Markov Chain Monte Carlo Applications in Bioinformatics and Astrophysics

A thesis presented

by

Hosung Kang

to

The Department of Statistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Statistics

Harvard University
Cambridge, Massachusetts

May, 2005

Thesis Advisors: Professors Jun S. Liu and David A. van Dyk Hosung Kang

Markov Chain Monte Carlo Applications in Bioinformatics and Astrophysics

Abstract

This thesis comprises three applications of Markov chain Monte Carlo methods in astrophysics and bioinformatics. The recent development of high resolution satellite telescopes and the technological advances in genotyping technologies for Single-Nucleotide Polymorphisms (SNPs) have given us a wealth of data in astrophysics and genetics and tremendous opportunities for statisticians to develop complex models and computation techniques. In recent years, thanks to the the powerful Markov chain Monte Carlo (MCMC) sampling method and improvement in computing speed, one can handle the complicated models with large data sets efficiently.

The main statistical framework for the applications in this thesis is data augmentation, which constructs iterative optimization or sampling algorithms via the introduction of unobserved data or latent variables. For deterministic algorithms, the EM (Expectation-Maximization) algorithm is generally used for maximizing a likelihood function or a posterior density. For stochastic algorithms, data augmentation and Gibbs sampling algorithms are popular for posterior sampling.

The first paper describes a new and powerful method for estimating the distribution of the temperature of matter in the outermost layer of the atmosphere of a star using data augmentation and Bayesian hierarchical modeling technique. This new method enables us to fit to either a selected subset of emission lines with measured fluxes or to perform a global fit to the full wavelength range of the instrument, to

obtain error bars to determine the significance of features seen in the estimation, and to directly incorporate prior information such as known atomic data errors, systematic effects due to calibration uncertainties, etc.

The second paper proposes a novel genotype clustering algorithm, based on a bivariate t-mixture model, which assigns a set of probabilities for each data point belonging to the candidate genotype clusters. Furthermore, the model allows us to use the probabilistic multi-locus genotype matrices as inputs for haplotype phasing. Combining the genotyping and phasing steps, we can perform haplotype inference directly on raw readouts from a genotyping machine such as the Taqman assay, with less error than other competing methods.

The third paper develops a Bayesian Linkage-Disequilibrium mapping model for complex diseases. Haplotype analysis of disease chromosomes allows us to localize disease mutations as well as to identify historical recombination events descending from founder haplotypes. The primary improvement of this model over previous ones is to discern the locations of two disease mutations as well as their interaction effects.

Contents

List of Figures

List of Tables

Acknowledgments

This work would not have been possible without the support and guidance of my advisors, Professor Jun Liu and Professor David van Dyk.

Professor Liu has provided me with excellent guidance and wisdom. I have been inspired by his intuition and understanding of a wide range of topics and motivated by his passion for statistics. I am very honored to have him as my advisor.

I can not thank enough Professor David van Dyk, who gave me an opportunity to work with him from my first year in the program and continuous support since then. I could not have had a better mentor for my research and various other aspects of my academic life at Harvard.

I would like to thank Professor Xiao-Li Meng for reading the draft of my thesis and providing his valuable comments. I am very grateful to Vinay Kashyap, Alanna Connors, Aneta Siemiginowska and other members of Astro Statistics Research Group and Tim Niu and Steve Qin for excellent collaboration experiences. I would like to thank Betsey Cogswell, Dale Rinkel, Vanessa Malcolm-Garcia and Maureen Stanton for their administrative support.

I would like to thank my colleagues past and present for making my graduate school life less miserable (or more livable): Claudia Pedroza, Shane Jensen, Sam Cook, Liz Stuart, Nondas Sourlas, Taeyoung Park, Jim Greiner, Charity Morgan, and my classmates, Gopi Goswami and Byron Ellis. I am so grateful to them for listening to my outbursts of anger and complaints at times. I would like to thank my friends Juyoung and Sangmin for their true friendship.

Most importantly, I would like to thank my family for their endless support and love. They taught me perseverance and modesty by example, which helped to make my Ph.D. life such a pleasant and rewarding experience.

A Bayesian Approach to the Reconstruction of Differential Emission Measure

Abstract

Astrophysicists are interested in studying the physical properties of the environment of a star such as its composition and temperature structure. This distribution of the temperature, known as the differential emission measure (DEM) provides a powerful tool for the characterization of the temperature structure of a stellar corona. Data are collected using state-of-the-art space-based telescopes such as the Chandra X-ray Observatory; Chandra can register the energy of each incoming X-ray from the source star. An ion at a given temperature in the stellar corona produces X-rays with a given energy following a certain conditional distribution of energy. These conditional distributions is computed using detailed quantum mechanical computations and ground-based laboratory measurements. Given this set of conditional distributions and the the observed marginal distribution of X-ray energies, we aim to reconstruct the marginal distribution of the temperature and the elemental abundances of the corona. Numerically, this involves a difficult inverse problem, which we accomplish using the method of data augmentation. Specifically we treat the photon count in each of a number of temperature bins and the count made up of each of the elements and other variables as missing data. Under this construction, we are able to use statistical methods that are designed to handle missing data problems. In particular, we use the EM Algorithm and Markov

chain Monte Carlo to compute the maximum a posteriori estimates and the highest posterior density intervals, respectively. We implement a Bayesian multi-scale (wavelet-like) prior distribution to smooth the DEM distribution, which gives us the flexibility to overcome the lack of information especially with low count data and high proportion of missing data. This approach allows for global spectral modeling, with the ability to include prior information in the form of known sequences and relative strengths of spectral lines; the inclusion and propagation of errors in atomic data; and a proper accounting of the uncertainties in the reconstructed DEM. We provide several simulation studies with both high-count and low-count data to evaluate the proposed method and Capella data DEM reconstruction results.

1.0 Preface

This work is a joint work with David A. van Dyk, Department of Statistics, University of California at Irvine, Vinay Kashyap, Harvard-Smithsonian Center for Astrophysics, and Alanna Connors, Eureka Scientific. This project is also a product of collaborative effort of the the California-Harvard astrostatistics collaboration (CHASC) whose members include J. Chiang, A. Connors, D. van Dyk, D. Esch, P. Freeman, H. Kang, V. L. Kashyap, X.-L. Meng, A. Siemiginowska, E. Sourlas, T. Park, Y. Yu, and A. Zezas. The funding for this project partially provided by NSF grant DMS-01-04129 and by NASA Contract NAS8-39073 (Chandra X-ray Center).

1.1 Introduction

The corona is the outermost layer of a stellar atmosphere. Figure 1.1 shows an optical image of the solar corona. Astrophysicists have been interested in studying the physical environment of stellar coronae such as its temperature and the elemental composition. The distribution of the amount of emission at different temperatures in a stellar corona is called the differential emission measure (DEM) and the fractions of each element compared to hydrogen are called the elemental abundances. In recent years, a number of new space telescopes have produced a tremendous amount of new data and allowed us to collect data from a wide range of energy range, from ultra-violet to γ-ray, with very high resolution. The *Chandra X-ray Observatory*, for example, produces energy spectra at least thirty times sharper than any previous X-ray telescope. However, even for the best X-ray telescope such as *Chandra*, there are some measurement errors that can cause a biased estimation if the error structure is not correctly accounted for.

The complexity of available data requires us to take a new and sophisticated statistical approach to estimate the DEM and the elemental abundance. Corona emits its

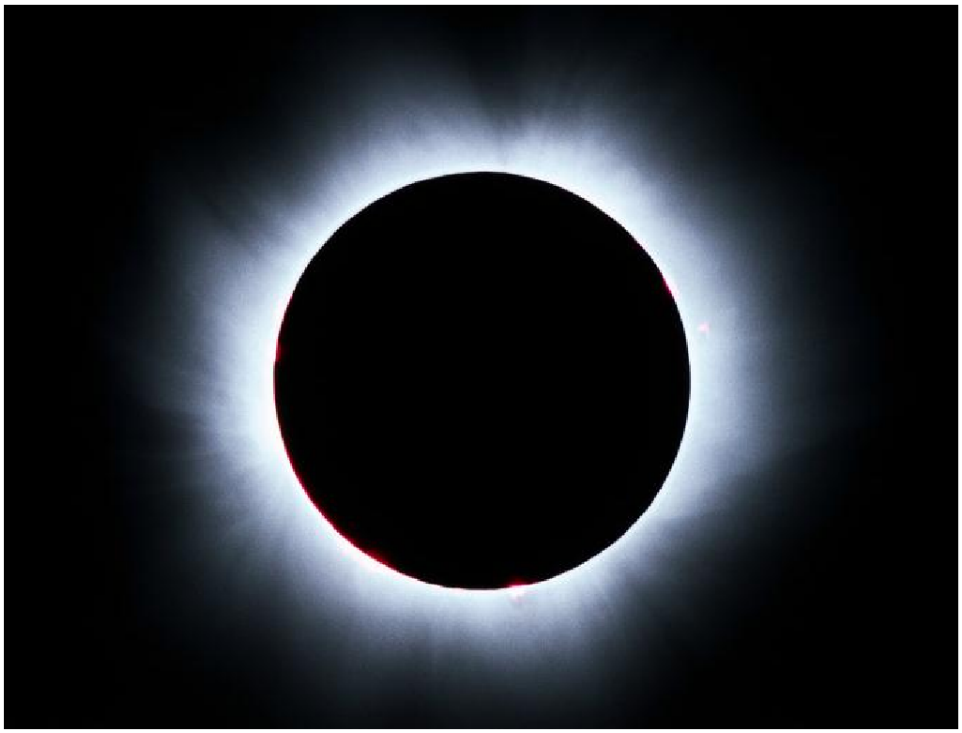

Figure 1.1: The Solar Corona.
The Sun's corona, photographed on August 11th, 1999 during a total solar eclipse. The corona is invisible during normal conditions because the surface of the Sun is much brighter than the mostly X-ray-emitting corona. The coronae of other stars can be detected with X-ray telescopes. (Image Credit: The European Southern Observatory PR Photo 24a/01)

photons. Because their energies follows a stochastic process that is involved with the DEM and the abundances, the DEM and abundances are indirectly observed by us in the form of a stellar spectrum (i.e., the photon counts according to the energies of the photon). Therefore, we can reconstruct the DEM and the elemental abundances from the photon count collected by the telescope given the quantum mechanics of the atoms. Moreover, the instrumental measurement error adds another level of the complexity. We describe the complex models that are designed to account for both the physics involved in the creation of electro-magnetic emissions at the astronomical source and the complexities of the data collection mechanisms inherent in astronomical instruments. This is similar to a well-known Poisson inverse problem. Previous approaches (Dupree *et al.* (1993), Kashyap and Drake

(1998)) focused on examining a limited number of (manually) selected influential emission lines to compute the DEM. This strategy can result in selection biases and large uncertainties in the estimates. In this paper, we apply a new method of DEM reconstruction to stellar extreme-ultra-violet(EUV) and X-ray data within a model based Bayesian framework. This method allows us to find a set of DEM parameters that describe the observed data best in terms of the Bayesian posterior distribution and simultaneously determine the element abundances, to incorporate atomic and calibration errors as prior information, and to produce error estimates on the fitted parameters.

In Section 1.2, we explore the scientific background of high-energy electromagnetic emission of stellar corona. In particular, we describe the relationship between the temperature and the composition of the source and the observed energy spectrum and in Section 1.2.3, we introduce the spectral model and its parameters. In Section 1.3, we explain the data collection process and instrumental effects. In Section 1.4, we introduce Bayesian missing data formulation to solve the numerically challenging inverse problem and their posterior distributions of the model parameters and missing data and we implement a Bayesian multi-scale (wavelet-like) model to smooth the DEM distribution, which gives us the flexibility to overcome the lack of information especially with low count data. Finally in Section 1.6, simulated data examples and Capella DEM reconstruction examples will be provided.

1.2 Scientific Background

1.2.1 Energy Spectra

The basic goal of high-energy spectral modeling from a statistical perspective is to model the distribution of the energy of high-energy photons (EUV, X-ray, or γ-ray)

from a particular astronomical source. Such a spectral model typically contains several additive components which can be formulated as a finite mixture model. Roughly speaking, the components can be split into two groups: *continuum* terms, which describe the distribution over the entire energy range of interest and *emission lines*, which are local positive aberrations from the continuum.

Continuum. The center of the star is composed of very hot gas, which produces copious photons that random walk their way to the surface of the star. This process creates a continuous spectrum, or continuum, of radiated energy and is known as blackbody emission. As another example, consider a high-temperature low-density plasma where photons are not thermalized by repeated collisions with ions in the plasma; the transitions between levels in free electron, induced by electrostatic interactions with ionized nuclei result in a so-called thermal Bremsstrahlung continuum. Among other things, the shape of the continuum indicates the temperature of the source. The shape of the continuum, as the name suggests, is smooth and smooth compared to emission lines. An example of a simple continuum appears in Figure 1.2.

Emission Lines. Emission lines are local features added to the continuum and represent extra emission of photons in narrow bands of energy. Such extra emission is due to photons that are emitted when an electron falls to a lower energy shell of a particular ion; the abundance of the extra emission indicates the abundance of the ion in the source. Thus, analysis of emission lines is informative as to the chemical composition of the surface of the astronomical source. Statistically the emission lines are represented by adding Gaussian, Lorentzian (i.e., a t-density with one degree of freedom), or delta functions to the continuum. An example of emission lines appears in Figure 1.2.

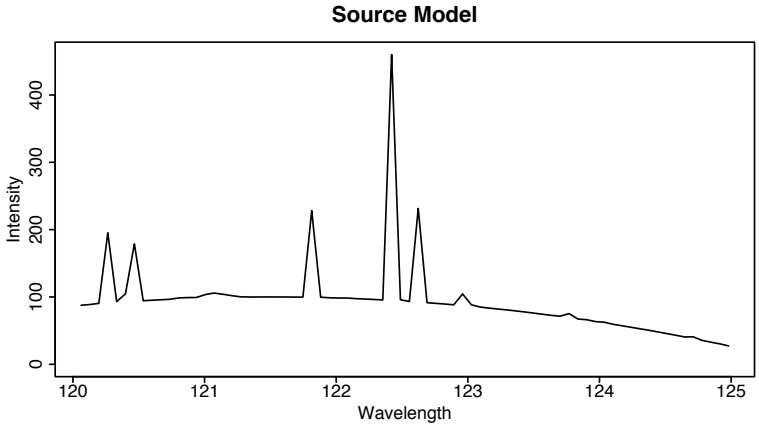

Figure 1.2: The Model Energy Spectrum.
The figure illustrates an artificial source model energy spectrum. The continuum is the smooth curve near the bottom of the plot and the sharp peaks represent the emission lines. x-axis represents wavelength ($\mathring{A}$), and y-axis represents expected counts.

1.2.2 Differential Emission Measure

The corona is the outermost layer of a stellar atmosphere and contains very low-density (about 10^9 particles/cm^3) and very hot ($> 10^6$ K) plasma. The DEM is a measure of the distribution of the emission at different temperatures in a stellar corona. Figure 1.3 illustrates the solar corona by imaging the Sun in three wavelengths[1]. The first panel is an optical image taken on March 29, 2001 of a portion of the Sun and illustrates the largest sunspot group to appear in a decade; at its peak this group was over ten times the size of the Earth. The second and third panels illustrate an EUV image and an X-ray image of the same region of the Sun, respectively. Although in visible light the sunspots appear as dark areas against the bright surface of the Sun, they are bright in the EUV and X-ray. The X-ray image shows large loops of glowing plasma arching above the sunspot group. The reason that the images look so different is that they are actually revealing different layers of the Sun's atmosphere. The visible photons originate from the photosphere, the

[1]URL: `http://antwrp.gsfc.nasa.gov/apod/ap010419.html`

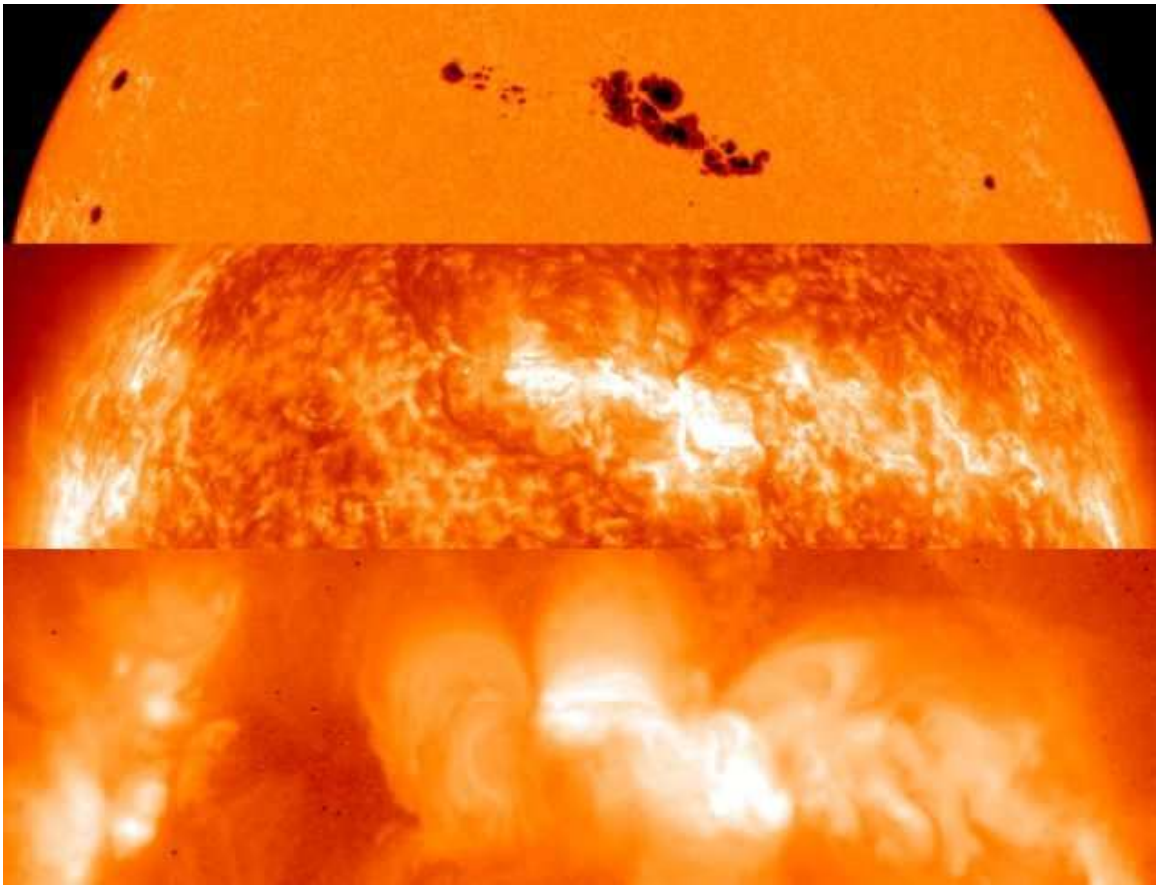

Figure 1.3: The Sun.
The three images are, from top to bottom, optical, EUV, and X-ray images. A stellar DEM is a representation of the distribution of the relative emissions at different temperatures in a stellar corona. (Image Credit: SOHO - MDI / EIT Consortium, Yohkoh / SXT Project.)

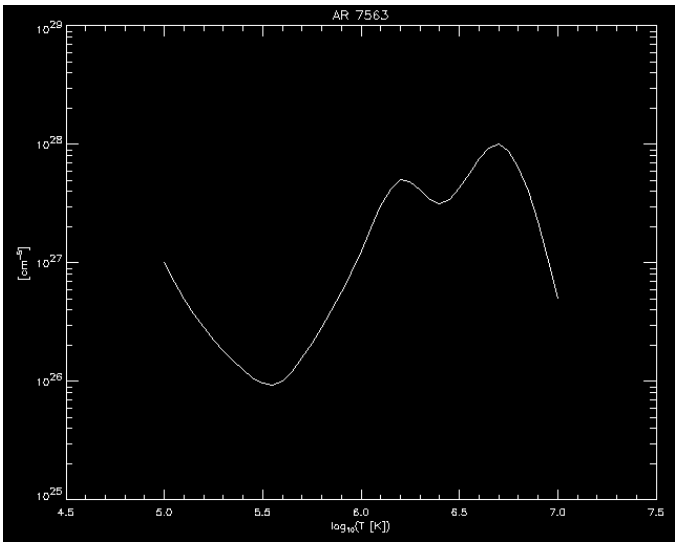

Figure 1.4: The Solar DEM in an Active Region.
This is a plot of the relative emission in a region of the solar corona with high sunspot activity as a function of the temperature of the plasma. The plot can be compared with that in Figure 1.5, which plots the solar DEM in a quiet region of the Sun. (Brosius *et al.*, 1996)

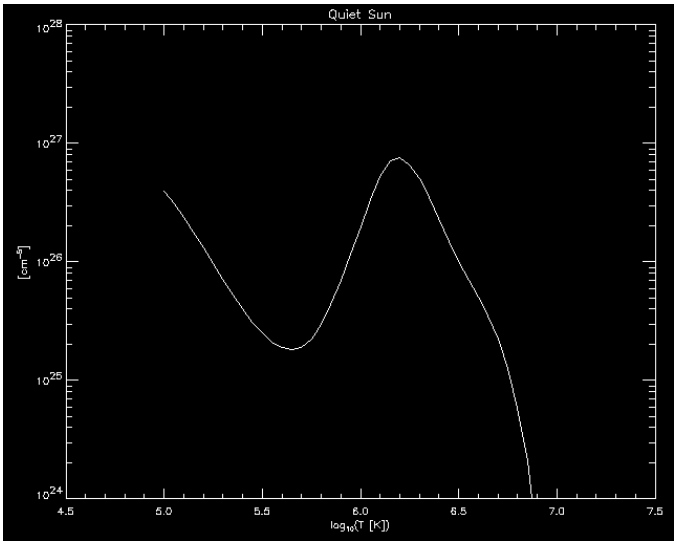

Figure 1.5: The Solar DEM in an Quiet Region.
This is a plot of the relative emission in a region of the solar corona with no sunspot activity as a function of the temperature of the plasma. The plot can be compared with that in Figure 1.4, which plots the solar DEM in an active region of the Sun. (Brosius *et al.*, 1996)

lowest and coolest layer at about 5000 degrees Kelvin, the EUV image reveals the chromosphere/transition region which is above the photosphere and hotter at 10–100 thousand degrees Kelvin. Finally, the X-rays originate from the solar corona that is even higher and is even hotter, at least a million degrees Kelvin.

The X-ray image in Figure 1.3 illustrates the complex structure in the intensity of the X-ray emission across the solar corona. The structure of the emission is a tracer of temperature and density in the corona. (The visible image also reveals temperature structure, the sunspots are much cooler than their surroundings. X-ray images, however, are not useful for viewing the temperature structure of the relatively cool photosphere.) This is illustrated in Figure 1.4 that plots the relative emission from the solar coronal region with high sunspot activity as a function of temperature; this is plot of the solar DEM in an active region. Figure 1.5 is the same plot but in a region with no sunspot activity. Notice that there is relatively less very hot plasma in the quiet region of the solar corona.

Although impressive images of the solar corona are available from the Solar & Heliospheric Observatory (SOHO)[2] and other solar observatory telescopes, very little is known about the temperature structure of stellar coronae. The stars being very distant, their disks cannot be resolved with existing resolutions of even the best telescopes. We can, however, infer their structures indirectly from examining their spectra. There are clues in the emission lines of stellar X-ray spectra that can be unlocked using prior information obtained from detailed quantum mechanical computations and ground-based laboratory measurements. A stellar corona plasma is made up of various ions which can be recognized in a spectrum from their identifying emission lines. Thus, relative strengths of the emission lines corresponding to different ions carry information about the temperature of the source.

[2]URL: `http://sohowww.nascom.nasa.gov`

1.2.3 Building the Spectral Model

Our spectral model can be split into two broad groups, *the continuum term* and *emission lines*. The number of photons from emission lines, so called spectral line fluxes, and continuum fluxes depend on the amount of the elements (relative to the amount of hydrogen) in the corona, so called element abundance and the relative emission in coronal region at different temperature (DEM). We emphasize that we are only interested in the *"relative"* magnitude of our parameters (the element abundances and the DEM).

Emission Line Spectral Model

The photon intensity corresponding to spectral line l generated in element k is the volume integrated product of the line contribution function, the elemental abundance γ_k, and the DEM at a given temperature T:

$$\lambda_l^{L,k} = A \int \gamma_k G_l^{L,k}(T) \mathrm{DEM}(T) d\log T, \tag{1.1}$$

where A is a known constant which includes the wavelength of the transition and the stellar distance, γ_k is the elemental abundance of element k, $G_l^{L,k}(T)$ is the contribution function of the line for element k, and $\mathrm{DEM}(T)$ is the DEM at temperature T (Kashyap and Drake, 1998). This function works as a relative probability that an emitted photon at a given temperature falls into each emission lines. Here, we use the term *element* k to refer the chemical composition of the corona such as hydrogen(H) and we use superscript $'L'$ to refer the emission lines. We define the elemental abundance γ_k as the relative abundance of element k to hydrogen(H) and Helium(He)'s abundance. We fix both $\gamma_{\mathrm{H}}, \gamma_{\mathrm{He}}$ to 1. Other abundances are given relative to solar photospheric abundances. For example, the iron(Fe) abundance, $\gamma_{\mathrm{Fe}} = 2$ for a corona implies that the corona has twice as many iron atoms relative to hydrogen(H) atoms in its corona compared with the Sun. The energy of an emission line, i.e, the location of the line on the spectrum originating from a particular

ion is predetermined because the energy emitted corresponds to the difference in the discrete levels of energy. Although it is technically possible to compute exact values of the constant term A and $\mathrm{DEM}(T)$ in equation (1.1), we can only compute their relative values because we do not take the exposure time into account in our model. In other words, when the data set has longer or shorter exposure time, our method produces higher or lower DEM values, respectively, but the absolute magnitude of the values are meaningless.

Thanks to quantum physics calculations, we are given the contribution function, $G_l^{L,k}(T)$. The contribution function is obtained from the Atomic Database (ATOMDB)[3] and it provides $G_l^{L,k}(T)$ at 50 discrete temperatures, which are $10^{4.1}, 10^{4.2}, \ldots, 10^{9.0}$ Kelvin. Since, $G_l^{L,k}(T)$ is not fully specified parametrically for every temperature value, but only available at selected temperatures, assuming that $G_l^{L,k}(T)$ is smooth enough to be linearly approximated, we interpolate the values of the contribution function on evenly spaced temperature values in log-scale ($T_1, T_2, \ldots,$ and T_T). Then, we can express the vector of the emission line intensities originating from element k by

$$\begin{aligned}
\lambda_l^{L,k} &= A \int \gamma_k G_l^{L,k}(T) N_e^2(T) dV(T) \\
&= A \int \gamma_k G_l^{L,k}(T) \left(N_e^2(T) \frac{dV(T)}{d\log T} \right) d\log T \\
&= A \int \gamma_k G_l^{L,k}(T) \mathrm{DEM}(T) d\log T \\
&\approx A(\Delta \log T) \sum_{t=1}^{T} \gamma_k \mathbf{G}_{lt}^{L,k} \mu_t, \quad \text{or equivalently} \\
\boldsymbol{\lambda}^{L,k} &= A(\Delta \log T) \gamma_k \mathbf{G}^{L,k} \boldsymbol{\mu} \;\propto\; \gamma_k \mathbf{G}^{L,k} \boldsymbol{\mu}, \qquad (1.2)
\end{aligned}$$

where N_e is the electron number density and $V(T)$ is the volume of the plasma at temperature T and $\mathbf{G}_{lt}^{L,k} = G_l^{L,k}(T_t)$ and $\mathrm{DEM}(T_t) = \mu_t, t = 1, \ldots, T$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_T)'$, and $\boldsymbol{\lambda}^{L,k} = (\lambda_1^{L,k}, \ldots, \lambda_{N_k}^{L,k})'$. We call $\mathbf{G}^{L,k}$ the *emissivity matrix*. (See Kashyap and Drake (1998) for rigorous derivation of the DEM)

[3]URL: `http://cxc.harvard.edu/atomdb/`(Smith *et al.*, 2001)

The dimension of $\mathbf{G}^{L,k}$ is $N_k \times T$, where N_k is the number of emission lines within the energy range of the data made up of element k. For example, between $80\mathring{A}$ and $140\mathring{A}$, there are no H-emission lines, but there are 3,270 Fe-emission lines listed in ATOMDB. Each row of $\mathbf{G}^{L,k}$ corresponds to the location of an emission line, and each column corresponds to a temperature point.

Continuum Spectral Model

The continuum[4] corresponding to a energy bin j is defined in a similar way to equation (1.1):

$$\begin{aligned} \lambda_j^{C,k} &= A \int \gamma_k G_j^{C,k}(T) \mathrm{DEM}(T) d\log T \\ &\approx A(\Delta \log T) \sum_{t=1}^{T} \gamma_k \mathbf{G}_{jt}^{C,k} \mu_t, \quad \text{or equivalently} \\ \boldsymbol{\lambda}^{C,k} &= A(\Delta \log T) \gamma_k \mathbf{G}^{C,k} \boldsymbol{\mu} \ \propto \ \gamma_k \mathbf{G}^{C,k} \boldsymbol{\mu}, \end{aligned} \tag{1.3}$$

where $\mathbf{G}_{jt}^{C,k} = G_j^{C,k}(T_t)$ and $\boldsymbol{\lambda}^{C,k} = (\lambda_1^{C,k}, \ldots, \lambda_J^{C,k})'$. The term *energy bin* refers to an (artificial) energy range corresponding the source energy spectrum. The dimension of $\mathbf{G}^{C,k}$ is $J \times T$, where J is the number of energy bins in the data and J is same for all elements. Each row of $\mathbf{G}^{C,k}$ corresponds to an energy bin, and each column corresponds to a temperature point.

