APPLICATIONS OF A NEW SAMPLING ALGORITHM FOR A BAYESIAN FINITE MIXTURE MODEL

By

Lin Xue

A practicum
Presented to the University of Manitoba
In partial fulfillment of the
Requirements for the degree of
Master of Science
In
Statistics

Winnipeg, Manitoba, Canada, May 2003

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION PAGE

APPLICATIONS OF A NEW SAMPLING ALGORITHM FOR A BAYESIAN FINITE MIXTURE MODEL

BY

LIN XUE

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirements of the degree of

Master of Science

Abstract

The Changjiang (Yangtze River) is the longest river in China and the third longest river in the world. The Changjinag drainage basin covers nearly one fifth of the total area of China. About 400 million people live in this area, which is about one third of the total population of China and about one fifteenth of the total population of the world. The Changjiang plays an important role in the lives of people and in the industries of China. The chemical elements and compounds are important indexes for the water quality in the Changjiang. The spatial distributions of major chemical elements in the river water along the Changjiang drainage basin remain an issue for geostatistics. In this work, we use a Bayesian finite mixture model to analyze the distribution of each major chemical elements. The traditional numerical method to deal with a Bayesian finite mixture model is the Markov Chain Monte Carlo method. But this method has difficulties dealing with finite mixture models with an unknown number of components, so a new sampling algorithm is used. This algorithm provides a valid and practical solution to a Bayesian finite mixture model.

Three variables are analyzed in this work, i.e., the concentrations of calcium (Ca), bicarbonate (HCO_3) and TDS (total dissolved solid). The multi-year averages of observations from 191 sampling stations in the Changjiang drainage basin over the period 1958-1990 are used for this study, so that 191 observations are available for each variable. Two subpopulations are identified in the distribution of each variable studied. Various marginal posterior distributions are given for the parameters in the mixture model. Furthermore, using the classifying probabilities, all 191 stations are classified into two groups.

Contents

List of Figures

List of Tables

Acknowledgements

First, I would like to express my sincere gratitude to my advisor Dr. Liqun Wang for the role he has played in the preparation of my practicum. I thank him for helping me choose this topic and for his expert counsel and patient encourgement. I also thank him for being an excellent teacher.

I also want to thank a member of my examining committee, Dr. Feiyue Wang of the Environmental Science Program & Department of Chemistry, who provided the Changjiang data sets used and gave me the guidance for my study of the geological setting in the Changjiang basin.

I am also thankful to committee member, Dr. James Fu, who patiently helped me resolve some issues relating to this research.

Finally, I wish to thank the Department of Statistics, for giving me the opportunity to pursue my M.Sc. degree in Statistics and for the financial support over the past two years.

Chapter 1

Introduction

1.1 Overview

The Changjiang (Yangtze River) is the longest river in China and the third longest river in the world. The Changjiang drainage basin covers an area of 1.8 million square kilometers. There are about 400 million people living in the Changjiang basin. This river has a significant influence on the lives of people and industries in China. The river originates in western China, and flows through the entire central region before it empties into the Pacific Ocean on the east coast. The geographic setting along the river basin is very complicated. There are plateaus, valleys, basins and plains. The river is joined by a large number of tributaries.

Despite all of its significance in the world, the major element chemistry of the Changjiang has not been well studied (Chen at al., 2002). The chemical elements in the river water are important indexes for the water quality and weathering processes in the basin. The spatial distributions of major chemical elements represent a gap in our understanding of the distributions of chemical elements in the river water. To understand the distributions

of major chemical elements in the river basin and tributaries is of great importance to human lives and industries. In addition, it is an interesting topic in geography, geology and environmental science. Figure 1.1 shows a map of the Changjiang drainage basin and its location in China.

From a geographical point of view, the basin can be divided into 10 sub-basins. The number of sub-populations for the underlying distribution of each chemical element is however unknown. The main goal of this practicum is to study the distributions of major chemical elements in the river basin and identify the number of sub-populations in these distributions. We study this question using a finite mixture model with Bayesian approach.

Two sub-populations are identified for the two major chemical elements and the total dissolved solid concentration of chemical elements. The marginal posterior distributions of the paramenters in the mixture model are also given, and estimations of the parameters are summarized in tables.

1.2 The Changjiang Data

The data were collected in the Hydrological Yearbooks of China, and have been used by Chen et al. (2002). Their research revealed the variability of major chemical elements and the underlying mechanisms of the variabiltiy. The spatial distributions of major chemical elements remained an issue for further study.

In this work, we will not discuss the data from a chemical or geological point of view, but from a statistical aspect. We are more interested in the spatial distributions of these chemical elements in the Changjiang basin. For example, we would like to know how many sub-

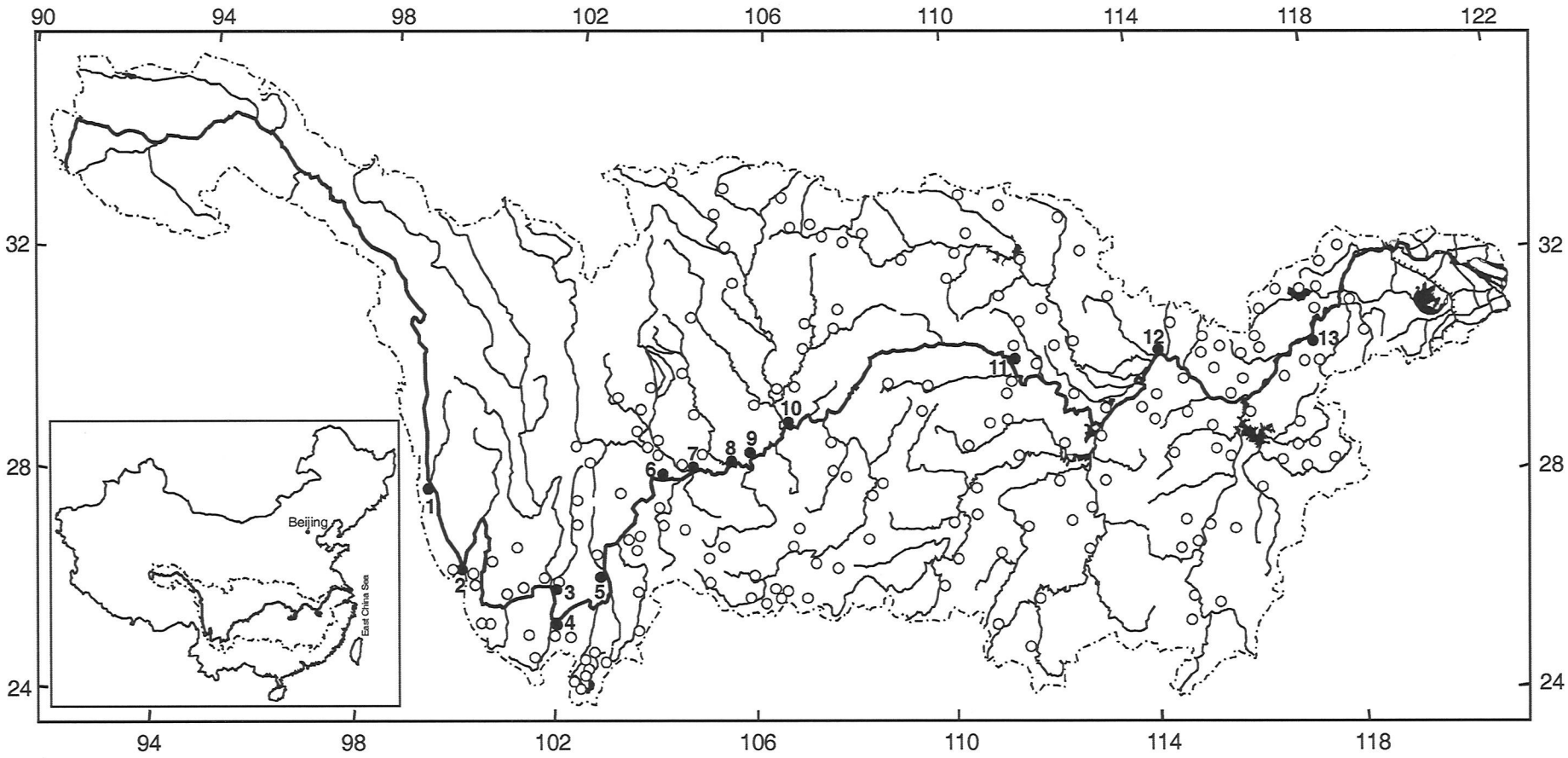

Figure 1.1 The Changjiang drainage basin and the sampling stations. The bold line is the main channel. The filled and open circles are sampling stations along the main channel and tributaries, respectively.

populations there are in the Changjiang basin and what the distributions of the parameters look like. There are 191 sampling stations in the Changjiang drainage basin. These sampling stations are monitored monthly over the period 1958-1990 for major chemical elements and compounds (e.g., Ca, Mg, Na, K, HCO_3, SO_4, Cl, and Si). Figure 1.1 shows the map of the Changjiang drainage basin and the locations of sampling stations. The collection and analytical methods of these data can be found in Chen et al. (2002). In this work, we use the multi-year averages over the period 1958-1990 for these 191 stations and analyze three variables:

TDS (mg/l): the total dissolved solid concentration, in miligram per liter.

Ca (mg/l): calcium concentration, which constitutes, on average, 16% of the TDS, in miligram per liter.

HCO_3(mg/l): bicarbonate concentration, which constitutes, on average, 64% of the TDS, in miligram per liter.

In order to examine what distributions these random variables might have and how many sub-populations there are, we plotted their histograms in Figure 1.2, 1.3 and 1.4, respectively.

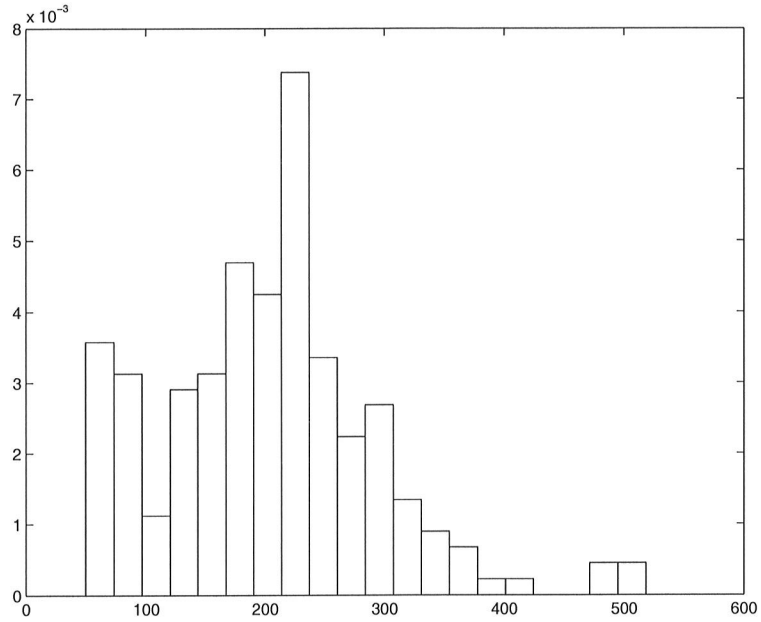

Figure 1.2: Histogram of TDS

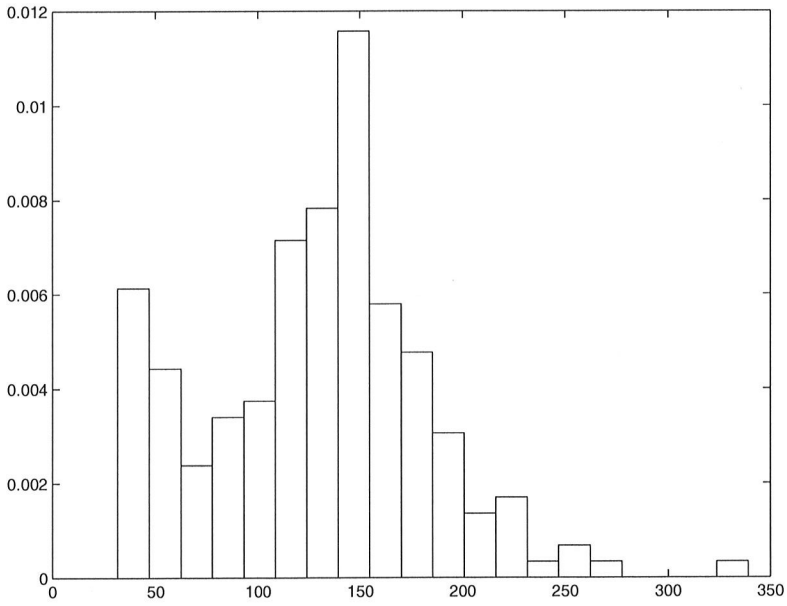

Figure 1.3: Histogram of HCO_3

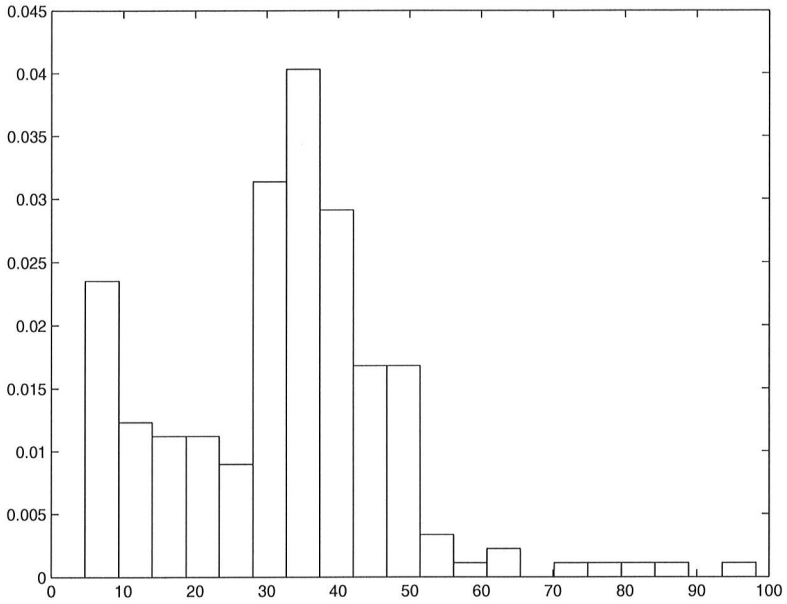

Figure 1.4: Histogram of Ca

From these histograms it is difficult to say what distributions the variables might have. We use the logarithm transformation technique to transform the original data. Figures 1.5, 1.6, and 1.7 show the histograms of the logarithm of TDS, HCO_3 and Ca, respectively. Now, all these histograms show a clear pattern of a mixture of two or more symmetric distributions. This indicates that, first of all, each variable has more than one underlying sub-population; and second, each sub-population has a symmetric distribution. So we choose a mixture model to describe the structure of the distribution.

