta)p(\tau^2)N(\mu; \theta, \tau^2)N(\mu; \bar{x}, \sigma^2/n)
\end{aligned}$$

We can see that this is a distribution which is not familiar and cannot be associated with any known distribution.

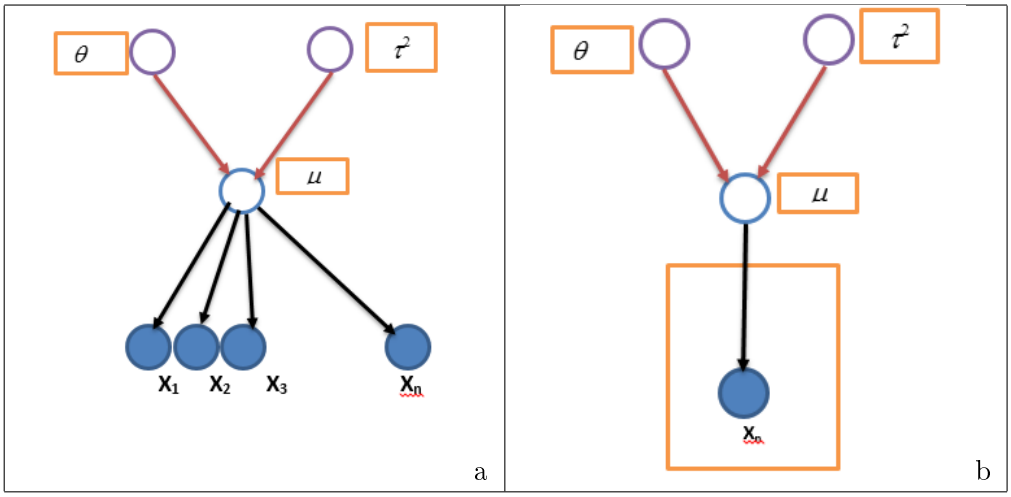

Figure 5.2: Graphical and Plate representation

5.1.2 Statistical and Posterior Inference

Goal:Draw samples of (θ, μ, τ) given observed data X

For this sampling to be done using Gibbs sampler we need to compute the full conditional distributions and hope these distributions are nice closed form. From this joint posterior distribution and the Normal distribution equalities proved above, we can easily compute the full conditional distributions of each parameter by just considering the expression containing that particular parameter and treating the rest as constants.

5.2 Finding Full Conditionals of the Parameters in The Simple Model

The probability of a random variable given everything else is called the full conditional distribution. In Bayesian hierarchical modeling full conditional is the conditional distribution of a parameter given everything else. We can use Gibbs sampler to sample from the joint distribution if we knew the full conditional for each parameter. For each parameter, the full conditional distribution is the distribution of the parameter conditional on all the other parameters and the evidence (observations). The full conditionals

needed for implementation of the Gibbs sampler are basically easy to find. From the joint distribution of all variables, only expression that contain the particular variable (parameter) are considered the rest are treated as constant. The difficulty (is always) in finding normalizing constants which we do not necessarily need to compute.

5.2.1 Full conditional Distribution of μ

Finding full conditionals is a requirement in Gibbs sampling. A wise choice of model will result in nice standard full conditional distribution. The full conditional of μ is easier as it involves taking products of normal distributions and we easily know how to do that. Like other full conditionals, we will only be interested on parts of the joint posterior equation above that involve μ

$$
\begin{aligned}
p(\mu|\tau^2,\theta,\sigma^2,x) &\propto p(\mu,\theta,\tau^2,\sigma^2|x) \\
&\propto p(\theta)p(\tau^2)N(\mu;\theta,\tau^2)\prod_{i=1}^{n}N(x_i;\mu,\sigma^2) \\
&\propto N(\mu;\theta,\tau^2)\prod_{i=1}^{n}N(x_i;\mu,\sigma^2) \\
&\propto N(\mu_j;\theta,\tau^2)\prod_{i=1}^{n}N(x_i;\mu_j,\sigma^2) \\
&\propto p(\theta)p(\tau^2)N(\mu;\theta,\tau^2)N(\bar{x};\mu,\sigma^2/n) \\
&\propto p(\theta)p(\tau^2)N(\mu;\theta,\tau^2)N(\mu;\bar{x},\sigma^2/n) \\
&= kN(\mu;\frac{\sigma^2/n\theta+\tau^2\bar{x}}{\sigma^2/n+\tau^2},\frac{\tau^2\sigma^2/n}{\sigma^2/n+\tau^2}) \\
&\propto N\left(\mu;\frac{\sigma^2/n\theta+\tau^2\bar{x}}{\sigma^2/n+\tau^2},\frac{\tau^2\sigma^2/n}{\sigma^2/n+\tau^2}\right)
\end{aligned}
$$

The full conditional of μ depends on (τ^2,σ^2,θ) and data .

5.2.2 Full conditional Distribution of θ

From the joint posterior equation above, we will only consider those expressions which involves θ

$$\begin{aligned}
p(\theta|\tau^2,\mu,\sigma^2,x) &\propto p(\mu,\theta,\tau^2,\sigma^2|x) \\
&\propto p(\theta)p(\tau^2)p(\mu|\theta,\tau^2)\prod_{i=1}^{n}p(x_i|\mu,\sigma^2) \\
&\propto N(\theta|a,b^2)p(\tau^2)N(\mu;\theta,\tau^2)\prod_{i=1}^{n}N(x_i;\mu,\sigma^2) \\
&\propto N(\theta|a,b^2)N(\mu|\theta,\tau^2) \\
&\propto N(\theta|a,b^2)N(\theta|\mu,\tau^2) \\
&\propto N\left(\theta;\frac{a\tau^2+\mu b^2}{\tau^2+b^2},\frac{b^2\tau^2}{b^2+\tau^2}\right)
\end{aligned}$$

The distribution is another Gaussian and this makes us happy.