Mixture of Emission Lines and Continuum

Each element has a continuum emissivity matrix and a line emissivity matrix. There are 28 emissivity matrices in total. Because the energy bin and emission line location do not correspond to each other, we need to identify which bin an emission line belongs to, when adding the two components of the fluxes to an

[4]Continuum can be modeled by a parametric curve and fit via generalized linear models. See van Dyk *et al.* (2001) and van Dyk and Kang (2004) for example. However, in this paper, the continuum curve is determined by contribution function and the DEM.

energy bin. For example, the flux from the emission line located at 6.690$\mathring{A}$ is added to the energy bin (6.6875$\mathring{A}$, 6.7$\mathring{A}$). We denote this matching process by binning$(\cdot)$, which matches the dimension of the line emissivity matrix to the continuum emissivity matrix, so that we can express the *total emissivity matrix* as element abundance-weighted sum of emissivity matrices, which equals to $\sum_{\text{element } k} \gamma_k\{\mathbf{G}^{C,k} + \text{binning}(\mathbf{G}^{L,k})\}$. Therefore, the expected photon counts at energy bin j, λ_j has typical Poisson image process equation:

$$\begin{aligned} \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)' &= \sum_{k=1}^{K} \{\boldsymbol{\lambda}^{C,k} + \text{binning}(\boldsymbol{\lambda}^{L,k})\} \\ &\propto \left(\sum_{k=1}^{K} \gamma_k \{\mathbf{G}^{C,k} + \text{binning}(\mathbf{G}^{L,k})\} \right) \boldsymbol{\mu}, \end{aligned} \tag{1.4}$$

where K is the total number of elements, i.e., $K = 14$.

Note that the column sums of the total emissivity matrix,

$$\mathbf{G}_{\text{total}} = \left(\sum_{k=1}^{K} \gamma_k \{\mathbf{G}^{C,k} + \text{binning}(\mathbf{G}^{L,k})\} \right)$$

are not equal. Thus, photons originating from different temperatures have different rates to be emitted as if there were different censoring probability for different temperatures. We call this property *emissivity matrix censoring*. We will discuss this issue later in Section 1.4.

1.3 Data Collection and Instrumentation

1.3.1 Detectors and Gratings

There are two detectors aboard *Chandra*; one is a high spatial resolution microchannel plate detector (the High Resolution Camera, or the HRC), and the other is an imaging spectrometer with higher spectral resolution (the Advanced CCD(Charge-Coupled Device) Imaging Spectrometer, or ACIS). Both instruments

are essentially photon counting devices, and register the arrival time, the energy, and the (two-dimensional) direction of arrival of incoming photons. Because of instrumental constraints, each of the four variables is discrete; the high resolution of *Chandra* means that the discretization is much finer than was previously available. The ACIS detector is composed of 10 CCDs, each of which has 1024×1024 pixels for spatial data. Because the data are discrete, they are compiled into a four-way table of photon counts. In our spectral application, we only focus on the energy variable, hence the time and the two special coordinates are marginalized out to produce one-way table of photon counts at each energy. Due to its digital nature, *Chandra* records the photon counts in a number of pre-specified energy bins.

It is possible to use one of two diffraction gratings with either of the two detectors. A diffraction grating is placed in the beam of X-rays and diffracts the photon by an angle that depends on the photon wavelength. (The wavelength of a photon is proportional to the reciprocal of its energy.) One of the two gratings, the High-Energy Transmission Grating Spectrometer (HETGS), is designed for high-energy X-rays, the other, the Low-Energy Transmission Grating Spectrometer (LETGS), is designed for low-energy X-rays. If *Chandra* is focused on a point source, such as a star, and a grating is in place, the energy of the photons can be recovered from the locations where they are recorded on the detector. Thus, the gratings greatly increases the spectral resolution of both of the detectors. Because the spectral resolution obtained with gratings is dominated by the size of the image, however, the advantage of the grating for spectral analysis is diminished for more extended source, such as nebula. Because the gratings also refract about 90% of the photons away from the detector, they are ordinarily only used with bright sources.

1.3.2 Stochastic Censoring

Photons arriving at the detector are not always recorded by the detector; a photon has a certain energy dependent probability of being recorded by the detector. This relative efficiency is called *effective area*. The mirrors on *Chandra* reflect the X-rays to focus them on the detector. Unfortunately, photons do not reflect uniformly. Each X-ray has a certain probability of being reflected away from the detector or being absorbed by the telescope mirrors. Since this probability depends on the energy of the photon, the probability that a photon is recorded by the detector depends on its energy. This process results in non-ignorable missing data mechanism and this should be accounted for to avoid biases in model fitting. See the change from the the third plot to the fourth plot in Figure 1.6. The height of the curve, i.e. the photon intensity, decreases because some of the photons are not registered by the detector.

1.3.3 Measurement Errors - Blurring

Chandra focuses X-rays with mirrors. Because the mirrors do not focus perfectly, energy spectra are blurred. The so-called line-spread function characterizes the probability distribution of a photon's recorded energy location relative to its true energy. The shape of the distribution varies with the energy of the incoming photon. Fortunately, in spectral analysis, well-known distributions such as t-distribution or normal distribution is accurate enough for data within a small energy range. For example, for *Chandra* HRC-S/LETG data with wavelength range about 1-180$\mathring{A}$(1$\mathring{A}$= 1.0 $\times 10^{-10}$ meters), t-distribution with 4 degree of freedom serves as a good approximation. Data from *EUVE (the Extreme Ultraviolet Explorer)* have a Gaussian line-spread function with standard deviation equal to 0.17$\mathring{A}$. The change from the fourth plot to fifth plot in Figure 1.6 illustrates how a delta-function-shaped emission line become a bell-shaped curve due to the blur-

ring.

1.3.4 Background Contamination

The photon counts detected by the telescope do not directly correspond to the source of interest. They are the mixture of the source photon counts and counts originating from other celestial objects (background counts) that are near the line of sight of the the source of interest. After adjusting for exposure time and the area in which the background counts are collected relative to that in which the source counts are collected, it is standard practice to directly subtract the counts observed in the background exposure from those observed in the source exposure, and treat the resulting counts as if it were source only counts. This procedure is problematic, because it can result in negative bin counts when the source counts are weak. We will describe a better strategy to model the counts in the two observation as independent Poisson random variables, one with only background intensity and the other with intensity equal to the sum of the background and source intensities (Loredo (1992), van Dyk (2003)). See the transition from the fifth plot to sixth plot in Figure 1.6. Background contamination adds another level of complexity.

1.3.5 Data Distortion Model

There are two major data distortion effects due to the instrument: *effective area*(or stochastic censoring) and *blurring* as described in the previous sections. The effective area of bin j is the probability d_j that an X-ray is not refracted off the detector and it varies with the photon energy. A blurring of the photon energy occurs because a photon that arrives with energy corresponding to bin j has probability M_{ij} of being recorded in detector channel i, where the channel refers to the energy ranges corresponding the observed data. Note that the matrix $\mathbf{M} = \{M_{ij}\}$ for a "perfect" resolution detector would be an identity matrix provided that the bins

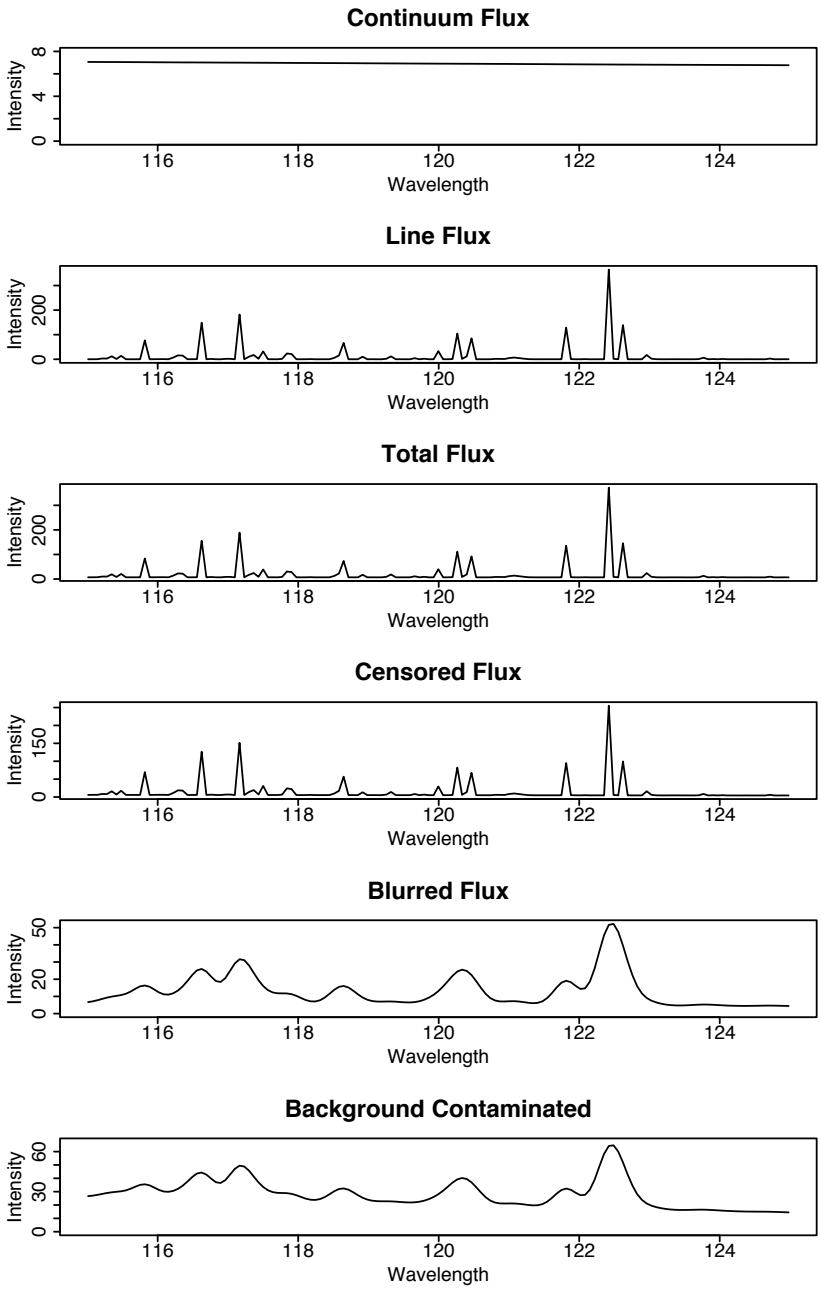

Figure 1.6: Degradation of Model Intensity.

The figure illustrates the various stochastic processes that significantly degrade the source model and result in the Poisson intensities for the observed counts. (An artificial source model is used.) For all plots, x-axis represents wavelength ($\mathring{A}$), and y-axis represents expected counts. The first plot shows the source intensity due to the continuum, the second plot shows the source intensity due to the emission lines and the third plot is the sum of the two components. The fourth plot illustrates *stochastic censoring* due to non-constant effective area and the fifth plot shows the *blurring* effect. Note that the three Fe-lines around $122\mathring{A}$ become indistinguishable after the blurring. The last plot shows background contamination; photon counts from the background are added to the counts from the source.

and channels coincide. Therefore, we model the observed counts at channel i with background contamination as independent Poisson variables with intensity

$$\xi_i = \sum_{j=1}^{J} M_{ij}\lambda_j d_j + \lambda_i^B, \;\; i = 1, \ldots, I, \tag{1.5}$$

where λ_i^B is the Poisson intensity of the background at channel i. The equation (1.5) is equivalent to the following matrix representation

$$\boldsymbol{\xi} = \mathbf{MD}\boldsymbol{\lambda} + \boldsymbol{\lambda}^B, \tag{1.6}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_I)$ is the vector of expected detector counts corresponding to channels, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)$ is the vector of the expected source counts corresponding to energy bin, $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_j)$ is a diagonal $J \times J$ matrix of the effective area, $\mathbf{M} = \{M_{ij}\}$ is the $I \times J$ blurring matrix, and $\boldsymbol{\lambda}^B = (\lambda_1^B, \ldots, \lambda_I^B)$ is the vector of expected background counts. We call $\mathbf{M}$ *spectral response matrix.* If the probability that a photon is recorded at a distant energy bin from its true energy is small, i.e. the tail probability of the line spread function is small, then we can save computational time by considering only near-diagonal elements of the spectral matrix.

In our model, we are provided with a known background channel intensity c_i up to a normalizing constant, that is, $\lambda_i^B = \beta c_i, \;\; i = 1, ..., I$, where β is a parameter for the background normalizing constant.

We can re-write the observed photon intensities as a function of emissivity matrices, DEM, and abundances:

$$\begin{aligned} \boldsymbol{\xi} &= \mathbf{MD}\boldsymbol{\lambda} + \boldsymbol{\lambda}^B \\ &\approx (A\Delta \log T)\mathbf{MD}\left(\sum_{k=1}^{K} \gamma_k \{\mathbf{G}^{C,k} + \mathrm{binning}(\mathbf{G}^{L,k})\}\right)\boldsymbol{\mu} + \boldsymbol{\lambda}^B. \end{aligned} \tag{1.7}$$

Given the channel intensity $\boldsymbol{\xi}$, the observed channel counts follow independent Poisson distribution. Therefore, the observed data likelihood is:

$$L(\theta|\mathbf{Y}^{\mathrm{obs}}) = \prod_{i=1}^{I} \frac{e^{-\xi_i}\xi_i^{Y_i^{\mathrm{obs}}}}{Y_i^{\mathrm{obs}}!}, \tag{1.8}$$

where θ denote the collection of model parameters, $\theta = (\boldsymbol{\mu}, \boldsymbol{\gamma}, \beta)$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)$. Note that other parameters are functions of θ ($\boldsymbol{\lambda} = \boldsymbol{\lambda}(\theta), \boldsymbol{\lambda}^B = \boldsymbol{\lambda}^B(\theta), \boldsymbol{\xi} = \boldsymbol{\xi}(\theta)$), but the notation θ is omitted for notational convenience in the following.

1.3.6 Prior Specification

We use the conjugate prior for $\boldsymbol{\gamma}$

$$\gamma_k \sim \mathbf{Gamma}(a_\gamma, b_\gamma),$$

for all k and the conjugate prior for β

$$\beta \sim \mathbf{Gamma}(a_\beta, b_\beta).$$

These prior distributions are chosen simply for the theoretical justification and not designed to be informative. For non-informative prior distributions, one can choose $a_\gamma = a_\beta = 1$ and $b_\gamma = b_\beta \approx 0 > 0$. For $\boldsymbol{\mu}$, however, we choose the conjugate prior distributions for multi-scale smoothing, to smooth the DEM estimates. The details of this prior are discussed in Section 1.4.2. All the parameters are independent a priori.

1.4 Bayesian Deconvolution Methods

1.4.1 Hierarchical Missing Data Structuring

Mathematically, if we were provided with the expected channel counts $\boldsymbol{\xi}$, finding the maximum likelihood estimate of the model parameters ($\boldsymbol{\mu}$, $\boldsymbol{\gamma}$, β) can be computed by solving equation (1.7), which involves multiple matrix inversions. In practice, $\boldsymbol{\lambda}$ has tens of thousands of components, while $\boldsymbol{\mu}$ has fewer than 100. This may seem to be a simple task, since there are far more data points than unknown

parameters, but the information regarding the emission line counts and the continuum counts is very sparse: $\mathbf{G}^{L,k}$ and $\mathbf{G}^{C,k}$ are nearly singular and the ML estimate can be very poorly behaved especially when the relative information of observed data to the augmented data is very low as we will show later in Section 1.6.2 with Chandra data example. We suggest Bayesian methods and informative prior information for $\boldsymbol{\mu}$.

As described in the previous sections, the model has multiple components (e.g., background, continuum and emission lines), high-dimensional parameters, and complex structure due to large dimensional matrices. We can formulate the model in terms of a set of layers of missing data. For example, data augmentation methods such as the EM (Dempster, Laird, and Rubin, 1977) or DA algorithm (Tanner and Wong, 1987) can be used to effectively fit highly structured models that are formulated in terms of missing data or latent variables. Thus, we impute the missing data at each level and find the conditional posterior distribution of the next level of missing data given the current level. This will eventually lead us to estimating the ideal photon counts emitted from each temperature value, which makes the estimation of θ straightforward.

To formulate the model in terms of missing data, we can state four major levels of missing photon counts: *channel level, bin level, temperature level* and *element level*. For notational convenience, we distinguish these levels by using the notation $\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{U}$ and $\mathbf{V}$, respectively. Recall that the energy channel refers to the energy ranges corresponding the observed data and the energy bin refers to an (artificial) energy range corresponding the source energy spectrum. In other words, channel level photon counts can be interpreted as a histogram whose bins are broken by the energy channel ranges. We also use index $i = 1, \ldots, I$ for channel level, $j = 1, \ldots, J$ for energy bin level, $t = 1, \ldots, T$ for temperature level, and $k = 1, \ldots, K$ for element level consistently throughout the paper. We also use superscript $'-'$ to denote a lower level of augmentation within a level. For example, at energy bin j, the con-

ditional distribution of the photon counts recorded by the detector Z_j^- given the total photon counts Z_j arriving is formulated with a binomial distribution with a success probability, d_j; $Z_j^-|Z_j \sim \mathbf{Binomial}(Z_j, d_j)$. We outlined the hierarchy of augmented data structures in Table 1.1 and in Figure 1.7.

This hierarchical structuring is desirable especially when the joint distribution of the missing data and the parameters are factorized into several terms, i.e., independent conditional distributions. In particular, the joint posterior distribution of our overall model has the following conditional distributions corresponding to each level of the missing data:

$$p(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \beta, \mathbf{V}, \mathbf{U}, \mathbf{U}^-, \mathbf{Z}, \mathbf{Z}^-, \mathbf{Y}|\mathbf{Y}^{\text{obs}}) \propto \tag{1.9}$$
$$p(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \beta)p(\mathbf{V}|\boldsymbol{\gamma}, \boldsymbol{\lambda})p(\mathbf{U}|\boldsymbol{\gamma}, \boldsymbol{\lambda})p(\mathbf{U}^-|\mathbf{U})p(\mathbf{Z}|\mathbf{U}^-)p(\mathbf{Z}^-|\mathbf{Z})p(\mathbf{Y}|\mathbf{Z}^-)p(\mathbf{Y}^{\text{obs}}|\mathbf{Y}, \beta),$$

where $p(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \beta)$ is the prior distribution of the model parameters. We will discuss the choice of prior distribution in Section 1.4.2. We only included relevant variables in the condition of the conditional distributions in equation (1.9) to show the conditional independent structure among the missing data.

In general, augmenting a higher level of missing data given a lower one within a major level involves separating components from the mixture of the components, and augmenting higher level missing data given lower ones between major levels corresponds to "inverting" the emissivity matrices. The model describes the relationships among the variables from top to bottom, while the data augmentation algorithm finds the posterior distribution of missing data from bottom to top, as in Table 1.1.

Note the similarity in construction in the emissivity matrix (equation (1.4)) and the spectral response matrix (equation (1.6)). To solve inverse problems for both equations, we need two major processes involved with missing data construction; the *stochastic censoring* and the *multinomial dispersion process* according to the line spread function or the contribution function.

Level	Variable	Notation
1.	Count of photons originating from element k,	V_k
$\downarrow$	*Contribution Function*	
2a.	Ideal photon count in temperature bin t,	U_t
$\downarrow$	*Stochastic Censoring*	
2b.	Stochastically censored photon count at temperature bin t	U_t^-
$\downarrow$	*Emissivity Matrices*	
3a.	Ideal bin count at energy bin j	Z_j
$\downarrow$	*Stochastic Censoring*	
3b.	Stochastically censored energy bin count at energy bin j	Z_j^-
$\downarrow$	*Line Spread Function*	
4a.	Source count at energy channel i	Y_i
4b.	Background photon count at energy channel i	Y_i^B
$\downarrow$	*Background contamination*	
4c.	**Observed** count for energy channel i, $Y_i^{\text{obs}} = Y_i + Y_i^B$	Y_i^{obs}

Table 1.1: Hierarchical Missing Data Variables for Data Augmentation.
We use index $i = 1, \ldots, I$ for channel level missing data, $j = 1, \ldots, J$ for energy bin level missing data, and $t = 1, \ldots, T$ for temperature level missing data, and $k = 1, \ldots, K$ for element level missing data throughout the paper. Augmenting higher level missing data given lower ones within a level involves separating components from the mixture of the components and augmenting higher level missing data given lower ones between levels corresponds to "inverting" the emissivity matrices or response matrix.

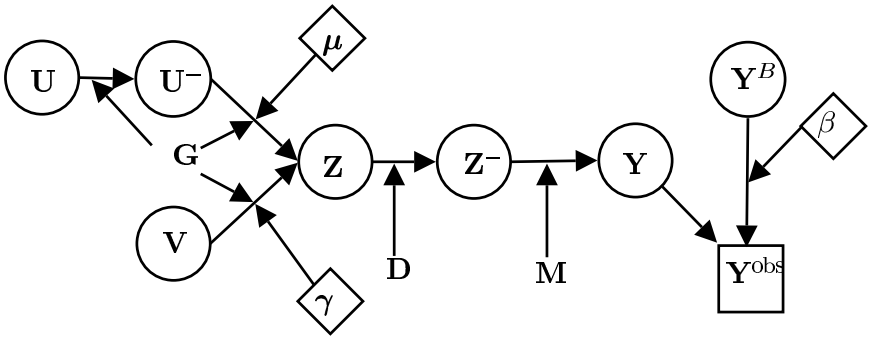

Figure 1.7: Graphical Representation of the Data Augmentation Scheme
The figure illustrates the conditional dependency of the various missing data and parameters. The circles represent the missing data and the diamonds represent the model parameters. Letters without boundaries are known constants. Arrows connecting missing data represent the conditional dependency from a higher level of missing data to a lower level of missing data.

Stochastic Censoring and Efficient Computation

Since the multiplicative constant term for the emissivity matrix does not affect the analysis on the relative DEM, we can freely re-normalize all the emissivity matrices. The normalizing constant is determined to achieve a fast algorithm, in other words, to reduce the augmented information for θ using conditional augmentation (Meng and van Dyk, 1997; van Dyk and Meng, 2001) by reducing the counts attributed to the censored photons. In general, the EM rate of convergence depends on relative information of observed data to the augmented data and the conditional augmentation approach is to minimize the geometric rate of convergence of the DA algorithm (Liu *et al.*, 1994).

Recall that the emissivity matrix censoring does not occur uniformly across the temperature, and the energies of the observed photons are biased toward areas of low absorption (i.e., the temperatures with large column sums of the emissivity matrix), complicating parameter estimation. It is important to note, however, that we need not account for (i.e. augment) all of the censored photons, but rather we only need the censoring probability to be uniform across the range of temperature. In particular, we normalize the total emissivity matrix such that the maximum column sum of the matrix is equal to 1 and we act as if the censoring rate were (1−"column sum of the normalized total emissivity matrix"). We define the normalizing constant $\mathbf{norm}$ by the maximum of the column sum of the total emissivity matrix, i.e.,

$$\mathbf{norm} = \max_{t=1,\ldots,T} \left\{ \sum_{j=1}^{J} \mathbf{G}_{\mathrm{total}\,j,t} \right\},$$

and let $\mathbf{G}^{\star}$ be the normalized total emissivity matrix, $\mathbf{G}^{\star} = \frac{1}{\mathbf{norm}} \mathbf{G}_{\mathrm{total}}$. By this construction, we create no augmentation situation for the temperature bin with the maximum column sum, and the least augmentation for other temperature bins, without distorting the different censoring rates on different temperatures. Note that $\mathbf{norm}$ changes its value at each iteration of the MCMC algorithms (see Sec-

tion 1.4.2), because of its dependency on $\boldsymbol{\gamma}$. This means that the data augmentation scheme is in a sense changing from iteration to iteration.

Let $g^{\star}_{+,t}$ be the t-th column sum of $\mathbf{G}^{\star}$. By treating $g^{\star}_{+,t}$ as the probability that a photon originating in temperature bin t is not censored, we have the following conditional relation for the missing data at the temperature level:

$$U_t^{-}|U_t, \theta \sim \mathbf{Binomial}(U_t, g^{\star}_{+,t}). \tag{1.10}$$

Similarly for the counts in energy bin level

$$Z_j^{-}|Z_j, \theta \sim \mathbf{Binomial}(Z_j, d_j). \tag{1.11}$$

Line Spread Function and Contribution Function

Given the photon counts at a certain temperature, energies of the photons are dispersed according to the column vector of the emissivity matrix. Thus, the effect of the emissivity matrix is modeled as a multinomial distribution:

$$\mathbf{Z}|\mathbf{U}^{-}, \theta \sim \sum_{t=1}^{T} \mathbf{Multinomial}\left(U_t^{-}, \frac{\mathbf{G}^{\star}_{\bullet t}}{\sum_{t=1}^{T} \mathbf{G}^{\star}_{\bullet t}}\right),). \tag{1.12}$$

where $\mathbf{G}^{\star}_{\bullet t}$ is the t-th column of $\mathbf{G}^{\star}$. Similarly, the energy channel counts given the energy bin counts are dispersed according to the column vector of the line-spread function matrix:

$$\mathbf{Y}|\mathbf{Z}^{-}, \theta \sim \sum_{j=1}^{J} \mathbf{Multinomial}\left(Z_j^{-}, \frac{\mathbf{M}_{\bullet j}}{\sum_{j=1}^{J} \mathbf{M}_{\bullet j}}\right), \tag{1.13}$$

where $\mathbf{M}_{\bullet j}$ is the j-th column of $\mathbf{M}$.

For the multinomial model involved with the element level of missing data and energy bin level of missing data, we take a similar step as described in equation (1.12). Note that $\mathbf{G}^{k}_{j,t}$ is a three-dimensional array (element k, energy bin j, and temperature bin t). $\mathbf{G}^{\star}$ is a weighted sum with respect to element with the weights equal to

the elemental abundances and is reduced to two-dimensional array. In similar pattern, we can take a weight sum with respect to temperature level with the weights equal to the DEM. Define such matrix $\mathbf{E}^{\star}$, where its element is

$$\mathbf{E}^{\star}_{j,k} = \frac{1}{\mathbf{norm}} \sum_{t=1}^{T} \mathbf{G}^{k}_{j,t}.$$

As in equation (1.12), $\mathbf{Z} = (Z_1, ..., Z_J)'$ also follows a multinomial distribution given the element counts $\mathbf{V} = (V_1, ..., V_K)'$:

$$\mathbf{Z}|\mathbf{V}, \theta \sim \sum_{k=1}^{K} \mathbf{Multinomial}\left(V_k, \frac{\mathbf{E}^{\star}_{\bullet k}}{\sum_{j=1}^{J} \mathbf{E}^{\star}_{j,k}}\right),$$

where $\mathbf{E}^{\star}_{\bullet k}$ is the k-th column of $\mathbf{E}^{\star}$.

1.4.2 DA Implementation

Data augmentation methods can simplify this convolved structure with a hierarchical formulation. Statistical inference for the unknown model parameters given the augmented data sets described in Table 1.1 is straightforward and the higher levels of missing data follow simple standard distributions given the model parameters and the lower level of missing data. The two conditional distributions are:

$$p(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \beta, |\mathbf{V}, \mathbf{U}, \mathbf{U}^{-}, \mathbf{Z}, \mathbf{Z}^{-}, \mathbf{Y}, \mathbf{Y}^{\text{obs}}), \text{ and} \tag{1.14}$$

$$p(\mathbf{V}, \mathbf{U}, \mathbf{U}^{-}, \mathbf{Z}, \mathbf{Z}^{-}, \mathbf{Y}, |\boldsymbol{\gamma}, \boldsymbol{\lambda}, \beta, \mathbf{Y}^{\text{obs}}). \tag{1.15}$$

When implementing EM, we take the expected value of the missing data under equation (1.15) and maximize equation (1.14) with respect to the parameters. When implementing MCMC sampler, we iteratively sample the missing data and the parameters under equation (1.15) and equation (1.14), respectively.

We provide the missing data distributions conditional on model parameters in Appendix.

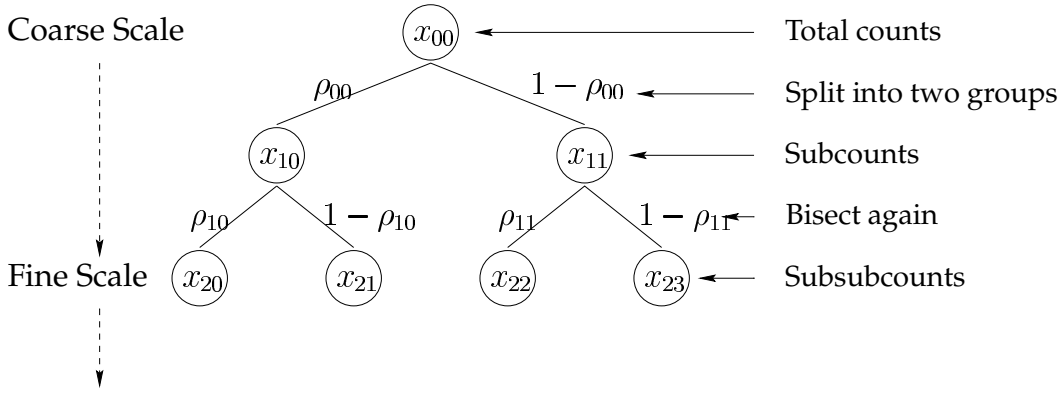

Figure 1.8: Multi-Scale Analysis and Modeling Represented on Binary Tree Graph. The Poisson intensity of a "parent" node is a sum of the Poisson intensity of the two "child" nodes. The smoothness of the intensities is controlled by the splitting factors $\{\rho_{r,k}\}$.