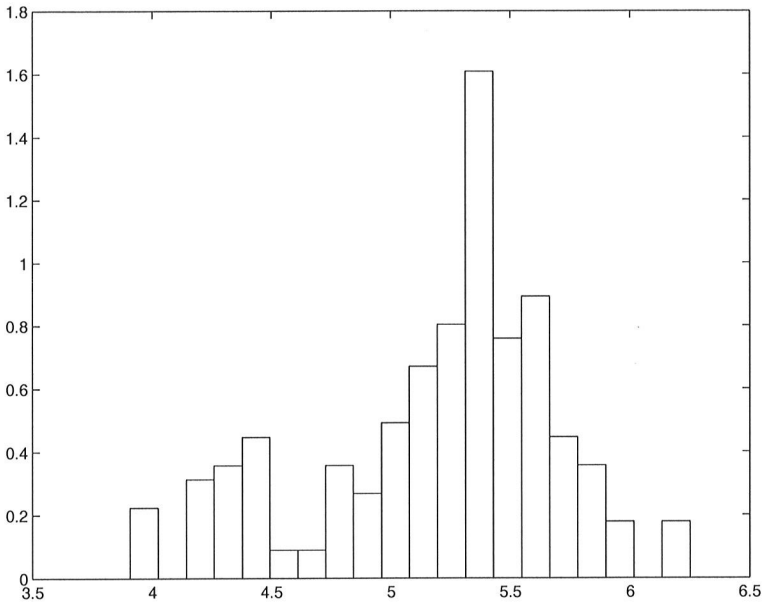

Figure 1.5: Histogram of logarithm of TDS

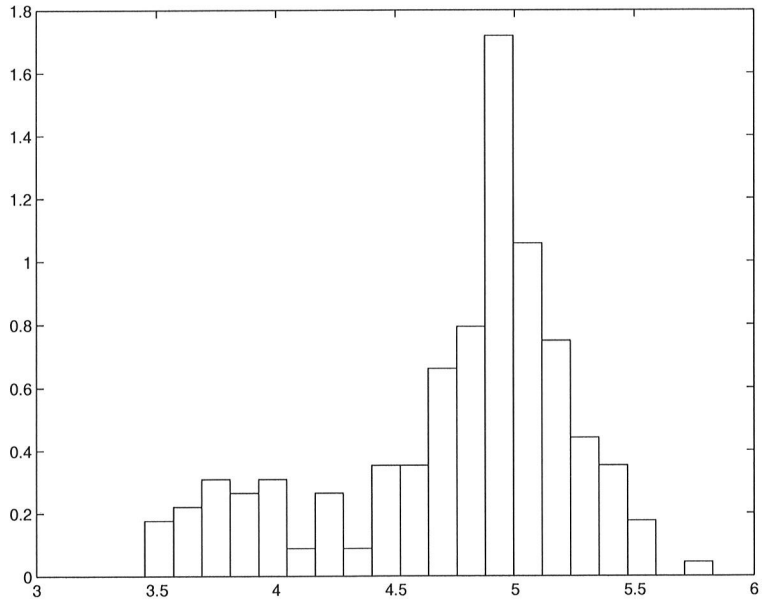

Figure 1.6: Histogram of logarithm of HCO_3

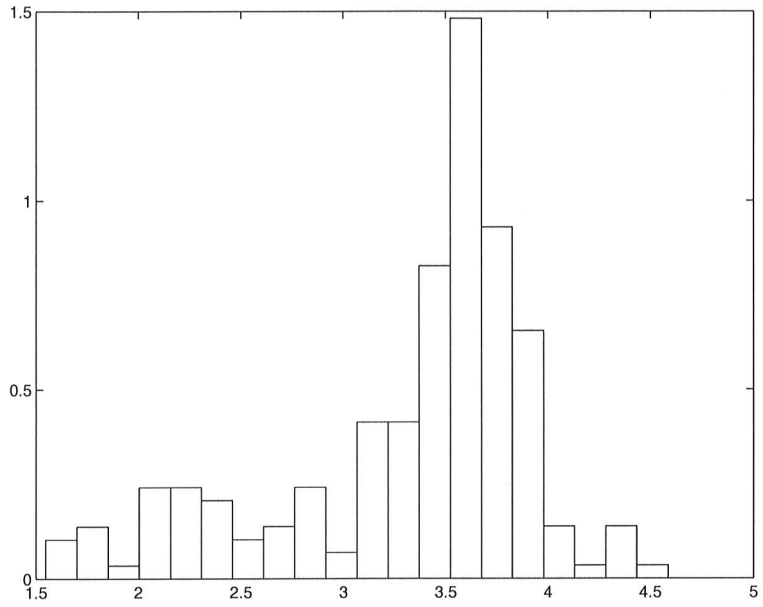

Figure 1.7: Histogram of logarithm of Ca

1.3 Methodology

Statistical models are very useful tools for studying random phenomena. The models provide probability distributions for random variables and which allow us to draw conclusions about the variables. A random variable is one whose value is a numerical outcome of a random phenomenon, e.g., rolling dice, tossing a coin. There are many simple statistical models, e.g., binomial, Poisson, normal, gamma, etc. These models provide probability distributions for some random variables. Based on these models, we can tell what values the random variables can take and how to assign probabilities to those values or ranges of values.

Although many phenomena allow simple, direct probability modelling, there are some observed phenomena which can be too intricate to be modelled by these simple forms. For example, the data are collected from an underlying distribution which has more than one sub-population. A mixture model is a more flexible statistical model for the complex data. We

can use a mixture of a number of simple distributions to estimate the underlying distribution from which we collected our data.

The histograms of major chemical elements show that there is more than one mode existing in the (logrithm of) data. So, a simple model is not suitable for these data, and a mixture model is more appropriate. The mixture model has the flexibility of catching the structure of the underlying distribution with more than one mode. In the mixture model, uncertainty may come from the unknown number of components, the unknown mixing weights and the unknown distribution parameters. Unknown mixing weights and distribution parameters represent the most common situations in applications. Determining the number and nature of the components is very challenging in applications. There is a remarkable variety of estimation methods that have been applied to finite mixture models. In particular, the method of moments, the maximum likelihood, and the Bayesian approach are often used.

In this work, we use the Bayesian approach to the finite mixture model. In Bayesian approach, one can draw inferences about parameters, by reference to their (marginal) posterior densities. The statistical analysis of such a mixture distribution has been difficult. This is due primarily to the fact that closed-form formulas generally do not exist for estimators of the various parameters of the mixture model. Hence, numerical methods are required for such situations. There are several numerical methods recently developed for statistical inferences, e.g., the EM algorithm, data augmentation and the Markov Chain Monte Carlo method. Comprehensive descriptions of these techniques are provided in Besag and Green (1993), Smith and Roberts (1993). Each of these numerical methods has advantages and disadvantages (Diebolt and Robert, 1994; Besag and Green, 1993).

The most commonly used numerical method for the Bayesian mixture model is the

Markov Chain Monte Carlo (MCMC) method. Smith and Roberts (1993) described the advantages of MCMC over traditional forms of Bayesian computation. In particular, MCMC invites one to go beyond simple point and interval estimates. But this method has a number of limitations. First, the rate of convergence can be slow. Second, this method has the weakness dealing with a multi-modal distribution (Besag and Green, 1993). In this paper, we introduce a new sampling algorithm developed by Fu and Wang (2002) to estimate the distributions of parameters in the mixture model and apply this method to our data sets. This method overcomes the disadvantages of MCMC. It is non-iterative, easy to implement and efficient in computation. This method is based on the concept of random discretization of the density function and it requires only the knowledge of the density function up to a normalizing constant. This algorithm has performed very well for many benchmark examples (see Fu and Wang, 2002).

The results give us the posterior distributions of the parameters in the mixture model and the predictive density for the future observations. Furthermore, marginal posterior densities and point estimates of the parameters of interest can be obtained. As a by-product, this algorithm also gives us the approximate maximum likelihood estimates (AMLE) of the parameters. Two sub-populations are identified in the distributions of two chemical elements (HCO_3, Ca) and total dissolved solid concentration (TDS) of chemical elements. Moreover, the 191 sampling stations for each variable are classified into two groups according to the classifying probabilities.

1.4 The Scope of This Work

The paper is divided into five parts: In Chapter 1 (this chapter), we briefly introduce the issue of geostatistics, the datasets and the method we use for this study. In Chapter 2, we give the background knowledge of the finite mixture model and the Bayesian framework. Chapter 3 gives the general computational methods for the Bayesian mixture model and an introduction of Fu and Wang's new sampling method. In Chapter 4, the Changjiang data are analyzed using the Bayesian mixture model and Fu and Wang's sampling method. Posterior distributions and classification of 191 sampling stations are presented there. Finally, conclusions are given in Chapter 5.

Chapter 2

A Bayesian Finite Mixture Model

2.1 A Finite Mixture Model

Finite mixture distributions have been used as models of distributions in modern statistics for more than a century (Titterington et al., 1985). A finite mixture model is a flexible model in practice. It is especially useful for data sets coming from a population consisting of many sub-populations. For example, if blood pressures are taken from a population including both cancer patients and non-cancer patients, the distribution of blood pressures reflects the mixture of distributions of both groups. The distuibution of heights of students at the University of Manitoba is a mixture representing the heights of both male and female students. A mixture model can be used in problems of this type, where the population of sampling units consists of a number of sub-populations, each of which has a relatively simple distribution, e.g., has a simple model.

Suppose $f_1(x)$, $f_2(x)$, $\cdots$, $f_k(x)$ are k probability density functions defined on the same

sample space $\mathbb{R}$. The mixture distribution of $f_1(x)$, $f_2(x)$, $\cdots$, $f_k(x)$ is defined as

$$f(x) = w_1 f_1(x) + w_2 f_2(x) + \cdots + w_k f_k(x), x \in \mathbb{R}. \tag{2.1}$$

where the mixing weights $w_j \geq 0$, $j = 1, 2, ..., k$, and $\sum w_j = 1$. k is the number of components. Functions $f_j(x)$ are called the component densities. It is straightforward to verify that (2.1) does define a probability density function because, first, $f(x) \geq 0$ for all x, and second,

$$\begin{aligned} \int_{\mathbb{R}} f(x)dx &= \int_{\mathbb{R}} (\sum_{j=1}^{k} w_j f_j(x))dx \\ &= \sum_{j=1}^{k} w_j \int_{\mathbb{R}} f_j(x)dx \\ &= \sum_{j=1}^{k} w_j = 1. \end{aligned} \tag{2.2}$$

In many situations, $f_1(x), f_2(x), \cdots, f_k(x)$ will have specified parametric forms and the right-hand side of (2.1) will have the more explicit representation,

$$f(x) = w_1 f_1(x|\theta_1) + w_2 f_2(x|\theta_2) + \cdots + w_k f_k(x|\theta_k). \tag{2.3}$$

where θ_j denotes the unknown parameter (vector) occurring in $f_j(x)$. It is common to assume that the mixture components are all from the same parametric family, such as the family of normal distributions (Leonard and Hsu, 1999; Gelman et al., 1995; Titterington et al., 1985). In practice, the weights w_j are usually unknown as well. Moreover, in many applications, including this work, even the number of components K is unknown. Objectives in density estimation include the assessment of the number of components of a finite mixture model and inference about the number of modes of a population distribution.

2.2 Bayesian Approach for The Finite Mixture Model

2.2.1 Bayesian Framework

As we mentioned above, there is a variety of estimation methods that have been applied to the finite mixture model problems, but we only pursue Bayesian approach in this work. The advantage of Bayesian inference is that one can treat every unknown parameter (e.g., parameters, mixing weights, and the number of components) as a random variable and assign a probability distribution to it. This method gives us the most flexibility to infer conclusions about the unknown parameters in the model, given the observables. In other words, model performance is addressed through features of posterior distributions of unknown variables, given the observable ones. In the Bayesian framework, the parameters occurring in (2.3) are regarded as random variables, which have their own probability distributions.

In this work, we use normal distribution components for the mixture model. One way of seeing that the class of normal mixture densities is a very broad one comes from the fact that any density can be approximated arbitrarily closely in a certain sense by a normal mixture density (Gelman et al., 1995). The unknown parameters in this mixture model include the mixing weights, the unknown number of components and the parameters in the normal distributions, e.g., the means and variances.

In the following, let X be the vector of random variables (the sample) of interest and $p(X|\theta)$ be the density of X given the vector of unknown parameters θ. First, we start with a joint probability distribution for θ and X. According to Bayesian rule, the joint density can be written as a product of two densities that are often referred to as the prior distribution

$p(\theta)$ and the sampling distribution $p(X|\theta)$ respectively,

$$p(\theta, X) = p(\theta)p(X|\theta). \tag{2.4}$$

Given the observations of X, the sampling distribution $p(X|\theta)$ is called the likelihood function of θ. Statistical inference is based upon features of the conditional distribution of the unknowns in the model, given the observables, using Bayesian theory,

$$p(\theta|X) = \frac{p(\theta, X)}{p(X)} = \frac{p(\theta)p(X|\theta)}{p(X)}. \tag{2.5}$$

$p(\theta|X)$ is called the posterior distribution, and $p(X) = \sum_\theta p(\theta)p(X|\theta)$ for discrete θ and $p(X) = \int p(\theta)p(X|\theta)d\theta$ for continuous θ. The denominator of (2.5) does not depend on θ and, with fixed X, can be considered as a constant. As such (2.5) is often written as

$$p(\theta|X) \propto p(\theta)p(X|\theta). \tag{2.6}$$

Since, $p(\theta|X)$ is a probability distribution of θ, given X,

$$\int \frac{p(\theta)p(X|\theta)}{p(X)} d\theta = \frac{1}{p(X)} \int p(\theta)p(X|\theta)d\theta = 1. \tag{2.7}$$

Hence,

$$\int p(\theta)p(X|\theta)d\theta = p(X). \tag{2.8}$$

This implies that the integral of $p(\theta)p(X|\theta)$ with respect to θ does not equal to 1, but to the constant that we omit from the posterior distribution. Given X, the value of the integral in (2.8) is called the normalizing constant. Equation (2.6) is the technical core for Bayesian inference. The primary task of any specific applications of Bayesian approach is to develop the prior probability distribution $p(\theta)$ and likelihood function $p(X|\theta)$, then perform the necessary computations to summarize $p(\theta|X)$ in appropriate ways.

2.2.2 Prior Distribution

The prior distribution $p(\theta)$ generally represents the uncertainty about θ before data are examined. There are two interpretations about prior distribution, one being that the prior distribution represents a population of possible parameter values, from which the values of parameters θ of interest have been drawn. In such a case, the prior distribution is called the formative prior distribution. This is the way in a more subjective sense to express our knowledge about θ as if its values could be thought of as random realizations from the prior distribution. The other interpretation of prior is that there is no logical basis for assigning one prior distribution to θ as opposed to any other. In such situations, we simply assign θ an uniform prior distribution which is called the informative prior distribution. This way, the prior distribution plays a minimal role in the posterior distribution.

In this study, we need to assign prior distributions to mixing weights, the unknown number of components and the means and variances in the component distributions. It is a common practice that we assume the normal components have different means and the same variances. We also assume the means come from the same distribution. Specifically, there will be a prior distribution for the vector of means and a prior distribution for the variance. According to the nature of the means, we assign normal distribution to the vector of means in the mixture model (Leonard and Hsu, 1999; Gelman et al., 1995). Other priors will be described in Chapter 4.

2.2.3 Posterior Distribution

Mathematically, the posterior distribution is the product of prior distributions and the likelihood function (up to a constant). It represents a compromise between the data and the prior distributions. The prior distributions represent our prior belief about θ, while the likelihood function represents information about θ provided by the data. The posterior distribution contains all the current information for the parameters about which we wish to draw conclusions. The posterior distribution in this work is the product of the likelihood function multiplied by the prior distributions for mixing weights, the number of components and the prior distributions of the vector of means and variance. The means and variance are the unknown parameters in the normal components. It is easy to see that the posterior distribution comes out with a very complicated form.

In many situations, it is not feasible to perform calculations on the posterior density function directly. In such cases simulation method is particularly useful to obtain inferences from the posterior distribution. The flexibility of Bayesian inference is reflected in that the posterior distribution can be summarized by simulation. In this study, various posterior distributions of the parameters and predictive density are estimated using simulation method of Fu and Wang (2002).