5.2.3 Full conditional Distribution of τ^2

Similarly, to obtain the full conditional distribution of τ^2 we will only consider the expressions in the Joint posterior above which involves τ^2 and treat everything else as

constant.

$$\begin{aligned} p(\tau^2|\theta,\mu,\sigma^2,x) &\propto p(\mu,\theta,\tau^2,\sigma^2|x) \\ &\propto p(\theta)p(\tau^2)p(\mu|\theta,\tau^2)\prod_{i=1}^{n}p(x_i|\mu,\sigma^2) \\ &\propto p(\theta|a,b^2)IG(\tau^2|c,d)N(\mu;\theta,\tau^2)\prod_{i=1}^{n}N(x_i;\mu,\sigma^2) \\ &\propto IG(\tau^2|c,d)N(\mu|\theta,\tau^2) \\ &\propto IG\left(c+1/2, d+(\mu-\theta)^2/2\right) \end{aligned}$$

This full conditional distribution of parameter τ^2 is another standard distribution. The full conditional distributions of the parameters are well known distribution and can easily we simulated.

5.3 A Hierarchical Normal Model for a Data from Several Groups

In this model data is observed from more than one populations.

Assume we have random samples from M populations, having sample sizes $n_1, n_2, .., n_m$ We specify the hierarchical data model:

$$X_{1j}, ..., X_{n_j j}|\mu_j, \sigma^2 \backsim N(\mu_j, \sigma^2)$$

$$\mu_j|\theta, \tau^2 \backsim N(\theta, \tau^2)$$

$$\tau^2 \backsim IG(a, b)$$

$$\theta \backsim N(\theta_0, \kappa^2)$$

$$\sigma^2 \backsim IG(c, d)$$

This can be represented graphically and using distribution as shown in Figure 5.3 and the graphical and plate is shown in Figure 5.4 and Figure 5.5 respectively

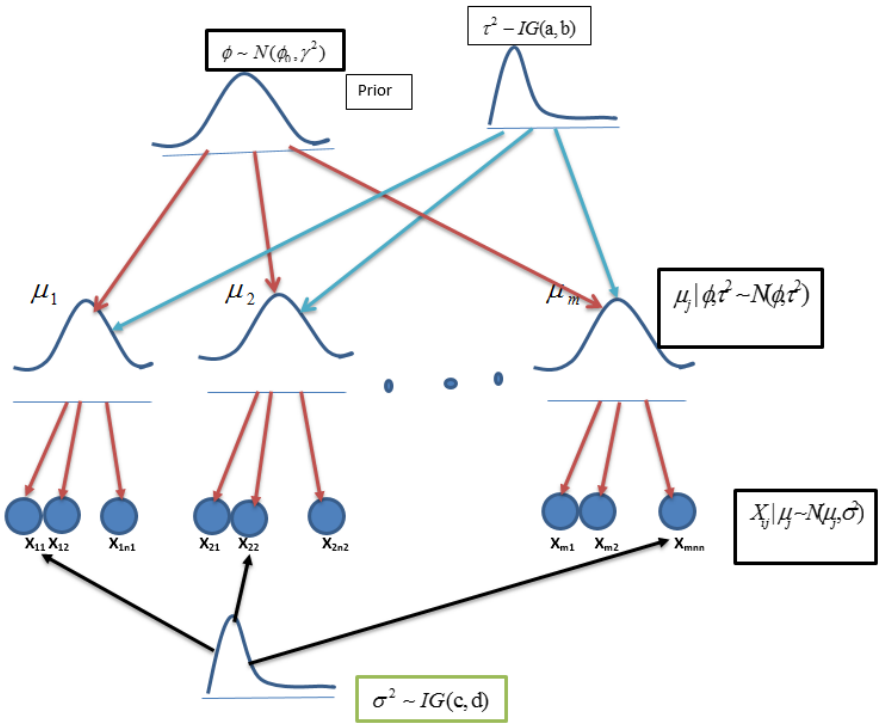

Figure 5.3: Data from several Normal Distributions

5.3.1 Joint Posterior

This model assumes variability across group means, but group variances are assumed to be constant $= \sigma^2$ across groups. It assumes that the the data and prior parameters are independent and identically independent.

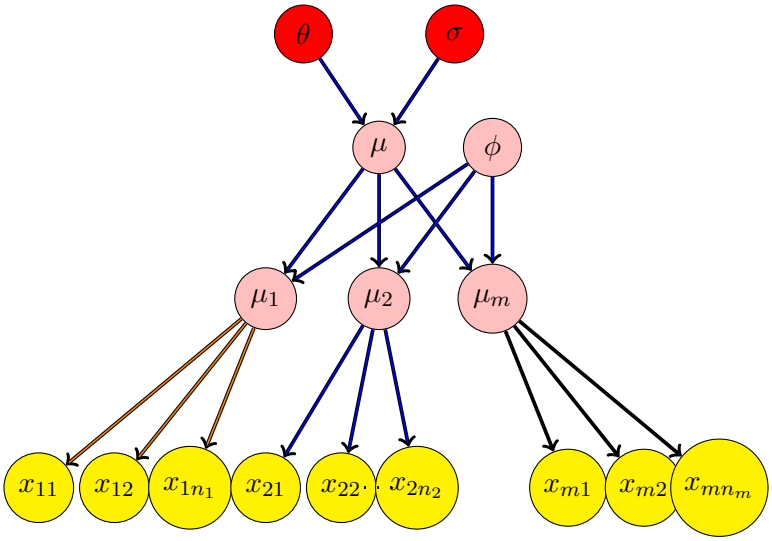

Figure 5.4: Data from several Normal Distributions

5.3.2 Posterior Inference

Goal: Draw samples of $(\mu_1, ..., \mu_m, \theta, \tau, \sigma)$ conditional on X