Model Parameter Distribution Conditional on Missing Data

We consider the imputed missing data as if they are observed from $U_t \sim \textbf{Poisson}(\mu_t)$, which leads us to directly apply Bayesian multi-scale analysis for the Poisson estimation problem. Multi-scale analysis gives us a nice tool for smoothing the underlying Poisson intensity. We briefly introduce the algorithm in this paper. See Nowak and Kolaczyk (2000) for more details.

Let us introduce a new data structure $\mathbf{x}$ generated with temperature bin counts $\mathbf{U}$ as illustrated in Figure 1.8:

$$\begin{aligned} x_{R,n} &:= U_{n+1}, \ \ n = 0, \ldots, 2^R - 1, \ T = 2^R, \\ x_{r,n} &= x_{r+1,2n} + x_{r+1,2n+1}, \ n = 0, \ldots, 2^r - 1, \ 0 \leq r \leq R-1, \end{aligned}$$

where the total level of scale, R. (In particular, we choose to use $R = 6$ and $T = 2^6 = 64$.) This produces the same binary tree structure for the DEM parameters:

$$\begin{aligned} \mu_{R,n} &:= \mu_{n+1}, \ \ n = 0, \ldots, 2^R - 1, \ T = 2^R, \\ \mu_{r,n} &= \mu_{r+1,2n} + \mu_{r+1,2n+1}, \ \ n = 0, \ldots, 2^r - 1, \ 0 \leq r \leq R-1. \end{aligned}$$

We use the following conjugate prior distributions,

$$\begin{aligned} \mu_{0,0} &\sim \mathbf{Gamma}(a_\mu, b_\mu), \\ \rho_{r,n} &\sim \mathbf{Beta}(\rho_{rn}|\alpha_r, \alpha_r). \end{aligned}$$

Note that we use symmetric beta priors of mean $1/2$. This prior distribution shrinks the DEM parameters toward equality, i.e., a smooth DEM reconstruction. The larger α_r is the smoother the reconstruction is. Sampling or finding MAP (maximum a posteriori) estimate of $\boldsymbol{\mu}$ is equivalent to sampling or finding MAP estimate of $\mu_{0,0}$ and $\rho_{r,n}$'s. because we can reconstruct DEM $\boldsymbol{\mu}$ via

$$\begin{aligned} \mu_{r+1,2n} &= \mu_{r,n}\rho_{r,n} \\ \mu_{r+1,2n+1} &= \mu_{r,n}(1-\rho_{r,n}), \;\; n = 0, \ldots, 2^r - 1, \; 0 \le r \le R-1, \\ \mu_{n+1} &= \mu_{R,n}, \;\; n = 0, \ldots, 2^R - 1, \; T = 2^R. \end{aligned}$$

For sampling the abundance parameters, we treat the imputed elemental counts as if they are observed. Under the conjugate gamma prior distribution $\gamma_k \sim \mathbf{Gamma}(a_\gamma, b_\gamma)$, for all the abundance parameters, we obtain the following posterior distributions:

$$\gamma_k \sim \mathbf{Gamma}\left(V_k + a_\gamma - 1, \mathbf{E}^\star_{+,k} + b_\gamma\right). \tag{1.16}$$

It is believed that some elements have similar abundance values due to empirical or theoretical reasons. For example, one can group elements based on observed similarity of behavior among different elements or group elements with low First Ionization Potential (FIP) into one group (Al, Ca, Ni, Mg, Si, Fe = Fe) and elements with high FIP into another (S, C, N, O, Ar, Ne = Ne). This grouping can also help computationally to speed up the algorithm especially when the elements with low counts is grouped with the elements with high counts. Thanks to the independence and conjugate priors for $\boldsymbol{\gamma}$, the posterior distributions of the grouped abundances are simply defined by the total photon counts and total Poisson intensities of the

group. For example, as in equation 1.16, the abundance for Ne-group (Ne and Ar) has the following posterior distribution:

$$\gamma_{(\mathrm{Ne-group})} \sim \mathbf{Gamma}\left(V_{\mathrm{Ne}} + V_{\mathrm{Ar}} + a_\gamma - 1, \mathbf{E}^{\star}_{+,\mathrm{Ne}} + \mathbf{E}^{\star}_{+,\mathrm{Ar}} + b_\gamma\right),$$

where $\mathbf{E}^{\star}_{+,\mathrm{Ne}}, \mathbf{E}^{\star}_{+,\mathrm{Ar}}$ is the column sum of $\mathbf{E}^{\star}$ corresponding to Ne and Ar.

The posterior distribution for the normalizing constant for the background counts also follows a gamma distribution under the conjugate gamma prior distribution $\beta \sim \mathbf{Gamma}(a_\beta, b_\beta)$. The posterior distribution for β is

$$\beta \sim \mathbf{Gamma}\left(\sum_{i=1}^{I} Y_i^B + a_\beta - 1, \sum_{i=1}^{I} c_i + b_\beta\right).$$

1.5 Atomic Data Errors

Even the best atomic emissivity databases have missing or misplaced lines and incorrect emissivities. Our method allows known information on these issues to be directly incorporated into the analysis.

Due to incomplete atomic data measurements as well as detector non-linearities, the observed locations of lines does not in general match the theoretical locations that we obtain from ATOMDB. We compensate for this effect by allowing the strongest lines in the spectral region to be shifted during the fit, and then move the remaining weaker lines accordingly. The fitting is done again in data augmentation fashion. We separate the channel counts coming from the strong lines and then impute the energy of the photons by randomly jittering each photon's energy within the energy channel range according to the probabilistic model for line spread function. Given the imputed locations (i.e. energies) of the photons coming from the strong lines, we can easily compute the posterior distributions of the fitted locations. This process requires two major imputation steps. Firstly, we need

to identify where a observed photon counts is attributed to. Secondly, we impute the energies of all photons originating from a particular emission line.

We outline the data augmentation scheme in the following. Given the emission line intensities $\boldsymbol{\lambda}^{L,k}$ and the continuum intensities $\boldsymbol{\lambda}^{C,k}$, we can compute the contribution of the intensity from each emission line and continuum component. For example, an emission line l with model intensity λ_l which belongs to energy bin j contributes its intensity to energy channel i by $M_{ij}\lambda_l d_j$. In the same way, we can compute the contribution from other continuum and emission lines. In practice, we can limit the number of components contributing to a certain energy channel by considering only a few emission lines and continuum components near by the energy channel, because the line spread function is ignorable when the energy is far away from the center of the function. After computing the amount of the contribution of the components under consideration, we can separate the source counts in energy channel i, Y_i into each component via multinomial distribution. We repeat this process on the other energy channels near emission line l to generate a histogram of photon counts originating from emission line l. Given this information, we impute the energies of each photon in the energy channel independently sampling from truncated line spread function, because the energy of the photon should be inside of the energy channel range. Given the energies of all the photons from the emission line, it is straightforward to sample or maximize the center of the line-spread function; Gaussian or t-distribution.

1.6 Results and Model Checking

1.6.1 DEM Reconstruction of Simulated Data

We assume a nominal form for the DEM, generate fluxes for a set of non-standard abundances, and reconstruct the DEM from the spectrum binned over the Chan-

dra's energy bin specification. We take the subset ranging from 3 to 30$\mathring{A}$; this corresponds to 2,161 wavelength bins of equal width of 0.0125$\mathring{A}$. There are 10,589 emission lines listed in ATOMDB within the range. The DEM reconstruction results are summarized in Figure 1.9, and the elemental abundance reconstructions are summarized in Table 1.2 and the results are consistent with the input values of the parameters for generating the simulated data set. The prior $\mathbf{Beta}(\alpha_r, \alpha_r),\ \alpha_r = 2, r = 0, \ldots, R-1$ are used for low degree of smoothing. We tested the algorithm with various sets of DEM and abundance input. The reconstruction results were very good in all simulation studies (Other simulation study results are not shown).

1.6.2 Capella DEM Reconstruction

Capella is a very bright X-ray source and for this reason it has been observed by every X-ray and EUV space-borne observatory. Capella is a close spectroscopic binary, at a distance of 12.93 pc [5], consisting of a G8 and G1 giant starts, [6] with masses of 2.7$M_\odot$ and 2.6$M_\odot$, and radii of 12$R_\odot$ and 9$R_\odot$, respectively. The binary has an orbital period of 104 days, not synchronous with the rotational period of the components, whose separation is 160$R_\odot$[7]. The raw spectrum of Capella collected with *Chandra* appears in Figure 1.10.

[5] 1 pc = 3.26 light year

[6] Temperature of stars are classified with the following characters.: 30,000 K is an 'O' star, 20,000 K is a 'B', 10,000 K is an 'A', 7,000 K is a 'F', 6,000 K is a 'G', 4,000 K is a 'K', 3,000 K is a 'M'. Each class is broken down into 0 to 9. The sun is a G2 star. A G1 star is a little warmer. A G3 star is a little cooler. Giant stars are 10 - 100 times bigger than the Sun.

[7] $M_\odot$ = Mass of the Sun, $R_\odot$ = Radius of the Sun

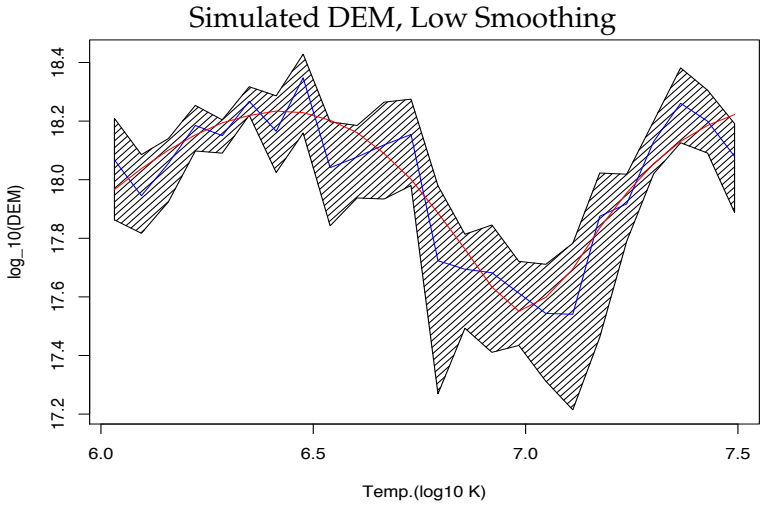

Figure 1.9: Error Bars for the Fitted DEM from the Simulation Study.
The plot draws DEM in log scale. The zigzag line is the posterior mean of the DEM and the shaded area represents component-wise 95% posterior intervals under low smoothing ($\alpha = 2$); the smooth line is the input DEM; Note that all the intervals contain the input DEM.

Element	Input Value	Mean	95% Interval
C	0.8	0.77	(0.70, 0.84)
Si	0.8	0.80	(0.74, 0.87)
N	2	2.00	(1.92, 2.10)
S	0.8	0.93	(0.75, 1.11)
O	0.5	0.50	(0.48, 0.52)
Ar	2.8	2.90	(2.68, 3.12)
Ne	5	5.06	(4.90, 5.22)
Ca	3.8	3.82	(3.45, 4.23)
Mg	3	2.99	(2.86, 3.12)
Fe	2	2.01	(1.95, 2.08)
Al	2.5	2.37	(1.57, 3.17)
Ni	2	2.03	(1.82, 2.26)

Table 1.2: Error Bars for the Fitted Abundances from the Simulation Study.
Posterior Mean and the 95% posterior interval for element abundance. Note that all the intervals contain the input abundance which we used for data generation.

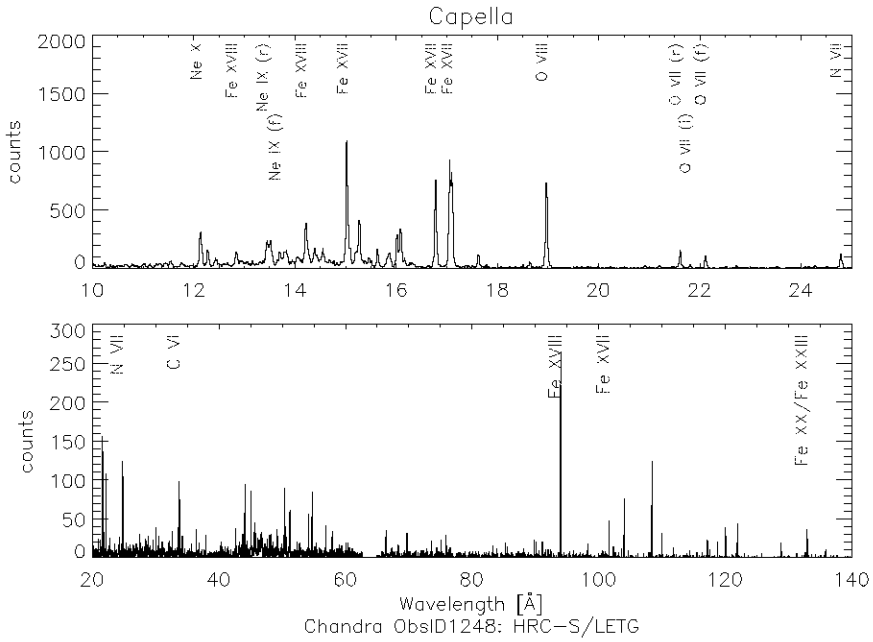

Figure 1.10: The Raw Spectrum of Capella (α Aur).
This high resolution spectrum of Capella was collected using *Chandra*'s HRC-S with the LETGS. The first panel magnifies the short-wavelength end of the spectrum. Notice the numerous emission lines that compose the spectrum; several important emission lines are labeled.

EUVE Data

Several authors (Brickhouse *et al.*, 2000; Dupree *et al.*, 1993) have reconstructed the DEM, based on data from *EUVE (the Extreme Ultraviolet Explorer)*[8]. We apply our methods using the same data set used in the previous studies to make a fair comparison with our methods. EUVE has a lower resolution and detects higher wavelength photons (i.e., lower energy photons) than Chandra. Line-spread function of EUVE data is about eight times wider than the *Chandra*'s HRC-S data. EUVE data's wavelength bins are more coarsely discretized that Chandra data; EUVE data's energy bin size is $0.0674\mathring{A}$, whereas Chandra's is $0.0125\mathring{A}$.

We select 4,704 emission lines from ATOMDB by taking the subset ranging from 80 to $140\mathring{A}$; this corresponds to 889 bins of equal width of $0.0674\mathring{A}$. Among the 4,704 emission lines, there are 3,270 Fe-emission lines. There are, however, very few other emission lines (the next frequent lines are 390 Mg-lines) and photon counts originating from the non-Fe elements are too few to be informative. We decided to group the abundances based on observed similarity of behavior in other stars. We group O,C,N, and S to O-group, Ne and Ar to Ne-group, Mg,Al, and Si to Si-group, and Fe,Ca, and Ni to Fe-group. Recall that we always fix the abundances of H and He to 1. (There is no single way to group the elements. For example, one can group all of elements with low First Ionization Potential (FIP) into one group (Al, Ca, Ni, Mg, Si, Fe = Fe) and elements with high FIP into another (S, C, N, O, Ar, Ne = Ne); we implemented the code in a way that the grouping can be user-defined.)

We computed the MAP estimate of the DEM using the EM algorithm, starting from a flat DEM and run until convergence, as measured by the increase in the posterior density evaluated at two consecutive iterates relative to the value evaluated at the first of the two iterates; convergence was called when this quantity was less

[8]EUVE was Launched in June, 1992, conducted the first extreme ultraviolet (70-760 Angstroms) survey of the sky.

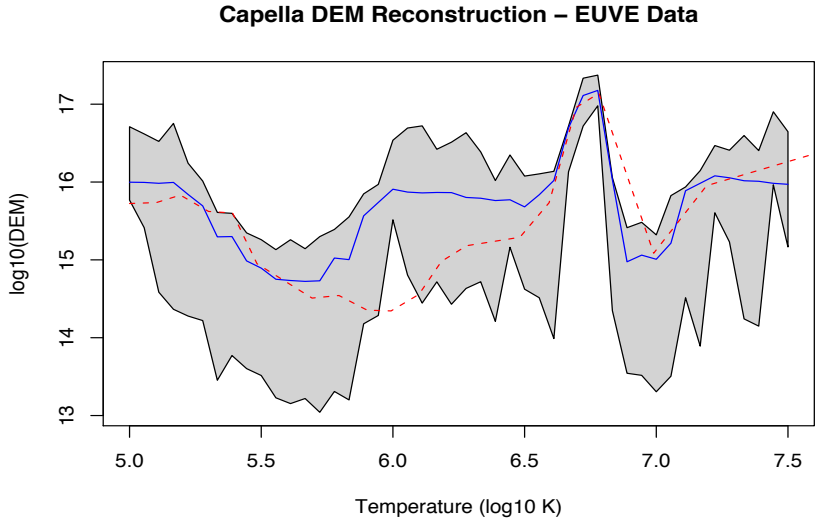

Figure 1.11: The Fitted DEM of Capella Using EUVE Data Set.
The plot draws DEM in log scale. The solid line is the posterior mode of the DEM via and the shaded area represents component-wise 95% posterior intervals for the DEM via MCMC sampling under a beta prior distribution with $\alpha_r = 2^{R-r-1}$, $r = 0, \ldots, R-1$. The dotted line is another DEM reconstruction by Dupree *et al.* (1993) with the same data set. Dupree's result is shifted along y-axis to match to the range of our result, because we only fit the relative values of the DEM. Overall shape of our result agrees with Dupree's result. Narrow confidence intervals around $\log_{10} T = 6.8$ provides us very strong information about Capella's temperature.

Element	Mean	95% Interval
O-group (O,C,N,S)	0.173	(0.005, 0.645)
Ne-group (Ne,Ar)	2.500	(1.350, 3.824)
Si-group (Mg,Al,Si)	4.772	(2.031, 8.128)
Fe-group (Fe,Ca,Ni)	5.863	(3.153, 8.398)

Table 1.3: Group Abundances of Capella Using the EUVE Data Set.
Posterior Mean and the 95% posterior interval for grouped elemental abundances.

than 10^{-6}. We used a multi-scale prior distribution with conjugate $\mathbf{Beta}(\alpha_r, \alpha_r)$ prior distribution on the split probabilities at r-th level of resolution, where $\alpha_r = 2^{R-r-1},\ r = 0, \ldots, R-1$. We used a flat prior for the background normalizing constants and the elemental abundances.

We fit the model via MCMC to compute the posterior mean of $\boldsymbol{\mu}$ and component-wise 95% posterior intervals using the same prior distributions and starting values. The result appears in Figure 1.11. Narrow confidence intervals around $\log T = 6.8$ provides us very strong information about Capella's temperature.

The 95% confidence intervals for the grouped elemental abundances are summarized in table 1.3. The mean estimate for the elemental abundance for Fe-group is 5.8 and this implies that Capella relative iron abundance is 5.8 times larger than the one of the Sun.

Chandra Data

The high resolution spectrum of Capella (see Figure 1.10) was collected using *Chandra*'s HRC-S with the LETGS diffraction grating. We select the energy range from 3 to 30$\mathring{A}$, corresponding to 2,160 energy bins of equal width of 0.0125$\mathring{A}$ and it contains 10,589 emission lines within the range. Like in the EUVE example, we computed the MAP estimate of the DEM using the EM algorithm, starting from a flat DEM and run until convergence and we fit the model via MCMC to compute the posterior mean and component-wise 95% intervals for model parameters. Starting from a flat DEM for we ran multiple Markov chains for 15,000 iterations and discarded the first 5,000 draws, for a burn-in period. We also used wavelength correction for the emissivity matrix for both EM and MCMC methods; We will discuss the atomic data error correction in the next section.

Figure 1.12 compares the DEM reconstruction results between a low smoothing prior and a high smoothing prior. In the first plot of Figure 1.12, the posterior

intervals at several temperature points do not include the posterior mode values. This is not due to a mistake in our algorithm but due to the slow convergence of the DEM parameters on the region where the ratio of missing data to observed data is too large. Consequently, the EM estimates stop before they reach the modes according to the stopping criteria and our MCMC run is not long enough to produce reliable confidence intervals. This is no longer a problem when we use strong smoothing prior as shown in the second plot of Figure 1.12. The narrow confidence interval around $\log_{10} T = 6.8$ provides a strong evidence of the existence of the sharp peak at $\log_{10} T = 6.8$. Recent works (Ness *et al.* (2001), Argiroffi *et al.* (2003)) report the results obtained from the analysis of the high resolution spectra of Capella gathered with *Chandra* and showed similar results that there is a peak around $\log_{10} T = 6.8$. In addition, the result shows a strong evidence of the second mode around $\log_{10} T = 6.4$. The temperature region, $\log_{10} T < 6.0, \log_{10} T > 7.0$ where the DEM estimates behave poorly under the low smoothing, shows the consistency between the EM and MCMC results nicely under the high smoothing. Note that this region also corresponds to the temperatures with very low column sums in the emissivity matrix. Especially, the flat DEM on $\log_{10} T < 6.0$ results mainly from the smoothing prior distribution rather than from the information in the data.

We summarized estimates for the elemental abundances in Table 1.4. The table corresponds to the second plot in Figure 1.12. Like the DEM reconstruction results, the EM and MCMC outputs for abundances are consistent under the high smoothing.

1.6.3 Model Checking

Model diagnostics are an important part of any model-based statistical analysis, and especially so in the context of complex models of the sort described in our pa-

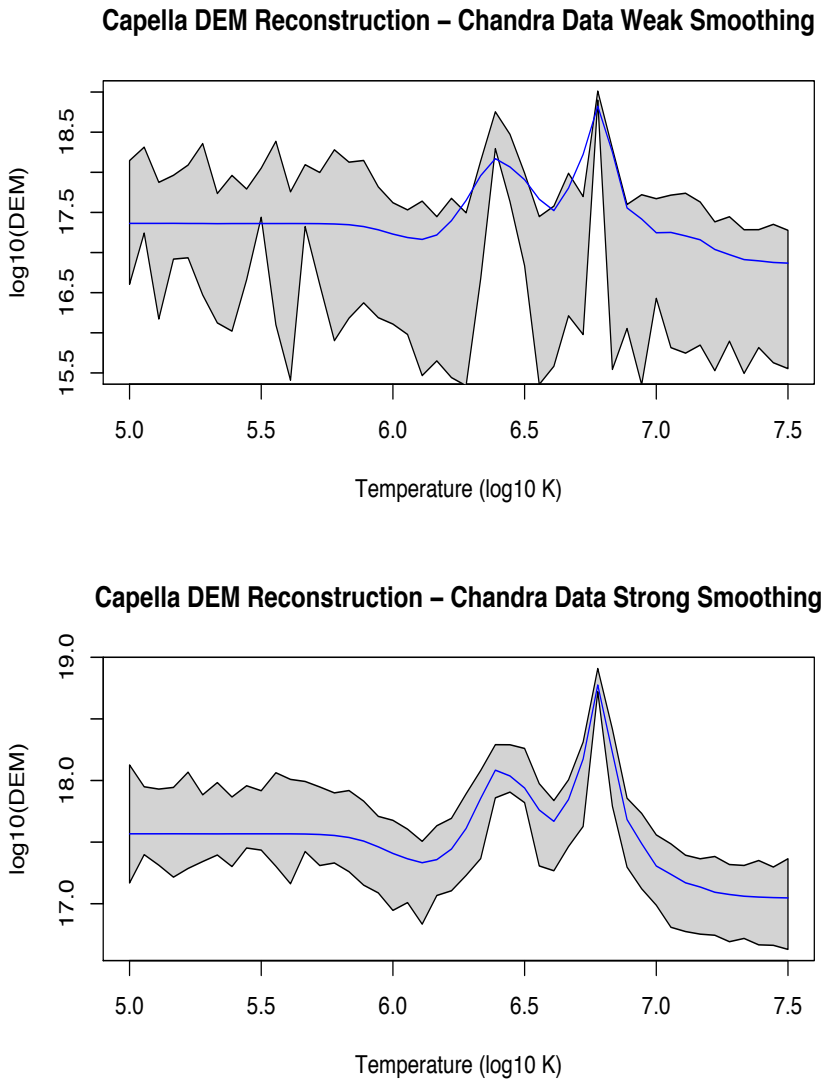

Figure 1.12: Comparisons of Error Bars for the Fitted DEM of Capella from Chandra Data between Different Smoothing Strategy.
The plot draws DEM in log scale. In the top plot, the solid line is the posterior mode of the DEM computed by EM algorithm and the shaded area represents component-wise 95% posterior intervals under low smoothing ($\alpha_r = 2 \times 2^{R-r-1}$, $r = 0, \ldots, R-1$). In the second plot, the solid line is the posterior mode of the DEM computed by EM algorithm and the shaded area represents component-wise 95% posterior intervals under high smoothing ($\alpha_r = 8 \times 2^{R-r-1}$, $r = 0, \ldots, R-1$). Wavelength correction are used for both results.

Element	Mode	Mean	95% Interval
C	0.155	0.149	(0.097, 0.205)
Si	0.266	0.255	(0.227, 0.286)
N	0.122	0.118	(0.110, 0.126)
S	0.300	0.293	(0.273, 0.315)
O	0.542	0.533	(0.492, 0.577)
Ar	0.235	0.251	(0.025, 0.555)
Ne	0.599	0.591	(0.540, 0.644)
Ca	0.362	0.356	(0.206, 0.517)
Mg	0.177	0.168	(0.085, 0.256)
Fe	0.303	0.295	(0.190, 0.405)
Al	0.428	0.422	(0.403, 0.442)
Ni	0.707	0.688	(0.616, 0.767)

Table 1.4: Posterior Mean, Mode and the 95% Posterior Interval for Element Abundance of Capella Using Chandra Data Set.
The wavelength correction and the strong smoothing corresponding to the second plot in Figure 1.12 are used. The table corresponds to the second plot in Figure 1.12.

per. Ideally such diagnostics should investigate both internal consistency and objective outside evaluation of the results. Outside evaluations might compare predictions under the model with data not used to fit the model as in cross-validations or when comparable data is available from other sources. In a Bayesian data analysis, internal consistency is often investigated by comparing the observed data with the posterior predictive distribution. Gelman *et al.* (1996) describe how one can quantify and assess discrepancies between the two. Such posterior predictive checks are a standard component of our methodology for the parameterized spectral analysis described in Section 1.2.3. (See van Dyk and Kang (2004) and van Dyk and Park (2004) for more details in the context of spectral analysis.) We will show how the posterior predictive distribution can be used to assess the magnitude of the inherently heteroskedastic residuals under Poisson models.

The models for DEM reconstruction described in Sections 1.2.3 and 1.3 rely more heavily on blurring matrices (the emissivity matrix and line-spread, respectively) than does the parametric spectral model. Thus, we expect our results to be more sensitive to misspecification of these matrices. To explore this, we generated sev-

eral replicate data sets from the posterior predictive distribution under the DEM reconstruction model. Five replicate data sets are compared with the observed Capella data in Figure 1.13. The basic structures and line locations of the the five replicate data sets appear to be very similar to those of the observed data. Thus, in general terms our model seems appropriate for the data.

A higher resolution diagnostic can be constructed by looking at a residual plot. Figure 1.15 plots the difference between the observed count in each channel and the expected count under the model evaluated at the posterior mean of the model parameters. The heteroskedastic nature of the residuals is evident. To assess the magnitude of the residuals, we sampled 1000 replicate data sets from the posterior predictive distribution. We used the replicates to construct 95% Monte Carlo prediction intervals for each of the channel count. The vertical range of the shaded area corresponding to each channel in Figure 1.15 represents the prediction interval for that channel. Ideally, we would expect about 5% of the observed counts to fall outside the shaded area. Unfortunately, the plot indicates that the residuals tend to be more dispersed than we would expect under the model: about 15% of the observed counts fall outside the shaded area.