2.2.4 Conjugate Family

Prior distributions represent the prior beliefs for the parameters before the data are examined, and there are many distributions which can be used. However, there are some families of distributions which may be more desirable than others for computational reason, such as

conjugate families. The formal definition is as follows:

If $\mathcal{S}$ is a class of sampling distributions $p(X|\theta)$, and $\mathcal{P}$ is a class of prior distribution for θ, then the class $\mathcal{P}$ is conjugate for $\mathcal{S}$ if

$$p(\theta|X) \in \mathcal{P} \text{ for all } p(X|\theta) \in \mathcal{S} \text{ and } p(\theta) \in \mathcal{P}. \tag{2.9}$$

Simply stated, conjugacy is the property that the posterior distribution follows the same parametric form as the prior distribution. Conjugacy has practical advantage in some cases, especially in computation, but this is not always the case. A non-conjugate prior distribution can make interpretation less transparent and computation more difficult, but a non-conjugate does not pose any conceptual restrictions. One advantage of Fu and Wang's (2002) algorithm is that no conjugate restriction is needed, which gives us more flexibility in analyzing practical problems.

2.2.5 Bayesian Hierarchical Model

In the Bayesian finite mixture model, we assign a joint probability distribution for the parameters in the model. For example, in this study, the parameters include the mixing weights, the means and variance in the components distributions and the number of components. For conceptual simplicity, we assume that given the number of components, the other sub-vectors are independent. So the joint probability would be the distribution of the number of components times the conditional distributions of other parameters given the number of components. This is a hierarchical structure.

The mixing weights, the means and the variance can be thohght of as sub-vectors. Let θ_j represents a sub-vector in the vector θ. We assume that the θ_j has the same distribution. In

other words, we are thinking of the parameters θ_j as independent samples from a population distribution governed by some unknown parameter (vector) ϕ_j;

$$p(\theta|\phi) = \prod_{j=1}^{J} p_j(\theta_j|\phi_j). \tag{2.10}$$

In general, ϕ_j is unknown and we must assign a prior distribution to ϕ_j. Which is called the hyper-prior distribution, $p_j(\phi_j)$, and ϕ_j is called a hyper-parameter (vector). The appropriate Bayesian joint prior distribution is

$$p_j(\phi_j, \theta_j) = p_j(\phi_j)p_j(\theta_j|\phi_j), \tag{2.11}$$

and the joint posterior distribution is

$$p_j(\phi_j, \theta_j|x) \propto p_j(\phi_j, \theta_j)p(x|\phi_j, \theta_j) = p_j(\phi_j, \theta_j)p(x|\theta_j). \tag{2.12}$$

In most real problems, in order to constrain the hyper-parameter into a finite region, the common strategy is to assign an informative distribution to ϕ_j or estimate ϕ_j from the data sets.

For example, in this work, we assign a normal distribution to the vector of means, the mean and variance in this prior will be hyper-parameters. We also need to assign hyper-prior distributions to them. In order to simplify the computations, we extract the information from the data to constrain the hyper-paramters into a finite region.

2.2.6 Predictive Distribution

Inference about future values of a random variable X, denoted by $\tilde{X}$, is often called predictive inference. In Bayesian framework, this is done through the density

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

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

$$= \int p(\tilde{X}|\theta)p(\theta|X)d\theta. \tag{2.13}$$

Equation (2.13) displays the posterior predictive distribution as an average of conditional predictions over the posterior distribution of θ.

The predictive distribution gives us a probability description about the future values of observations, given current observations. For example, what kinds of values might the future observations take and what probabilities are associated with them.

Chapter 3

Computation of The Bayesian Mixture Model

While the Bayesian mixture model provides a feasible and coherent description of data coming from a population with more than one sub-population, the numerical computation of this model remains a challenging task. Markov Chain Monte Carlo (MCMC) method is a commonly used tool for Bayesian computations (Gilks et al., 1998). When applied to real problems, major difficulties arise from multi-modality of the underlying distribution and ill-shaped sample space, especially in dealing with a mixture model with an unknown number of components. In this chapter, we introduce a new discretization-based numerical algorithm developed by Fu and Wang (2002). This method has the advantages of simplicity in concept and efficiency in computation. It is an effective method of simulation, especially for a Bayesian mixture model.

Posterior distribution is the major component in Bayesian inferences. In practice, except for some very special cases, the posterior distribution and predictive distribution do not

usually come out with simple and standard forms, so that direct computation and interpretation are not feasible. In such cases numerical methods are required to analyze the target distribution. Simulation is a numerical method used to perform statistical inference about the distribution. It forms a central component of applied Bayesian analysis.

Generally, simulation is used to generate samples from the target distribution. The computer program utilizes random numbers to generate the values of random variables having the assumed probability distribution, which are used to estimate the distribution of parameters of interest. In other words, we draw a sample according to the target distribution, and then use the simulated data to construct a histogram to display the sample distribution. It is a relatively easy way to get an idea of the form of the distribution when analytic difficulties arise. In performing simulation, it is helpful to think that, given a large enough sample, the histogram provides practically complete information about the density function. Also, certain quantities of the distribution, such as the mean and variance, can be estimated using the drawn sample.

3.1 The Inverse Transform Method for Random Number Generation

As an introduction to the idea of simulation, let us first briefly review the inverse transform method, which is the foundation for generating random numbers from a distribution of either a discrete or a continuous type (Ross, 1997).

Let U be a uniform random variable over interval (0,1), denoted by $U \sim U(0,1)$. For any

given cumulative distribution function F(x), the random variable X defined by $X = F^{-1}(U)$, has distribution F.

The method shows that we can generate a random number X from the continuous distribution F by generating a random number U from the uniform distribution $U(0,1)$ and then setting $X = F^{-1}(U)$.

For a discrete distribution, the procedure is as follows. Suppose a discrete random variable X has probability mass function

$$P(X = x_j) = P_j, j = 0, 1, \cdots, \sum_j P_j = 1,$$

To generate a random value of X, we generate a random number from $U(0,1)$ and set

$$X = \begin{cases} x_0, & \text{if } U < p_0, \\ x_1, & \text{if } p_0 \leq U < p_0 + p_1, \\ \cdots & \\ x_j, & \text{if } \sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i, \\ \cdots & \end{cases} \tag{3.1}$$

3.2 A Discretization-Based Sampling Method

This discretization-based sampling algorithm has been developed by Fu and Wang (2002) and has been applied successfully to some benchmark examples of Bayesian models and also in some real applications. A good review of this algorithm and its applications can be found in Fu and Wang (2002). This algorithm requires only the knowledge of the distribution up to a normalizing constant. It is dimension-free and non-iterative in contrast to Markov Chain Monte Carlo method. It can be used in the computation of a high dimension-multivariate

distribution, and in a Bayesian mixture model with an unknown number of components. To illustrate this algorithm, let us consider the case where the posterior density function $p(\theta)$ is continuous. We only need the knowledge of $p(\theta)$ up to a normalizing constant, which means we may ignore any constant in $p(\theta)$. Suppose θ is a vector of d-dimension and $p(\theta)$ has a compact support $S(p) = [a, b]^d$, where $-\infty \leq a \leq b \leq \infty$ are known. The algorithm consists of three steps: discretization, contourization and sampling.

3.2.1 Discretization

The first step of Fu and Wang's algorithm is to create a discrete set of $S_n(p)$, which approximate the sample space $S(p)$. This can be done by first generating n independent random points from the uniform distribution on $[a, b]^d$, $\theta_1, \theta_2, \cdots, \theta_n \sim U[a, b]^d$. There are thus n random vectors. These n random vectors are independently, identically and uniformly distributed on $[a, b]^d$. Define $S_n(p) = (\theta_j, j = 1, 2, \cdots, n)$. Then $S_n(p)$ is the discretized version of $S(p)$. When n is large, $S_n(p)$ is an approximation of $S(p)$.

3.2.2 Contourization

Second, we group $S_n(p)$ into l groups, E_1, E_2, $\cdots$, E_l. Each group has u points, where $u = n/l$. E_1 contains the points of θ_j, at which $p(\theta_j)$ have the highest values, E_2 has the points at which $p(\theta_j)$ have the second highest values, and so on. We call E_i the contours. These contours are mutually disjoint, and the union of these contours forms $S_n(p)$. Further, we define a discrete distribution on the contours E_i, which is proportional to the average

heights of the posterior densities on each contour E_i:

$$P(E_i) = \frac{1}{u} \sum_{\theta_j \in E_i} p(\theta_j), i = 1, 2, \cdots, l. \tag{3.2}$$

3.2.3 Sampling

Suppose we would like to generate m independent and identically distributed observations from $S_n(p)$ according to the posterior distribution $p(\theta)$. We use the inverse transform method described earlier. First, we randomly sample with replacement m contours according to the discrete distribution (3.2). Suppose m_i is the number of occurance of E_i in these random draws, $\sum m_i = m$. Then within contour E_i, we sample with replacement m_i points randomly. All points thus obtained forms the desired sample. A detailed description of this method can be found in Fu and Wang (2002).

Chapter 4

The Changjiang (Yangtze River) Data

In this section, we apply the sampling algorithm of Chapter 3 to the Changjiang data. Two sub-populations are identified in the data sets and various marginal posterior distributions are given. Furthermore, we calculate the classifying probabilities to classify the 191 stations into two groups.

4.1 The Bayesian Mixture Model

As mentioned earlier in section 1.2, we study mainly three variables: TDS, HCO_3 and Ca. Histograms of the logarithm of the samples are shown in Figures 1.5, 1.6 and 1.7, respectively. The histograms show that there is more than one mode existing in all the three data sets. So a Bayesian finite mixture model is suitable for the distribution of the data.

A Bayesian finite mixture model with normal density components is used for each of these three variables. From the histograms it is difficult, however, to tell how many components there are. It is therefore realistic and desirable, that we treat the number of components as

an unknown paremeter (a random variable).

Suppose the data x_i, $i = 1, 2, \cdots, N$ represent an independently and identically distributed random sample from a finite normal mixture distribution. We apply a normal mixture model with an unknown number of components, K, to each of the variables, so that the likelihood function is given by

$$\prod_{i=1}^{N} f(x_i|\mu^k, \sigma_k^2, w^k, k) = \prod_{i=1}^{N} \sum_{j=1}^{k} \frac{w_{kj}}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{(x_i - \mu_{kj})^2}{2\sigma_k^2}\right] \tag{4.1}$$

where $\mu^k = (\mu_{k1}, \mu_{k2}, \cdots, \mu_{kk})$ is the vector of means, given K, and σ_k^2 is the variance, given K. As mentioned earlier, we assume that different components have different means but the same variance. $w^k = (w_{k1}, w_{k2}, \cdots, w_{kk})$ are the mixing weights. As explained earlier, the number of components K is unknown and is treated as a random variable.

Now let us consider the prior distributions. Note that the distributions of all other parameters depend on the value of K. This is the hierarchical structure as we mentioned in section 2.2.5. Specifically, given different K, there will be different set of parameters:

$$k = 1,\ w_{11} \equiv 1,\ \mu_{11},\ \sigma_1^2;$$

$$k = 2,\ w_{21},\ w_{22}\ (w_{21} + w_{22} = 1),\ \mu_{21},\ \mu_{22},\ \sigma_2^2;$$

$$k = 3,\ w_{31},\ w_{32},\ w_{33}\ (w_{31} + w_{32} + w_{33} = 1),\ \mu_{31},\ \mu_{32},\ \mu_{33},\ \sigma_3^2.$$

The above parameters are all unknown and we assume that, given K, μ^k, σ_k^2 and w^k are conditionally independent (Fu and Wang, 2002). Thus, the full posterior distribution is

$$\prod_{i=1}^{N} f(x_i|\mu^k, \sigma_k^2, w^k, k) p\left(\mu^k|k\right) p\left(\sigma_k^2|k\right) p\left(w^k|k\right) p(k) \tag{4.2}$$

As mentioned earlier, according to the nature of the location parameters, it is common to choose the normal distribution for them. In order to avoid labelling problems, there is a restriction for μ^k (Fu and Wang, 2002), so the prior distribution is:

$$p\left(\mu^k|k\right) = k! \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{(\mu_{kj}-\mu_0)^2}{2\sigma_0^2}\right], \mu_{k1} < \mu_{k2} < \cdots < \mu_{kk} \tag{4.3}$$

A simple prior for σ_k^2 is to assume σ_k^2 is known. Such situation will seldom arise in practice (Leonard and Hsu, 1999). The common prior for σ_k^2 is the inverse-gamma distribution (Ibrahim et al., 2002; Escobar and West, 1995). Invers-χ^2 is also used as a prior of σ_k^2 (Belisle et al., 2002; Gelman et al., 1995). In this work, we use the inverse-gamma prior distribution

$$p\left(\sigma_k^2|k\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\sigma_k^2\right)^{-\alpha-1} e^{-\beta/\sigma_k^2}, \sigma_k^2 > 0 \tag{4.4}$$

For the weights w^k, a natural choice of prior distribution is the Dirichlet distribution (Diebolt and Robert, 1994):

$$p\left(w^k|k\right) = \frac{\Gamma(k\gamma)}{\Gamma(\gamma)} \prod_{j=1}^{k-1} w_{kj}^{\gamma-1} \left(1 - \sum_{j=1}^{k-1} w_{kj}\right)^{\gamma-1}, 0 \leq \sum_{j=1}^{k-1} w_{kj} \leq 1. \tag{4.5}$$

Since the K has a discrete distribution, we assign a discrete uniform distribution over the set $\{1, 2, \cdots, k_{max}\}$. This is an informative prior distribution for K. So, $p(K = k) = 1/k_{max}$. This provides equal support for K between 1 and k_{max}.

Note that the parameter space is the space of all parameters $w^k, \mu^k, \sigma_k^2, K = 1, 2, \cdots, k_{max}$, where k_{max} is given. From the histograms, we can see that there are probably 2 or 3 modes existing, so we choose k_{max}=3.

All parameters in the prior distributions are hyper-parameters. We also need to assign distributions or values to the hyper-parameters in the prior distributions of paramters. There

are five parameters (μ_0, σ_0^2, α, β and γ) to which we need to assign distributions or values to. In this work, we use a similar strategy as in Richardson and Green (1997) and Fu and Wang (2002). The μ_0 is the mean of the distribution of μ^k and σ_0^2 is the variance of μ^k. We extract this information from the data as follows. Let M_x and R_x denote the midrange and the range of the data respectively. Then, we set $\mu_0 = M_x$ and $\sigma_0^2 = R_x^2$. For the inverse-gamma prior, we choose $\alpha = 2$ and $\beta = 1$ to prevent the value of σ_k^2 to be close to zero (Escobar and West, 1995). In the Dirichlet distribution, we set γ=1. This corresponds to an uniform distribution over the range of the values of the weights.

Note that the choice of values for hyper-parameters will have an influence on the posterior distributions and estimations. Such an influence can be examined by a sensitivity analysis. Richardson and Green (1997) have carried out a sensitivity analysis for their Bayesian models and they found that the posterior estimates are not very sensitive to the values of hyper-parameters chosen according to this strategy.

4.2 Distribution of TDS

We present the computational results in this section. The algorithm is introduced in Chapter 3. All numerical computations in this work are carried out using MATLAB in an Unix environment.

Initial values of hyper-parameters and other parameters are specified as follows.