We will use Gibbs sampler to make dependent, approximate draws from this target distribution. To perform Gibbs sampling we need the full conditionals for each parameter. We now approximate the joint posterior as via the following useful factorization:

$$
\begin{aligned}
p(\mu_1, ..., \mu_m, \theta, \tau^2, \sigma^2 | x_1, .., x_m) &= \frac{p(\mu_1, ..., \mu_m, \theta, \tau^2, \sigma^2, x_1, .., x_m)}{p(x_1, .., x_m)} \\
&\propto p(\mu_1, ..., \mu_m, \theta, \tau^2, \sigma^2, x_1, .., x_m) \\
&\propto p(x_1, .., x_m | \mu_1, ..., \mu_m, \theta, \tau^2, \sigma^2) * p(\mu_1, ..., \mu_m, \theta, \tau^2, \sigma^2) \\
&\propto p(X | \mu_1, ..., \mu_m, \theta, \tau^2, \sigma^2) P(\mu_1, ..., \mu_m | \theta, \tau^2, \sigma^2) p(\theta, \tau^2, \sigma^2) \\
&= \prod_{j=1}^{m} \prod_{i=1}^{n_j} p(x_{ij} | \mu_j, \sigma^2) \prod_{j=1}^{m} p(\mu_j | \theta, \tau^2) p(\theta | \tau^2, \sigma^2) p(\tau^2, \sigma^2) \\
&= \prod_{j=1}^{m} \prod_{i=1}^{n_j} p(x_{ij} | \mu_j, \sigma^2) \prod_{j=1}^{m} p(\mu_j | \theta, \tau^2) p(\theta) p(\tau^2) p(\sigma^2)
\end{aligned}
$$

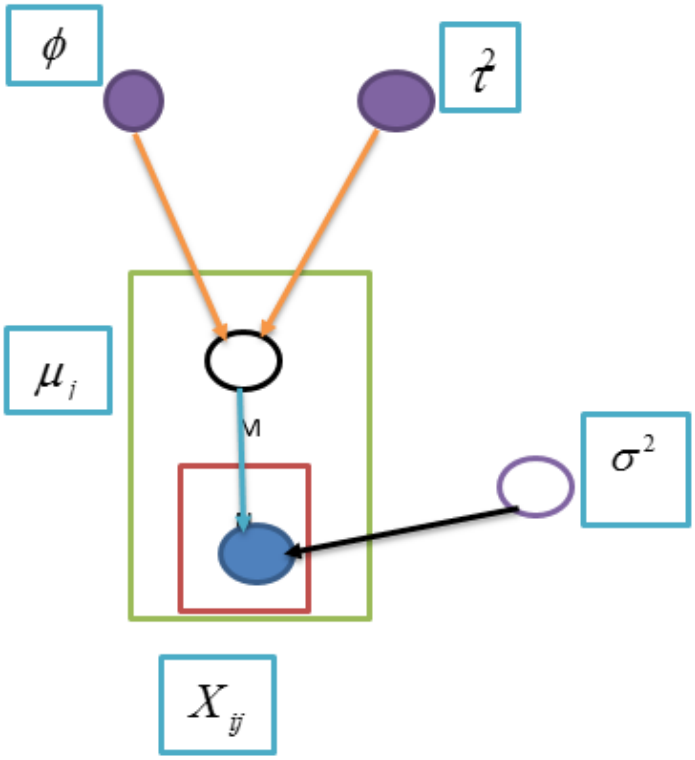

Figure 5.5: Data from several Normal Distributions: Plate

From the joint posterior distribution only expression that contain the particular variable are considered the rest is treated as constant which do not affect the distribution.

5.4 The full conditionals of the parameters

5.4.1 Full conditional of μ_j

The full conditional distribution of μ_j is easier as we are taking products of normal and we easily know how to do that. Like other full conditionals, we will only be interested

on parts of the joint posterior equation above that involve μ_j

$$
\begin{aligned}
p(\mu_j|\mu_1,...,\mu_m,\tau^2,\theta,\sigma^2,X) &\propto p(\mu_1,...,\mu_1,\theta,\tau^2,\sigma^2|X=(x_1,..,x_m)) \\
&\propto \prod_{j=1}^{m}\prod_{i=1}^{n_j} p(y_{ij}|\mu_j,\sigma^2)\prod_{j=1}^{m} p(\mu_j|\theta,\tau^2)p(\theta)p(\tau^2)p(\sigma^2) \\
&\propto p(\mu_j|\theta,\tau^2)\prod_{i=1}^{n_j} p(x_{ij}|\mu_j,\sigma^2) \\
&\propto N(\mu_j;\theta,\tau^2)\prod_{i=1}^{n_j} N(x_{ij};\mu_j,\sigma^2) \\
&\propto N(\mu_j;\theta,\tau^2)N(\bar{x_j};\mu_j,\frac{\sigma^2}{n_j}) \\
&\propto N(\mu_j;\theta,\tau^2)N(\mu_j;\bar{x_j},\frac{\sigma^2}{n_j}) \\
&\propto N\left(\mu_j;\frac{\theta\frac{\sigma^2}{n_j}+\bar{x_j}\tau^2}{\tau^2+\frac{\sigma^2}{n_j}},\frac{\tau^2\frac{\sigma^2}{n_j}}{\tau^2+\frac{\sigma^2}{n_j}}\right)
\end{aligned}
$$