The most likely explanation for this lack of fit is that the precision of the emissivity matrix is inadequate. Indeed, it is well known that this matrix is recorded with error. In an attempt to quantify the extent of the error, error bars are computed on the elements of the matrix. (See the Atomic Database (ATOMDB).) Unfortunately, it is also known that the errors in the matrix are highly correlated and these correlations are not readily available. Nonetheless, we have made some progress in accounting for imprecisions in the emissivity matrix by fitting the location (in wavelength) of some of the stronger spectral lines. (See Section 1.5.) The emissivity matrix provides a strong prior distribution for these locations and the data provides enough information to tweak the locations enough to improve the fit. This innovation to the model was inspired by residual plots of the sort illustrated in Figure 1.14 and

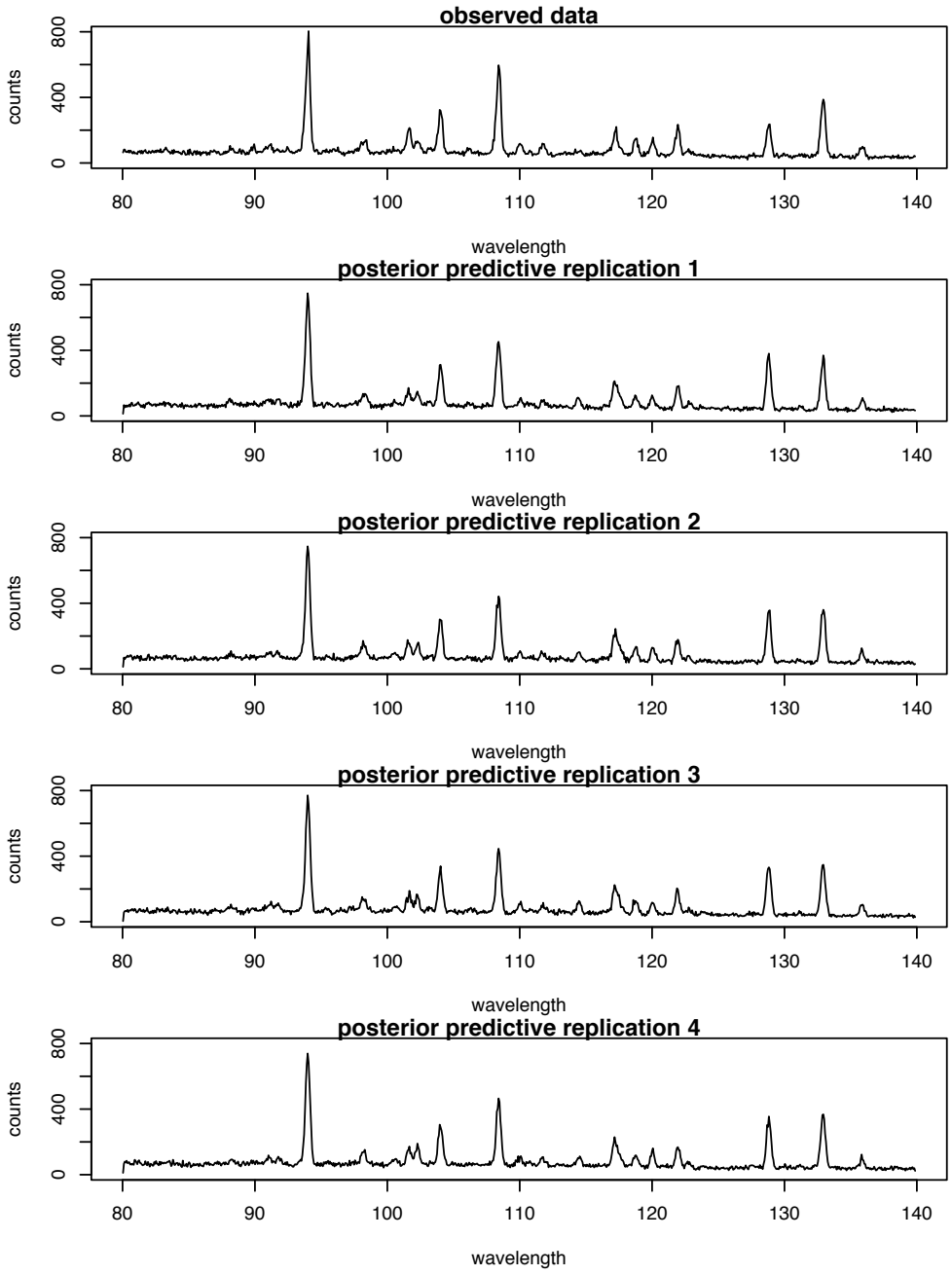

Figure 1.13: Comparing the Posterior Predictive Distribution with the Observed Capella Data.
The first panel shows the observed X-ray photon count data for Capella. The remaining five panels illustrate replicate data sets sampled from the posterior predictive distribution under a DEM reconstruction model. The basic structures and line locations of the the five replicate data sets appear to be very similar to those of the observed data.

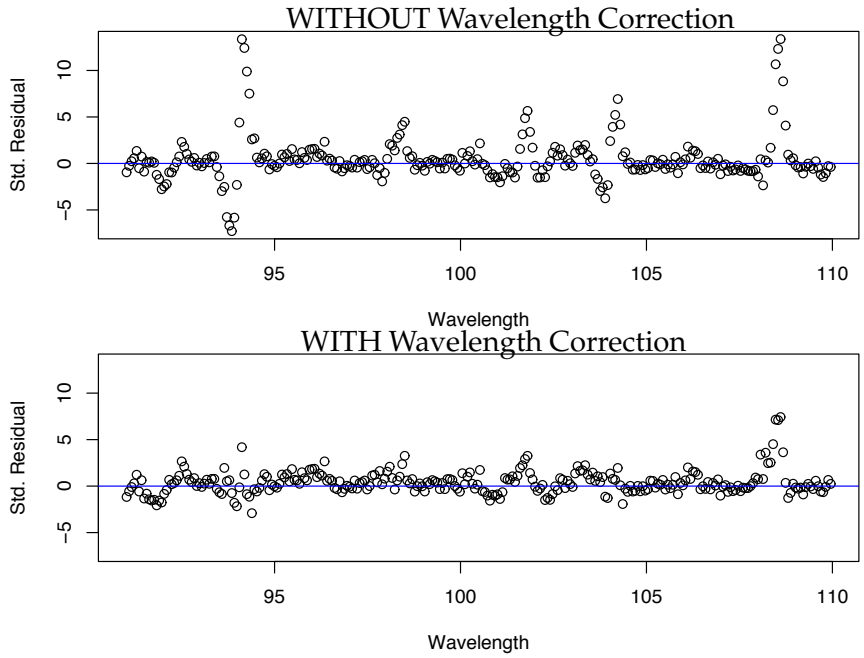

Figure 1.14: Residual Plots With and Without ATOMDB Error Correction for the Reconstructed DEM of Capella.
We plot the difference between the observed channel counts and the expected count under the model evaluated at the MAP estimates from EM output with and without correcting the wavelength error at ATOMDB. y-axis values are standardized error with Poisson variance: $\frac{Y_i^{\text{obs}} - \xi_i(\hat{\theta})}{\sqrt{\xi_i(\hat{\theta})}}$, where $\hat{\theta}$ is the MAP estimates. Due to over-dispersion and errors in the emissivity matrices, much more than 5% of the residuals are greater than +2 or less than -2. Note that the range of the y-axis is the same for both plots for comparisons. Big negative residuals and big positive residuals immediately after that around 93$\mathring{A}$ area suggest that there is a mismatch between the measured wavelength grid and the expected locations of the lines.

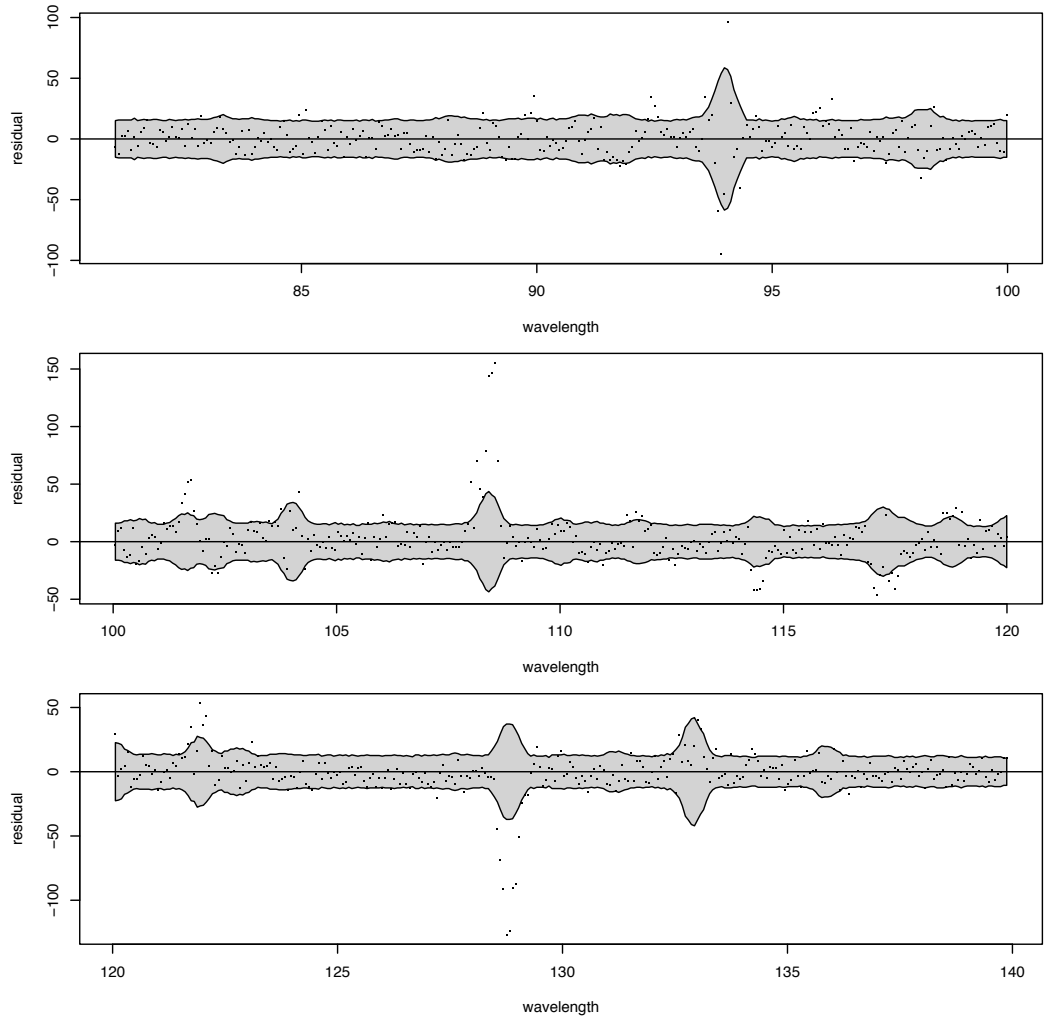

Figure 1.15: Residual Plot for the Reconstructed DEM of Capella.
We plot the difference between the observed channel counts and the expected count under the model evaluated at the posterior mean of the model parameter as a function of channel wavelength. These residuals are compared with the posterior predictive variability of the channel counts. The vertical range of the shaded area corresponds to 95% Monte Carlo prediction intervals for each of the channel counts. Ideally, we would expect about 5% of the observed counts to fall outside the shaded area. That about 15% of the counts fall outside the shaded area indicates that the model does not fully account for the observed variability in the data.

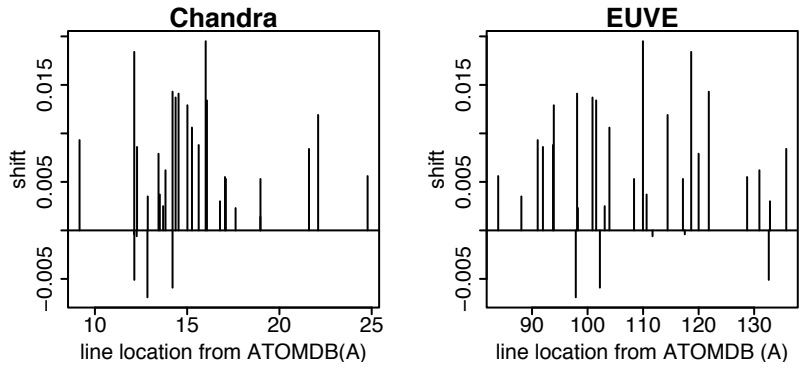

Figure 1.16: Emission Line Position Shift Due to Atomic Error.
This plots the differences between the fitted line locations and theoretical line locations ((fitted-theoretical) vs. (theoretical)) for 30 strong emission lines in the data range of interest from Chandra data and EUVE data.

1.15.

Figure 1.14 plots the difference between the observed channel counts and the expected count under the model evaluated at the MAP estimates of the model parameters from EM output with or without correcting the wavelength error at ATOMDB. It is evident that the fit improves with the wavelength error correction and suggests that more precise atomic database should be built. See Figure 1.16 for the differences between the fitted line locations and theoretical line locations.

Judging from Figure 1.15 and Figure 1.14, there is still room for improvement in our models ability to account for the variability in the observed photon counts. There are two basic strategies for accomplishing this improvement. First, we can attempt to further model uncertainties in the emissivity matrix (or elsewhere in the model) by including additional free parameters perhaps with highly informative prior distributions. We expect that adding flexibility to the model in this way also adds variability to the predictive distributions. The second strategy is to add an omnibus over-dispersion component to the model. Our colleagues in astronomy have a strong preference for the former strategy because it fosters understanding of the physical processes that give rise to the data. Although our primary goal is to

reconstruct the DEM, the emissivity matrix is of interest in and of itself. If we are able to pin down the emissivity matrix or any physical component of the model by reconstructing the DEM, this would be an important scientific contribution.

Nonetheless, in light of the over-dispersed residuals, allowing for over-dispersion in the data seems a reasonable strategy. A standard conjugate strategy for Poisson models is to mix the Poisson parameter over its conjugate prior gamma distribution. If the gamma parameters are viewed as model parameters to be fit to the data, this strategy replaces the one parameter Poisson sampling distribution with a two parameter negative binomial sampling distribution. In our case, we have multiple conditionally independent Poisson distributions that we parameterize in terms of nested binomial or multinomial distributions. The probability parameters are in turn modeled using conjugate beta distributions. We impose structure on the parameters of the conjugate distributions in order to favor smooth reconstructions.

1.7 Discussion

We developed a novel Bayesian DEM reconstruction method based on hierarchical missing data structuring and data augmentation method. We demonstrated the robustness of our method using simulations, and we showed the consistency of the DEM structure between our results and the other previous results. Some previous methods took a non-parametric approach by fitting algebraic polynomials or splines to a set of measured line fluxes, whereas our method is completely probabilistic and imposes no pre-specified structure on the DEM except for the multi-scale smoothing method by use of prior distribution. Kashyap and Drake (1998) was the first to use MCMC and derived point estimates at different temperatures, but their method is limited to measured fluxes in lines or integrated passband. However, our method uses the full detailed emissivity matrix to do a global fit to the spectrum. Consequently, it is flexible to handle atomic data errors

(errors in wavelength, ion balance, emissivities, missing lines, etc.), and easier to incorporate prior distribution such as multiscale smoothing.

Future works includes more of the correction of the atomic data errors. The main source of missing lines in current emissivity database tables is the absence of di-electronic recombination (DR) lines associated with weaker resonance lines. It is believed that such lines can be guessed from the structure of the strong resonance lines. We believe that our method can also serve as a tool for finding other atomic errors and improving the ATOMDB for future high-energy research for astrophysicists.

We would like to thank Professor Loh, Professor Gelman and Dr. Brickhouse for their thoughtful comments and discussions.

Appendix

Missing Data Distributions Conditional on Model Parameters

We describe the conditional distributions of the higher level of missing data given the lower level of missing data and the model parameters.

1. Independently separate the background counts from the observed counts:

$$Y_i^B|Y_i^{\text{obs}}, \theta \sim \textbf{Binomial}(Y_i^{\text{obs}}, \lambda_i^B/\xi(\theta)),\ i = 1, \dots I.$$

2. Restore the blurred photons:

$$\mathbf{Z}^-|\mathbf{Y}, \theta \sim \sum_{i=1}^{I} \textbf{Multinomial}\left(Y_i, \frac{(M_{1i}d_i\lambda_i, \dots, M_{Ji}d_J\lambda_J)'}{\sum_j M_{ji}d_j\lambda_j}\right).$$

3. Independently restore the absorbed counts due to the effective area:

$$Z_j|Z_j^-, \theta \sim Z_j^- + \textbf{Poisson}((1-d_j)\lambda_j), j = 1, \dots J.$$

4. Restore $\mathbf{U}^-$ given $\mathbf{Z}, \theta$:

$$\mathbf{U}^-|\mathbf{Z}, \theta \sim \sum_{j=1}^{J} \textbf{Multinomial}\left(Z_j, \frac{(G_{j1}^\star\mu_1, \dots, G_{jT}^\star\mu_T)'}{\sum_t G_{jt}^\star\mu_t}\right).$$

5. Independently restore the censored temperature counts:

$$U_t|U_t^-, \theta \sim U_t^- + \textbf{Poisson}((1-g_{+,t}^\star)\mu_t), t = 1, \dots, T.$$

6. Restore the element counts from the augmented energy counts:

$$\mathbf{V}|\mathbf{Z}, \theta \sim \sum_{j=1}^{J} \textbf{Multinomial}\left(Z_j, \frac{\mathbf{E}_{j\bullet}^\star}{\sum \mathbf{E}_{j\bullet}^\star}\right),$$

where $\mathbf{E}_{j\bullet}^\star$ is the j-th column of $\mathbf{E}^\star$.

References

Argiroffi, C., Maggio, A., and Peres, G. (2003). On coronal structures and their variability in active stars: The case of capella observed with chandra/letgs. *Astronomy and Astrophysics* **404**, 1033–1049.

Brickhouse, N. S., Dupree, A. K., Edgar, R. J., Liedahl, D. A., Drake, S. A., White, N. E., and Singh, K. P. (2000). Coronal structure and abundances of capella from simultaneous euve and asca spectroscopy. *The Astrophysical Journal* **530**, 387–402.

Brosius, J., Davila, J., and Thomas, R. (1996). Measuring active and quiet-sun coronal plasma properties with extreme-uv spectra from serts. *The Astrophysical Journal Supplement Series* **106**, 143–164.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **39**, 1–37.

Dupree, A. K., Brickhouse, N. S., Doschek, G. A., Green, J. C., and Raymond, J. C. (1993). The extreme ultraviolet spectrum of alpha aurigae (capella). *Astrophysical Journal Letters* **418**, 115–126.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness (with discussion). *Statistica Sinica* **6**, 733–807.

Kashyap, V. and Drake, J. J. (1998). Markov-chain monte carlo reconstruction of emission measure distributions: Application to solar extreme-ultraviolet spectra. *The Astrophysical journal* **503**, 450–466.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27–40.

Loredo, T. J. (1992). Promise of Bayesian inference for astrophysics. In *Statistical Challenges in Modern Astronomy* (Editors: E. Feigelson and G. Babu), 275–306. Springer-Verlag, New York.

Meng, X.-L. and van Dyk, D. A. (1997). The EM algorithm – an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **59**, 511–567.

Ness, J.-U., Mewe, R., Schmitt, J. H. M. M., Raassen, A. J. J., Porquet, D., Kaastra, J. S., van der Meer, R. L. J., Burwitz, V., and Predehl, P. (2001). Helium-like triplet density diagnostics. applications to chandra-letgs x-ray observations of capella and procyon. *Astronomy and Astrophysics* **367**, 282–296.

Nowak, R. D. and Kolaczyk, E. D. (2000). A bayesian multiscale framework for poisson inverse problems. *IEEE Transactions on Information Theory* **46**, 1811–1825.

Smith, R. K., Brickhouse, N. S., Liedahl, D. A., and Raymond, J. C. (2001). Spectroscopic challenges of photoionized plasmas. In *ASP Conference Series,* (Editors: Gary Ferland and Daniel Wolf Savin), vol. 247, 159. San Francisco: Astronomical Society of the Pacific.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528–550.

van Dyk, D. and Park, T. (2004). Efficient EM-type algorithms for fitting spectral lines in high-energy astrophysics. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (Editors: A. Gelman and X.-L. Meng), 285–296. Wiley & Sons, New York.

van Dyk, D. A. (2003). Hierarchical models, data augmentation, and mcmc. In *Statistical Challenges in Modern Astronomy III* (Editors: G. J. Babu and E. D. Feigelson), 41–56. Springer, New York.

van Dyk, D. A., Connors, A., Kashyap, V., and Siemiginowska, A. (2001). Analysis of energy spectra with low photon counts via Bayesian posterior simulation. *The Astrophysical Journal* **548**, 224–243.

van Dyk, D. A. and Kang, H. (2004). Highly structured models for spectral analysis in high energy astrophysics. *Statistical Science* **19**, 275–293.

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *The Journal of Computational and Graphical Statistics* **10**, 1–111.

Incorporating Genotyping Uncertainty in Haplotype Inference for Single-Nucleotide Polymorphisms

Abstract

The accuracy of the genotypic information generated by high-throughput genotyping technologies is crucial in haplotype analysis and linkage-disequilibrium mapping for complex diseases. To date, most automated programs lack quality measures for the allele calls; therefore, human interventions, which are both labor intensive and error prone, have to be performed. Here, we propose a novel genotype clustering algorithm based on a bivariate t-mixture model, which assigns a set of probabilities for each data point belonging to the candidate genotype clusters. Furthermore, we describe an expectation-maximization (EM) algorithm for haplotype phasing, which can use probabilistic multi-locus genotype matrices as inputs. Combining these two model-based algorithms, we can perform haplotype inference directly on raw readouts from a genotyping machine, such as the TaqMan assay. By using both simulated and real data sets, we demonstrate the advantages of our probabilistic approach over the current genotype scoring methods, in terms of both the accuracy of haplotype inference and the statistical power of haplotype-based association analysis.

2.0 Preface

This paper is a joint work with Zhaohui S. Qin, Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Tianhua Niu, Division of Preventive Medicine, Department of Medicine, Brigham and Women s Hospital, Harvard Medical School, Boston, and Jun S. Liu, Department of Statistics, Harvard University, Cambridge, MA. The related work to this paper is published in Kang *et al.* (2004).

2.1 Introduction

A Single Nucleotide Polymorphism or SNP is a DNA sequence variation, occurring when a single nucleotide (adenine (A), thymine (T), cytosine (C) or guanine (G)) in the genome is altered. Since, SNPs make up 90% of all human genetic variations, they have been widely used in genetics for disease association studies and linkage-disequilibrium (LD) mapping. Haplotype is a set of DNA polymorphism markers physically located on a single chromosome. Haplotype analysis provides greater statistical power than a single marker analysis, therefore, haplotype reconstruction based on SNP genotype data has become a very important task. However, direct laboratory haplotyping assays are expensive and low-throughput (Michalatos-Beloin *et al.*, 1996), and it is necessary to develop a high-throughput automatized genotyping and haplotyping method.

Current high-throughput genotyping technologies such as the 5 nuclease assay (TaqMan), oligonucleotide ligation assay (OLA) and Sequenom's matrix assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry assay give genotype information on each marker for each individual. However, these methods are prone to errors due to the experimental artifact and the misjudgment on the genotype scoring when genotype clusters are not sufficiently separated. In

that case, genotype scoring is usually performed manually or by inferior clustering methods such as the K-means algorithm (Hartigan and Wong, 1979). In section 2.2, we propose a powerful and flexible clustering method using the mixture of bivariate t-distributions. This algorithm avoids the error-prone deterministic calling, and instead, it calculates the probability for a marker to be a certain genotype, which, in turn is used to generate probabilistic multi-locus genotype matrices. We also compare the new clustering algorithm to the conventional K-means algorithm and

An array of *in silico* haplotype inference algorithms given genotype information have been developed and improved over the past decade in terms of the accuracy and speed of the algorithms. (See Clark (1990), Excoffier and Slatkin (1995), Hawley and Kidd (1995), Long *et al.* (1995), Stephens *et al.* (2001), Niu *et al.* (2002), Lin *et al.* (2002), Qin *et al.* (2002).) Stephens and Donnelly (2003) is a good review article in the literature. Several recent studies have demonstrated that even the slightest amount of genotyping error can lead to serious consequences with regard to haplotype reconstruction and frequency estimation (Kirk and Cardon, 2002). In section 2.3, we present a new haplotype inference method for unrelated individuals which takes the probabilistic multilocus genotype matrices computed from the clustering methods introduced in section 2.2 as inputs for accurate estimation.

In section 2.5, we illustrate the advantages of the new algorithm by various simulation studies and apply the new method to real-data example.

2.2 Genotype Scoring

For fluorescence-based genotyping assays such as TaqMan and OLA, the reactions are assessed by a fluorescent reader. The two different alleles are labeled with two different dyes. For each dye used, the reader produces a fluorescent intensity (FI)

value. Each pair of FI readouts, denoted as $(x_i, y_i),\ i = 1, ..., n$, forms a point on the scatter plot (Figure 2.1) indicating the quantitative intensities of the two SNP alleles for a given individual.

As depicted in Figure 2.2.A, a typical SNP scatter plot normally has four distinct clusters (or "groups"), representing the "no fluorescence signal" (NFS) cluster, "wild-type allele homozygote" (AA) cluster, the heterozygote (Aa) cluster, and "the variant allele homozygote" (aa) cluster. The NFS cluster is always located in the lower left corner close to the origin, the AA, aa, and Aa clusters in the upper left, lower right, and upper right corners, respectively (Figure 2.2.A). Ideally, if the NFS, AA, Aa, and aa clusters have distinct boundaries, and visual inspection is sufficient to make the genotype call (e.g. Figure 2.1.A). However, due to various artifacts, segregation can be poor with points lying between groups (e.g. Figure 2.1.B), which often results in ambiguous genotype calls.

When the genotype clusters do not segregate sufficiently from each other (Figure 2.2.B, medium- and high-ambiguity cases), one manually makes deterministic calls such that any data point is assigned to the *visually closest* cluster. The K-means algorithm has been widely used as an alternative to the manual clustering. We will discuss the disadvantages with K-means algorithm in section 2.2.2. To make quantitative genotype calls, we compute the probability that a data point belongs to each cluster. Probabilistic scoring is particularly attractive when genotype clusters are not well segregated.

2.2.1 A Fast-Convergent Clustering Algorithm based on the t-Mixture Model

This new clustering algorithm uses a mixture of four bivariate t-distributions to fit the observed pairs of FI readouts, where the four distributions represent clusters of heterozygotes, major allele homozygotes, minor allele homozygotes, and NFS,

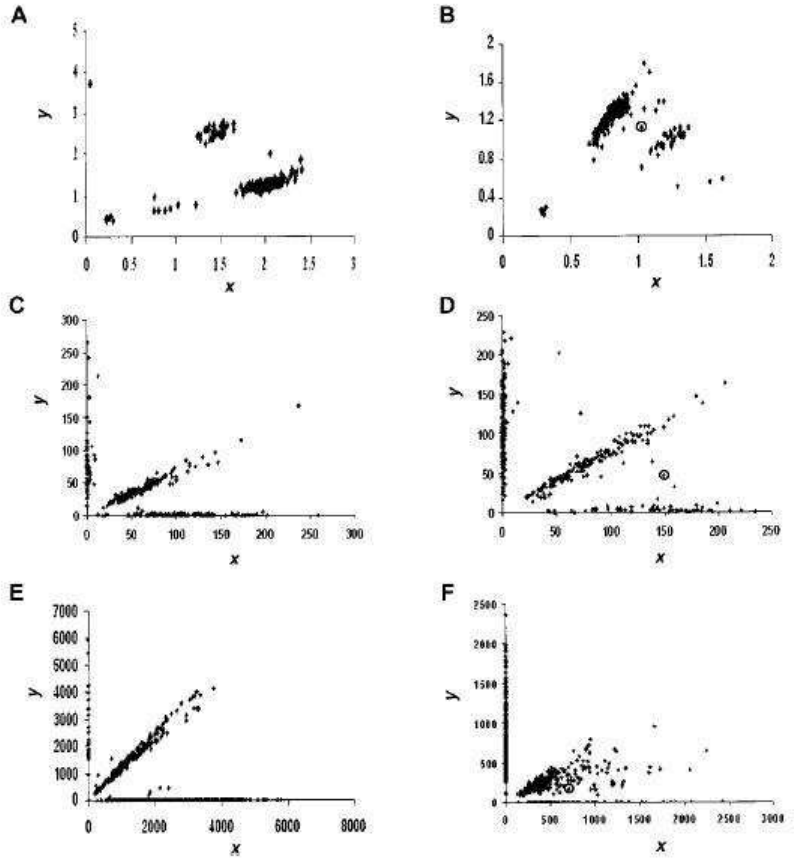

Figure 2.1: Scatterplots of FI Readouts from Genotyping Markers by Use of Various Assays.
Each point (x, y) represents the genotype of an individual, where x and y denote the FI values for the two alleles, respectively. (A), A typical good result from the TaqMan assay. Four distinct clusters are shown, corresponding to major-allele homozygotes, minor-allele homozygotes, heterozygotes, and NFS. (B), A typical but not ideal result from the TaqMan assay. It is difficult to separate all points into distinct clusters. The point in a circle is located between two groups of dense points, demonstrating the case in which a clear-cut genotype call is difficult to make. (C), A typical good result from the OLA. The three genotype clusters are in the form of three straight lines: the one close to the x-axis and the one close to the y-axis correspond to major and minor homozygotes respectively, and the center line corresponds to heterozygotes. The points near the origin indicate experimental failures, resulting in NFS. (D), A typical but not ideal result from the OLA. The points located between line patterns demonstrate the cases in which a clear-cut genotype call is difficult to make. (E), A typical good result from the MassARRAY assay. The scatterplot looks similar to the ones obtained from the OLA. (F), A typical but not ideal result from the MassARRAY assay. The points that are located between the genotype line patterns are the cases in which a clear-cut genotype call is difficult to make.

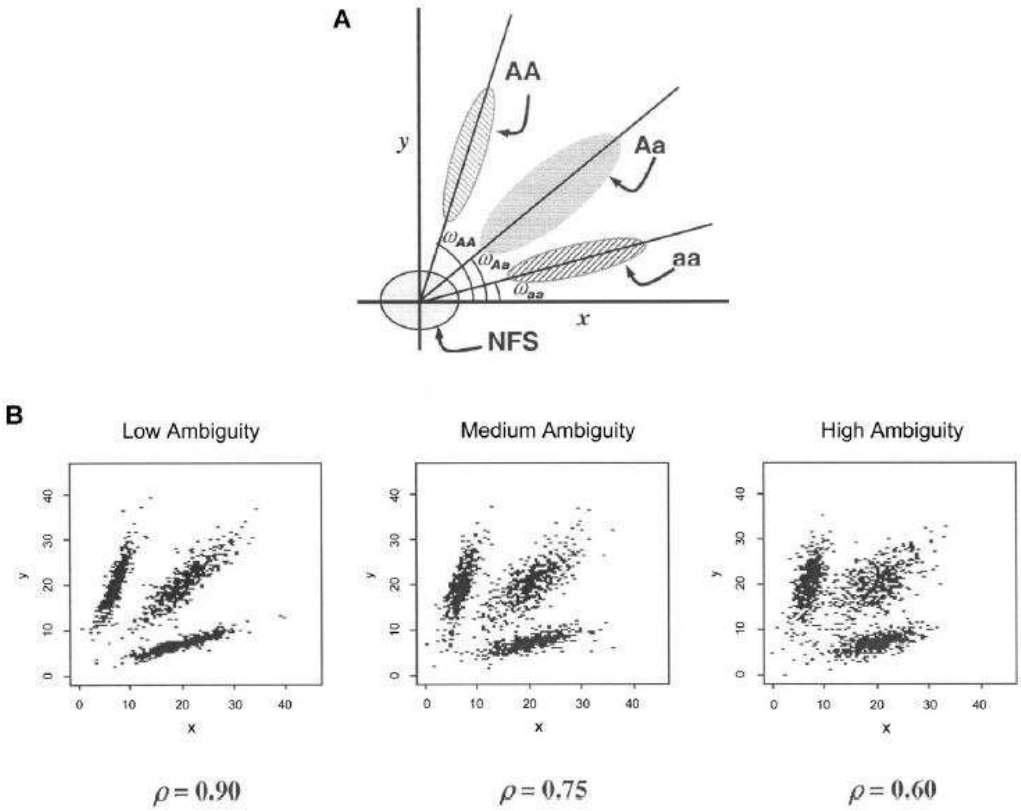

Figure 2.2: Illustration of the Genotype Clusters and their Ambiguity Levels on 2-D Fluorescent
'A' represents wild-type allele and 'a' represents variant allele. 'AA', 'Aa' and 'aa' represent wild-type allele homozygote, the heterozygote, and the variant allele homozygote, respectively. ω_{AA}, ω_{Aa}, and ω_{aa} represent the angle between the x-axis and the three clusters. Plots in the second row illustrate the the simulated FI scatterplots at low, medium, and high ambiguity levels.

respectively. We chose t-distribution instead of popular Gaussian distribution, because the t-distribution has a heavier tail than the Gaussian distribution, hence the t-mixture model is less sensitive to the outlying points and more robust than a Gaussian mixture model. Note that the Gaussian mixture model can be viewed as a t-mixture model with infinite degrees of freedom. Although t-distributions have various desired properties, they have not been broadly used in practice because of the computational difficulties in parameter estimation.