- Hyper-parameters for TDS:

$$\mu_0 = 5.0781, \sigma_0^2 = 5.4951$$

$$\alpha = 2, \beta = 1, \gamma = 1, k_{max} = 3$$

- The compact intervals for the parameters:

$$k \in (1:3),\ \mu^k \in [3,7]^k,\ \sigma_k^2 \in [0.03, 0.3] \text{ and } w^k \in [0,1]^k.$$

- The prior distributions for μ^k, σ_k^2 and w^k, given K:

$$p\left(\mu^k|k\right) = k! \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi \times 5.4951}} \exp\left[-\frac{(\mu_{qj} - 5.0781)^2}{2 \times 5.4951}\right], \mu_{k1} < \mu_{k2} < \cdots < \mu_{kk} \quad (4.6)$$

$$p(\sigma_k^2|k) = (\sigma_k^2)^{-3} e^{-1/\sigma_k^2}, \sigma_k^2 > 0. \quad (4.7)$$

$$p(w^k|k) = (k-1)!, \text{ for } 0 \leq w_{k1} + \cdots + w_{kk-1} \leq 1, w_{kk} = 1 - w_{k1} - \cdots - w_{kk-1}. \quad (4.8)$$

This implies that we have no previous knowledge about the distributions of w_k's, and we let the data speak for themselves. This is an informative prior distribution as mentioned in Chapter 2. It is easy to see that the posterior distribution has a dimension d=13. The log-posterior distribution has the form

$$\sum_{i=1}^{N} \log f(x_i|\mu^k, \sigma_k^2, w^k, k) + \log p\left(\mu^k|k\right) + \log p\left(\sigma_k^2|k\right) + \log p\left(w^k|k\right) + \log p(k) \quad (4.9)$$

The simulation processes are based on the log-posterior distribution. Samples are drawn from the compact intervals. First, $n = 5 \times 10^6$ uniform base points are simulated from the compact intervals to form $S_n(p)$. Secondly, we group the points into contours and we assign each contour with ten points. We compute the discrete distribution on each contour according to the log-posterior probabilities. Third step, a sample of size $m = 3000$ is drawn using the discrete distribution. This sample is the desired sample.

The marginal posterior distribution of K is given in Table 4.1. The posterior for K clearly favors 2 modes. Because the prior provides equal support for K between 1 and

Table 4.1: Prior and posterior distributions of the number of components K, for TDS.

K	1	2	3
Prior	1/3	1/3	1/3
Posterior	0	0.6337	0.3663

Table 4.2: AMLE of σ_k^2, μ^k and w^k, $k = 1, 2, 3$, for TDS

	$k=1$	$k=2$		$k=3$		
σ_k^2	0.2537	0.0856		0.0736		
μ^k	5.1964	4.3854	5.4022	4.3419	5.3463	5.7898
w^k	1	0.2009	0.7991	0.1926	0.7192	0.0883

3, the likelihood function puts most of its weight on $k = 2$. As is typical with inference about overlapping mixtures, there is clearly a great deal of uncertainty about the number of components. But unlike traditional approaches to density estimation, the computations here provide a formal assessment of such uncertainty. We present in Table 4.2 the approximate maximum likelihood estimations of σ_k^2, μ^k and w^k's for $k = 1, 2, 3$, respectively.

This algorithm also gives us numerical output of means, standard deviations, minimums and maximums of the parameters. Because the posterior probability of $k = 1$ is 0, we do not report the estimates for the case of one component. The results are presented in Tables 4.3 and 4.4 with $k = 2$ and $k = 3$, respectively. All the marginal posterior distributions of the parameters are computed. Since the posterior favors 2 modes, then let us examine the marginal posterior distributions of the parameters for $k = 2$. Figure 4.1 shows the marginal

posterior distributions of the sub-population means and weights for $k = 2$. We can see that the distributions of μ_{21} and μ_{22} are roughly symmetric. The location of μ_{21} is near 4.4 and the range is from about 4.2 to 4.6. The location of μ_{22} is near 5.4 and range is from about 5.3 to 5.5. Thus, based on the logarithm of original data, the marginal posterior distribution of first sub-population mean most likely has a value of about 4.4 and standard deviation around 0.07. The marginal posterior distribution of second sub-populatin location most likely has a value of about 5.4 and standard deviation about 0.03. The marginal posterior distributions for the weights are almost symmetric. The w_{21} has a location around 0.2 and a range from about 0.1 to 0.3, and the w_{22} has a location around 0.8 and a range from about 0.7 to 0.9. Figure 4.3 shows the marginal posterior distribution for the population variance σ_2^2. This distribution has a small variance and a location around 0.1. Figure 4.4 shows us, given the existing observations, what values the future observations might take and what kinds of probabilities are associated with them. Two modes are presented, one around 4.4 and the other around 5.4.

Table 4.3: Posterior means, standard deviations, minimums and maximums of σ_k^2, μ^k and w^k, $k = 2$, for TDS.

	mean	st.d	minimum	maximum
σ_2^2	0.0990	0.0120	0.0641	0.1535
μ^2	4.3823	0.0674	4.1431	4.6019
	5.4006	0.0275	5.3021	5.4880
w^2	0.2015	0.0330	0.1049	0.3101
	0.7985	0.0323	0.6900	0.8951

Table 4.4: Posterior means, standard deviations, minimums and maximums of σ_k^2, μ^k and w^k, $k = 3$, for TDS.

	mean	std	minimum	maximum
σ_3^2	0.0969	0.0130	0.0624	0.1491
μ^3	4.3457	0.1215	3.1527	4.5910
	5.1617	0.3356	4.1599	5.4678
	5.5178	0.2154	5.3136	6.9312
w^3	0.1780	0.0553	0.0002	0.3271
	0.3754	0.2715	0.0004	0.8621
	0.4466	0.2929	0.0003	0.8781

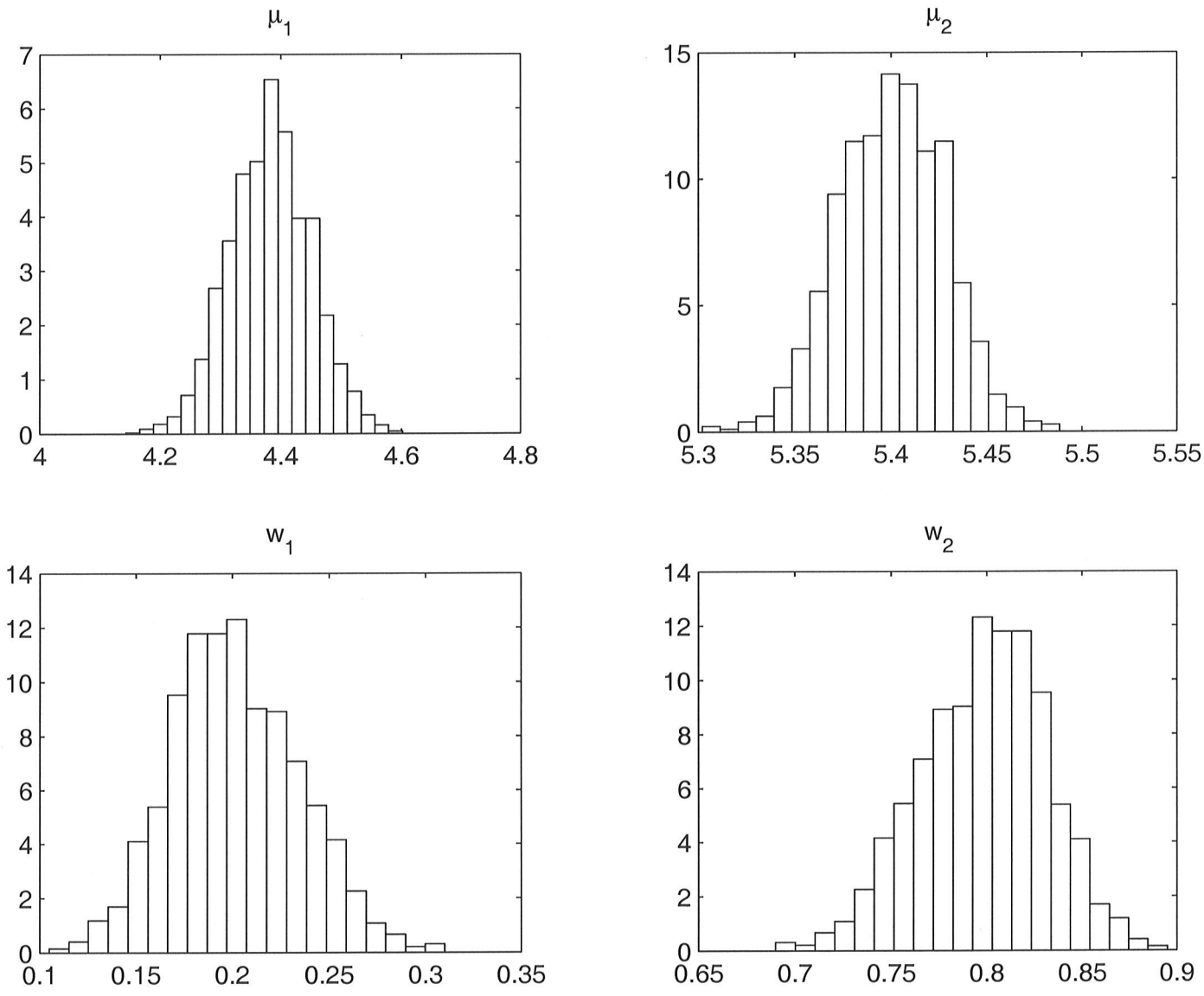

Figure 4.1: Marginal Posterior distributions of μ^k,w^k, $k = 2$, for TDS.

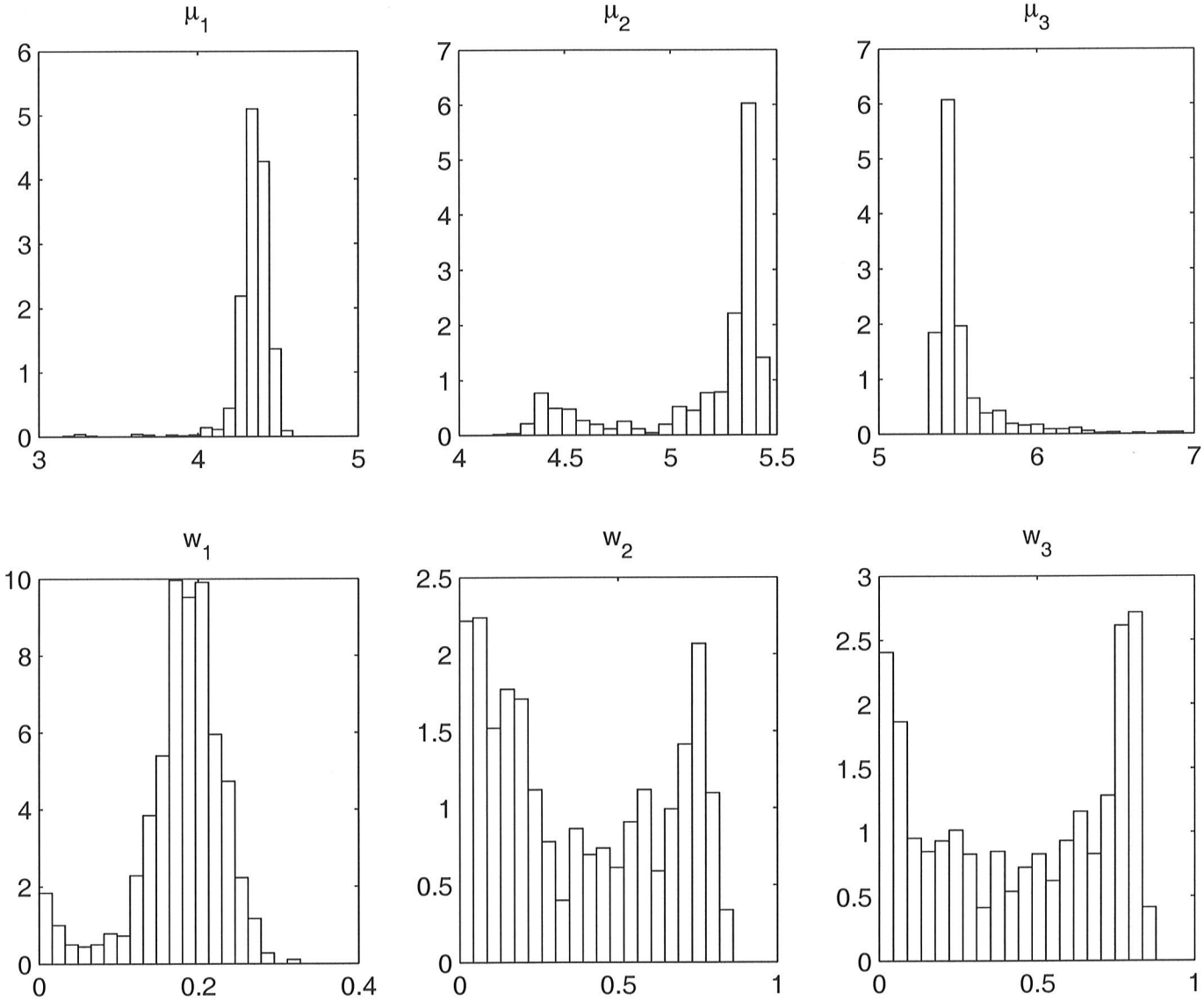

Figure 4.2: Marginal Posterior distributions of μ^k,w^k, $k = 3$, for TDS.

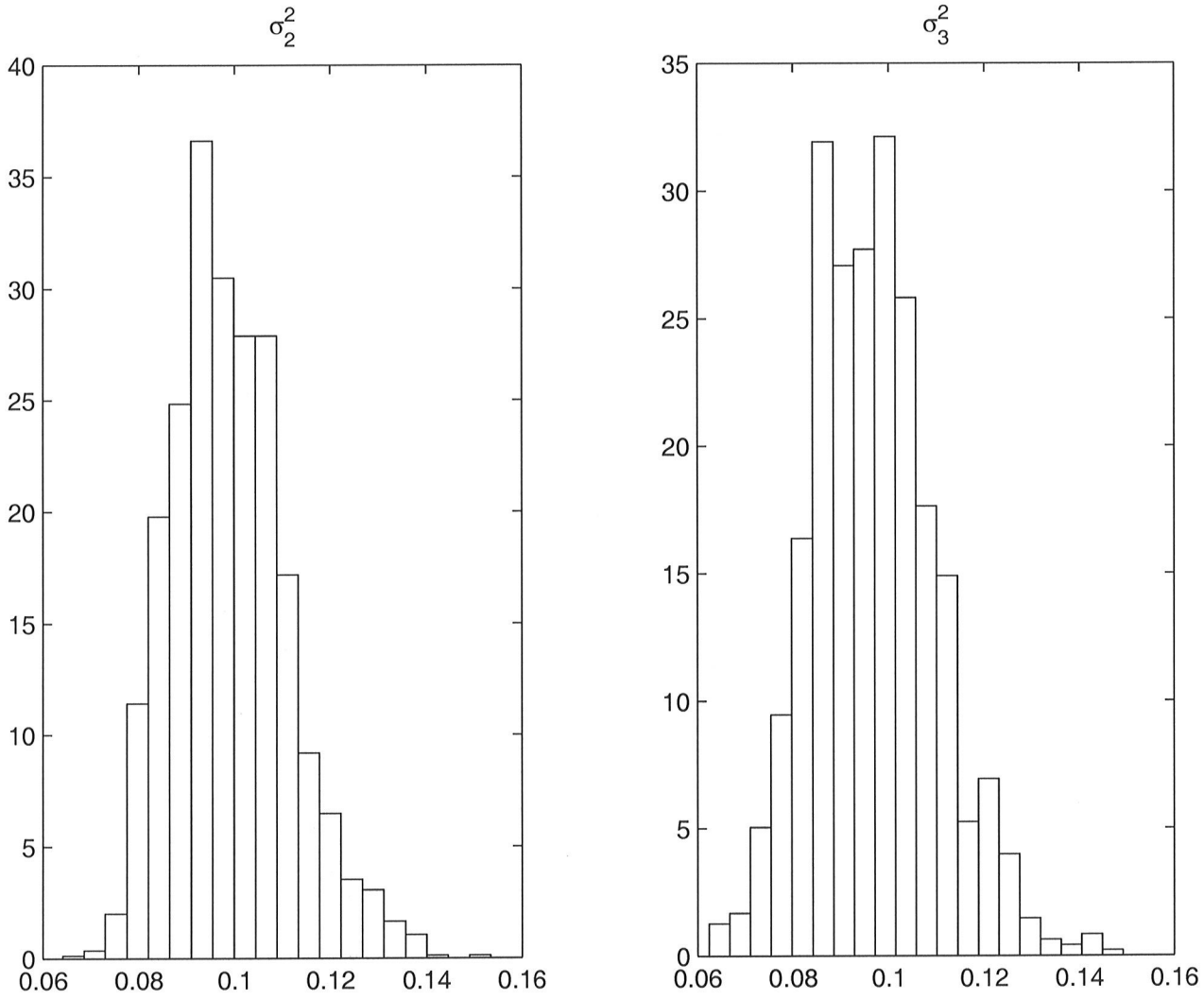

Figure 4.3: Marginal Posterior distributions of σ_k^2, $k = 2$, 3, for TDS.