The full conditional of μ_j depends on (τ,σ^2,θ) and data from group j. It is a univariate normal distribution (closed form)

5.4.2 Full conditional distribution of θ

$$
\begin{aligned}
p(\theta|\mu_1,...,\mu_m,\tau^2,\sigma^2,X) &\propto p(\mu_1,...,\mu_m,\theta,\tau^2,\sigma^2|x_1,..,x_m) \\
&\propto \prod_{j=1}^{m}\prod_{i=1}^{n_j} p(x_{ij}|\mu_j,\sigma^2)\prod_{j=1}^{m} p(\mu_j|\theta.\tau^2)p(\theta)p(\tau^2)p(\sigma^2) \\
&\propto p(\theta)\prod_{j=1}^{m} p(\mu_j|\theta,\tau^2) \\
&\propto N(\theta;\theta_0,\kappa^2)N(\bar{\mu};\theta,\frac{\tau^2}{m}) \\
&\propto N(\theta;\theta_0,\kappa^2)N(\theta;\bar{\mu},\frac{\tau^2}{m}) \\
&\propto N\left(\theta;\frac{\theta_0\frac{\tau^2}{m}+\bar{\mu}\kappa^2}{\kappa^2+\frac{\tau^2}{m}},\frac{\kappa^2\frac{\tau^2}{m}}{\kappa^2+\frac{\tau^2}{m}}\right)
\end{aligned}
$$

As we can see by Markov blanket property this full conditional is independent of the data X.It depends on the other variables (parameters) and is a standard univariate distribution

5.4.3 Full conditional distribution of τ^2

We now compute the full conditional distribution for τ^2

$$
\begin{aligned}
p(\tau^2|\mu_1,...,\mu_m,\theta,\sigma^2,X) &\propto p(\mu_1,...,\mu_1,\theta,\tau^2,\sigma^2|X) \\
&\propto \prod_{j=1}^{m}\prod_{i=1}^{n_j} p(x_{ij}|\mu_j,\sigma^2)\prod_{j=1}^{m} p(\mu_j|\theta,\tau^2)p(\theta)p(\tau^2)p(\sigma^2) \\
&\propto p(\tau^2)\prod_{j=1}^{m} p(\mu_j|\theta,\tau^2) \\
&\propto IG(a,b)\prod_{j=1}^{m} N(\mu_j|\theta,\tau^2))
\end{aligned}
$$

$$
\begin{aligned}
p(\tau^2|\mu_1,...,\mu_m,\theta,\sigma^2,X) &\propto IG(a,b)\prod_{j=1}^{m} N(\mu_j|\theta,\tau^2)) \\
&\propto (\tau^2)^{-a-1}e^{\frac{-b}{\tau^2}}\prod_{j=1}^{m} N(\mu_j|\theta,\tau^2)) \\
&\propto (\tau^2)^{-a-1}e^{\frac{-b}{\tau^2}}\prod_{j=1}^{m}(\tau^2)^{-1}e^{\frac{-(\mu_j-\theta)^2}{2\tau^2}} \\
&\propto (\tau^2)^{-a-1}e^{\frac{-b}{\tau^2}}(\tau^2)^{-1/2}e^{\frac{-\sum_{j=1}^{m}(\mu_j-\theta)^2}{2\tau^2}} \\
&\propto (\tau^2)^{-a-1-1/2}e^{\frac{-\sum_{j=1}^{m}(\mu_j-\theta)^2-2b}{2\tau^2}} \\
&\propto (\tau^2)^{-a-1/2-1}e^{\frac{-\frac{\sum_{j=1}^{m}(\mu_j-\theta)^2-2b}{2}}{\tau^2}}
\end{aligned}
$$

Clearly this is a kernel of inverse Gamma. That is

$$
\tau^2|\mu_1,...,\mu_m,\theta,\sigma^2,X \backsim IG\left(a-1/2,\frac{\sum_{j=1}^{m}(\mu_j-\theta)^2-2b}{2}\right)
$$

5.4.4 Full conditional distribution of σ^2

Last but by no means the least we compute the full conditional for σ^2. This is a bit intense but tractable. The trick is recalling the relationship between gamma inverse gamma distribution. That is if X has a $gamma(\alpha,\beta)$ distribution then $Y = 1/X$ has