The clustering algorithm is based on typical Gibbs sampling framework, which samples parameters given missing data and missing data given the parameters and iterates. The missing data are the cluster indexes and the parameters are the location, the scale and the weight of the t-distributions. For fast convergence, we apply parameter-expanded data augmentation(PXDA) algorithm (Liu and Wu, 1999) when sampling the paramters of t-distributions given the missing cluster index.

After the convergence, we use the mean of the sampled values for estimating t-distributions to compute the likelihood values required by our probabilistic allele-calling scheme. Note that we can also use the t-mixture clustering algorithm to make deterministic calls by assigning individuals to their most probable clusters (i.e., the ones with the highest posterior probabilities).

Clustering Model

The likelihood function of the bivariate t-mixture model is

$$p(\mathbf{x}|\mu,\Sigma) = \prod_{i=1}^{n}\left(\sum_{c=1}^{C} w_c t_2(\mathbf{x}^i;\mu_c,\Sigma_c,\nu)\right),$$

where $\mathbf{x} = \{\mathbf{x}^i = (x_i, y_i)'; i = 1, ..., n\}$ is the set of observed pairs of FI values for a SNP location, C is the number of mixture components, the w_c's are the mixture weights (i.e., $0 <$ $w_c <$ for all $c = 1, ..., C$ and $\sum_{c=1}^{C} w_c = 1$), and $t_2(\mathbf{x}^i;\mu_c,\Sigma_c,\nu)$ is the probability density function of the bivariate t-distribution with location parameter

μ_c, scale parameter Σ_c, and known degrees of freedom ν. Since the choice of ν is not critical to the analysis, we set $\nu = 7$ as a default choice. In practice, lower value of degrees of freedom is especially desirable when there are many ambiguous points in the scatter plot the FI values. The number of mixture components C is fixed at 4 to represent four clusters: AA, Aa, aa, and NFS. See Figure 2.2.B.

Our algorithm iterates the following two steps and outputs Markov chain samples of the model parameters and the cluster indicator (for each individual) from the desired posterior distribution. First, given current values of the parameters, $w_c^{(t)}, \mu_c^{(t)}$, and for $\Sigma_c^{(t)}$, for $c = 1, ..., C$, we sample the unobserved mixture indicator $J^{i,(t)} = (j_1^{i,(t)}, ..., j_C^{i,(t)})$ for each $\mathbf{x}_i$ from $\mathbf{Multinomial}(1; q_1^{i,(t)}, ..., q_C^{i,(t)})$, where $j_c^{i,(t)}$ is equal to one if $\mathbf{x}^i$ is assigned to c-th cluster and zero otherwise, and

$$q_c^{i,(t)} = \frac{w_c^{(t)} t_2(\mathbf{x}^i; \mu_c^{(t)}, \Sigma_c^{(t)}, \nu)}{\sum_{c=1}^{C} w_c^{(t)} t_2(\mathbf{x}^i; \mu_c^{(t)}, \Sigma_c^{(t)}, \nu)},$$

the probability that $\mathbf{x}_i$ belongs to c-th cluster at the t-th iteration. Second, given the current mixture indicator, $J^{i,(t)} = (j_1^{i,(t)}, ..., j_C^{i,(t)})$, we sample the parameters, $w_c^{(t+1)}, \mu_c^{(t+1)}$, and for $\Sigma_c^{(t+1)}$, for $c = 1, ..., C$ from their posterior distribution.

Note that given the mixture index, model fitting is straightforward because the parameters follow a series of standard distributions. We assume the natural conjugate proper prior on the mixture weights,

$$(w_1, ..., w_C) \sim \mathbf{Dirichlet}(1, ..., 1),$$

which result in the conjugate posterior distribution

$$(w_1^{(t+1)}, ..., w_C^{(t+1)}) \sim \mathbf{Dirichlet}\left(1 + \sum_{i=1}^{n} I(j_1^{i,(t)} = 1), \ldots, 1 + \sum_{i=1}^{n} I(j_C^{i,(t)} = 1)\right).$$

For each cluster, given that we know which cluster each point belongs to from $J^{i,(t)}$, the sampling of $(\mu_c^{(t+1)}, \Sigma_c^{(t+1)})$ is equivalent to fitting a multivariate t-distribution,

which can be achieved efficiently using a parameter-expanded data augmentation (PXDA) scheme (Liu and Wu (1999), van Dyk and Meng (2001)) shown at the next section.

Parameter-Expanded Data Augmentation(PXDA) for Multivariate t-distribution

To illustrate the PXDA scheme, we let $t_p(\mu, \Sigma, \nu)$ denote the p-dimensional t-distribution with center μ, covariance matrix Σ, and known degrees of freedom ν. Note the fact that $\mathbf{x}^i|\mu, \Sigma \sim t_p(\mu, \Sigma, \nu)$ is equivalent to

$$\mathbf{x}^i|\tau_i, \mu, \Sigma \sim \mathbf{N}_p(\mu, \alpha\Sigma/\tau_i), \text{ and}$$

$$\tau_i|\mu, \Sigma \sim \frac{\alpha\chi^2_\nu}{\nu}, \ i = 1, ..., n.$$

The auxiliary scale parameter α is incorporated here in order to derive a fast-converging Gibbs sampling algorithm. To avoid an improper posterior distribution, we use the conjugate prior distribution for (μ, Σ), which can be parameterized in terms of hyperparameters $(\mu_0, \Lambda_0/\kappa_0; \nu_0, \Lambda_0)$,

$$\Sigma \sim \mathbf{Inv-Wishart}_{\nu_0}(\Lambda_0^{-1}), \text{ and}$$

$$\mu|\Sigma \sim \mathbf{N}(\mu_0, \Sigma/\kappa_0).$$

Jointly, we have a prior distribution:

$$p(\mu, \Sigma) \propto |\Sigma|^{-((\nu_0+p)/2+1)} \exp\left(-\frac{1}{2}\text{tr}(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\mu_0 - \mu)'\Sigma^{-1}(\mu_0 - \mu)\right).$$

According to Liu and Wu (1999), we used Jeffreys' prior for the auxiliary variable. Under this prior specification, we obtain the following iterative sampling scheme:

- Draw
$$\tau_i|\mathbf{x}, \mu, \Sigma, \alpha \sim \frac{\alpha\chi^2_{\nu+p}}{(\mathbf{x}^i - \mu)'\Sigma^{-1}(\mathbf{x}^i - \mu) + \nu},$$
independently for $i = 1, ..., n$.

- Draw

$$\Sigma^{-1}|\mathbf{x}, \tau, \mu, \alpha \sim$$
$$\mathbf{Wishart}_{n+\nu_0}\left[\left(\Lambda_0 + \alpha^{-1}\sum_{i=1}^{n}\tau_i(\mathbf{x}^i - \hat{\mu})(\mathbf{x}^i - \hat{\mu})' + \frac{\kappa_0\sum_{i=1}^{n}\tau_i}{\alpha\kappa_0 + \sum_{i=1}^{n}\tau_i}(\mu_0 - \hat{\mu})(\mu_0 - \hat{\mu})'\right)^{-1}\right],$$

  and $\alpha|\mathbf{x}, \tau \sim \frac{\nu\sum_{i=1}^{n}\tau_i}{\chi^2_{n\nu}}$, where $\mathbf{Wishart}_k(A)$ denotes the Wishart distribution with scale matrix A and degrees of freedom k.

- Draw $\mu|\mathbf{x}, \Sigma, \tau, \alpha \sim \mathbf{N}_p\left(\hat{\mu}, \frac{\Sigma}{\sum_{i=1}^{n}\tau_i/\alpha+\kappa_0}\right)$, where $\hat{\mu} = \frac{\sum_{i=1}^{n}\tau_i\mathbf{x}^i/\alpha+\kappa_0\mu_0}{\sum_{i=1}^{n}\tau_i/\alpha+\kappa_0}$.

Liu and Wu (1999) showed that the scheme converges to the correct posterior distribution for (μ, Σ), although the posterior distribution of α is still improper. They also proved that the PXDA converges faster than the standard data augmentation scheme and attains the optimal convergence speed when Jeffrey's prior on α is used.

Stabilizing the t-mixture Clustering Algorithm

Some difficulties in mixture modeling include the label switching problem (Stephens, 2000), the incorrect specification of the cluster numbers, and the occurrences of clusters of small sizes. To make the algorithm stable, we use our prior knowledge of the well-known structure of the FI value scatter plot. First, we use proper priors for the parameter μ_c and Σ_c. They prevent the posterior distribution from being improper even when the data set has an empty cluster. We let the prior distribution of μ_c conditional on Σ_c be $\mathbf{N}(\mu_{c0}, \Sigma_c/\kappa_0)$, where μ_{c0} can be either input by the user or defaulted at one of the four "corners" of the data scatterplot, and κ_0 can be chosen by the user (default at 1). The prior for Σ_c is taken as $\mathbf{Inv-Wishart}_{\nu_0}(\Lambda_0^{-1})$, where Λ_0^{-1} is the sample covariance matrix based on all the FI values, and $\nu_0 = p+1$, where $p = 2$ is the dimension of the data point. Second, we

impose an identifiability constraint on the parameter space of μ_c. Since the general pattern of the scatterplot of FI values contains three clusters away from the origin and one close to the origin, we impose a constraint such that $|\mu_c| > |\mu_{\text{NFS}}|$, $c = \text{AA}$, Aa, and aa, and $|\cdot|$ denotes the distance from the origin to the vector. Furthermore, for non-NFS clusters, we impose another constraint that $\omega_{AA} > \omega_{Aa} > \omega_{aa}$ (Figure 2.2), where ω_c is the angle of between the vector μ_c and the x-axis. The subscripts indicate the heterozygote cluster (Aa), the homozygote cluster near the x-axis (aa), and the homozygote cluster near the y-axis (AA), respectively.

After the Markov chain of the above posterior sampling scheme has converged, we estimate the likelihood for the i-th individual's FI values at this marker, given that it is in cluster c by $p_c^i \approx t_2(\mathbf{x}^i; \bar{\mu}_c, \bar{\Sigma}_c, \nu)$, where $\bar{\mu}_c$ and $\bar{\Sigma}_c$ are posterior means for the location and scale parameter of the cluster 'c'. The reason for using only this value instead of the cluster membership posterior probability is because of the need of computing $P(\mathbf{x}^i|Y^{i,j})$ in the new EM algorithm proposed in section 2.3.2. We also compute the posterior mean of the mixture weights, $\bar{w}_c$ to compute the cluster membership posterior probabilities for deterministic calls. We repeat this process for all the SNP markers to obtain the multilocus genotype matrix of the i-th individual:

$$\begin{array}{c} \\ \text{genotype}'0' \\ \text{genotype}'1' \\ \text{genotype}'2' \end{array} \begin{array}{c} \begin{array}{cccc} \text{SNP1} & \text{SNP2} & \cdots & \text{SNP}m \end{array} \\ \begin{pmatrix} p^i_{0,1} & p^i_{0,2} & \cdots & p^i_{0,m} \\ p^i_{1,1} & p^i_{1,2} & \cdots & p^i_{1,m} \\ p^i_{2,1} & p^i_{2,2} & \cdots & p^i_{2,m} \end{pmatrix} \end{array},$$

where $p^i_{c,k}$ is the probability that the genotype of the k-th SNP of individual i equals to c.

2.2.2 Comparing the K-means and the t-mixture Model for Genotype Scoring

We compared the accuracies of the K-means algorithm and the t-mixture model under low, medium, and high ambiguity levels. The ambiguity level is con-

trolled by changing the correlation coefficient (ρ) of the covariance matrix, such that ρ =0.9, 0.75, and 0.6 correspond to low, medium, and high ambiguity levels, respectively. To reduce the complexity of the simulation study, we focused on a three-cluster model without the NFS cluster for the FI outputs. In our simulation, bivariate Gaussian distributions were used to generate 100 FI data points with fixed location parameters of the distributions for AA, Aa and aa clusters at those estimated from a true dataset and the scale parameters depending on the ambiguity level (we also used t-distributions for simulating the FI scatter plots and the results were similar). We clustered all the 100 points using both the K-means algorithm and the t-mixture model. The K-means algorithm was implemented using the K-means function in software package R v.1.5.0. We also implemented the K-means algorithm using Splus v.5.1 and found that the results were comparable to the R implementation.

To make a fair comparison, we assumed that the number of clusters was three and we gave the same starting points for the centers of clusters for both algorithms. In the t-mixture model, we picked the cluster with the highest probability. We counted the number of erroneous calls (defined as the calls different from the true calls) in each simulation and repeated this procedure 100 times. At every ambiguity level, the t-mixture model outperformed the K-means algorithm (Table 2.1) by a large margin. One of the reasons for the poor performance of the K-means method is that it had difficulty accommodating the elongated shapes of the FI clusters because of its use of the standard Euclidean distance. In contrast, the mixture t-model can utilize the shape information by updating the covariance matrix of its each component.

Besides its poor performance, the K-means algorithm also requires correct specification of the number of clusters, which requires the human judgment of eyeballing the scatter plot to determine the proper number of clusters before running the program. In contrast, the t-mixture model is not sensitive to the input cluster number

Algorithm	Miscalls (%) for Scenario		
	Low Ambiguity	Medium Ambiguity	High Ambiguity
K-means	9.59	8.82	8.82
t-mixture	0.03	0.30	0.72

Table 2.1: Comparison of Clustering Accuracy between the K-means Algorithm and the t-mixture Model in Making Deterministic Genotype Calls.
For comparison purposes, we generated 100 data sets for each of low, medium, or high ambiguity scenarios. In each data set, the Gaussian mixture model was used in generating 100 data points forming three genotype clusters. For each algorithm, the percentage of miscalls was defined as $100\times$ (the number of miscalls) / (total genotype calls).

as long as it matches or exceeds the true number (at most 4 in this case). The use of informative priors ensures that the t-mixture model is not sensitive to empty clusters. Examples of obvious mistakes made by the K-means algorithm in the clustering are shown in Figure 2.3.

2.3 Haplotype Phasing Methods

2.3.1 Conventional EM with Deterministic Inputs (EM-I)

For deterministic inputs for multiple linked SNPs, the conventional EM algorithm has been applied successfully both to construct individual haplotype phases and to estimate population haplotype frequencies from deterministic multi-locus genotype data due to its stable convergence (Excoffier and Slatkin (1995), Hawley and Kidd (1995), Long *et al.* (1995), Niu *et al.* (2002), Qin *et al.* (2002)).

Let $Y = (Y^1, ..., Y^n)$ denote the genotypes of a sample with n individuals, let $Z = (Z^1, ..., Z^n)$ denote the unobserved haplotype configuration, where $Z^i = (z_1^i, z_2^i)$ represents the haplotype pairs for the i-th individual, and let $\Theta = (\theta_1, ..., \theta_s)$ denote the population haplotype frequencies, where s is the total number of existing haplotypes. In data augmentation framework, Z's are missing data, and Θ is pa-

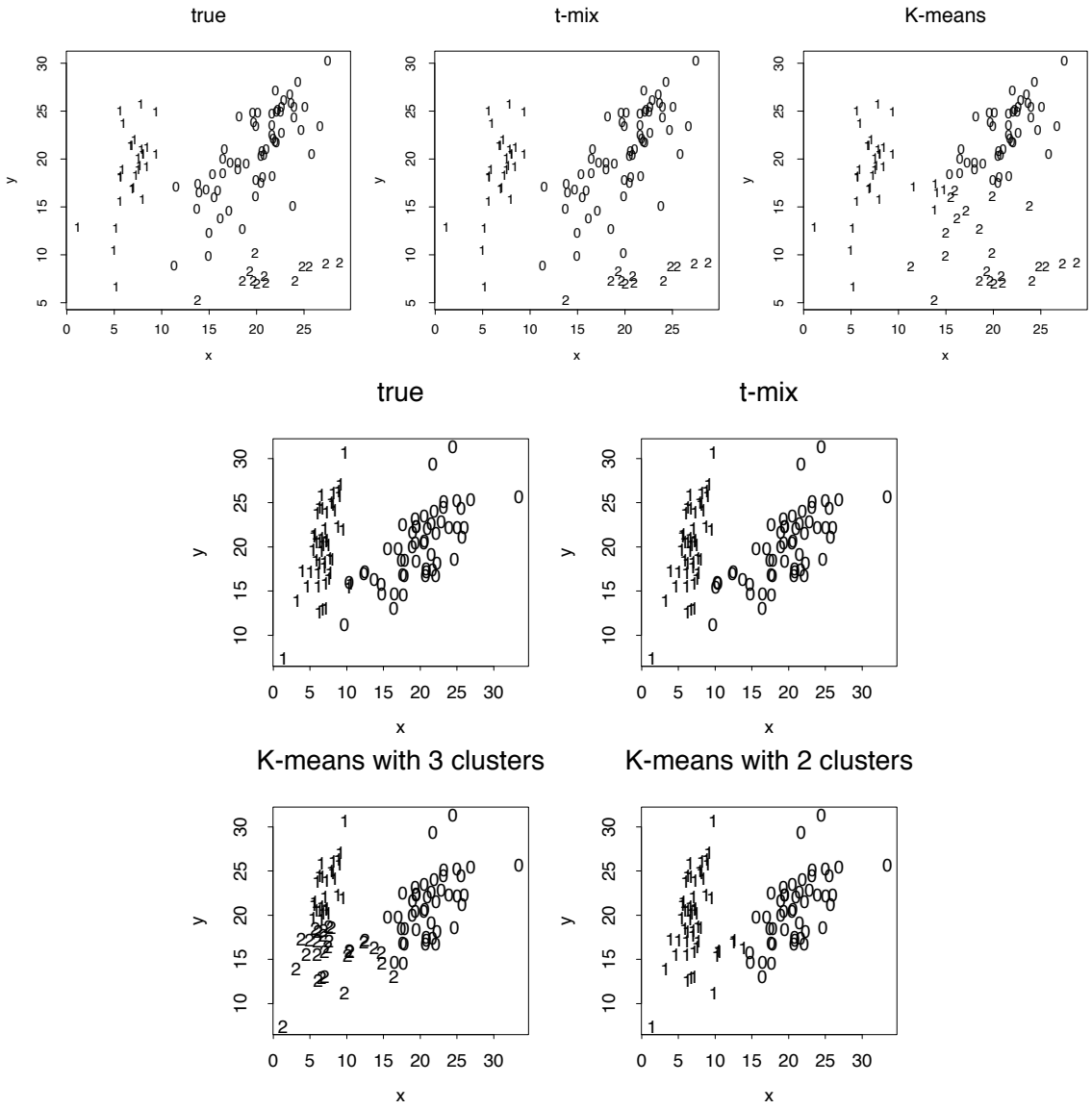

Figure 2.3: Comparisons of the K-means Algorithm and the t-mixture Algorithm. Each point (x, y) represents the genotype of an individual, where x and y denote the FI values for the two alleles, respectively. The cluster label is shown for each data point for the ground truth, as well as clustering results of the bivariate t-mixture model (t-mix), and the K-means algorithm. Three plots in the first row illustrate three-cluster example; and the other four plots in the second and third row illustrate a two-cluster example. Note that the K-means algorithm requires the user to pre-specify the number of clusters, whereas the t-mixture algorithm can determine the number of clusters automatically. In both examples, the performance of t-mixture clustering is superior to the K-means algorithm.

rameter. We use the notation $z_1^i \oplus z_2^i = Y^i$ to denote that the two haplotypes are compatible with genotype Y^i. The likelihood function can be written as:

$$P(Y|\Theta) = \prod_{i=1}^{n} P(Y^i|\Theta) = \prod_{i=1}^{n} \left(\sum_{(g,h):g\oplus h=Y^i} \theta_g \theta_h \right).$$

If we observed the missing phase configuration Z, the maximum-likelihood estimate (MLE) of Θ should satisfy $\theta_g = \frac{n_g}{2n}$, where n_g is the count of occurrence of haplotype g in a particular phase configuration Z (M-step). Then, we replace n_g with $E_\Theta(n_g)$, where $E_\Theta(\cdot)$ represents to expected value over Z under the distribution $P(Z|\Theta, Y)$ (E-step). With $\Theta^{(t)}$ denoting the frequency estimation at the t-th iteration, the EM iterates as:

$$\theta_g^{(t+1)} = \frac{E_{\Theta^{(t)}}(n_g|Y)}{2n} = \frac{1}{2n} \sum_{i=1}^{n} \frac{\theta_g^{(t)} \theta_{Y^i \backslash g}^{(t)} \{1 + I(g = Y^i \backslash g)\}}{\sum_{(g',h'):g'\oplus h'=Y^i} \theta_{g'}^{(t)} \theta_{h'}^{(t)}},$$

where $Y^i \backslash g$ denotes the complement haplotype that pairs with to make up the genotype and is an indicator function. Given the final estimate $\hat{\Theta}$, we phase the i-th individual's genotype Y^i by finding a compatible haplotype pair $(g, h) : g \oplus h = Y^i$ that maximizes $\hat{\theta}_g \hat{\theta}_h$.

2.3.2 An EM Algorithm with Probabilistic Inputs (EM-II)

For probabilistic inputs of multiple linked SNPs such as those resulting from the t-mixture algorithm, the conventional method (EM-I) can no longer be applied. Here, we introduce a new EM algorithm, which can handle such inputs. Let $p_{0,k}^i$, $p_{1,k}^i$, and $p_{2,k}^i$be the likelihood of the i-th individual's FI readouts at marker k given that its genotype at this marker is heterozygote (Aa, denoted by '0'), wild-type homozygote (AA, denoted by '1'), and variant homozygote (aa, denoted by '2'), respectively. That is, $p_{c,k}^i = P\{\mathbf{x}_k^i | y_k^i =' c'\} = t_2(\mathbf{x}_k^i; \mu_{c,k}, \Sigma_{c,k}, \nu)$, where $\mathbf{x}_k^i$ represents the FI values of the k-th SNP of the ith individual, y_k^i represents the genotype at the k-th SNP, and $t_2(\cdot; \mu_{c,k}, \Sigma_{c,k}, \nu)$ is the density function of the bivariate t-distribution,

with mean $\mu_{c,k}$, scale $\Sigma_{c,k}$, and known degrees of freedom ν, for cluster 'c' (c=0,1,2) at the kth SNP. Note the distinction between the likelihood of the FI values given a cluster, $p^i_{c,k} = P\{\mathbf{x}^i_k | y^i_k =' c'\}$, and the posterior cluster (membership) probability, $P\{y^i_k =' c' | \mathbf{x}^i_k\} = w_c p^i_{c,k} / \sum_d w_d p^i_{d,k}$, where w_c is the mixture weight for cluster 'c'. These likelihood vectors for both markers form a $3 \times m$ matrix:

$$\begin{array}{c} \\ \text{genotype}'0' \\ \text{genotype}'1' \\ \text{genotype}'2' \end{array} \begin{array}{c} \begin{array}{cccc} \text{SNP1} & \text{SNP2} & \cdots & \text{SNP}m \end{array} \\ \begin{pmatrix} p^i_{0,1} & p^i_{0,2} & \cdots & p^i_{0,m} \\ p^i_{1,1} & p^i_{1,2} & \cdots & p^i_{1,m} \\ p^i_{2,1} & p^i_{2,2} & \cdots & p^i_{2,m} \end{pmatrix} \end{array},$$

where m is the total number of markers in consideration. From this matrix, we can obtain the likelihood of any m-SNP genotype of this individual by multiplying the corresponding single-marker genotype likelihoods under the assumption that the SNPs' FI readouts are mutually independent. For example, the likelihood for a 3-SNP genotype $Y^i = (1, 0, 2)$ is:

$$\begin{aligned} P\{\mathbf{x}^i | Y^i = (1,0,2)\} &= P\{\mathbf{x}^i_1 | y^i_1 =' 1'\} \times P\{\mathbf{x}^i_2 | y^i_2 =' 0'\} \times P\{\mathbf{x}^i_3 | y^i_3 =' 2'\} \\ &= p^i_{1,1} p^i_{0,2} p^i_{2,3} \\ &\approx t_2(\mathbf{x}^i_1; \bar{\mu}_{1,1}, \bar{\Sigma}_{1,1}, \nu) t_2(\mathbf{x}^i_2; \bar{\mu}_{0,2}, \bar{\Sigma}_{0,2}, \nu) t_2(\mathbf{x}^i_3; \bar{\mu}_{2,3}, \bar{\Sigma}_{2,3}, \nu), \end{aligned}$$

where $\mathbf{x}^i$ represents the FI values of the i-th individual, and $\bar{\mu}_{c,k}$ and $\bar{\Sigma}_{c,k}$ are the estimated location and scale parameters of cluster 'c' at the k-th SNP. Note that this equation is an approximation because the estimated (as opposed to the true but unknown) values of the location and scale parameters are used. We order all m-SNP genotypes of the i-th individual as $Y^{i,1}, \ldots, Y^{i,l_i}$, with their associated likelihoods $\pi^{i,1}, \ldots, \pi^{i,l_i}$, and generate the list of the possible genotype for individual i, $(Y, \Pi)^i = \begin{pmatrix} Y^{i,1}, & \pi^{i,1} \\ Y^{i,2}, & \pi^{i,2} \\ \vdots & \vdots \\ Y^{i,l_i}, & \pi^{i,l_i} \end{pmatrix}$, where l_i is the number of possible genotypes for the i-th individual, i.e., those with $\pi^{i,j} > 0$. Although there are a total of 3^m possible genotypes for m closely linked bi-allelic SNP markers, we usually only need to list a small number of the genotypes with non-zero likelihood values.

As in typical data augmentation algorithms, we treat $(Y, \Pi)^i$ as observed data, phase configuration Z as missing data, and the population haplotype frequencies Θ as parameter. The likelihood function for haplotype frequencies can be computed as:

$$\begin{aligned} P[\mathbf{x}|\Theta] &= \prod_{i=1}^{n} P(\mathbf{x}^i|\Theta) \\ &= \prod_{i=1}^{n} \left[\sum_{j=1}^{l_i} \left(P(\mathbf{x}^i|Y^{i,j}) P(Y^{i,j}|\Theta) \right) \right] \\ &= \prod_{i=1}^{n} \left[\sum_{j=1}^{l_i} \left(\left\{ \prod_{k=1}^{m} P(\mathbf{x}_k^i|y_k^{i,j}) \right\} P(Y^{i,j}|\Theta) \right) \right] \\ &\approx \prod_{i=1}^{n} \left[\sum_{j=1}^{l_i} \left(\pi^{i,j} \sum_{(g,h):g\oplus h=Y^{i,j}} \theta_g \theta_h \right) \right]. \end{aligned}$$

From this expression, we are able to obtain the following EM iteration for Θ:

$$\theta_g^{(t+1)} = \frac{E_{\Theta^{(t)}}(n_g|Y,\Pi)}{2n} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{l_i} \frac{\pi^{i,j} \theta_g^{(t)} \theta_{Y^{i,j}\backslash g}^{(t)} \{1 + I(g = Y^{i,j}\backslash g)\}}{\sum_{j=1}^{l_i} \pi^{i,j} \sum_{(g',h'):g'\oplus h'=Y^{i,j}} \theta_{g'}^{(t)} \theta_{h'}^{(t)}},$$

where $Y^{i,j}\backslash g$ denotes the complement haplotype that pairs with g to make up the genotype $Y^{i,j}$. Note that the EM-I algorithm is a special case of the EM-II with $l_i = 1$. The i-th individual's genotype is phased given the final estimate $\hat{\Theta}$ by finding a compatible haplotype pair (g, h) that maximizes $\pi^{i,j} \hat{\theta}_g \hat{\theta}_h$.