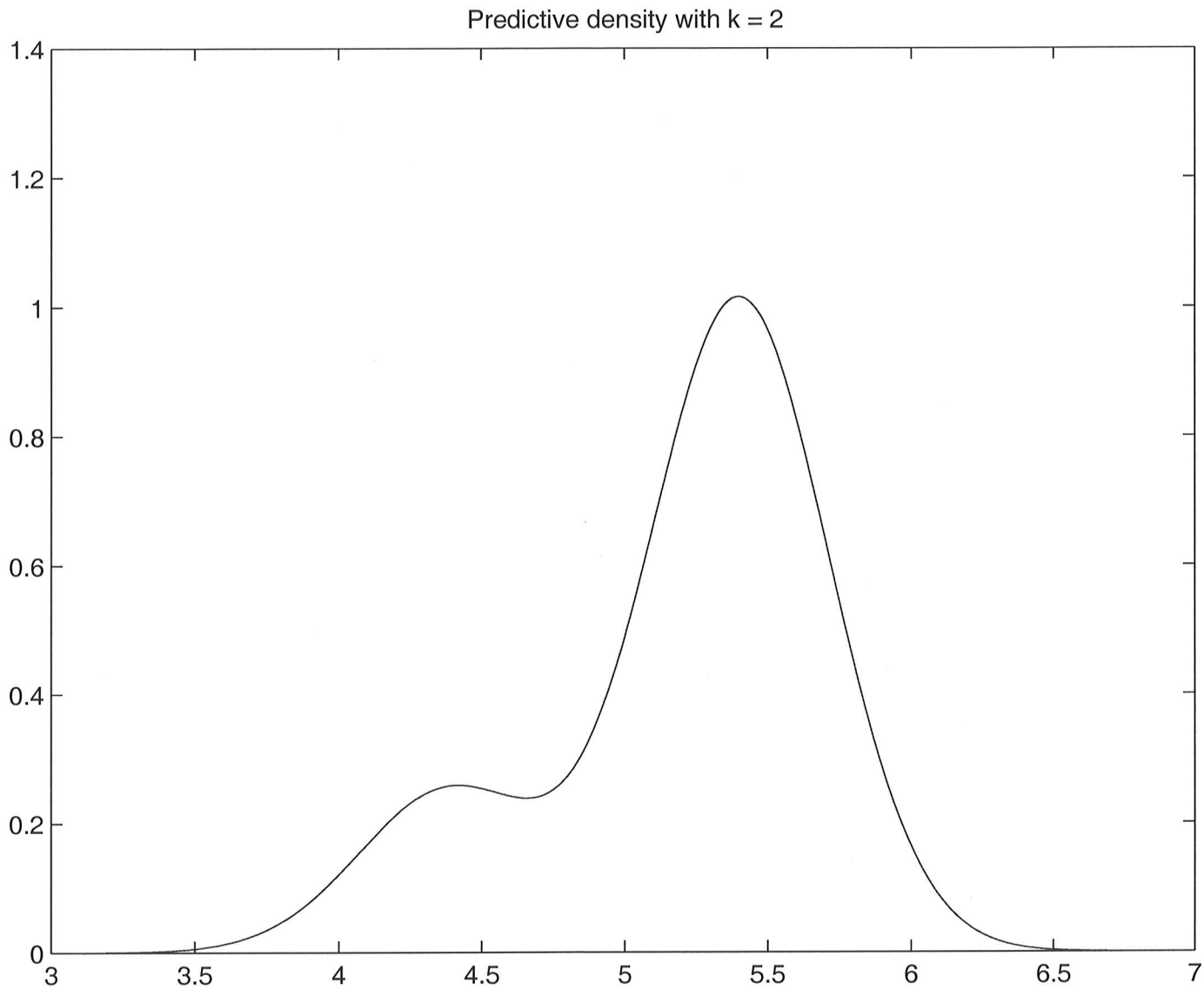

Figure 4.4: Predictive density, for TDS.

4.3 Distribution of HCO_3

Initial values of hyper-parameters and other parameters are specified as follows.

- The hyper-parameters:

$$\mu_0 = 4.6381, \sigma_0^2 = 5.6468$$

$$\alpha = 2, \beta = 1, \gamma = 1, k_{max} = 3.$$

- The compact intervals for the parameters:

$$k \in (1:3),\ \mu^k \in [3,6]^k,\ \sigma_k^2 \in [0.03, 0.3] \text{ and } w^k \in [0,1]^k.$$

- The prior distributions for μ^k, σ_k^2 and w^k, given K are:

$$p\left(\mu^k|k\right) = k! \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi \times 5.6468}} \exp\left[-\frac{(\mu_{kj} - 4.6381)^2}{2 \times 5.6468}\right], \mu_{k1} < \mu_{k2} < \cdots < \mu_{kk} \tag{4.10}$$

$$p(\sigma_k^2|k) = (\sigma_k^2)^{-3} e^{-1/\sigma_k^2}, \sigma_k^2 > 0. \tag{4.11}$$

$$p(w^k|k) = (k-1)!, \text{ for } 0 \leq w_{k1} + \cdots + w_{kk-1} \leq 1, w_{kk} = 1 - w_{k1} - \cdots - w_{kk-1}. \tag{4.12}$$

Except for some hyper-parameters, the prior distributions are of the same forms as those of TDS. The posterior distribution also has a dimension d=13. The log-posterior distribution has the form

$$\sum_{i=1}^{N} \log f(x_i|\mu^k, \sigma_k^2, w^k, k) + \log p\left(\mu^k|k\right) + \log p\left(\sigma_k^2|k\right) + \log p\left(w^k|k\right) + \log p(k) \tag{4.13}$$

We simulate the parameters using the log-posterior distribution. First, $n = 5 \times 10^6$ uniform base points are simulated from the compact intervals to form $S_n(p)$. Then we group the

Table 4.5: Prior and posterior distributions of the number of components K, for HCO_3.

k	1	2	3
Prior	1/3	1/3	1/3
Posterior	0	0.6263	0.3737

Table 4.6: AMLE of σ_k^2, μ^k and w^k for $k = 1, 2, 3$, for HCO_3.

	$k=1$	$k=2$		$k=3$		
σ_k^2	0.2560	0.0715		0.0667		
μ^k	4.7476	3.8697	4.9693	3.8857	4.9456	5.1335
w^k	1	0.2032	0.7968	0.2014	0.7360	0.0625

points into contours and each contour has ten points. We compute the discrete distribution on contours. Finally a sample of size $m = 3000$ is drawn. This sample is the desired sample. The marginal posterior distribution of K is given in Table 4.5. The prior for K equally support K between 1 to 3. The posterior strongly supports 2 modes. So, we assume that two sub-populations exist for the underlying distribution of this variable. Table 4.6 presents the approximate maximum likelihood estimates of σ_k^2, μ^k and w^k, for k=1, 2, 3, respectively.

This algorithm also gives us numerical output for the means, standard deviations, minimums and maximums of the parameters. The posterior probability for $k = 1$ is 0, we do not report the estimates for the case of one component. The results are presented in Tables 4.7, 4.8 with $k = 2$ and $k = 3$, respectively. All the marginal posterior distributions of the parameters are computed. Since the posterior distribution of K indicates that there are two

Table 4.7: Posterior means, standard deviations, minimums and maximums of σ_k^2, μ^k and w^k, $k = 2$, for HCO_3.

	mean	st.d.	minimum	maximum
σ_2^2	0.0828	0.0097	0.0573	0.1192
μ^2	3.8773	0.0578	3.6944	4.0854
	4.9594	0.0254	4.8680	5.0473
w^2	0.1976	0.0307	0.0964	0.2966
	0.8024	0.0307	0.7034	0.9036

modes, let us examine the marginal posterior distributions of parameters for $k = 2$. Figure 4.5 shows the marginal posterior distributions of the sub-population means. The distributions are near symmetric. The first sub-population mean has a location near 3.9 and the range is from about 3.7 to 4.1. The second sub-population mean has a location near 4.95 and range is from about 4.85 to 5.05. The marginal posterior distributions for the weights are almost symmetric. The w_{21} has a location around 0.2. The w_{22} has a location around 0.8. Figure 4.7 shows that the marginal posterior distribution for the variance of components has a location about 0.11. Figure 4.8 show us the predictive density of future observations given observations. Two modes are presented, one around 3.9 and the other around 5.

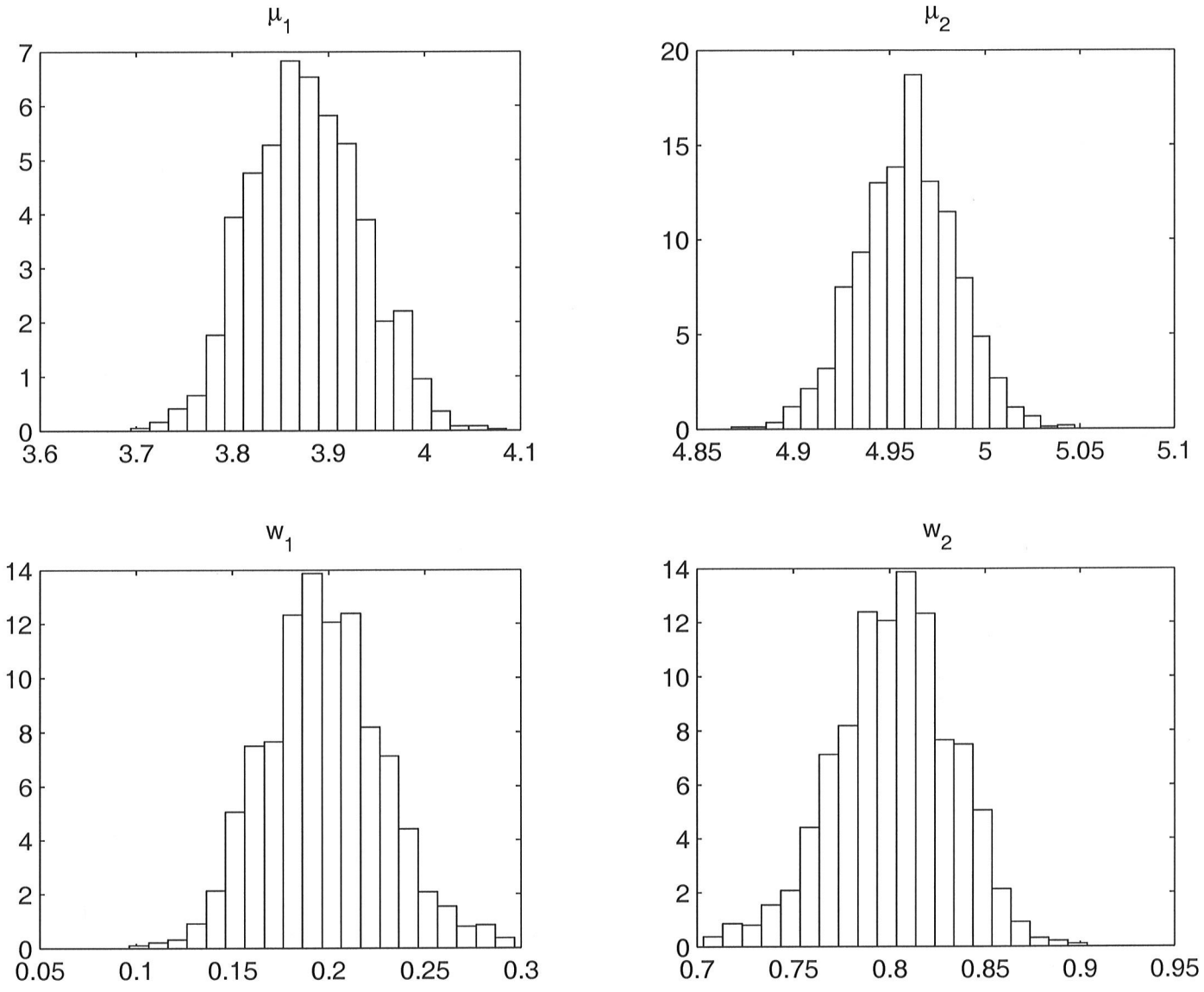

Figure 4.5: Marginal Posterior distributions of μ^k,w^k, $k = 2$, for HCO_3.