inverse gamma distribution $IG(\alpha, \beta)$

$$
\begin{aligned}
p(\sigma^2|\mu_1, ..., \mu_m, \tau^2, \theta, X) &\propto p(\mu_1, ..., \mu_1, \theta, \tau^2, \sigma^2|X) \\
&\propto \prod_{j=1}^{m}\prod_{i=1}^{n_j} p(x_{ij}|\mu_j, \sigma^2) \prod_{j=1}^{m} p(\mu_j|\theta, \tau^2)p(\theta)p(\tau^2)p(\sigma^2) \\
&\propto p(\sigma^2) \prod_{j=1}^{m}\prod_{i=1}^{n_j} p(x_{ij}|\mu_j, \sigma^2) \\
&\propto p(\sigma^2) \prod_{j=1}^{m}\prod_{i=1}^{n_j} N(x_{ij}|\mu_j, \sigma^2) \\
&\propto IG(c, d) \prod_{j=1}^{m}\prod_{i=1}^{n_j} N(y_{ij}|\mu_j, \sigma^2) \\
&\propto (\sigma^2)^{-c-1} e^{\frac{-d}{\sigma^2}} \prod_{j=1}^{m} \frac{1}{\sigma^{\sum n_j}} e^{-\frac{1}{2\sigma^2}\sum_i^{n_j}(x_{ij}-\mu_j)^2} \\
&\propto (\sigma^2)^{-c-1} e^{\frac{-d}{\sigma^2}} \frac{1}{\sigma^{\sum n_j}} \prod_{j=1}^{m} e^{-\frac{1}{2\sigma^2}\sum_i^{n_j}(x_{ij}-\mu_j)^2} \\
&\propto (\sigma^2)^{-c-1} e^{\frac{-d}{\sigma^2}} (\sigma^2)^{-\frac{\sum n_j}{2}} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij}-\mu_j)^2} \\
&\propto (\sigma^2)^{-(c+\frac{\sum n_j}{2})-1} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{m}\sum_{i=1}^{n_j}(ij-\mu_j)^2-\frac{d}{\sigma^2}} \\
&\propto (\sigma^2)^{-(c+\frac{\sum n_j}{2})-1} e^{-\frac{1}{\sigma^2}\left(\frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij}-\mu_j)^2+2d}{2}\right)}
\end{aligned}
$$

We recognize this to be the kernel of another inverse Gamma and hence

$$
\frac{1}{\sigma^2}|\mu_1, ..., \mu_m, \tau^2, \theta, X \propto gamma\left(c + \frac{\sum n_j}{2}, \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij}-\mu_j)^2 + 2d}{2}\right)
$$

Equipped with these full conditionals we can now use Gibbs sampling to generate MC random variable.

5.5 Gibbs Sampler in Action

The goal is to sample $\theta_1, \theta_2, ..., \theta_n$ given the joint distribution $p(\theta_1, \theta_2, ..., \theta_n)$. Due to the complexity of the joint posterior, we are instead going to sample from full conditional distributions which is the heartbeat of Gibbs sampling and assuming the full conditionals are standard univariate distributions. We kick off by by initializing $(\theta_1^0, \theta_2^0, ..., \theta_n^0)$ (can start from any reasonable starting point), the Gibbs sampler then draws variables in the following manner. Assuming we have the first k iteration then the k+1 iteration and thereafter will be obtained as:

$$\begin{aligned}
\theta_1^{k+1} &\sim p(\theta_1|\theta_2 = \theta_2^k, ..., \theta_n = \theta_n^k) \\
\theta_2^{k+1} &\sim p(\theta_2|\theta_1 = \theta_1^{k+1}, \theta_3 = \theta_3^k, ..., \theta_n = \theta_n^k) \\
&\vdots \quad \vdots \\
\theta_n^{k+1} &\sim p(\theta_n|\theta_1 = \theta_1^{k+1}, ..., \theta_{n-1} = \theta_{n-1}^{k+1})
\end{aligned}$$

As the number of iterations increases $(k \to \infty)$ [2] shows that the distribution of $(\theta_1^k, \theta_2^k, ..., \theta_n^k)$ converges to $p(\theta_1, \theta_2, ..., \theta_n)$. Which as k becomes large and large our kth sample comes from a distribution which is very close to the target distribution and it is as if we are sampling from the target distribution though not quite. The marginal distribution of $\theta_i^{(k)}$ converges to $p(\theta_i)$ for $(i = 1, 2, ..., n)$. Now that we know how to sample, let the sampling game begin.The sampling will done in matlab.

5.6 Matlab Simulations, Histograms, Run Charts and Inferences

Chemical treatment. The data is shown in the table below. This data is from four types of chemical treatment applied in variety of potatoes in Kenya. The treatments included Agrifos 110, Agrifos 65,untreated control and Ridomil alternated with Dithane (Conventional control) Treatments were applied at 7-day intervals during the crop season.

The sprayer was calibrated prior to commencement of fungicide application so as to deliver spray volume as per the recommendation of the product. Treatment applications were initiated at the onset of late blight symptoms. Fungicides were applied with lever-operated knapsack sprayers with maximum working pressures of approximately 300 kPa. To compare effectiveness of 4 types of Chemicals treatment the yields of equal size plots of the potatoes variety were collected

Table 5.1: Data of yields after treatment

Agri110	33.45	34.23	34.67	21.23	21.43	25.6	27.56	28.33
	30.34	36.95	39.05	35.33	33.00	37.85	34.55	33.45
	27.90	33.33	40.75	34.55	35.14	38.35	41.00	39.55
Agri65	29.86	35.96	32.91	22.56	19.67	19.57	27.77	25.57
	28.90	35.91	37.78	36.75	34.56	33.50	35.28	28.95
	25.14	28.95	36.87	33.15	31.00	38.90	37.98	38.95
Cont	26.50	29.43	27.32	11.24	13.68	12.21	17.22	16.47
	15.22	35.22	32.80	33.25	26.77	29.20	27.50	14.87
	17.45	18.45	32.00	29.07	26.85	25.70	29.56	31.00
Rid	31.58	33.54	35.05	24.95	24.23	23.85	29.85	26.40
	31.75	34.87	37.10	36.00	35.35	34.80	32.95	31.00
	30.05	34.50	38.72	32.80	35.87	39.45	39.56	40.66