2.4 Three Phasing Strategies Based on Raw FI Values

Three phasing strategies (denoted as "SCHEME 1", "SCHEME 2", and "SCHEME 3") have been used in our study (illustrated in Figure 2.4):

SCHEME 1 : clustering step uses the t-mixture model; phasing step uses EM-I algorithm;

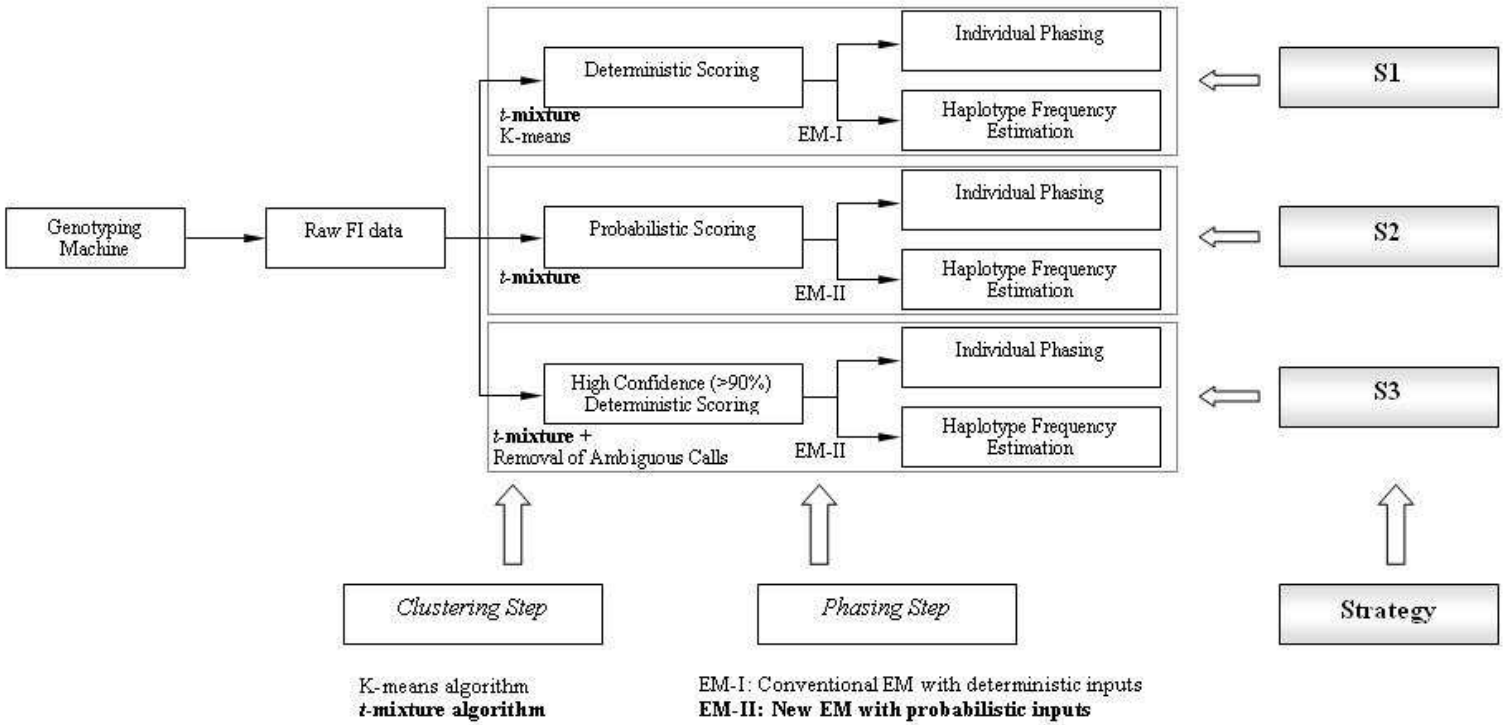

Figure 2.4: Schematic Diagram for Strategies SCHEME 1, SCHEME 2, and SCHEME 3.
Each strategy consists of two steps: a clustering step and a phasing step. For each strategy, the raw FI scatter data were used and both individual phasing and haplotype frequency estimation were achieved. SCHEME 3, mimics the human "best guess" strategy. SCHEME 1, and SCHEME 3, output deterministic calls, and S2 outputs probabilistic genotype calls. The new algorithms introduced in this paper were in bold face.

SCHEME 2 : clustering step uses the t-mixture model; phasing step uses EM-II algorithm;

SCHEME 3 : clustering step uses the t-mixture model with a removal of ambiguous points; phasing step uses EM-I algorithm.

SCHEME 1 uses the t-mixture model in the clustering step to make deterministic calls (assigning each individual to its most probable cluster) and uses the EM-I algorithm in the phasing step. For example, for a data point with cluster probabilities 0.51, 0.48, and 0.01 of belonging to the AA, Aa, and aa clusters, respectively, SCHEME 1 will still deterministically assign it to the AA cluster. Although the K-means algorithm can also be applied in the clustering step, we observed that the results obtained by the K-means algorithm were much worse than those based on the t-mixture model in our simulation comparisons (see Table 2.1). We thus drop the K-means algorithm from subsequent analyses. SCHEME 2 uses the t-mixture model in the clustering step in making probabilistic calls and uses EM-II in the following phasing step. SCHEME 3 is essentially the same as SCHEME 1, except that it attempts to simulate the human "best guess" strategy commonly practiced by laboratory technicians: when a data point cannot be assigned with a consensus call by two independent readers, it will be removed. Here, we assume that the independent human readers will not be able to make consensus calls for all ambiguous data points (i.e., a SNP with all the cluster probability values < 0.9 can not be assigned to any of the AA, Aa, or aa genotype clusters.). Thus, all such ambiguous data points of the raw FI data will be removed at this step and not used in the phasing step. For example, for a data point with cluster probabilities 0.51, 0.48, and 0.01 of belonging to the AA, Aa, and aa clusters, respectively, SCHEME 3 will toss it away. However, for a data point with cluster probabilities 0.045, 0.91, and 0.045 of belonging to the AA, Aa, and aa clusters, respectively, SCHEME 3 will assign it to the Aa cluster.

2.5 Results

2.5.1 Simulation Studies

Accuracy Comparison of the Three Phasing Strategies

We compare the accuracies of the haplotype phase and frequency reconstruction of SCHEME 1, SCHEME 2, and SCHEME 3 by simulation study.

Only two SNPs are considered for demonstration purposes, and both SNPs are assumed to have the same allele frequency distributions (three different minor allele frequencies were used: 0.1, 0.3, and 0.5), and haplotype frequencies are generated in such a way that low, medium and high LDs are found between the two markers (D' (Lewontin, 1964) ranges from 0 to 0.5; 0.5 to 0.75; and 0.75 to 0.95). According to the haplotype frequencies, 200 haplotypes are generated and we randomly pair them to generate 100 individuals' genotype. Given the genotype, we generate the FI-values for each SNP with three different levels of ambiguity. The ambiguity level is controlled by changing the correlation coefficient (ρ) of the variance matrix of the t-distribution, such that ρ =0.9, 0.75, and 0.6 correspond to low, medium, and high ambiguity levels, respectively. (Figure 2.2.B). For all the markers, we use $\mu_{\mathrm{AA}} = (7, 20)$, $\mu_{\mathrm{Aa}} = (20, 20)$, and $\mu_{\mathrm{aa}} = (20, 7)$, for the location parameter and $\Sigma_{\mathrm{AA}} = \begin{pmatrix} 2 & \sigma_{12} \\ \sigma_{12} & 16 \end{pmatrix}$ $\Sigma_{\mathrm{Aa}} = \begin{pmatrix} 16 & \sigma_{12} \\ \sigma_{12} & 16 \end{pmatrix}$ and $\Sigma_{\mathrm{AA}} = \begin{pmatrix} 16 & \sigma_{12} \\ \sigma_{12} & 2 \end{pmatrix}$ for the scale parameter of the t-distribution for each cluster, where σ_{12} is determined by the level of ambiguity, ρ. These values are based on those estimated from a real dataset in the XRCC1 gene study. Note that we do not include the NFS cluster when simulating the FI values. This is a legitimate assumption because in real experiments, most NFS points result from empty DNA samples that are artificially added for experimental convenience to serve as a negative control.

Overall, we have the 27 different cases (3 ambiguity levels $\times$ 3 allele frequencies $\times$

3 LD levels) and each of the 27 cases repeats 100 times generating the set of 2-SNP FI-values of 100 individuals.

For frequency estimates, we used the following discrepancy measure (Excoffier and Slatkin (1995), Stephens *et al.* (2001)): $D(\theta, \hat{\theta}) = \frac{1}{2} \sum_{g=1}^{s} \left| \theta_g^{\text{true}} - \hat{\theta}_g \right|$, where s denotes the total number of existing haplotypes and θ_g^{true}, and $\hat{\theta}_g$ denote the true haplotype frequency and the estimated haplotype frequency, respectively. The results are presented in Figure 2.5. At low ambiguity level, all three strategies perform similarly. At medium and high ambiguity levels, SCHEME 2 outperforms both SCHEME 1 and SCHEME 3. As we expected, SCHEME 2 was especially advantageous in high LD cases. For the phasing of each individual's haplotypes, SCHEME 1 and SCHEME 2 showed comparable accuracies, although in the case of high LD, SCHEME 2 outperformed SCHEME 1 slightly. This is consistent with the result from the frequency estimate. Both SCHEME 1 and SCHEME 2 were much more robust than SCHEME 3 in the high ambiguity case.

Power Comparison of the Three Phasing Strategies

To find out whether the power of detecting the disease-related haplotype in these tests can be enhanced by considering genotyping uncertainties, we conducted the following haplotype-based case-control association tests. Suppose that the haplotypes consist of two linked SNP markers which are associated with the disease (denoted as SNP-1 with alleles A and a, and SNP-2 with alleles B and b). The four haplotypes are: AB, Ab, aB and ab, with haplotype frequencies: $p_{\text{AB}}, p_{\text{Ab}}, p_{\text{aB}}$, and p_{ab}, respectively, which satisfies $p_{\text{AB}} + p_{\text{Ab}} + p_{\text{aB}} + p_{\text{ab}} = 1$. For the hypothetical case-control study, we considered three different models in our simulation experiment with the frequencies listed as: $p_{\text{AB}}, p_{\text{Ab}}, p_{\text{aB}}, p_{\text{ab}}$. These models are (1) case group: 0.4, 0.3, 0.2, 0.1; control group: 0.25, 0.25, 0.25, 0.25; (2) case group, 0.4, 0.4, 0.1, 0.1; control group: 0.25, 0.25, 0.25, 0.25; and (3) case group: 0.4, 0.3, 0.2, 0.1;

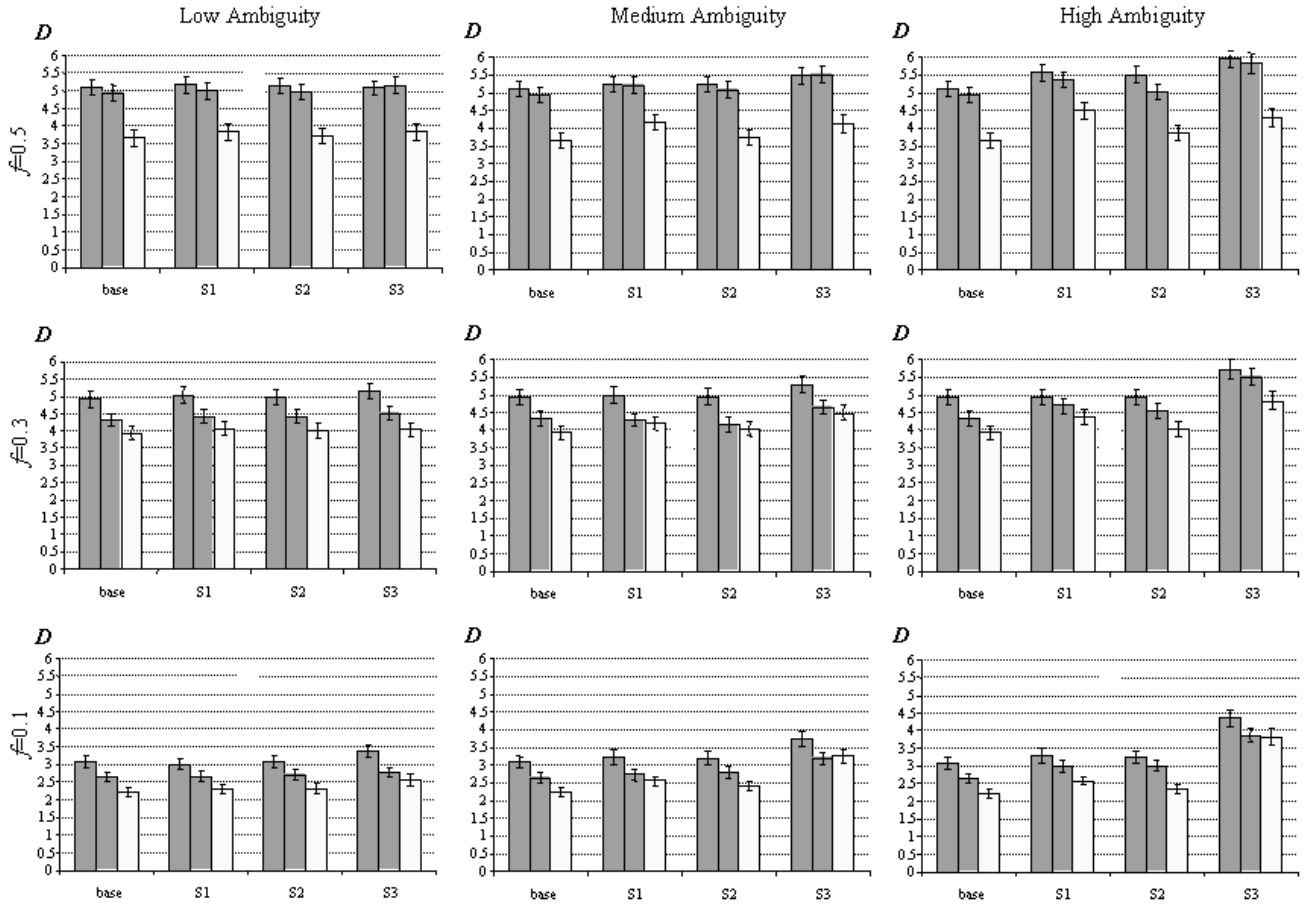

Figure 2.5: Performance Comparison of Haplotype Frequency Estimations of the Three Strategies
The vertical axis measures discrepancy $D(\theta, \hat{\theta}) = \frac{1}{2} \sum_{g=1}^{s} \left| \theta_g^{\text{true}} - \hat{\theta}_g \right|$, the scaled absolute difference between the estimated and the true haplotype frequencies. The error bars are shown as ± 1 standard error. S1, S2, and S3 represent competing strategies shown in Figure 2.4, and "base" refers to the use of true genotype calls to feed in the EM-based haplotype phasing algorithms. A total of 100 data sets were generated for each calculation, and each simulated data set contained 100 individuals. Left bar, low LD (D' ranges from 0 to 0.5); Middle bar, medium LD (D' ranges from 0.5 to 0.75); Right bar, high LD (D' ranges from 0.75 to 0.95). $f =$ minor allele frequency.

control group: 0.3, 0.2, 0.4, 0.3. The simulation proceeds as follows:

1. Simulate n=100 haplotypes and randomly pair them to obtain 50 individual genotypes in each of the case and control populations according to each group's haplotype frequencies.

2. Pool all 100 individuals (50 cases + 50 controls) and generate their FI values according to low, medium, and high ambiguity levels.

3. Cluster the 100 individuals using the t-mixture model; obtain the estimated cluster likelihoods $p^i_{0,k}$, $p^i_{1,k}$, and $p^i_{2,k}$ as well as the cluster posterior probabilities for each individual and SNP.

4. Phase the 100 genotypes using strategies SCHEME 1, SCHEME 2, and SCHEME 3 and count the number of times each of the four different haplotypes appears in the case and control populations. Record counts in each cell of the 2 (case/control) by 4 (AB/Ab/aB/ab) table. It is also possible to use the expected haplotype counts as in the EM algorithm.

5. Compute the χ^2-test statistic for the 2 by 4 table.

6. Randomize to obtain the critical values:

   (a) Assign individuals randomly into the control and case groups along with their $p^i_{0,k}$, $p^i_{1,k}$, and $p^i_{2,k}$ values obtained in step 3. Redo steps 4-5 for this randomly permuted data set.

   (b) Repeat step 6.(a) 500 times and obtain the 90th, 95th, and 99th percentiles of the test statistics, which serve as critical values for significance levels 0.10, 0.05 and 0.01, respectively.

7. Record whether the null hypothesis is rejected or accepted by comparing the test statistics of the original simulated data with the critical values from step 6.(b).

8. Repeat steps 1-7 500 times.

9. Compute the power of the test, i.e., the proportion of times the test was rejected.

Although the test statistic would have an approximate χ^2(d.f.=3) distribution, if we observed the haplotype counts, under the null hypothesis of no association in the standard situation, we can not use this property here because the haplotype counts in the table are not truly observed. Rather, these counts are estimated from the genotype data, which introduces additional uncertainty and may inflate the type-I error. As an alternative, we employed a randomization procedure to determine the critical values for a given significant level, as detailed in step 6 above.

The results of power comparison in association test are presented in Table 2.2. Models 1 and 2 assume that the two SNPs are in perfect linkage equilibrium among the controls; whereas among the cases they are in strong LD (Model 2 had a stronger LD than Model 1). Model 3 mimics a complex disease scenario when the case and the control haplotype distributions differed only slightly. Overall, the haplotype distribution differences are the greatest in Model 2. Thus, for each method considered, the power was always the greatest in Model 2. As we expected, the test using the true genotypes as inputs for the haplotype phasing has the largest power in every scenario, which is likely due to the fact that only phasing uncertainty, but no clustering uncertainty, is present. In low ambiguity cases, SCHEME 1, SCHEME 2, and SCHEME 3 yielded similar powers. In medium and high ambiguity cases, it can be seen that SCHEME 1 and SCHEME 2 always outperformed SCHEME 3 due to the obvious reason that in SCHEME 3 one throws away information (by removing ambiguous points). For Model 2, where the cases have a significant LD compared to the controls, SCHEME 2 had the greatest power among the three under all ambiguity and significance levels.

Model And α		Base	Power(%) Low Ambiguity S1	S2	S3	Med. Ambiguity S1	S2	S3	High Ambiguity S1	S2	S3
Model 1:											
	.10	55.6	55.2	56.2	55.4	55.2	56.8	51.4	57	58	54
	.05	45	43.6	44	43.4	44.4	44	42.4	46.6	48.2	41
	.01	29.2	29.8	29.6	30.4	29.4	29.4	25.6	30.4	28.4	24
Model 2:											
	.10	85.2	82.8	84.2	82.6	80.8	82	78.6	78.8	81	77.4
	.05	75	73.2	74	72.4	72.2	73.4	70.8	68.6	71.2	67.2
	.01	55.4	53	53.4	52.8	52.8	54.6	52.4	49.8	51.2	46.6
Model 3:											
	.10	71.8	68.2	68.8	67.4	67.4	68.4	64.4	64.2	65.2	60.2
	.05	56.8	55	55.2	54.4	55.8	54.6	51.2	49	50	49.2
	.01	32.6	31.4	32.2	29.8	30.2	28.6	25.6	26.8	26.4	26.2

Table 2.2: Comparison of Power to Detect Disease-Related Haplotype through Use of Different Haplotype Inference Strategies under Various Disease Models and Disease Prevalences at Different Type I Error Rates.
α = type I error. For the hypothetical case-control study, we considered three different models in our simulation experiment with the frequencies listed as $p_{\mathrm{AB}}, p_{\mathrm{Ab}}, p_{\mathrm{aB}}, p_{\mathrm{ab}}$. These models are (1) case group: 0.4, 0.3, 0.2, 0.1; control group: 0.25, 0.25, 0.25, 0.25; (2) case group, 0.4, 0.1, 0.1, 0.4; control group: 0.25, 0.25, 0.25, 0.25; and (3) case group: 0.4, 0.1, 0.2, 0.3; control group: 0.3, 0.1, 0.4, 0.2. S1, S2 and S3 represent SCHEME 1, SCHEME 2, and SCHEME 3, respectively.

2.5.2 A Real-Data Example

We applied SCHEME 1, SCHEME 2, and SCHEME 3 on a real genotype data set of four SNPs (C26304T, C26602T, G28152A, and G36189A) located on the XRCC1 gene using TaqMan assay (Han *et al.*, 2003) This data set results from a nested case-control study of breast cancer within the Nurses' Health Study. Among them, the genotypes of 2,244 individuals (a mix of both cases and controls) were used to derive the overall population haplotype frequencies. We applied SCHEME 1, SCHEME 2 and SCHEME 3 on a subset of 315 subjects (including both cases and controls). Haplotype inference was done using PLEM (Qin *et al.*, 2002). A bootstrap-like simulation study demonstrated that the haplotype frequencies estimated by the PLEM in the overall sample (with N >2,000) were very close to the "truth", and we thus used this estimate as the "benchmark." All the results were summarized in Table 2.3. The discrepancy rates (D) for SCHEME 1, SCHEME 2, and SCHEME 3 were 0.03215, 0.0284, and 0.03215, respectively, indicating that SCHEME 2 performed better than both SCHEME 1 and SCHEME 3 in this example.

2.6 Discussion

We developed a novel clustering algorithm based on the t-mixture model for making genotype calls. Using extensive simulations, we compared the performance of this new algorithm with that of the K-means algorithm. Our findings are in agreement with those of Olivier *et al.* (2002), who found that K-means algorithm often placed two centroids within one group of data that would be assigned manually to a single cluster (see Figure 2.3 for examples). As noted by Olivier *et al.* (2002), this is particularly apparent when one of the homozygote clusters had only a few data points. A reason why K-means performed poorly is that the K-means algorithm cannot incorporate information on the approximate locations of the genotype clus-

	Strategy			
Haplotype	SCHEME 1	SCHEME 2	SCHEME 3	Benchmark
0000	0.1378	0.1377	0.1347	0.1347
0001	0.3905	0.3948	0.3917	0.3917
0010	0.3591	0.3592	0.3637	0.3637
0011	0.0044	0.0000	0.0000	0.0000
0100	0.0557	0.0557	0.0574	0.0574
1000	0.0513	0.0507	0.0505	0.0505
1001	0.0001	0.0001	0.0001	0.0001
1010	0.0012	0.0018	0.0019	0.0019

Table 2.3: Comparison of Haplotype Frequency Estimates Using SCHEME 1, SCHEME 2 and SCHEME 3 for a Dataset Obtained Using TaqMan Assay.
In this study, 4 SNP markers (from left to right, C26304T, C26602T, G28152A, and G36189A) in the XRCC1 gene were typed using TaqMan assay for a subset of 315 individuals out of the overall sample (N=2,244). In the first column, "0" stands for major allele, "1" stands for minor allele. The weighted average (case plus control) of haplotype frequency estimates reported in Han *et al.* (2003). The haplotype frequency estimates of the benchmark were obtained by using PLEM.

ters, and cannot handle well the elongated shape of these clusters. The t-mixture clustering method addresses the inherent limitation of the K-means method using a Bayesian approach based on the mixture of t-distributions and can score genotypes probabilistically, which allows for the incorporation of genotyping uncertainties in subsequent analyses.

In our t-mixture clustering algorithm, users can either include or exclude the NFS cluster beforehand. The reasons for excluding the NFS cluster a priori are as follows: (1) blank control samples are often known to the laboratory technician in advance and there is no need to classify them (i.e. there is no "ambiguity"). (2) Genotyping assays for the vast majority of SNP assays typically have a success rate of greater than 98%, which results in a very small group size for assay failures of real samples, which are visually detectable as belonging to the NFS cluster. (3) The small cluster size of NFS may result in an unstable estimate of the variance-covariance matrix, which may compromise the performance in some cases. However, we overcome these problems by using informative priors and imposing an

identifiability constraint on the parameter space.

Poor separation between genotype clusters always constitutes a problem in genotype scoring. For those ambiguous data points, we demonstrated that one clearly loses information by throwing away ambiguous individuals and tends to result in reduced accuracy in haplotype frequency estimation when using deterministic calls. Probabilistic scoring gives rise to more quantitative information and flexibility in the haplotype phasing step and thus can improve the accuracy in haplotype phasing especially in high LD and high ambiguity situations.

The haplotype inference method presented here is formulated for unrelated individuals in random samples of case-control association studies or sib-pair studies without parental data. Although many genotyping errors can be directly resolved in light of parental genotype data, a substantial fraction of errors may still go undetected on the basis of inheritance checking (Douglas *et al.*, 2002). The strategies described here should be also applicable to pedigree data, but modifications of the haplotype inference procedure are necessary. Facing the same capacity problem as encountered by the EM algorithm for haplotype inference, the current approach is limited in the number of linked loci, especially when ambiguous marker loci are abundant. The Partition-Ligation strategy introduced in Niu *et al.* (2002) can be applied to solve this problem, where genotyping uncertainties can be addressed at each atomistic unit.

It is still an unsolved issue in case-control epidemiology studies how to best use the haplotype frequencies and phases inferred from the genotype data. The classical chi-square test is no longer valid because haplotype counts in both cases and controls are not observed, but rather inferred. We used a randomization procedure for the power comparison of the three phasing strategies in case-control studies (Table 2.2). The randomization procedure is a non-parametric means for deriving the threshold for a pre-specified type-I error and may thus be less powerful compared

to a valid parametric test. However, such permutation tests are guaranteed to have the stated significance level and have been a popular method in case-control studies for investigating haplotypic effects.

In sum, the statistical handling of uncertainties in genotype scoring merits more attention than they have received in the past. The use of formal statistical procedures like ours relieves geneticists of the responsibility of manually determining the correct values of doubtful genotypes, and is thus essential for an efficient analysis of high-throughput data. The statistical model presented here is formulated only for SNP markers and is not directly applicable to microsatellite genotyping. But our algorithms can be straightforwardly generalized to that situation or be used directly if the microsatellite alleles are binned into two categories using a reasonable allele size cut-off. Although we considered only Taqman, OLA and MassARRAY, the same strategies developed in this article can be extended to handle data from other experimental platforms such as florescence polarization-single base extension and Illumina's BeadArray technologies, Third Wave's Invader assay, rolling circle amplifications, and molecular beacons.

References

Clark, A. G. (1990). Inference of haplotypes from pcr-amplified samples of diploid populations. *Molecular Biology and Evolution* **7**, 111 – 122.

Douglas, J., Skol, A., and M, B. (2002). Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *The American Journal of Human Genetics* **70**, 487–495.

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution* **12**, 921–927.

Han, J., Hankinson, S. E., DeVivo, I., Spiegelman, D., Tamimi, R., Mohrenweiser, H. W., Colditz, G. A., and Hunter, D. (2003). A prospective study of xrcc1 haplotypes and their interaction with plasma carotenoids on breast cancer risk. *Cancer Research* **63**, 8536–8541.

Hartigan, J. A. and Wong, M. A. (1979). [Algorithm AS 136] A K-means clustering algorithm (AS R39: 81V30 p355-356). *Applied Statistics* **28**, 100–108.

Hawley, M. E. and Kidd, K. K. (1995). Haplo: a program using the em algorithm to estimate the frequencies of multi-site hap-lotypes. *Journal of Heredity* **86**, 409 – 411.

Kang, H., Qin, Z., Niu, T., and Liu, J. S. (2004). Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphisms. *The American Journal of Human Genetics* **74**, 495–510.

Kirk, K. M. and Cardon, L. R. (2002). The impact of genotyping error on haplotype reconstruction and frequency estimation. *European Journal of Human Genetics* **70**, 496–508.

Lewontin, R. C. (1964). The interaction of selection and linkage. i. general consideration; heterotic models. *Genetics* **49**, 49–67.

Lin, S., Cutler, D. J., Zwick, M. E., and Chakravarti, A. (2002). Haplotype inference in random population samples. *American Journal of Human Genetics* **71**, 1129–1137.

Liu, J. S. and Wu, Y. N. (1999). Parameter expansion scheme for data augmentation. *Journal of the American Statistical Association* **94**, 1264–1274.

Long, J. C., Williams, R. C., and Urbanek, M. (1995). An E-m algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics* **56**, 799–810.

Michalatos-Beloin, S., Tishkoff, S. A. Bentley, K. L., Kidd, K. K., and Ruano, G. (1996). Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range pcr. *Nucleic Acids Research* **24**, 4841–4843.

Niu, T., Qin, Z. S., Xu, X., and Liu, J. S. (2002). Bayesian haplotype in-ference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics* **70**, 157–169.

Olivier, M., Chuang, L., Chang, M., Chen, Y., Pei, D., Ranade, K., de Witte, A., Allen, J., Tran, N., Curb, D., Pratt, R., Neefs, H., de Arruda Indig, M., Law, S., Neri, B., Wang, L., and Cox, D. (2002). High-throughput genotyping of single nucleotide polymorphisms using new biplex invader technology. *Nucleic Acids Research* **30**, e53.

Qin, Z. S., Niu, T., and Liu, J. S. (2002). Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *American Journal of Human Genetics* 1242–1247.

Stephens, M. (2000). Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society, Series B* **62**, 795–809.

Stephens, M. and Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics* **73**, 1162 – 1169.

Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**, 978 – 989.

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *The Journal of Computational and Graphical Statistics* **10**, 1–111.

Bayesian Approach to Haplotype Linkage Disequilibrium Mapping for Complex Disease

Abstract

The BLADE algorithm (Liu *et al.*, 2001) introduced a Bayesian approach for utilizing full haplotype information to make an inference on the location of the disease mutation as well as the ancestral haplotypes, mutation rate, and the historical recombination events. One of the key model assumptions of the BLADE algorithm is that there is a single mutation in the disease gene and all mutations occur in the same location of the disease gene. We developed a new method relaxing the one-mutation-locus assumption to allow multiple mutations on different loci on the same chromosome. This approach allows one to discern the interaction effects of susceptible loci to the disease as well as the localization of the mutations and the identification of the historical recombination events descending from founder haplotypes. Using simulation data sets and the well-known Cystic Fibrosis data set, we prove that the new method (BLADE2) performs better the old BLADE method even when there is one mutation locus.

3.0 Preface

This paper is an extension of Liu *et al.* (2001).