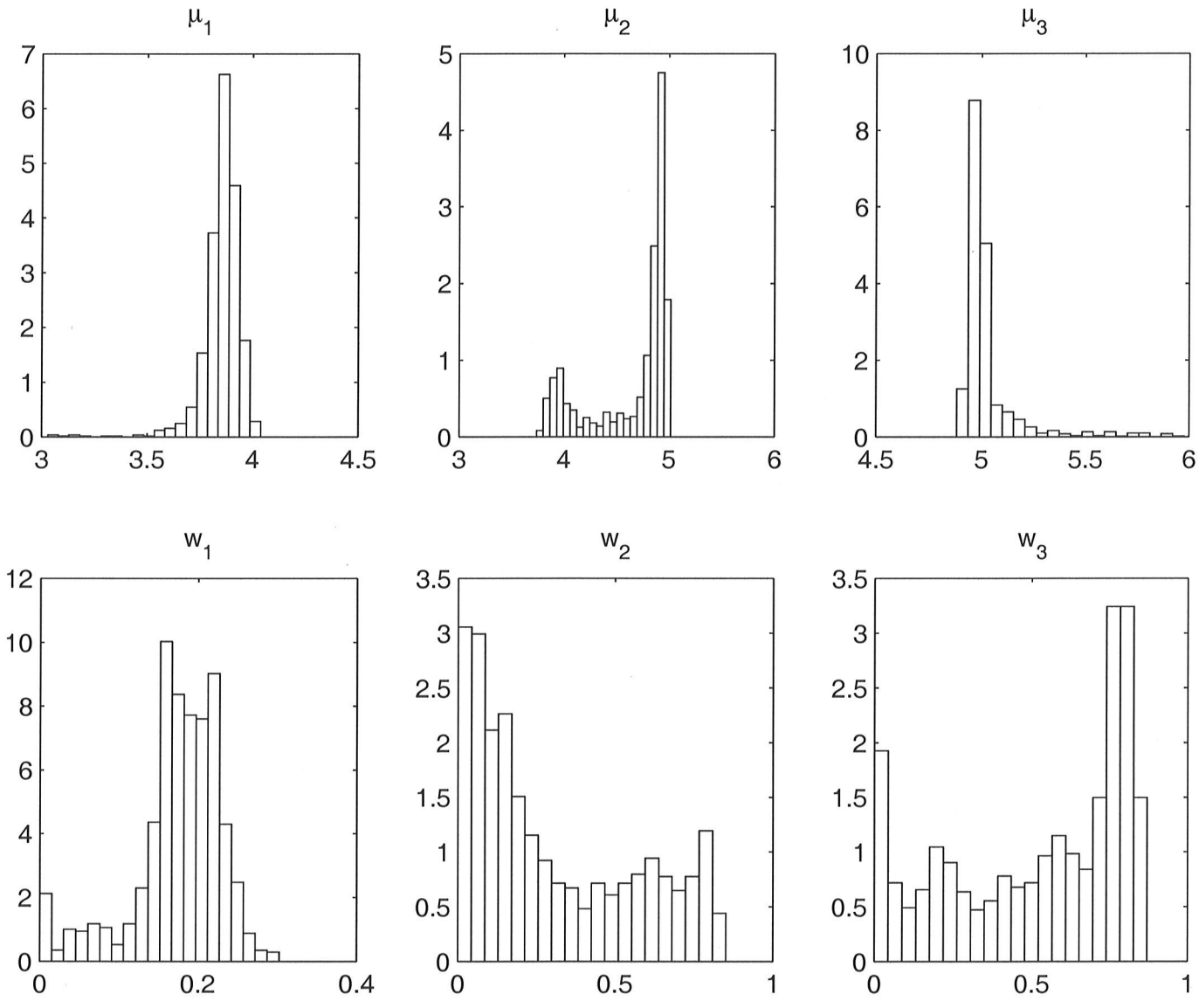

Figure 4.6: Marginal Posterior distributions of μ^k,w^k, $k = 3$, for HCO_3.

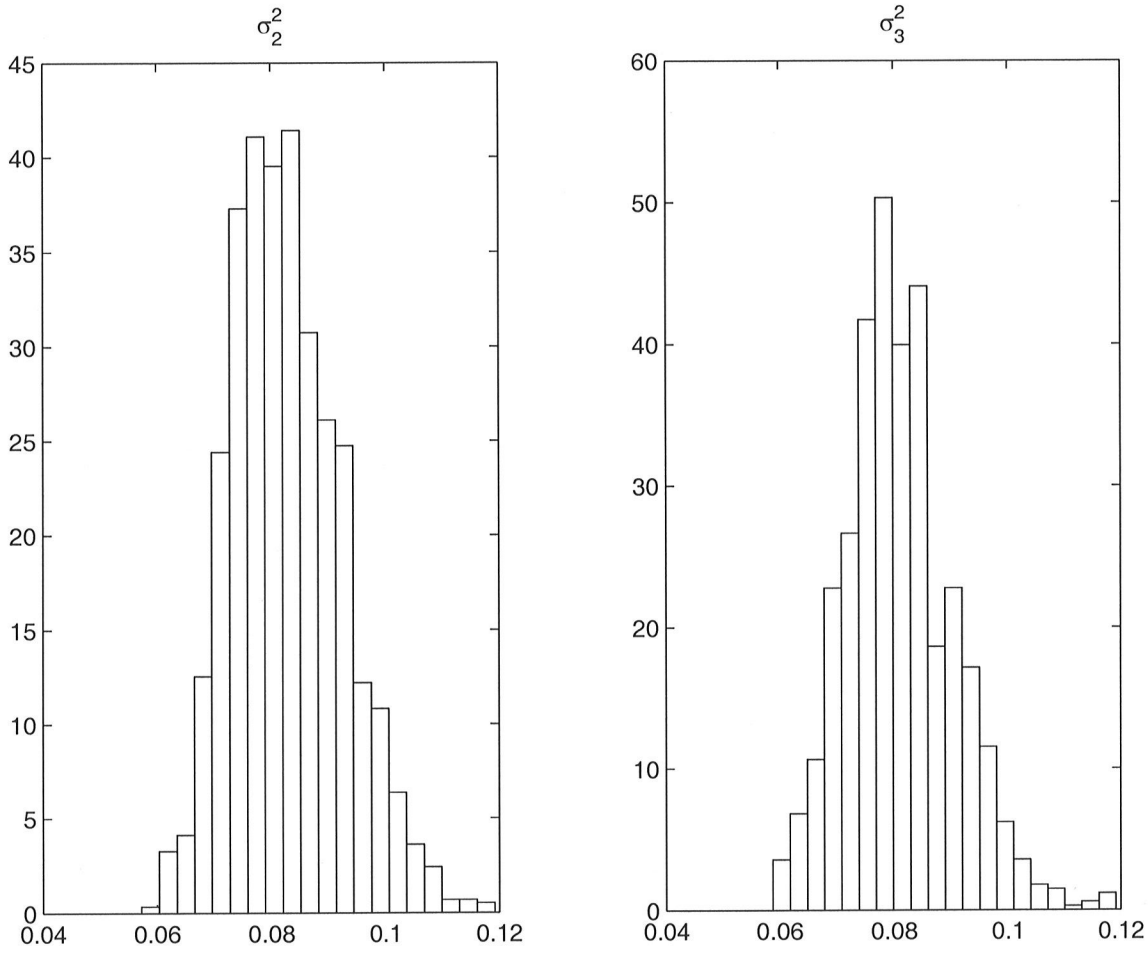

Figure 4.7: Marginal Posterior distributions of σ_k^2, $k = 2, 3$, for HCO_3.

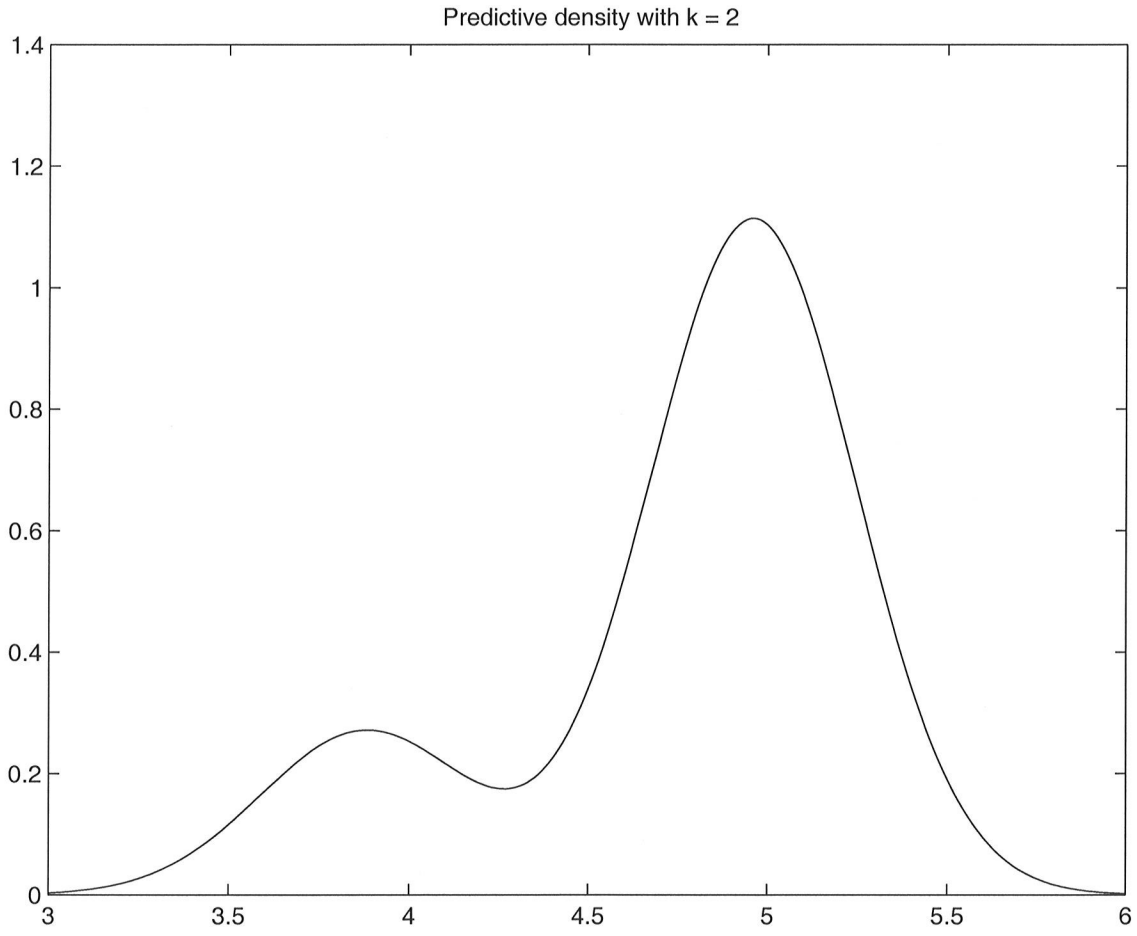

Figure 4.8: Predictive density, for HCO_3.

Table 4.8: Posterior means, standard deviations, minimums and maximums of σ_k^2, μ^k and w^k, $k = 3$, for HCO_3.

	mean	std	minimum	maximum
σ_3^2	0.0811	0.0095	0.0588	0.1191
μ^3	3.8527	0.0953	3.0308	4.0409
	4.6545	0.3870	3.7419	5.0181
	5.0273	0.1467	4.8824	5.9777
w^3	0.1740	0.0568	0.0005	0.3031
	0.3059	0.2587	0.0002	0.8519
	0.5201	0.2778	0.0002	0.8698

4.4 Distribution of Ca

Initial values of hyper-parameters and other parameters are specified as follows.

- The hyper-parameters:

$$\mu_0 = 3.0673, \sigma_0^2 = 9.2382$$

$$\alpha = 2, \beta = 1, \gamma = 1, k_{max} = 3$$

- The compact intervals for the parameters:

$$k \in (1:3),\ \mu^k \in [1,5]^k,\ \sigma_k^2 \in [0.04, 0.4] \text{ and } w^k \in [0,1]^k.$$

Table 4.9: Prior and posterior distributions of the number of components K, for Ca.

k	1	2	3
Prior	1/3	1/3	1/3
Posterior	0	0.6790	0.3210

- The prior distributions for μ^k, σ_k^2 and w^k, given K are

$$p\left(\mu^k|k\right) = k! \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi \times 5.6468}} \exp\left[-\frac{(\mu_{kj} - 3.0673)^2}{2 \times 9.2382}\right], \mu_{k1} < \mu_{k2} < \cdots < \mu_{kk} \tag{4.14}$$

$$p(\sigma_k^2|k) = (\sigma_k^2)^{-3} e^{-1/\sigma_k^2}, \sigma_k^2 > 0. \tag{4.15}$$

$$p(w^k|k) = (k-1)!, \text{ for } 0 \leq w_{k1} + \cdots + w_{kk-1} \leq 1, w_{kk} = 1 - w_{k1} - \cdots - w_{kk-1}. \tag{4.16}$$

Except some hyper-parameters values, the prior distributions are of the same forms as those of TDS and HCO_3. The posterior distribution also has a dimension d=13. The log-posterior distribution has the form

$$\sum_{i=1}^{N} \log f(x_i|\mu^k, \sigma_k^2, w^k, k) + \log p\left(\mu^k|k\right) + \log p\left(\sigma_k^2|k\right) + \log p\left(w^k|k\right) + \log p(k) \tag{4.17}$$

We simulate the parameters using the log-posterior distribution. Samples are drawn from the compact intervals. First, $n = 5 \times 10^6$ uniform base points are simulated from the compact intervals to form $S_n(p)$. Second, we group the points into contours according to the log-posterior probabilties. We assign ten points to each contour and compute the discrete distribution on contours. On third step, according to the discrete distribution, a sample of size $m = 3000$ is drawn. This is the desired sample.

Table 4.10: AMLE of σ_k^2, μ^k and w^k for $k = 1, 2, 3$, for Ca

	$k=1$	$k=2$		$k=3$		
σ_k^2	0.3875	0.1006		0.0831		
μ^k	3.3206	2.2227	3.5804	2.0492	2.5783	3.5950
w^k	1	0.1834	0.8166	0.1494	0.0840	0.7666

The marginal posterior distribution of K is given in Table 4.9. Two modes are strongly supported by the posterior of K. The results are presented in Table 4.10 for the approximate maximum likelihood estimates of σ_k^2, μ^k and w^k for k=1, 2, 3, respectively.

Numerical output of means, standard deviations, minimums and maximums of the parameters are also given. The posterior probability for $k = 1$ is 0, we do not report the case of one component. The estimates of the parameters are presented in Tables 4.11, 4.12 with $k = 2$ and $k = 3$, respectively. The marginal posterior distribution of the number of components favors two modes, so let us take a look at the marginal posterior distribution of sub-population means on Figure 4.9. The distributions are near symmetric. The marginal posterior distribution of μ_{21} has a location near 2.2 and the range is from about 2 to 2.4. The marginal posterior distribution of μ_{22} has a location near 3.6 and range is from about 3.5 to 3.7. The marginal posterior distributions for the weights are roughly symmetric. The w_{21} has a location around 0.2 and w_{22} has a location around 0.8. The marginal posterior distribution of variance has a location 0.1 from Figure 4.11. Figure 4.12 shows us the predictive density of future observations given observations. Two modes are presented, one around 2.2 and the other around 3.6.