We are going to simulate this in Matlab, run the MCMC and perform the inferences. In our simulation we are going to discard the first couple of runs (simulations) as burn in, then simulate the rest according our objective. We will then plot the run charts and histograms of all the parameters. We will conclude by finding the sample means and variances of the three parameters We will also try to change the starting points just to see the convergence behavior. In my simulation populations variance was assumed to vary The following are the charts and histograms of the parameters

The parameters estimates after 100000 simulation and 20000 burn-in, is shown in the Table 5.2

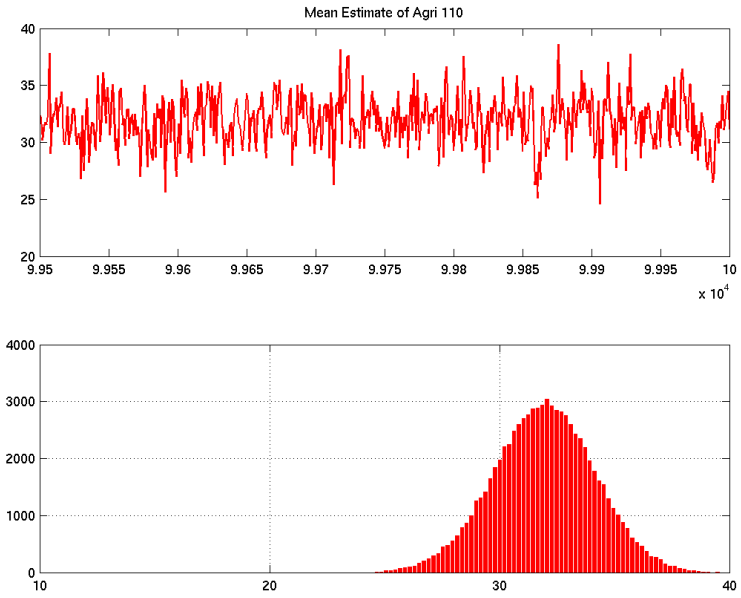

Figure 5.6: The Runchart and Histogram of Angril110

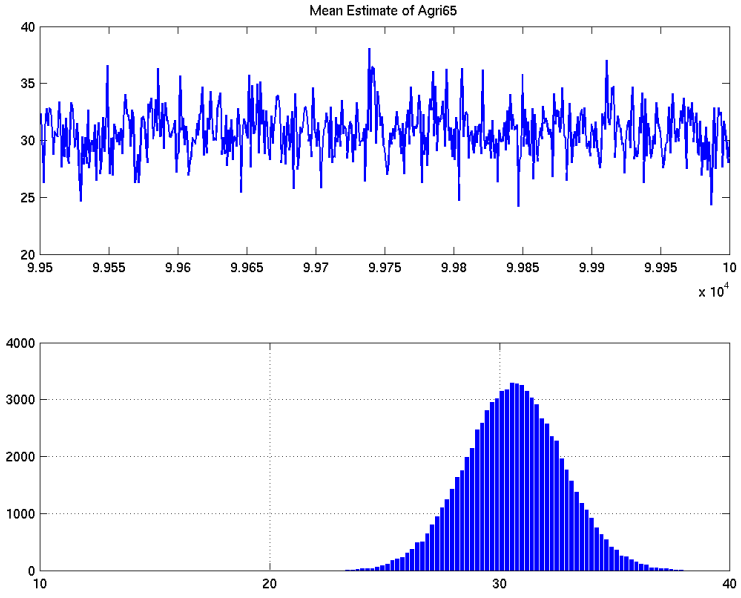

Figure 5.7: The Runchart and Histogram of Angril65

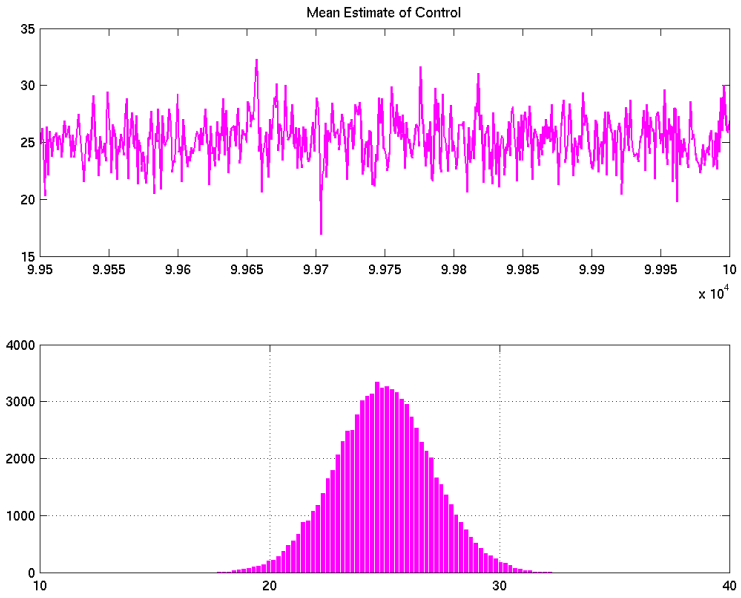

Figure 5.8: The Runchart and Histogram of Control Treatment

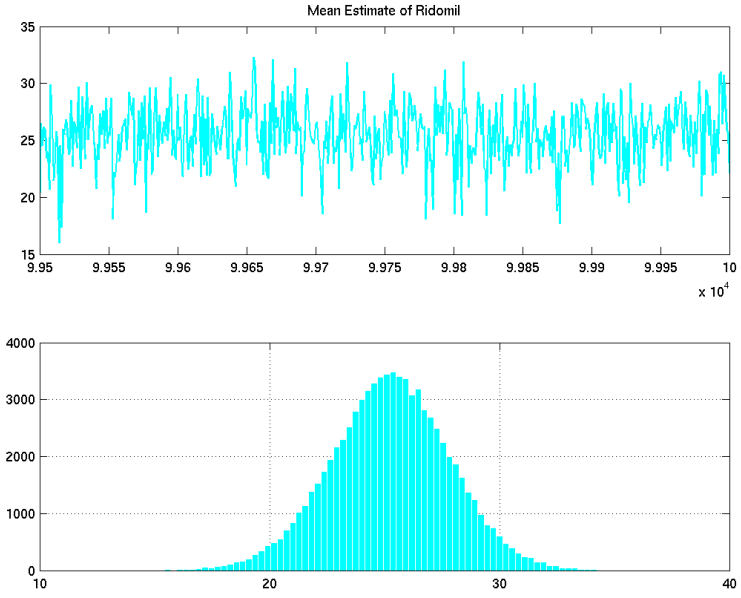

Figure 5.9: The Runchart and Histogram of Ridomil

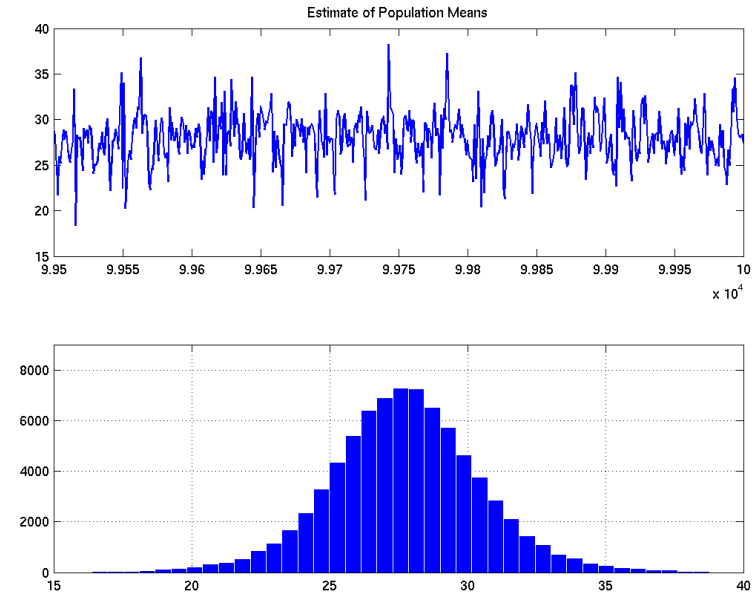

Figure 5.10: The Runchart and Histogram of μ^2

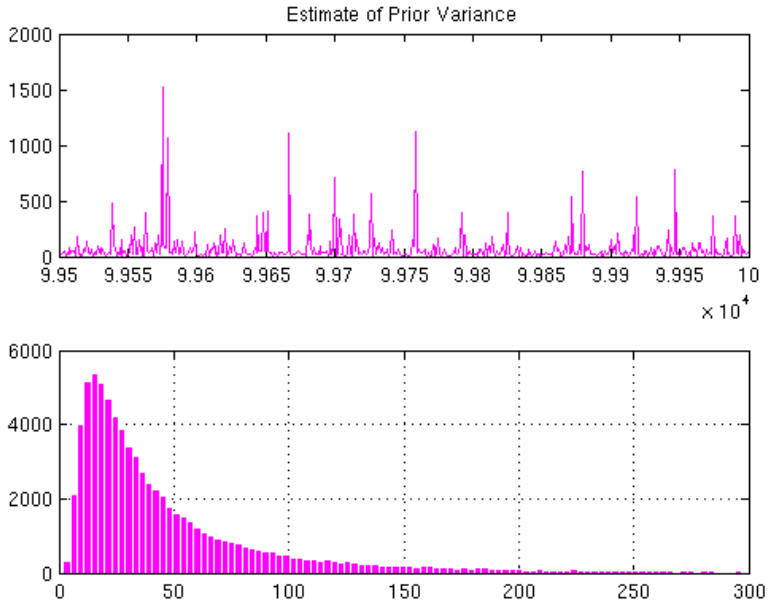

Figure 5.11: The Runchart and Histogram of τ^2

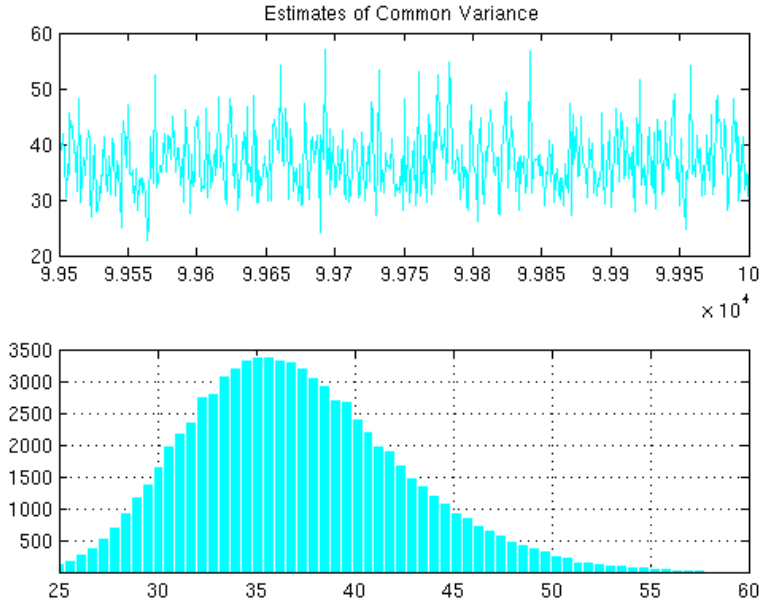

Figure 5.12: The Runchart and Histogram of σ^2

Table 5.2: Parameter Estimates

parameters	First Trial	Trial 2
σ^2	38.50	37.36
τ^2	77.63	79.12
θ	27.76	26.30
μ_1	31.93	31.96
μ_2	30.63	30.45
μ_3	24.96	24.93
μ_4	25.05	24.84

Chapter 6: Conclusion

In this thesis we proposed a Gaussian Hierarchical Bayesian model with a higher multidimensional joint posterior distribution which is not a standard distribution. The goal is to sample from this multivariate joint posterior but due to the curse of dimensionality and the nature of the distribution we could not easily draw samples from such a distribution .To safe the day we proved and showed techniques of computing full conditionals, we did this by presenting an MCMC technique and graphical inference algorithm for a Bayesian hierarchical model. The MCMC method we used in this thesis is Gibbs sampling. This sampler requires that we obtain