3.1 Introduction

For last two decades, disease association studies have made it possible to claim that a glitch in one or another gene raises the risk of diseases from cancer to heart disease, schizophrenia to depression. In many cases, disease gene carriers in the current population are descendant from a small number of "founders" in whose genomes the deleterious mutation appeared some generation ago. Because of the shared ancestry, in the haplotypes of the disease population, we observe the non-random assortment of alleles. This is the key idea of linkage disequilibrium (LD). (Ardlie *et al.*, 2002)

Most common methods for finding LD is to measure the discrepancy at each marker between a case and a control sample. However, this is not powerful when the markers are tightly linked, because it is believed that contiguous markers in a chromosome are not independent and haplotypes have a block structure (Gabriel *et al.*, 2002). McPeek and Strahs (1999) and Morris *et al.* (2002) introduced hidden Markov model incorporating the dependent structure of the contiguous markers (i.e., haplotype). Liu *et al.* (2001) introduced a Bayesian approach (BLADE algorithm), which explicitly model positions of the historical recombination and mutation events that produced the observed haplotypes from an initial set of founders.

The BLADE algorithm assumes that there is a single mutation in the disease gene and all mutations occur in the same location of the disease gene. However, it is believed that some diseases such as type 2 diabetes and Crohn disease have multiple disease susceptible loci (Rioux *et al.* (2001), Florez *et al.* (2004)) and our understand-

ing of the genetic association of such complex disorders is still mostly unknown. Hence, the main focus of this paper is to extend the BLADE algorithm to perform LD mapping for complex disease with two probable disease mutations and study the interaction effects of susceptible loci to the disease.

The primary goals of our algorithm are the localization of genes responsible for the disease within the considered set of markers, the determination of ancestral haplotypes, the construction of haplotypes from unphased chromosomes, the separation of distinct founders of the disease, and inference on the ages of the mutations causing the disease. Furthermore, we want to determine which of the two disease mutations is found in individuals' haplotypes who shares a founder. In our Bayesian model framework, the main parameters of interest are the locations of the disease mutations, the age of mutations, and ancestral haplotypes. To simplify the sampling steps, we also introduce the (unobservable) locations of recombinations as auxiliary variable. Like most of Bayesian data augmentation approaches, our algorithm is very flexible in treating various complications such as missing marker data, multiple founders, and unphased chromosomes. We call our new algorithm BLADE2.

In section 3.2, we present model assumptions for building the likelihood function and in section 3.4, we describe the iterative Monte Carlo sampling steps for each variables. In section 3.5, we test the new algorithm by various simulation studies and illustrate the advantage of the new method over the previous BLADE algorithm and apply the new method to real-data example.

3.2 Model Assumptions

The basic idea of the model is that the current disease haplotypes are the descendant from a few founders with disease mutations and the haplotype sharing

among the descendant from a common ancestor decays in succeeding generations. The decay is due to recombination events and mutations, which result in conservation of only a small region of the ancestral haplotype around the mutation.

We assume that there are $k+1$ clusters for the disease haplotype, corresponding to k founder haplotypes in the current disease population and one null cluster for all other disease chromosomes. The null clusters can also represent phenocopies of disease haplotypes because disease haplotypes could also carry none of the disease mutations at the disease loci. Each cluster except for the null cluster shares the haplotype with its own ancestor. Given the known ancestor, disease haplotypes are assumed to be from unrelated individuals.

The decay of haplotype sharing occurs mostly due to recombinations. A recombination frequency between markers is correlated with the distance between markers. Generally speaking, the greater the physical separation between genes along a chromosome, the greater chance for a recombination event to take place. Instead of the physical distance, we use the genetic distance (or map distance) to calculate the recombination frequency and it is known to us. Given the genetic distance d in unit of Morgan between two markers, the recombination probability is known as $\theta = (1 - e^{-2d})/2$ according to Haldane's mapping function (Hartl and Jones, 1998). We further define $\tau = -\log(1-\theta)$ so that $e^{-\tau}$ is the probability that there is no recombination between the two markers. When d is small, $\tau \approx d$ holds true, hence, we refer to τ as the genetic distance. We assume that recombinations occur as a homogeneous Poisson process disregarding a haplotype block structure. We assume that the ratio of physical to genetic distance is constant (we assume that 1cM≈1MB). To simplify the mutation process, we assume that there is a small probability r for each locus to mutate and r is assumed to be identical for each marker and each individual. When a marker mutates, it has a equal chance to to turn into other alleles.

To simplify the genealogy structure of the data, we assume that the coalescence process that generates the observed disease haplotypes within each cluster can be approximated by a star genealogy; that is, given the same ancestor, haplotypes in the cluster are mutually independent. The correlations among the disease haplotypes are partially accounted for by allowing for multiple founder haplotypes with different ages.

Finally, we assume that there are at most two loci with disease mutations on the ancestral haplotypes and all mutations occur in the same location of the disease gene for all the k ancestral mutations. This assumption nests the case where there is only one mutation locus as the previous BLADE algorithm, because having the two mutation loci on the same (or very close) location is equivalent to having a single location. By relaxing the one-locus assumption, we are allowed to study interactions of the disease susceptible mutations and determine which is the dominant mutation when there are two competing disease variants. This gives our model more flexibility to study different risks involved with each mutations. Figure 3.1 illustrates various cross-over scenarios and corresponding disease risks. If the interaction of the two genetic variant causes high risk of the genetic disease, then we find more disease haplotypes containing the both disease mutations in the current population. If one of the two mutations is dominant over the other for causing the disease, then we find the haplotypes with that mutation more than the one with the other in the current population. Especially when the two mutations are not too close to have a non-trivial probability for having recombination between the two mutations, this sub-clustering plays a key role to determine the genetic cause of the disease.

3.3 Model Formulation

Data Description

We denote a collection of disease haplotypes $\mathbf{H} = (H^1, ..., H^N)$, where N is the number of haplotypes. Each individual haplotype is a vector $H^t = (H^t_1, ..., H^t_m)$, where H^t_j is (a numeric code of) the allele at marker locus j for the individual t and m is the number of marker loci. The order and distance of marker loci are known a priori by the genetic distance d_i in unit of Morgan between the left most marker and i-th marker from the left. As we discussed before, we can compute that the probability of having no recombination between marker i and marker j as $e^{-(\tau_j - \tau_i)}, j > i$. At marker j, there are n_j different alleles, labeled as $\{1, 2, ..., n_j\}$. Depending on the inter-marker distances of the data, it is necessary to introduce a Markovian structure on the non-disease haplotype in which we need to estimate the allele frequencies conditional on the adjacent markers in stead of marginal frequencies.

3.3.1 Control Data Model

The data have control group and disease group. The control group haplotypes are used for estimating control data haplotype frequencies and they are not directly involved with the model parameters such as the disease locations and ancestral haplotypes. However, it is important to obtain accurate estimates for the control data haplotype frequencies for estimating the parameters accurately, because the inference on the parameters are based on the LD between the two group.

We need to specify the model for the control data haplotypes to compute the likelihood for case group when computing the probability of the segment of the disease haplotype replaced by non-disease haplotype through recombinations. However, we do not focus on making an inference on the frequencies of the non-disease hap-

lotypes, because it is not our main interest in this paper and usually we can obtain a large number of control data to have accurate estimates by simply using sample means. It is more important to note, however, that we need to choose which model to use to compute the probability of the haplotype given the estimated marginal or conditional frequencies of the alleles. We suggest two choices as follows.

Independent Marker Model

This model assumes that the genetic markers are in linkage equilibrium, that is, an allele frequency at a given marker is independent of the allele at the neighboring markers. This model is suitable for the markers whose genetic distances between them are large. The likelihood is

$$\Pr_0(\{H^t\}_S^T) = \prod_{S \le j \le T} p_{j,H_j^t},\ i = 1, 2, ..., n_j, \tag{3.1}$$

where $p_{j,i}$ is the estimated allele frequency of allele i at marker j, $\{H^t\}_S^T$ is the contiguous haplotype segment $(H_S^t, ..., H_T^t)$ and $\Pr_0(\cdot)$ is a notation for the control data likelihood. We estimate $p_{j,i}$ by using sample mean.

Markov Model

This model assumes that the genetic markers are in linkage disequilibrium, that is, an allele frequency at a given marker is dependent of the allele at the neighboring markers when the markers are close to each other. For simplicity of the model, we only consider lag-one dependence and assume that lag-one dependence construction takes account for the other information for the linkage disequilibrium, such as haplotype block structure or the genetic distances between markers. Again, we estimate the lag-one conditional allele frequencies by sample mean. We construct the joint distribution from a left marker to a right marker, since the direction of the

conditional distribution does not matter for computing the joint distribution:

$$\Pr_0(\{H^t\}_S^T) = p_{j,H_S^t} \prod_{S \le j \le T-1} \Pr(H_{j+1}^t | H_j^t), \tag{3.2}$$

where $\Pr(H_{j+1}^t | H_j^t)$ the estimated conditional allele frequency at marker $j+1$ given the allele information at marker j.

3.3.2 The Disease Data Likelihood

Model Parameters

Our main parameters of interest are the location of the two mutations represented by the genetic distances between leftmost marker and the disease mutations, $\boldsymbol{\tau} = (\tau_1, \tau_2)$, $\tau_1 < \tau_2$. The number of generations, referred as the age of mutations for the current sample is denoted by $\mathbf{G} = (G_1, ..., G_k)$, where G_k is the age for cluster k. For each cluster k, there is an ancestral haplotype $\mathbf{A} = (A_1, ..., A_k)$, where $A_c = (A_{c,1}, ..., A_{c,m})$, is the ancestral haplotype for founder k at each marker. Another model parameter is the mutation probability r. According to our model assumption, we compute the probability that an allele at locus i mutate from allele $'a'$ to allele $'b'$ in G generations:

$$r(i, G, a, b) = \begin{cases} (1-r)^G & : \quad a = b \\ \dfrac{1-(1-r)^G}{n_i - 1} & : \quad a \neq b. \end{cases}$$

Missing data

Missing data refer to the unobservable variables introduced for the computational convenience. In our model, we introduce a cluster indicator variable C^t for each disease haplotype H^t, $C^t = 0$ indicates that the haplotype H^t belongs to the null cluster (phenocopy) and $C_t = c \neq 0, c = 1, ..., k$ indicates that the disease loci are inherited from the founder haplotype c. That is, $C_t = c \neq 0$ implies that the

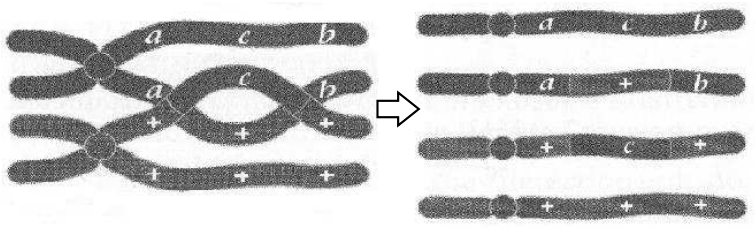

Figure 3.1: Diagram showing a double cross-over.
If the interaction of genetic variant 'a' and 'b' cause the genetic disease, then the next generation still contains the two mutations and is risky of the disease. If the interaction of genetic variant 'a' and 'c' cause the genetic disease, then the interaction effect is broken by the double cross-over and the next generation is less risky of the disease. (Image credit: Hartl and Jones (1998).)

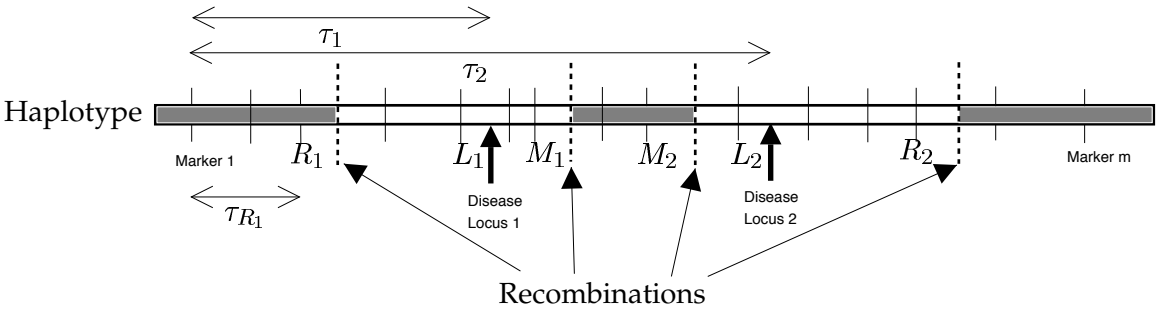

Figure 3.2: A graphical representation of the haplotype model.
There are a total of m markers. Each vertical solid line represents the location of a marker. The vertical dotted lines are the locations of recombinations. Parameter τ_1 and τ_2 are the "recombination distance" from the disease locus to the first marker. The gray area represents the chromosomal piece which was replaced by non-ancestral alleles by recombination events. Note that the middle gray area can result from a double cross-over. The white area is a chromosomal piece in the region that is identical by descent with the founder. The recombination event closest to the first disease locus occurred between markers R_1 and $R_1 + 1$ and that from the middle region occurred between markers M_1 and $M_1 + 1$, and M_2 and $M_2 + 1$, and that from the right arm occurred between markers R_2 and $R_2 + 1$.

haplotype of the t-th individual shares the alleles with the ancestor haplotype c in the small region around the mutations.

Given that $C_t \neq 0$, we can sub-classify the haplotype into three cases: the disease haplotype conserves the mutation only at the first mutation locus τ_1, the disease haplotype conserves the mutation only at the second mutation locus τ_2, and the disease haplotype conserves the mutation at both loci. Therefore, we define the index for identifying this disease mutation conservation status, $S_t = 0, 1, 2$, and 3, where $S_t = 0$ means that $C_t = 0$, i.e. the haplotype is a phenocopy and conserves no ancestral mutations, $S_t = 1$ suggests that the disease haplotype conserves the mutation at the first locus, $S_t = 2$ suggests that the disease haplotype conserves the mutation at the second locus, and $S_t = 3$ suggests that the disease haplotype conserves the mutation at both loci.

Depending on the sub-cluster information, we can specify the set of variables to locate the position of recombination events nearest to the disease loci. When there is only one disease locus, there is only one contiguous set of markers that conserve the alleles from the ancestor. When $S_t = 1, 2$, we refer R_1^t that the nearest recombination event on the left side of the haplotype from the mutation locus occurs between markers R_1^t and $R_1^t + 1$ Similarly, we define R_2^t on the right side.

When both mutations were conserved, one possible scenario is that a set of contiguous markers between τ_1 and τ_2 is replaced by non-ancestral haplotypes by a double cross-over between the two mutations. We add another set of variables (M_1^t, M_2^t) denoting that the two ends of the replaced markers such that the left end of the double cross-over occurs between markers M_1^t and $M_1^t + 1$, and the right end occurs between M_2^t and $M_2^t + 1$. Figure 3.2 illustrates the structure of the disease haplotype decayed by recombinations. Note that R_1^t, R_2^t, M_1^t and M_2^t may not exist provided that there is no recombination events in the region of the data since its founder.

Given the missing data $(C^t, S^t, R_1^t, R_2^t, M_1^t, M_2^t)$, it is simplified to write the conditional likelihood or the complete data likelihood and they can get easily marginalized out to compute the observed data likelihood, because $(C^t, S^t, R_1^t, R_2^t, M_1^t, M_2^t)$ are all discrete variables.

For notational convenience, we denote the two markers flanking the mutation locus at τ_1 by L_1 and L_1+1 and the two markers flanking the mutation locus at τ_2 by L_2 and L_2+1.

The Complete Data Likelihood

The complete data likelihood is the joint distribution of the observed data $\mathbf{H}$ and missing data $(C^t, S^t, R_1^t, R_2^t, M_1^t, M_2^t)$, given the parameters $(\mathbf{A}, \mathbf{G}, \tau_1, \tau_2, r)$. If the t-th individual haplotype is clustered as a phenocopy, i.e. $C^t = 0$, then $(R_1^t, R_2^t, M_1^t, M_2^t)$ are not defined and the likelihood is same as the control data likelihood, which simply is a population frequency the haplotype among the control group:

$$\Pr(H^t, C^t = 0 | \mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r) = \alpha_0 \Pr_0(\{H^t\}_1^m),$$

where α_0 is a priori probability of the cluster frequency for the null cluster. When $C^t \neq 0$, the complete data likelihood equation can be factorized into two or three parts depending on the sub-cluster data S^t. When there exists only one mutation either at τ_1 or at τ_2, the joint distribution is factorized into the left arm of the haplotype and the right arm:

$$\begin{aligned}
&\Pr(H^t, C^t = c, S^t, R_1^t, R_2^t | \mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r) \\
&\quad = \alpha_c \beta_l \Pr(H^t, R_1^t, R_2^t | A_c, G_c, \boldsymbol{\tau}, r, C^t = c) \\
&\quad = \alpha_c \beta_l \Pr(\{H^t\}_1^{L_l}, R_1^t | A_{c\cdot}, G_c, \boldsymbol{\tau}, r, C^t = c) \\
&\qquad \times \Pr(\{H^t\}_{L_l+1}^m, R_2^t | A_{c\cdot}, G_c, \boldsymbol{\tau}, r, C^t = c),\ l = 1, 2,
\end{aligned} \tag{3.3}$$

for $c \neq 0$, where A_c denotes the ancestral haplotype for cluster c, α_c is a priori probability of the cluster frequencies for cluster c, and β_l is a priori probability of the sub-cluster frequencies for sub-cluster l. We discuss this and other prior distributions in the next section.

When $S^t = 3$, the complete data likelihood is factorized into three parts, which correspond to the three subpieces of the haplotype, the left of τ_1, between τ_1 and τ_2, and the right side from τ_2, as illustrated in Figure 3.2:

$$\begin{aligned}
&\Pr(H^t, C^t = c, S^t = 3, R_1^t, R_2^t, M_1^t, M_2^t | \mathbf{A}, \mathbf{G}, \boldsymbol{\tau}) \qquad (3.4)\\
&\quad = \alpha_c \beta_3 \Pr(H^t, R_1^t, R_2^t, M_1^t, M_2^t | A_c, G_c, \boldsymbol{\tau}, C^t = c, S^t = 3)\\
&\quad = \alpha_c \beta_3 \Pr(\{H^t\}_1^{L_1}, R_1^t | A_c, G_c, \tau_1, C^t = c, S^t = 3)\\
&\qquad \times \Pr(\{H^t\}_{L_1+1}^{L_2}, M_1^t, M_2^t | A_c, G_c, \boldsymbol{\tau}, C^t = c, S^t = 3) \qquad (3.5)\\
&\qquad \times \Pr(\{H^t\}_{L_2+1}^{m}, R_2^t | A_c, G_c, \tau_2, C^t = c, S^t = 3).
\end{aligned}$$

Since the probability of having at least one combination in G generations between the a-th marker and b-th marker is $1 - e^{-|\tau_b - \tau_a| G}$, we have

$$\begin{aligned}
&\Pr(\{H^t\}_1^{L_1}, R_1^t | A_c, G_c, \boldsymbol{\tau}, C^t = c, S^t = l) = \qquad (3.6)\\
&\quad (1 - e^{-(\tau_{R_1^t+1} - \tau_{R_1^t}) G_c}) e^{-(\tau_1 - \tau_{R_1^t+1}) G_c} \Pr_0(\{H^t\}_1^{R_1^t}) \prod_{R_1^t < j < L_1} r(j, G_c, A_{c,j}, H_j^t)
\end{aligned}$$

and

$$\begin{aligned}
&\Pr(\{H^t\}_{L_2+1}^{m}, R_2^t | A_c, G_c, \boldsymbol{\tau}, C^t = c, S^t = l) = \qquad (3.7)\\
&\quad (1 - e^{-(\tau_{R_2^t+1} - \tau_{R_2^t}) G_c}) e^{-(\tau_2 - \tau_{R_2^t+1}) G_c} \Pr_0(\{H^t\}_{R_2^t+1}^{m}) \prod_{L_2 < j \leq R_2^t} r(j, G_c, A_{c,j}, H_j^t),
\end{aligned}$$

where $\Pr_0(\cdot)$ is the probability of the segment of the haplotype under control data model as described in Section 3.3.1.

The second term in the complete data likelihood (equation (3.5) is:

$$
\begin{aligned}
&\Pr(\{H^t\}_{L_1+1}^{L_2}, M_1^t, M_2^t | A_{c\cdot}, G_c, \tau, C^t = c) \\
&= \begin{cases}
(1-\phi^{G_c})\Pr_0(\{H^t\}_{M_1^t+1}^{M_2^t}) \displaystyle\prod_{L_1<j\le M_1^t, L_1<j\le M_1^t} r(j, G_c, A_{c,j}, H_j^t) \\
\qquad\qquad \times \dfrac{2(\tau_{M_1^t+1}-\tau_{M_1^t})(\tau_{M_2^t+1}-\tau_{M_2^t})}{(\tau_2-\tau_1)^2} & (*) \\
\phi^{G_c} \displaystyle\prod_{L_1<j\le L_2} r(j, G_c, A_{c,j}, H_j^t) & (**)
\end{cases}
\end{aligned}
$$

where $\phi = 1 - e^{-(\tau_2-\tau_1)}(1+(\tau_2-\tau_1))$ is the probability that no double cross-over occurs between the two disease mutations during a meiosis, which is an approximation under the assumption that the genetic distance between the two mutation locus is too close and the probability of a triple or more cross-over on the haplotype region between τ_1 and τ_2 is ignorable. This is a reasonable assumption, because we are considering haplotypes on a single chromosome. Equation$(*)$ corresponds to the case where at least double cross-over occurs between two disease mutations in G_c generations. This equation is an approximation under the assumption that $1-\phi$ is close to 0. Equation$(**)$ corresponds to the case where no double cross-over occurs in G_c generations.

Since all the missing data $(C^t, S^t, R_1^t, R_2^t, M_1^t, M_2^t)$ are discrete, we can marginalize them out by summations to compute the observed data likelihood of the single

haplotype. First, we marginalize out R_1^t, R_2^t, M_1^t and M_2^t

$$\Pr(H^t, C^t = c, S^t = l|\mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r) =$$

$$\begin{cases} \alpha_0 \Pr_0(\{H^t\}_1^m) & \text{if } c = 0 \\ \alpha_c\beta_l \sum_{R_1^t=1}^{L_l} \Pr(\{H^t\}_1^{\leq L_l}, R_1^t|A_{c\cdot}, G_c, \boldsymbol{\tau}, C^t = c, S^t = l) & \\ \quad \times \sum_{R_2^t=L_l+1}^{m} \Pr(\{H^t\}_{L_l+1}^m, R_2^t|A_{c\cdot}, G_c, \boldsymbol{\tau}, C^t = c, S^t = l) & \text{if } c \neq 0,\ l = 1, 2 \\ \alpha_c\beta_l \sum_{R_1^t=1}^{L_1} \Pr(\{H^t\}_1^{L_1}, R_1^t|A_{c\cdot}, G_c, \boldsymbol{\tau}, C^t = c, S^t = l) & \\ \quad \times \sum_{I_1 \leq M_1^t \leq M_2^t \leq I_2} \Pr(\{H^t\}_{L_1+1}^{L_2}, M_1^t, M_2^t|A_{c\cdot}, G_c, \tau, C^t = c, S^t = l) & \\ \quad \times \sum_{R_2^t=L_2+1}^{m} \Pr(\{H^t\}_{L_2+1}^m, R_2^t|A_{c\cdot}, G_c, \boldsymbol{\tau}, C^t = c, S^t = l) & \text{if } c \neq 0,\ l = 3 \end{cases} \tag{3.8}$$

then, we marginalize out C^t and S^t

$$\Pr(H^t|\mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r) = \sum_{c=0}^{k} \Pr(H^t, C^t = c|\mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r) \tag{3.9}$$

$$= \sum_{c=0}^{k} \sum_{l} \Pr(H^t, C^t = c, S^t = l|\mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r), \tag{3.10}$$

where $l = 0$ if $c = 0$, and $l = 1, 2, 3$ if $c \neq 0$.

Under the conditional independence assumption, the observed data likelihood of the disease haplotypes is obtained by the product of the single-observation likelihood:

$$\Pr(\mathbf{H}|\mathbf{A}, \mathbf{G}, \tau, r) = \prod_{t=1}^{N} \Pr(H^t|\mathbf{A}, \mathbf{G}, \tau, r). \tag{3.11}$$

Prior Distributions

For computational convenience and non-subjective estimation, we use uniform prior for our parameters. Since, τ_1 and τ_2 have a finite space, we can use proper uniform prior $\tau_1, \tau_2 \sim \mathbf{Uniform}(\mathbf{0}, \tau_\mathbf{m})$. We also assume uniform prior on the ancestral haplotype $\mathbf{A}$, i.e. the marginal allele frequency of a particular marker is

equal for all possible alleles for all cluster c. For G_c, we limit its value on finite discrete support with equal probability. For r, uniform prior on log-scale is used, $\log_{10} r \sim \textbf{Uniform}(-3, -4)$. These prior distributions simplify the implementation and our empirical study showed that the estimation on τ_1, τ_2 and $\mathbf{A}$ is not sensitive on the choice of the prior values on G_c and r. All the parameters are independent a priori.

3.4 The Sampling Algorithm

By Bayes rule, the posterior distribution of the parameters is

$$p(\mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r|\mathbf{H}) \propto p(\mathbf{H}|\mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r)p(\mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r).$$

However, it is easier to work with complete data likelihood. Data Augmentation algorithm (Tanner and Wong, 1987) makes model-fitting simpler especially in high-dimensional models. Thus, we impute the missing data at each iteration, treat them as if they were observed, and find the conditional posterior distribution of the parameters given the current values of the missing data. This iterative process generates the draws from the joint distribution of missing data and parameters:

$$\begin{aligned} &p(\mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r, \mathbf{R_1}, \mathbf{R_2}, \mathbf{M_1}, \mathbf{M_2}, \mathbf{C}, \mathbf{S}|\mathbf{H}) \propto \\ &\quad p(\mathbf{H}, \mathbf{R_1}, \mathbf{R_2}, \mathbf{M_1}, \mathbf{M_2}, \mathbf{C}, \mathbf{S}|\mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r)p(\mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r), \end{aligned}$$

where $\mathbf{R}_1, \mathbf{R}_2, \mathbf{M}_1, \mathbf{M}_2, \mathbf{C}$ and $\mathbf{S}$ are vector notations of $R_1^t, R_2^t, M_1^t, M_2^t, C^t$ and S^t, respectively.

Due to the high dimensionality of the posterior distribution, we decompose the sampler into the following conditional sampling steps. We use $[X|Y, .., Z]$ to denote the conditional distribution of X given $Y, ..., Z$ under the target posterior distribution. We denote the collection of missing data, $\mathbf{Y}_{\text{mis}} = (\mathbf{R}_1, \mathbf{R}_2, \mathbf{M}_1, \mathbf{M}_2, \mathbf{C}, \mathbf{S})$.

To speed up the convergence of the sampler, we marginalize out missing data, if necessary. Since all the conditional posterior distribution is conditioned on the observations, the notation $\mathbf{H}$ is omitted in the following formulas. The algorithm iterates through the following steps.

1. Draw $\mathbf{C}$ from $[\mathbf{C}|\mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r]$. Since C^t is discrete, a new C^t is sampled with the probability proportional to

$$[C^t = c|\mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r] \propto \Pr(H^t, C^t = c|\mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r).$$

2. Draw $\mathbf{S}$ from $[\mathbf{S}|\mathbf{C}, \mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r]$

$$[S^t = l|C_t = c, \mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r] \propto \Pr(H^t, C^t = c, S^t = l|\mathbf{A}, \mathbf{G}, \boldsymbol{\tau}, r).$$

3. Draw $\mathbf{R}_1, \mathbf{R}_2, \mathbf{M}_1, \mathbf{M}_2$ from $[\mathbf{R}_1, \mathbf{R}_2, \mathbf{M}_1, \mathbf{M}_2|\boldsymbol{\Theta}, \mathbf{C}, \mathbf{S}]$. In particular, Draw R_1^t, R_2^t according to equation (3.6) and (3.7), if $S^t = 1, 2, C^t \neq 0$. Draw $R_1^t, R_2^t, M_1^t, M_2^t$ according to equation (3.4), if $S^t = 3, C^t \neq 0$.

4. Draw $\mathbf{G}$ from $[\mathbf{G}|\mathbf{A}, \boldsymbol{\tau}, r, \mathbf{R}_1, \mathbf{R}_2, \mathbf{M}_1, \mathbf{M}_2, \mathbf{C}, \mathbf{S}]$. Since we use the discrete prior for $\mathbf{G}$, a new G_c will be sampled with the probability evaluated at each value of G_c:

$$\begin{aligned}&[G_c = g|\mathbf{A}, \boldsymbol{\tau}, r, \mathbf{R}_1, \mathbf{R}_2, \mathbf{M}_1, \mathbf{M}_2, \mathbf{C}, \mathbf{S}] \\ &\quad \propto \prod_{t:C^t=c} \Pr(H^t, C^t, S^t, R_1^t, R_2^t, M_1^t, M_2^t|G_c = g, A_{c\cdot}, \boldsymbol{\tau}, r)\end{aligned}$$

5. Draw $\boldsymbol{\tau}$ from $[\boldsymbol{\tau}|\mathbf{G}, \mathbf{A}, r]$. Since the value of $\boldsymbol{\tau}$ is highly correlated with the recombination positions, we need to sum out all the missing data to speed up the sampler. We draw $\boldsymbol{\tau}$ via Metropolis-Hastings step (Hastings, 1970; Metropolis *et al.*, 1953) by sampling candidate draw $\boldsymbol{\tau}'$ from uniform distribution on $(\tau_1, \tau_2), \tau_1 < \tau_2$:

$$q(\tau_1'|\boldsymbol{\tau}) = \frac{1}{\tau_{1,\text{upper}} - \tau_{1,\text{lower}}},$$

where $\tau_{1,\text{lower}} = \max(0, \tau_1 - s)$ and $\tau_{1,\text{upper}} = \min(\tau_m, \tau_1 + s)$. Subsequently,

$$q(\tau_2'|\tau_1', \boldsymbol{\tau}) = \frac{1}{\tau_{2,\text{upper}} - \tau_{2,\text{lower}}},$$

where $\tau_{2,\text{lower}} = \max(\tau_1', \tau_2 - s)$ and $\tau_{1,\text{upper}} = \min(\tau_m, \tau_2 + s)$. The candidate is accepted with probability

$$\alpha(\boldsymbol{\tau}, \boldsymbol{\tau}') = \min\left(1, \frac{\Pr(\mathbf{G}, \mathbf{A}, \boldsymbol{\tau}', r) q(\boldsymbol{\tau}|\boldsymbol{\tau}')}{\Pr(\mathbf{G}, \mathbf{A}, \boldsymbol{\tau}, r) q(\boldsymbol{\tau}'|\boldsymbol{\tau})}\right),$$

where $\Pr(\mathbf{G}, \mathbf{A}, \boldsymbol{\tau}, r)$ is from equation (3.11).