Table 4.11: Posterior means, standard deviations, minimums and maximums of σ_k^2, μ^k and w^k, $k = 2$, for Ca.

	mean	st.d.	minimum	maximum
σ_2^2	0.1116	0.0127	0.0758	0.1712
μ^2	2.2104	0.0661	1.9778	2.4361
	3.5787	0.0291	3.4593	3.6796
w^2	0.1923	0.0308	0.0989	0.3010
	0.8077	0.0308	0.6990	0.9011

Table 4.12: Posterior means, standard deviations, minimums and maximums of σ_k^2, μ^k and w^k, $k = 3$, for Ca.

	mean	std	minimum	maximum
σ_3^2	0.1079	0.0142	0.0651	0.1611
μ^3	2.1477	0.1438	1.0650	2.3882
	3.0570	0.5300	2.0997	3.6520
	3.6862	0.2442	3.4887	4.9773
w^3	0.1569	0.0616	0.0000	0.3223
	0.2864	0.2754	0.0022	0.8566
	0.5567	0.2982	0.0005	0.8760

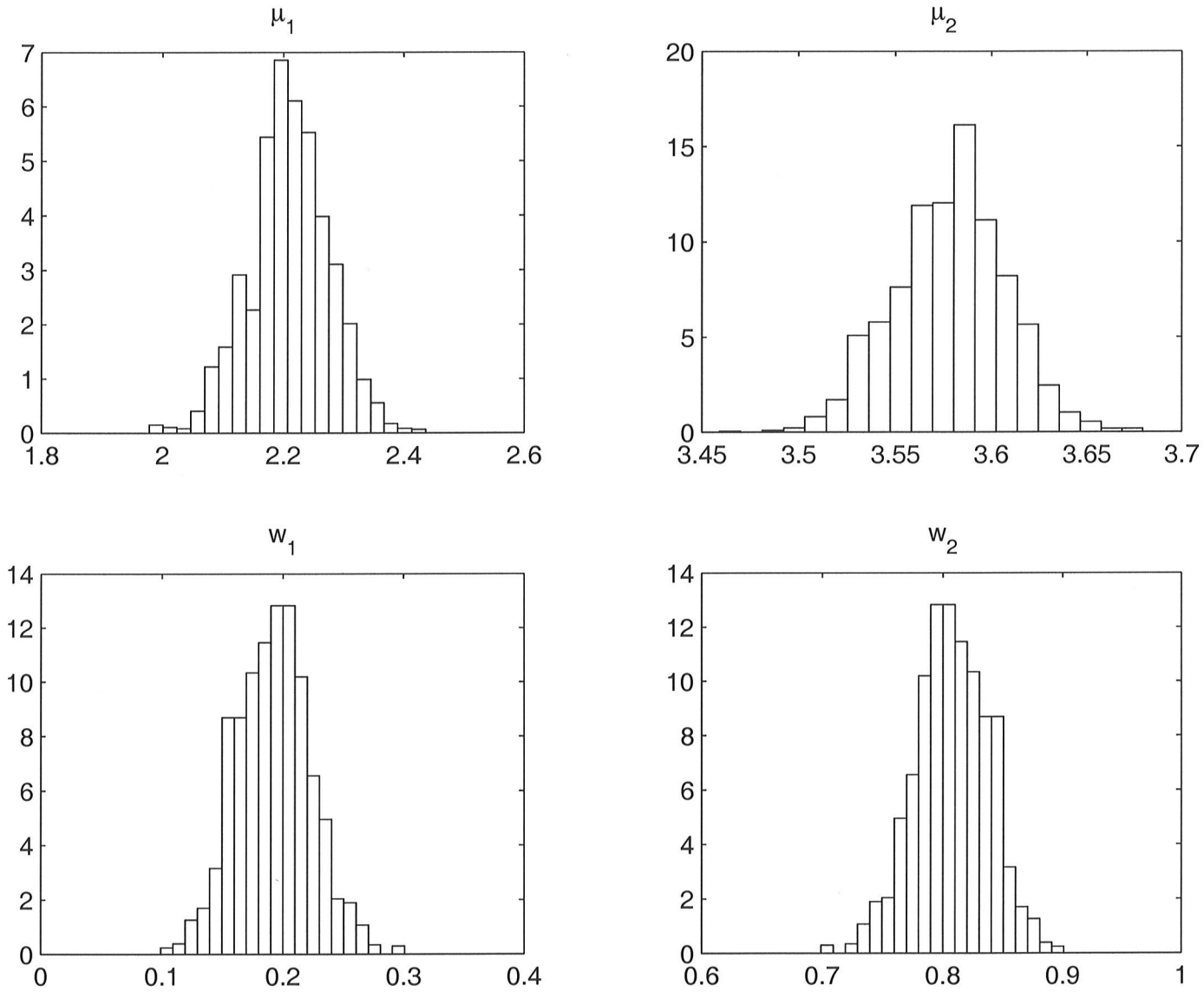

Figure 4.9: Marginal Posterior distributions of μ^k,w^k, $k = 2$, for Ca.

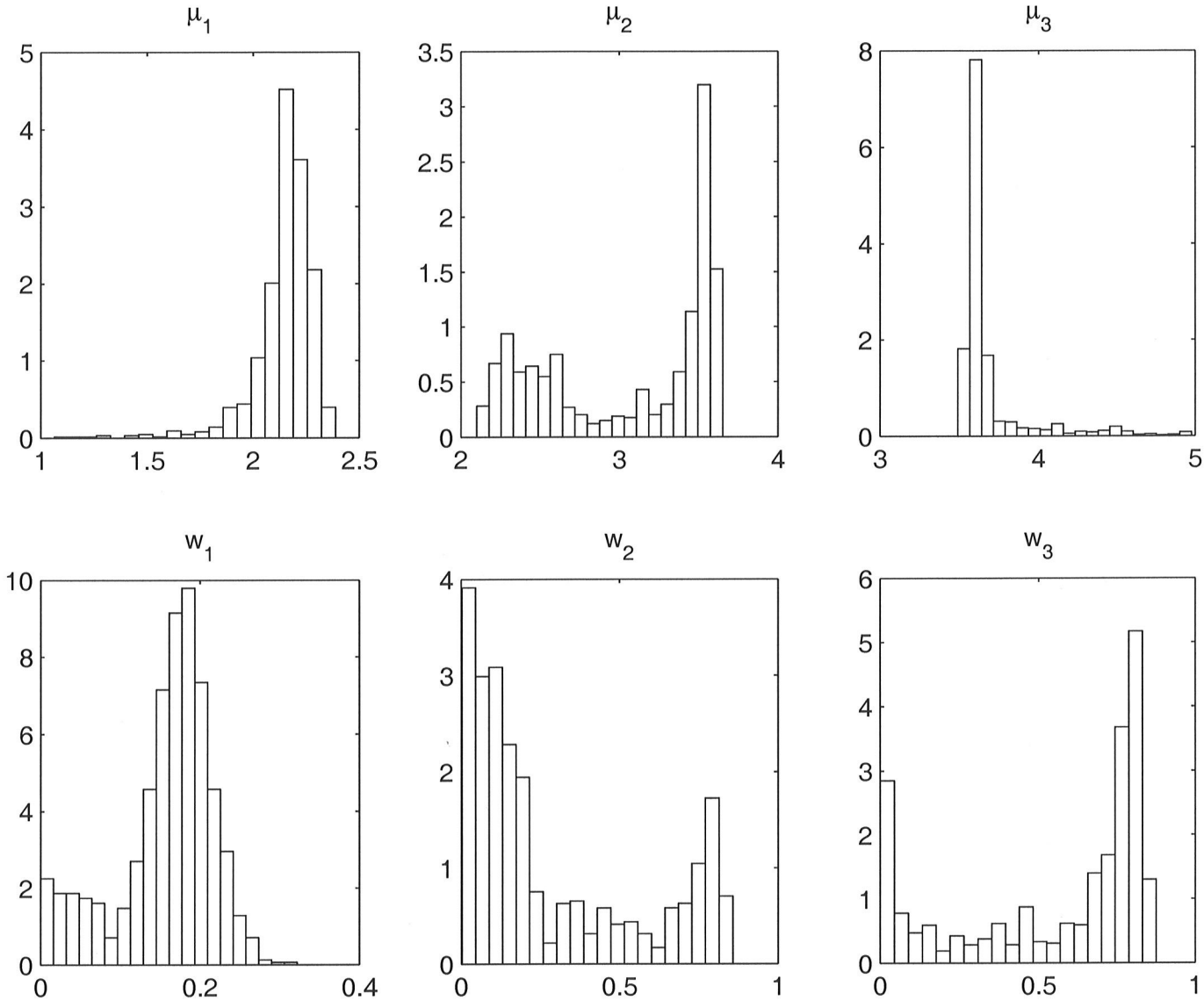

Figure 4.10: Marginal Posterior distributions of μ^k,w^k, $k = 3$, for Ca.

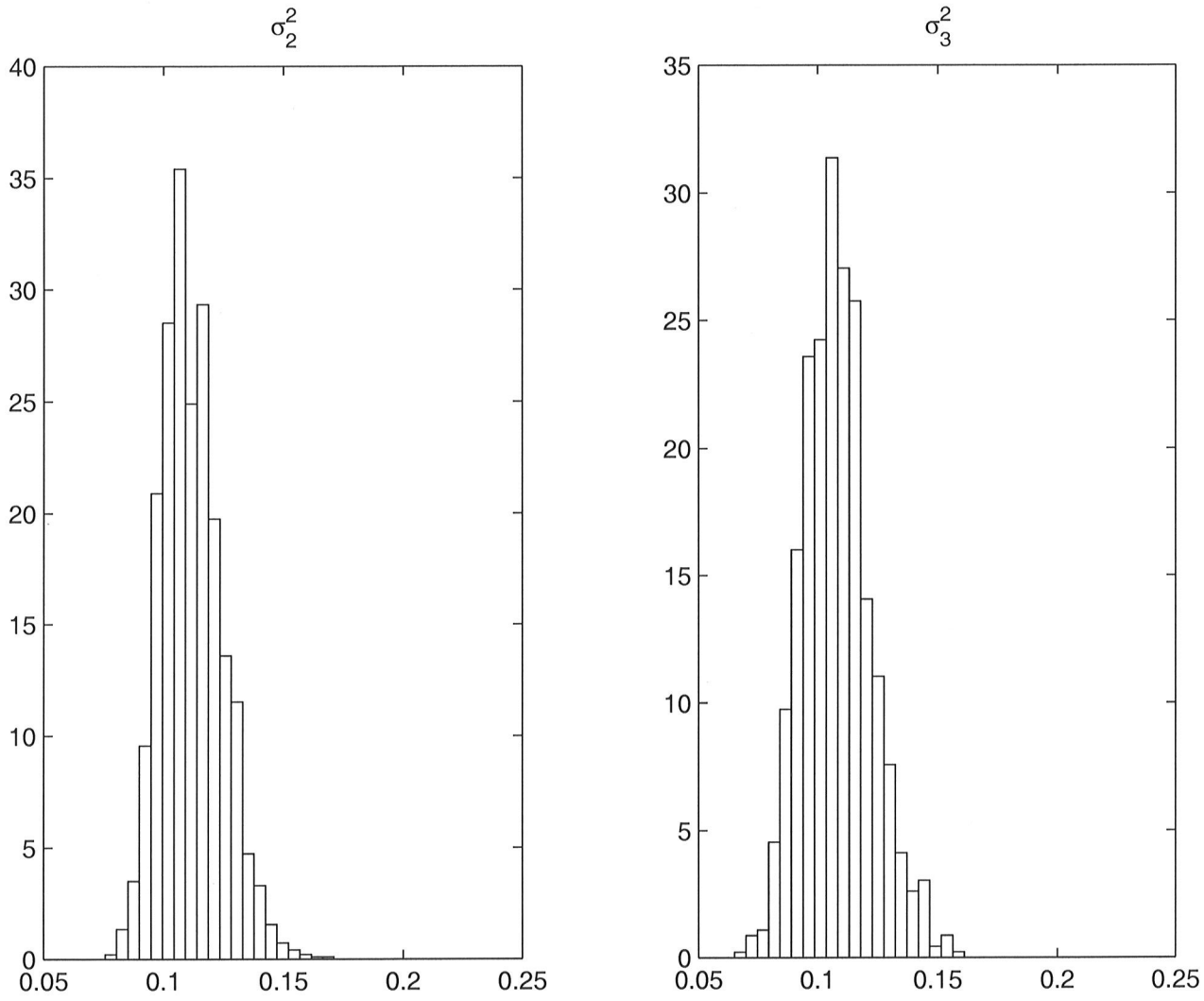

Figure 4.11: Marginal Posterior distributions of σ_k^2, $k = 2, 3$, for Ca.

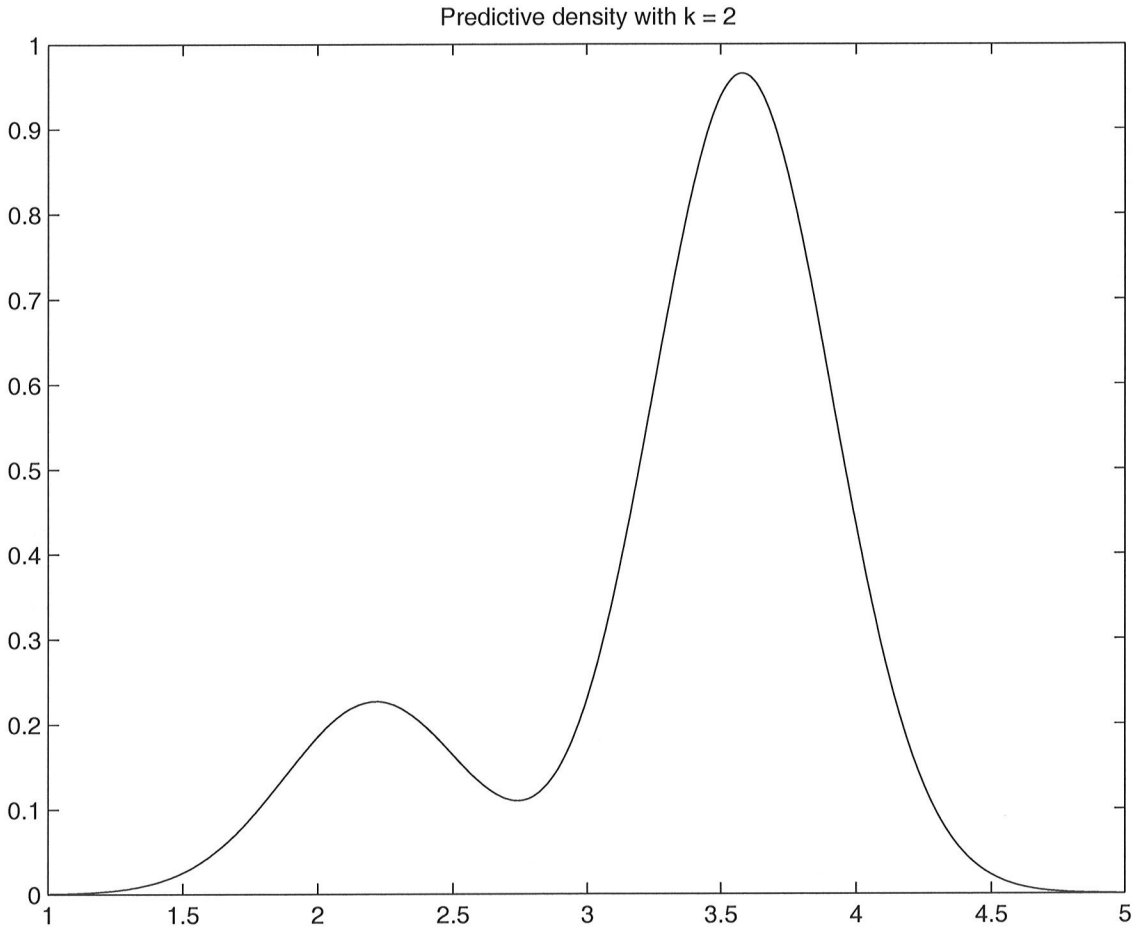

Figure 4.12: Predictive density, for Ca.

4.5 Classification of the Changjiang Basin

From the above computational results, the posterior distributions of K favor two modes for the three variables. Now we would like to know how these distributions are associated with geological locations.

According to Everitt et al. (2001), having estimated the parameters of the assumed mixture distribution, observations can be associated with particular groups on the basis of the maximum value of the following estimated probability:

$$p(x_i \in Class(j)|x) = \frac{f_j(x_i)}{\sum_{j=1}^{k} f_j(x_i)}, i = 1, 2, \cdots, 191; j = 1, 2; k = 2. \tag{4.18}$$

where

$$f_j(x_i) = \frac{w_{2j}}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{(x_i - \mu_{kj})^2}{2\sigma_k^2}\right] i = 1, 2, \cdots, 191; j = 1, 2; k = 2. \tag{4.19}$$

We call it classifying probability. For each station, we compute the classifying probability and assign it to class j, j=1 or 2, whichever has the higher probability value. The results are shown in Figures 4.13, 4.14 and 4.15.