closed form univariate full conditional distributions. We hereby showed that our Gaussian hierarchical model admits closed form full conditional distributions for all the parameters in the model. We showed and proved important probability results which came in handy when computing the full conditionals. Conjugate priors,though not highly mentioned in this thesis played a significant role when computing full conditionals, especially those that involved the product of normal and inverse gamma. We finally generated samples from these full conditionals on matlab using Gibbs sampling and given real data. From the simulated samples and after a burn in we estimated the parameters of our model. We have accomplished our mission for now but there is still more which can be done to improve Bayesian directed acyclic graphical model. This leads to a couple of interesting future ideas.

6.1 Future Work

. Since we considered a continuous distribution case model, a possible future research is to set up a similar model for discrete distribution or a mixture of continuous and non

continuous(discrete case). Every model need to be tested and checked, hence another possible research is to perform model checking and model fitting. It will be interesting to perform model transformation if need be and check how such transformation affect the multivariate joint posterior. It will also be interesting to use other inference techniques on this model and see how they compare with Gibbs Sampling.

There is still very little which has been done on attributes of hierarchical Bayesian model selection.

A parallel approach using other MCMC algorithms like Metropolis Hasting (Gibbs is a special case of MH) would be interesting to see how they compare.

Bibliography

[1] Casella G. and George E. , *Explaining the Gibbs Sampler*. Journal of the American Statistical Association, 46, 167-174, 1992

[2] Geman, S. ,and D.Geman *"Stochastic Relaxation, Gibbs Distributions and Bayesian restoration of Images"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 6, 1984

[3] Geman S, Geman D Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans Pattern Anal Machine Intelligence vol6, 721-741 1984