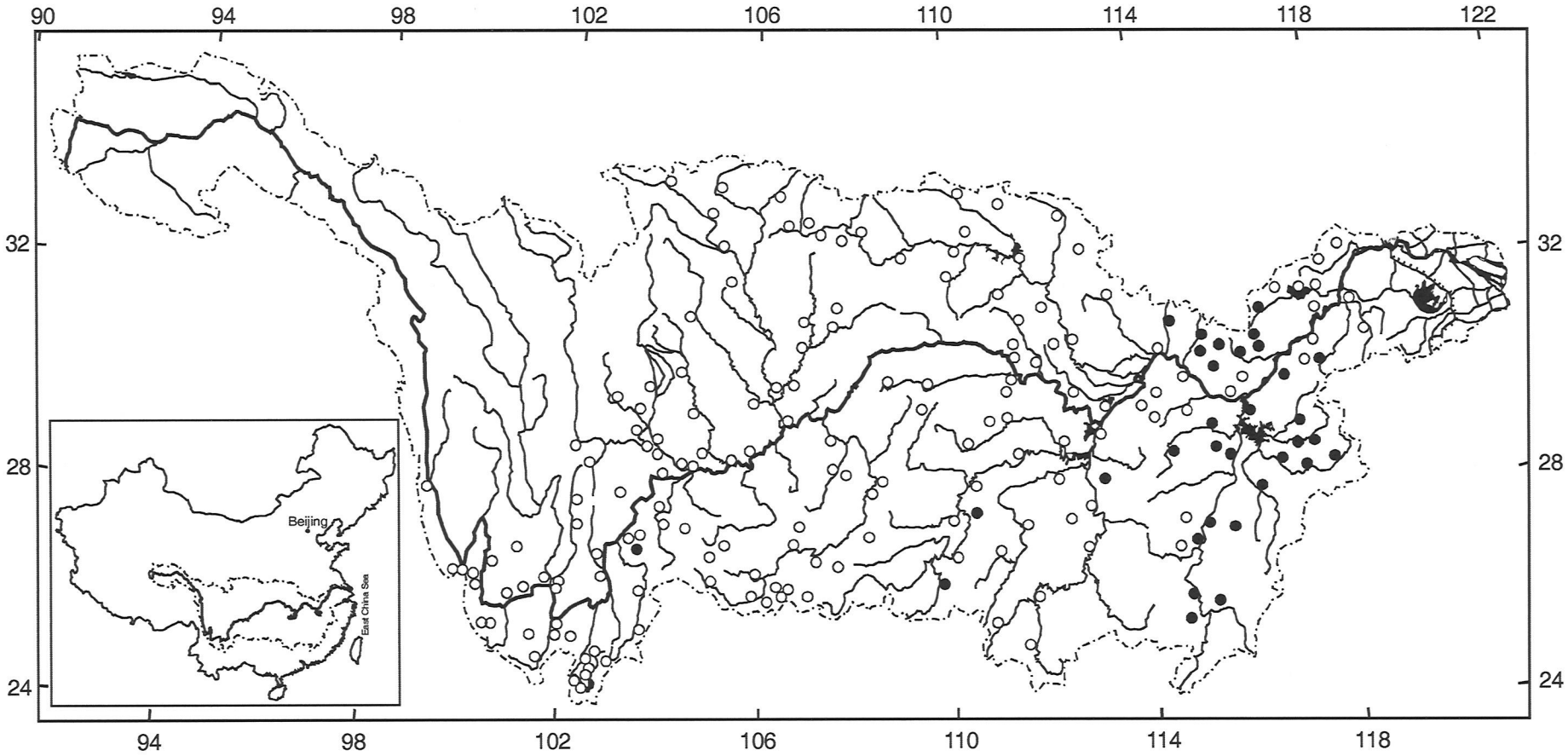

Figure 4.13 Classification of TDS. Black dots represent the group one and circles are the stations of group two.

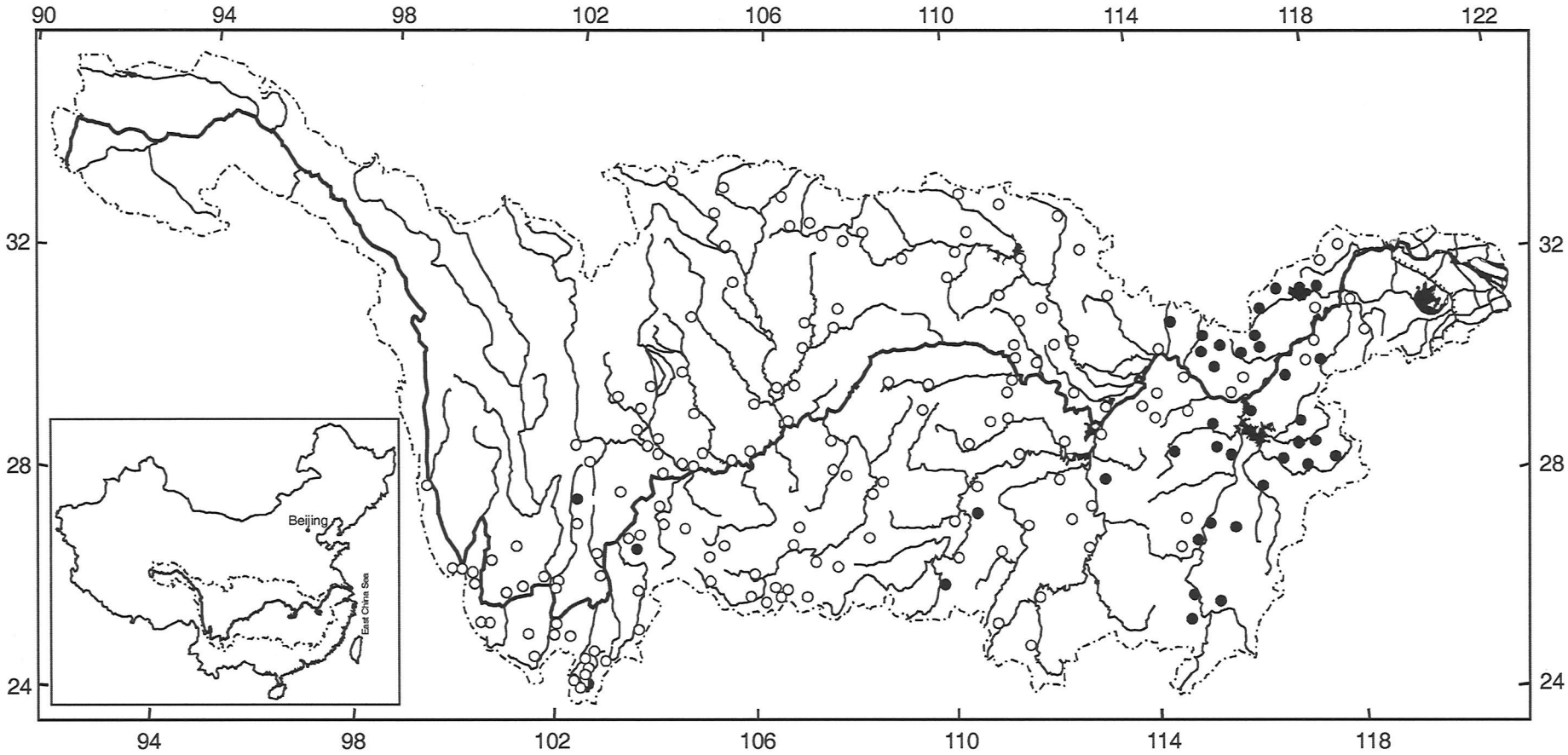

Figure 4.14 Classification of Bicarbonate (HCO_3). Black dots represent the group one and circles are the stations of group two.

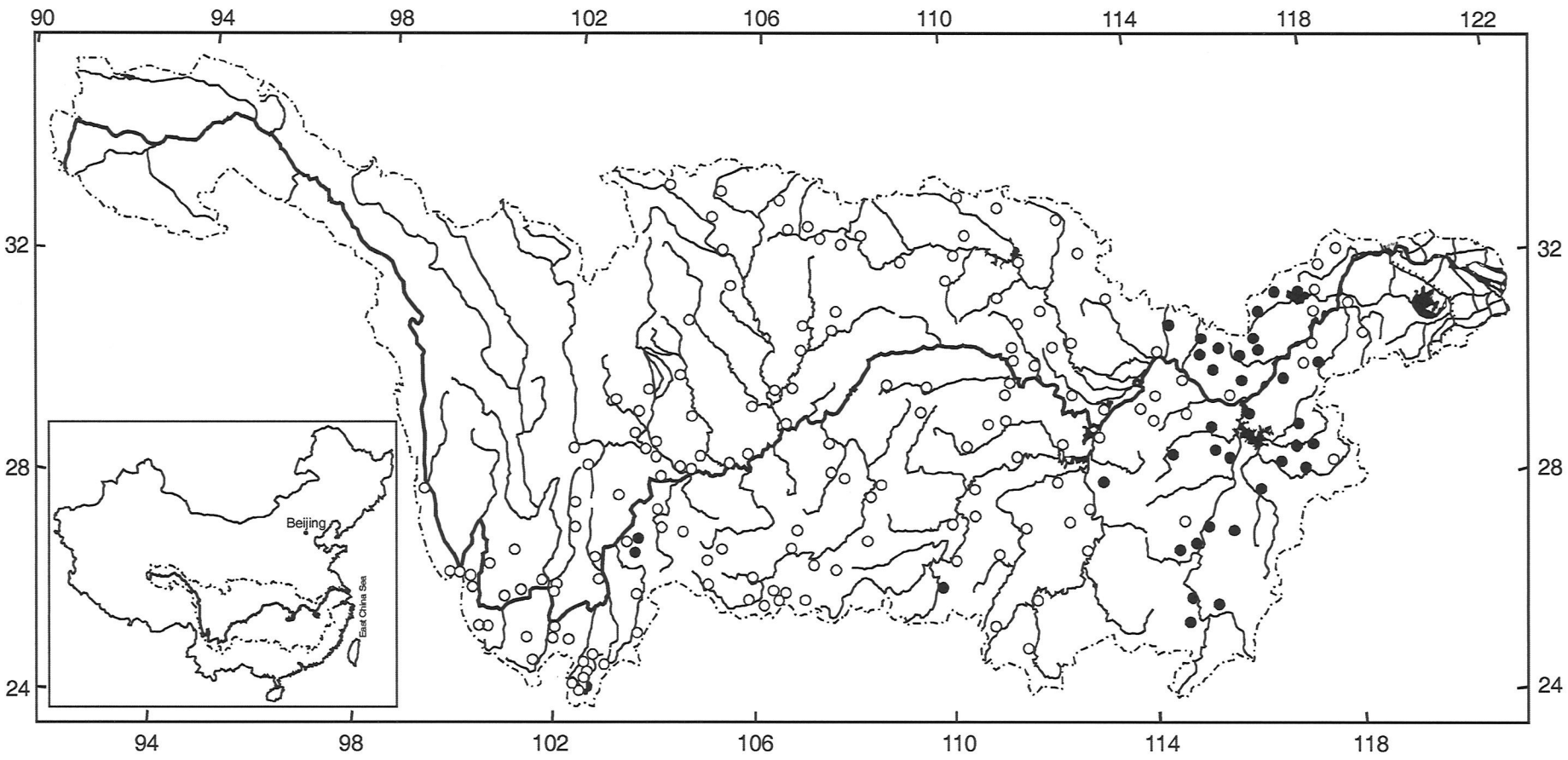

Figure 4.15 Classification of Calcium (Ca). Black dots represent the group one and circles are the stations of group two.

Chapter 5

Conclusions

We have studied the distribution of each variable defined earlier, using a Bayesian finite mixture model and a new sampling algorithm. The key contributions here lie in the understanding of the distribution of each chemical element studied in the Changjing basin and in the identifying of two sub-populations existing for the underlying distributions. The marginal posterior distributions of parameters in the mixture model provide us an insight into the underlying distributions. In addition to demonstrating the sub-populations, we classified the sampling stations into two groups, showing how the geological locations are associated with the distributions.

This work represents the first systematical study of the distributions of chemical elements in the Changjiang basin. The proposed algorithm provides an attractive approach to these types of problems. It gives us with much more flexibility for making inference for a mixture model with an unknown number of components. This is a big challenge in the computation of Bayesian finite mixture models. The algorithm allows us to overcome the computational inapplicability of formal Bayesian estimators, while maintaining the strength of the Bayesian

Table 5.1: Estimated sub-populatin means of TDS, HCO_3 and Ca

	TDS	HCO_3	Ca
$Gorup1$	79.84	48.42	9.12
$Group2$	221.41	142.59	35.87

approach.

According to the classifying results, most of the group 1 stations are located in the lower reaches of the Changjiang river and the stations of group 2 are located in the upper and middle reaches of the river area. The major chemical elements of the Changjiang are mainly controlled by chemical weathering, atmospheric precipitaition, and other natural processes as well as human activities (Chen et al., 2002). In the lower reaches of the river basin, there is a higher level of annual average precipitation than in the middle and upper reaches of the river. Carbonate rocks, the weathering of which produces Ca and HCO_3 in the river water, are also less abundant in the lower reaches. We think that these are the main causes as to why most of the stations in the lower reaches have a lower level of the chemical elements studied.

Since we transformed the data sets using the logarithm technique at the beginning, now we need to transform them back to their original scale. The estimated sub-population means are given in Table 5.1.

Further study is needed on the prediction part of the problem. Although we have a predictive distribution, which shows us that, given the sample x, what kinds of values future observations might take, and what probabilities are associated with these values. Efforts

are needed to focus on how to predict the chemical elements' values for specific locations without observations, as well as the estimation errors.

Bibliography

[1] Besag, J., and Green, P.J. (1993). Spatial Statistics and Bayesian Computation. *J.R. Statist. Soc. B, 55, 25-37.*

[2] Belisle, P., Joseph, L., Wolfson, D.B., and Zhou, X. (2002). Bayesian Estimation of Cognitive Decline in Patients with Alzheimer's Disease. *The Canadian Journal of Statistics, 30, 1-176.*

[3] Chen, J., Wang, F., Xie, X., and Zhang, L. (2002). Major Element Chemistry of the Changjiang (Yangtze River). *Chemical Geology, 187, 231-255.*

[4] Diebolt, J., and Robert, C.P. (1994). Estimation of Finite Mixture Distributions through Bayesian Sampling. *J.R. Statist. Soc. B, 56, 363-375.*

[5] Escobar, M.D., and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association, 90, 577-588.*

[6] Everitt, B.S., Landau, S., and Leese, M. (2001). *Cluster Analysis.* Fourth edition. London: Arnold.

[7] Fu, J., and Wang, L. (2002). A Random-Discretization Based Monte Carlo Sampling Method and its Application. *Methodology and Computing in Applied Probability, 4, 5-25.*

[8] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

[9] Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1998). *Markov Chain Monte Carlo in Practice.* London: Chapman and Hall.

[10] Ibrahim, J.G., Chen, M.H., and Gray, R.J. (2002). Bayesian Models for Gene Expression with DNA Microarray Data. *Journal of American Statistical Association, 97, 88-99.*

[11] Leonard, T., and Hsu, J.S.J. (1999). *Bayesian Methods (An analysis for Statistician and Interdisciplinary researchers).* U.K. Cambridge.

[12] Richardson, S., and Green, P. (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components. *J.R. Statist. Soc. B, 59, 731-792.*

[13] Ross, S.M. (1997). *Simulation* Second edition. New York: Acadimic Press. New York: Springer.

[14] Smith, A.F.M. and Roberts, G.O. (1993). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo methods. *J.R. Statist. Soc. B, 55, 3-23.*

[15] Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distribution.* U.K. John Wiley and Sons Ltd.