[4] Geman S. and Geman D., Stochastic relaxation, Gibbs distribution, and the Bayes restoration of images. IEEE transactions on pattern analysis and machine intelligence, 6, 721?74, 1984

[5] Gelfand A. E. and Smith F. M., 1990, Sampling-based approaches to calculating marginal densities. Journal of the America Statistical Association, 85, 398?409, 1990

[6] Gianola D, Foulley JL Variance estimation from integrated likelihood (VEIL). Genet Sel Evol 22, 403-417, 1990

[7] Neapolitan, R. *"Learning Bayesian Networks"*, Pearson Prentice Hall Series in Artificial Intelligence, 2004.

[8] Gelman, A., Carlin, J., et al. *Bayesian Data Analysis: Texts in Statistical Science*, Chapman and Hall, 1996.

[9] Borysiewicz M, Wawryzynczak A, et al *Stochastic algorithm for estimation of the model's unknown parameters via Bayesian inference*, FedCSIS, 2012.

[10] D. Koller and N. Friedman, *Probabilistic graphical models: Principles and techniques*, MIT Press, 2009.

[11] Hasting.w.k ,*Monte Carlo Sampling Methods Using Markov Chains and Their Applications*, Biometrika, Vol. 57, No. 1. (Apr., 1970), pp. 97-109

[12] Keats, A., Yea, E. and F.-S Lien *Bayesian inference for source determination with applications to a complex urban environment.* Atmos Environ., 41, 465-479, 2007.

[13] David M. Blei, Michael I. Jordan , et al. *Hierarchical Bayesian Models for Applications in Information Retrieval*,BAYESIAN STATISTICS 7, pp. 25?43, Oxford University Press, 2003.

[14] Dileep George, Jeff Hawkins, *A Hierarchical Bayesian Model of Invariant Pattern Recognition in the Visual Cortex*

[15] Gilks,W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, *Markov chain MonteCarlo in practice*, Chapman and Hall/CRC, 1996.

[16] Philip resnik, Eric hardisty, *Gibbs Sampling for the Uninitiated* UMIACS- TR2010

[17] Jordan M.I . *Learning in Graphical Models,* MIT Press, Cambridge, MA.

[18] Bolstand W. M. *Understsnding Computation Bayesian Statistics*, Wiley series in Computational Statistics. Inc, 2010

[19] Jayanta K. Gosh, Tapas Samanta , etl *An Introduction to Bayesian Analysis; Theory and Methods*, Springer Texts in Statistics, 2006