6. For cluster $c = 1, \ldots, k$, we update the ancestral haplotype A_c. one marker at a time. We draw from $[A_{c,i}|C^t = c, G_c, A_{c\cdot}, \boldsymbol{\tau}, r]$ for $i = 1, \ldots, m$. The probability $A_{c,i} = j$ is proportional to

$$[A_{c,i} = j|C^t = c, G_c, A_{c\cdot}, \boldsymbol{\tau}, r] \propto \prod_{t:C^t=c} \Pr(H^t, C^t|G_c, (A_{c,1}, ..., A_{c,i} = j, ..., A_{c,m}), \boldsymbol{\tau}, r).$$

7. Draw r from $[r|\mathbf{A}, \mathbf{G}, \boldsymbol{\tau}]$ via Metropolis step (Metropolis *et al.*, 1953) by sampling candidate draw r' from uniform distribution on log-scale,

$$\log_{10} r' \sim \textbf{Uniform}(-3, -4).$$

The candidate is accepted with probability

$$\alpha(r, r') = \min\left(1, \frac{\Pr(\mathbf{G}, \mathbf{A}, \boldsymbol{\tau}, r')r'}{\Pr(\mathbf{G}, \mathbf{A}, \boldsymbol{\tau}, r)r}\right),$$

where $\Pr(\mathbf{G}, \mathbf{A}, \boldsymbol{\tau}, r)$ is from equation (3.11).

3.5 Results

3.5.1 Simulation Study

We simulated three populations of the disease haplotypes originating from a single founder who had two mutation loci, 150 generations ago. Three populations are:

1. a population of haplotypes that conserve the fist mutation only, 2. a population of haplotypes that conserve the second mutation only, and 3. a population of haplotypes that conserve the both mutations from the founder. The growth rate of the populations is 1.031, except for the first eight generations where the expansion was doubled. These parameters were chosen to mimic the history of the European population. We set mutation rate for each marker to be 0.00002 per generation. When recombination occurs, a disease haplotype recombines with a random one generated by the Markov control model. We randomly generated the allele transition matrix for the Markov model in correlation with the inter-marker distance, i.e., we set the transition probability close to 0 or 1 for the markers close together, and set the transition probability close to 0.5 for markers far apart. We considered 20 bi-allelic markers, located at 0.00, 0.07, 0.24, 0.35, 0.42, 0.62, 0.69, 0.79, 0.99, 1.10, 1.29, 1.45, 1.51, 1.62, 1.71, 1.83, 1.96, 2.05, 2.18, and 2.35 cM from the left most marker, and they are 0.1175 cM apart on average. The first founder mutation is set to locate between markers 7 and 8, 0.73 cM away from the left most marker and the second founder mutation is set to locate between markers 14 and 15, 1.65 cM away from the left most marker. Under the Poisson process assumption, the probability that a triple or more cross-over occurs during a meiosis within this haplotype range is very low, 2.12×10^{-6}, which agrees with our assumption for the approximation of the complete data likelihood. The 100 control haplotypes were simulated from the Markovian model. We randomly selected 80 disease haplotypes from the final generation of the three populations and added 20 randomly generated independent haplotypes from the control data population as phenocopies. To test the algorithm in various situations, we had different combinations of the three kinds of disease haplotypes for each simulation study. In the following table, n_0 denotes the number of phenocopies in the disease haplotype data set, n_1 denotes the number of the disease haplotypes from the population 1, n_2 denotes the number of the disease haplotypes from the population 2, and n_3 denote the number of the disease haplotypes from the population 3.

	n_0	n_1	n_2	n_3
CASE 0	20	80	0	0
CASE 1	20	0	0	80
CASE 2	20	20	20	40
CASE 3	20	10	30	40
CASE 4	20	30	10	40
CASE 5	20	40	40	0

CASE 0 is the case that only the first mutation is the cause of the disease and the second one is irrelevant and CASE 1 occurs when the disease is caused only when the disease haplotypes have both mutations. Assuming that the simulated data set is randomly sampled from the disease population, CASE 2, 3, and 4 emulate the situation that the two mutations positively interact to increase the disease risk; when both mutations are equally risky, when the first mutation is less risky, and when the second mutation is less risky, respectively. CASE 5 represents the scenario that the two mutations interact negatively toward the risk in a way that two mutations completely erase the risk. We added 20 phenocopies for every case to test the robustness of the algorithm. Note that only CASE 0 is suitable for using the previous BLADE. Although CASE 0 has only one mutation, we expect the BLADE2 algorithm to draw τ_1 and τ_2 around the true location.

Firstly, we analyzed the simulation CASE 0 for direct comparisons between BLADE and BLADE2. We computed a single marker measure of disequilibrium, which is defined as

$$\delta = \max_A \left(\frac{P(A|D) - P(A|N)}{1 - P(A|N)} \right),$$

where A ranges over all of the possible alleles, D indicates the disease population, and N the normal population. Based on the δ-values (indicated as dots in Figure 3.3), we could roughly estimate that the disease locus was near marker 7, where δ value was the highest. We also drew 1,500 MCMC draws and discarded the first 5000 draws for burn-in period via BLADE and BLADE2. We summarized the results in Figure 3.3 and Table 3.1.

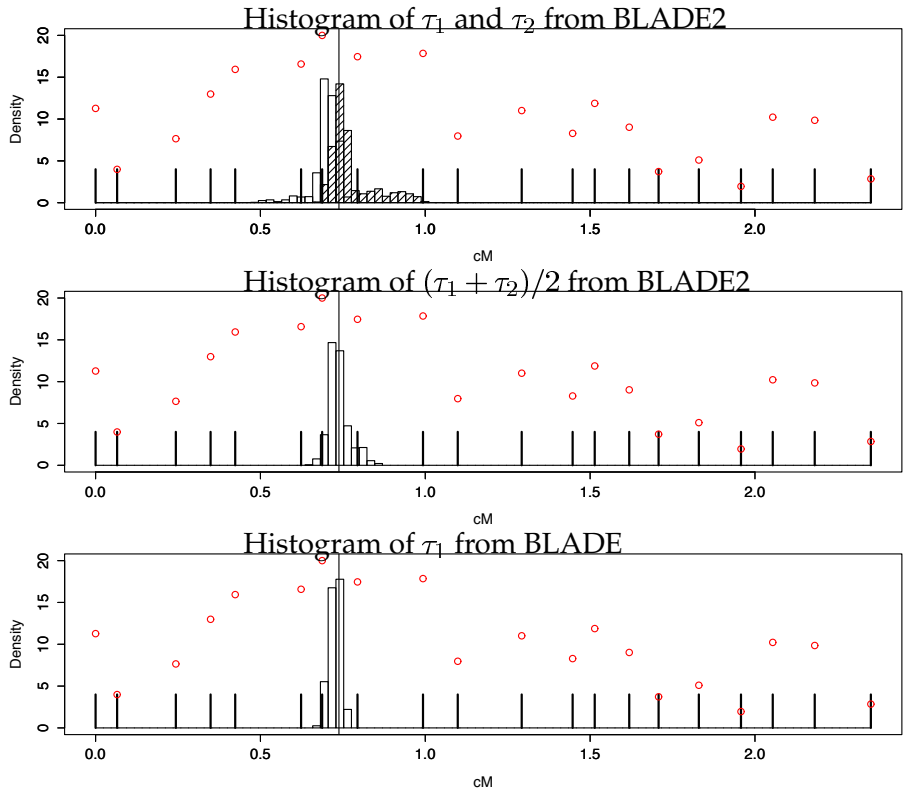

Figure 3.3: Histograms of the Mutation Loci for Simulation Study CASE 0.
The top plot is the histogram of τ_1 and τ_2 from the BLADE2 output. Note that the two variables are sampled around the input location marked by the vertical line at 0.738 cM. The empty bars are the histogram of τ_1 and the shaded bars are the histogram of τ_2. Remember the restriction that $\tau_1 < \tau_2$. The middle plot is the histogram of $(\tau_1+\tau_2)/2$ from the BLADE2 output. The bottom plot is the histogram of τ_1 from the BLADE output. The long vertical line at 0.738 cM is the input value of the mutation location for data generation. The thick short vertical lines at each histogram represent the locations of 20 markers in the data set. The dots are single marker measures of disequilibrium (δ) at all marker locations. One can roughly guess that the disease locus is around 7,8-th markers based on the δ-values.

Algorithm	Variable	95% Interval	Median	Mean	Root.Mse
BLADE2	τ_1	(0.584, 0.750)	0.704	0.701	0.055
	τ_2	(0.702, 0.958)	0.748	0.773	0.077
	$(\tau_1 + \tau_2)/2$	(0.690, 0.819)	0.731	0.737	0.032
BLADE	τ	(0.691, 0.756)	0.726	0.725	0.021

Table 3.1: Summary of the Posterior Draws for the Mutation Loci from Simulation Study CASE 0.
All intervals contain the input value for the mutation location, 0.738 cM away from the leftmost marker. Root.Mse is computed by $\sqrt{\sum(\{\tau_1\}_i - 0.738)^2}/(\text{the number of MCMC draws})$.

As illustrated in the first histogram in Figure 3.3, the posterior draws of τ_1 and τ_2 from the BLADE2 result are close to the true value, as we expected. No draws wander more than two markers away from the true location and in about 60% of the draws, the true location is contained between the two parameters. This is a very strong indication that the data set has only one disease mutation. Therefore, we decided to use $(\tau_1 + \tau_2)/2$ as a summary statistics and the histogram is shown at the second plot in Figure 3.3. The new statistics does a better job than using either τ_1 or τ_2 only. The Root MSE value is much smaller than the previous cases. We also applied BLADE to this data provided that we knew that there is only one mutation locus. The histogram is shown at the third plot in Figure 3.3. Among the three histrograms, BLADE did the best job according to Root MSE values, but only marginally better than the method by $(\tau_1 + \tau_2)/2$ from BLADE2. However, one should note that BLADE is applicable only under the assumption that the disease haplotypes have one mutation locus and the smaller variance (or MSE) on the estimate is benefiting by this assumption. We will examine the case when this assumption is not valid.

Now, consider CASE 1, where the disease occur only when the both mutations are conserved from the founder. Single marker measures of disequilibrium value δ predicts that the disease locus is near 11-th marker althogh its value is not significantly higher that the others. The single marker approaches does a poor job unlike CASE 0 Using BLADE for such data set can also be dangerous, especially when the sampler can not correctly estimate either of the mutation. Figure 3.4 shows the histograms of the posterior draws by the two algorithms. It is clear from Figure 3.4 that BLADE does a poor job at estimating the mutation loci, whereas BLADE2 does a superior job at finding the mutation loci. Similar results were obtained when we used the BLADE algorithm for two locus data sets. We summarized the posterior inferences of the other simulation cases in Table 3.2.

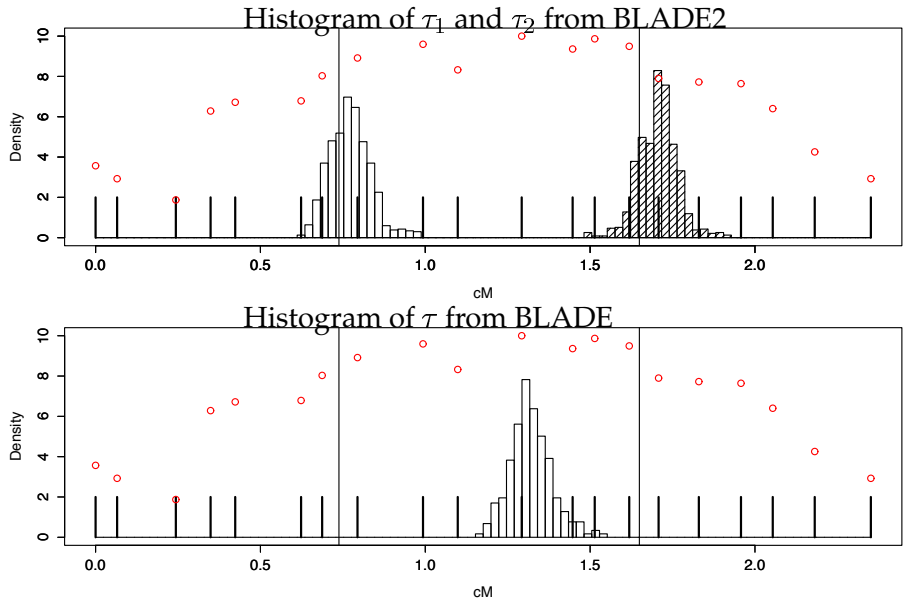

Figure 3.4: Histograms of the Mutation Loci for Simulation Study CASE 1.
The top plot is the histogram of τ_1 and τ_2 from the BLADE2 output. The empty bars are the histogram of τ_1 and the shaded bars are the histogram of τ_2. The bottom plot is the histogram of τ from the BLADE output. The long vertical lines at 0.738 cM and 1.649 cM are the input values of the mutation locations for data generation. The thick short vertical lines at each histogram represent the locations of 20 markers in the data set. It is clear that the estimate by BLADE is off the target, however, the BLADE2's posterior draws contain the true value in their 95% posterior intervals. The dots are single marker measures of disequilibrium (δ) at all marker locations, whose values are higher around 11-th marker. Using the single marker measure or the BLADE algorithm for two marker mutation data sets can be dangerous.

	τ_1 (0.738)				τ_2(1.649)			
Case	2.5%	97.5%	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	Mean	$\sqrt{\text{MSE}}$
CASE 1	0.669	0.916	0.771	0.069	1.589	1.825	1.706	0.082
CASE 2	0.564	0.885	0.731	0.073	1.559	1.855	1.711	0.099
CASE 3	0.733	1.094	0.877	0.165	1.590	1.858	1.701	0.086
CASE 4	0.477	0.797	0.663	0.113	1.466	1.880	1.663	0.112
CASE 5	0.731	1.009	0.882	0.159	1.555	2.029	1.655	0.103

Table 3.2: Summary of the Posterior Draws for the Mutation Loci from Various Simulation Studies.
All 95% intervals contain the input values for data generation (numbers in parentheses). $\sqrt{\text{MSE}} = \sqrt{\sum_i (\{\tau\}_i - \tau^{\text{true}})^2}/(\text{the number of MCMC draws})$.

Sensitivity to Genetic Mapping

The BLADE and BLADE2 algorithms uses homogeneous Poisson process for recombination counts, i.e., exponential decay of LD among the disease haplotypes. However, this may not be sufficient due to recent discovery of haplotype block structure. Recent studies even observed that male and female gametes have different recombination rates. (Kong *et al.*, 2004) A simple solution to this problem is to allow for inhomogeneous conversion rates between physical and genetic distances.

To evaluate the sensitivity of BLADE2 to a genetic map, we provide another case study example where the genetic map of the data is different from the one that generated the data. We analyze the same haplotype data set used for CASE 1 in the previous study with different intermarker distances. The order of the markers are same as before. The new intermarker distances are generated by (true distance)$\times \mathbf{Unif}(0, 4)$, which doubles the genetic mapping distances as well as the recombination rate on average. The top plot in Figure 3.5 shows the change from the true (data generating) genetic distance to the arbitrary genetic distance for each marker. In particular, the first mutation located between 7th and 8th marker is moved from 0.74 cM to 1.67 cM and the second mutation located between 14th and 15th marker is moved from 1.66 cM to 3.65 cM. We ran the Markov chain for 1,500 iterations and discarded the first 500 draws, for a burn-in period. The histogram of τ_1 and τ_2 appears in Figure 3.5. The result shows that the mean values of τ_1 and τ_2 miss the true value only by one marker and it still performs moderately well in terms of containing true value inside the 95% intervals even under the wrong set of genetic distances. The algorithm is not strongly sensitive to the conversion rates in this example.

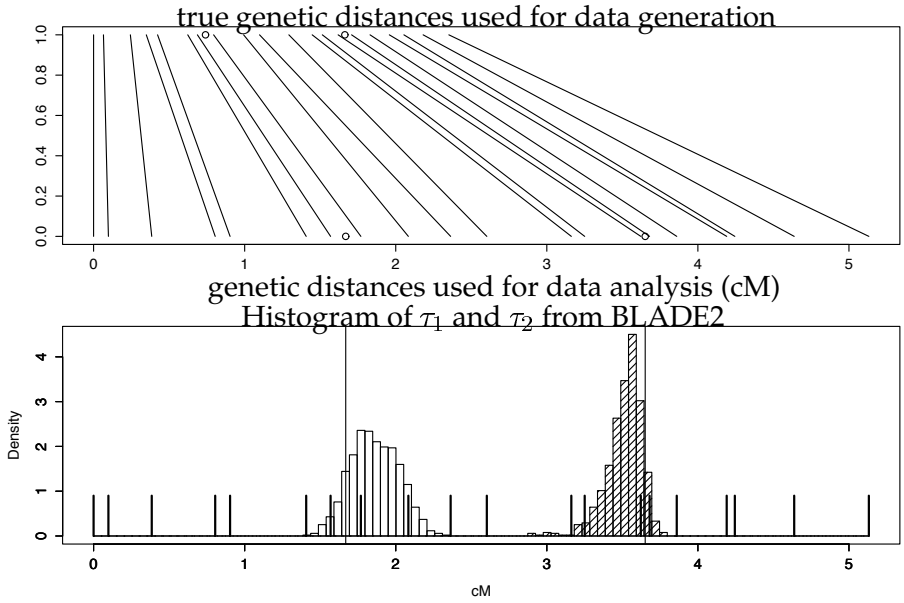

Figure 3.5: Histograms of Posterior Draws for the Location of the Disease Locus by BLADE2.
The top plot shows the change from the true genetic map to an arbitrary genetic map. Each line connects the marker location of each marker before and after the arbitrary conversion. In the first panel, the top ends of the lines represent the true marker distances used for data generation and the bottom ends of the lines represent the marker distances used for data analysis. The two dots on each end are the target disease mutation locations. In the bottom plot, the empty bars are for the histogram of τ_1 and the shaded bars are for the histogram of τ_2. The long vertical lines mark the target locations for the founder mutations. The thick short vertical lines at the bottom of histograms represent the locations of 20 markers in the simulated data set under the new genetic map. BLADE2 performs well in terms of locating nearby markers even under the wrong set of genetic distances in this example.

3.5.2 Real Data Example

Cystic Fibrosis Data Set.

The CF data set (Kerem *et al.*, 1989) contains haplotypes on 23 bi-allelic markers around the CF transmembrane conductance regulator gene on chromosome 7q31.2. The control group has 92 haplotypes and the disease group has 94. The founder mutation, ΔF_{508}, is located between markers 17 and 18, $\sim$0.88 cM away from the leftmost marker. By modeling the control haplotypes as an inhomogeneous Markov chain, BLADE gave a very accurate location estimate for the disease mutation. The posterior mean was 0.88 cM and the 95% posterior probability interval for the location was [0.82, 0.93] cM (Liu *et al.*, 2001). We apply the BLADE2 algorithm to the same data set and expect to see the draws of τ_1 and τ_2 are practically identical.

As illustrated in the top plot of Figure 3.6, about 75% of the overall range of the draws of τ_2 overlaps with the range of the draws of τ_1 and about 72% of the overall range of the draws of τ_1 overlaps with the range of the draws of τ_2. This agrees with the well-known fact that there is a single founder mutation at $\sim$0.88 cM in the Cystic Fibrosis data set. For more detailed analysis of CF data by BLADE, see Liu *et al.* (2001).

3.6 Discussion

LD mapping provides a powerful method for fine-structure localization of rare disease genes, but has not yet been widely applied to common diseases. It has become increasingly recognized that due to a low penetrance, a single genetic variant is often insufficient to lead to the manifestation of a common disease. Previous studies (Liu *et al.*, 2001; McPeek and Strahs, 1999; Morris *et al.*, 2002) have shown that con-

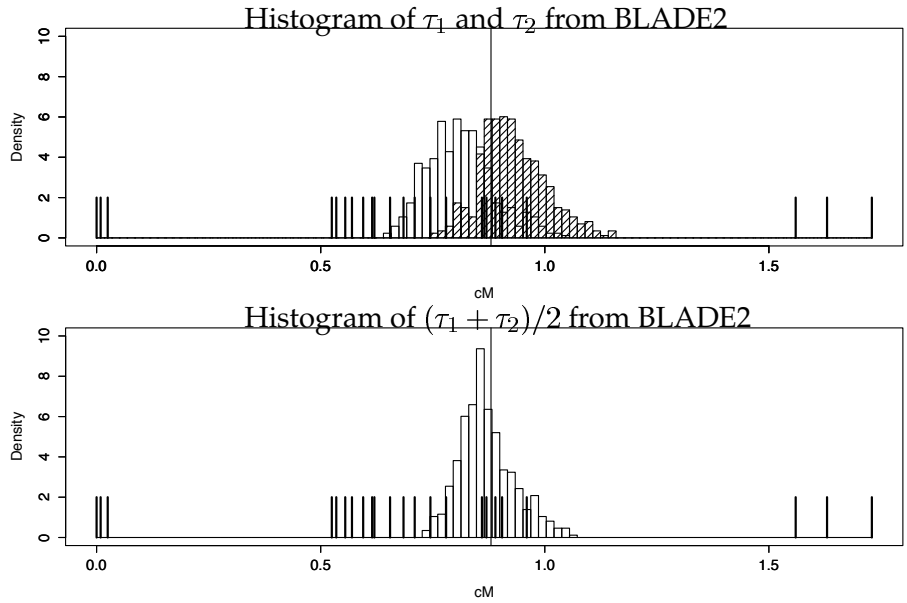

Figure 3.6: Histograms of Posterior Draws for the Location of the Disease Locus for Cystic Fibrosis by BLADE2.
The haplotype The top plot is the histogram of τ_1 and τ_2 from the BLADE2 output. The empty bars are the histogram of τ_1 and the shaded bars are the histogram of τ_2. There is a wide range of overlap between the two histogram, which indicates that there is one mutation locus in the data set. The bottom plot is the histogram of $(\tau_1 + \tau_2)/2$. The long vertical marks the known location for the founder mutation. The thick short vertical lines at the bottom of histograms represent the locations of 23 markers in the data set.

sidering the entire haplotype leads to more robust estimates for a *single* disease location. However, two genetic variants residing within the same genomic region that occur simultaneously can have a significant impact on either a continuous or a dichotomized trait. This is precisely what BLADE2 is targeting at, which has never been addressed by any previous fine-mapping methodologies. By simulation studies, we showed that considering two locus model as a preliminary analysis can help users to decide whether to use one-locus model (BLADE) or two-locus model (BLADE2), and prevent users from being misled by inappropriate model. Indeed, when one-locus model is used for a data set with two mutation loci, the estimate of the mutation location is not close to either of the two loci but somewhere in the middle of the two locations, and it is even more dangerous that the posterior probability is quite high at this wrong location. See the second histogram in Figure 3.4. Furthermore, we showed that BLADE2 did a robust estimation when the disease haplotypes have one mutation locus.

Future works includes applying BLADE2 to real disease study where the disease is believed to be caused by the two variants. An example for a continuously distributed trait affected by multiple variants is the plasma triglyceride (TG) level. It is shown that the simultaneous occurrence of two variant alleles for a SNP pair (-93T/G and D9N) located at the lipoprotein lipase gene is associated with significantly lower plasma TG levels, whereas the simultaneous presence of two variant alleles for two linked loci (-1208-1209TTdel and A455V) in the thrombomodulin gene are associated with significantly higher plasma TG levels (Konstantoulas *et al.*, 2004; Talmud *et al.*, 1998). An intriguing example for a dichotomized trait is Crohn disease (Peltekova *et al.*, 2004) - The locus was first mapped to a large region spanning 18 cM of chromosome 5q31 by a genome-wide scan (LOD score=3.90) (Rioux *et al.*, 2000), which was further narrowed down to a 983-kb region by fine-scale LD mapping (Rioux et al. 2001). Although 11 SNPs bounded within a 250-kb region appear to be associated with the disease, Rioux *et al.* (2001) could not further

pinpoint where the disease-associated mutations are located. Later, Peltekova *et al.* (2004) found that L503F (located in a transmembrane domain) in SLC22A4 gene exon 9 and G-207C (located within a heat-shock transcription factor-binding element) in the SLC22A5 gene promoter, which are 50 kb (≈0.05 cM) apart, confer an increased risk for Crohn disease. Besides Crohn disease, concomitant occurrence of two linked variants, TNF-alpha-308*2 and LT-alpha-NcoI*1 polymorphisms, is associated with atopic asthma (Wang *et al.*, 2004), and concomitant presence of Pro12Ala and C1431T variants of PPARG is significantly associated with type 2 diabetes (Doney *et al.*, 2004). Given that all the above continuous and binary disease examples conform to the double-mutation genetic model, BLADE2 will have a significant implication for localizing the positions for dual causal mutations in real world datasets.

Currently, BLADE2 requires multiple (at least 10) genetic loci across the region of interest to take full advantage of linked haplotypes. Thus, one caveat for the application of BLADE2 is that the practitioners are pressed to both identify and genotype a dense map of markers in both diseased and non-diseased samples within the boundaries of the genomic region of interest. With the advent of an ultra-fine-scale human HapMap (The International HapMap Consortium, 2003), such dense marker map is becoming available and once its value is fully acknowledged, we may witness a surging need for our proposed approach described in this paper in genetic association studies of complex human diseases.

Finally, the homogeneous Poisson process assumption should be examined further in light of recent discussions of haplotype blocks and recombination hotspots in which crossing-over events cluster (McVean *et al.*, 2004). Larger genetic to physical distance conversion, i.e., greater than 1cM / 1MB, should be applied when recombination hotspots are believed to exist between markers. Estimating accurate conversion rate and sensitivity of BLADE2 to different conversion rate deserve more attention.

We would like to thank Dr. Tianhua Niu for his thoughtful comments and discussions.

References

Ardlie, K. G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics* 299–309.

Doney, A., Fischer, B., Cecil, J., Boylan, K., McGuigan, F., Ralston, S., Morris, A., and Palmer, C. (2004). Association of the pro12ala and c1431t variants of pparg and their haplotypes with susceptibility to type 2 diabetes. *Diabetologia* **47**, 555–558.

Florez, J., Burtt, N., de Bakker, P., Almgren, P., Tuomi, T., Holmkvist, J., Gaudet, D., Hudson, T., Schaffner, S., Daly, M., Hirschhorn, J., Groop, L., and Altshuler, D. (2004). Haplotype structure and genotype-phenotype correlations of the sulfonylurea receptor and the islet atp-sensitive potassium channel gene region. *Diabetes* **53**, 1360–1368.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgin, s. J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229.

Hartl, D. L. and Jones, E. W. (1998). *Genetics*. Jones and Bartlett Publishers, Sudbury.

Hastings, W. K. (1970). Monte Carlo sampling methods usings Markov chains and their applications. *Biometrika* **57**, 97–109.

Kerem, B., Rommens, J., Buchanan, J., Markiewicz, D., Cox, T., Chakravarti, A., Buchwald, M., and Tsui, L. (1989). Identification of the cystic fibrosis gene: Genetic analysis. *Science* **245**, 1073–1080.

Kong, A., Barnard, J., Gudbjartsson, D., Thorleifsson, G., Jonsdottir, G., Sigurdardottir, S., Richardsson, B., Jonsdottir, J., Thorgeirsson, T., Frigge, M., Lamb, N., Sherman, S., Gulcher, J., and Stefansson, K. (2004). Recombination rate and reproductive success in humans. *Nature Genetics.* **36**, 1203–1206.

Konstantoulas, C., Cooper, J., Warnock, G., Miller, G., Humphries, S., and Ireland, H. (2004). A combination of two common thrombomodulin gene variants (-1208-1209ttdeltt and a455v) influence risk of coronary heart disease: a prospective study in men. *Atherosclerosis* **177**, 97 – 104.

Liu, J. S., Sabatti, C., Teng, J., Keats, B. J. B., and Risch, N. (2001). Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Research* **11**, 1716–1724.

McPeek, M. S. and Strahs, A. (1999). Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *American Journal of Human Genetics* **65**, 858–875.

McVean, G., Myers, S., Hunt, S., Deloukas, P., Bentley, D., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Morris, A. P., Whittaker, J. C., and Balding, D. J. (2002). Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *American Journal of Human Genetics* **70**, 686–707.

Peltekova, V., Wintle, R., Rubin, L., Amos, C., Huang, Q., Gu, X., Newman, B., Van, O. M., Cescon, D., Greenberg, G., Griffiths, A., St George-Hyslop, P., and

Siminovitch, K. (2004). Functional variants of octn cation transporter genes are associated with crohn disease. *Nature Genetics* **36**, 471–475.

Rioux, J., Daly, M., Silverberg, M., Lindblad, K., Steinhart, H., Cohen, Z., Delmonte, T., Kocher, K., Miller, K., Guschwan, S., Kulbokas, E., O'Leary, S., Winchester, E., Dewar, K., Green, T., Stone, V., Chow, C., Cohen, A., Langelier, D., Lapointe, G., Gaudet, D., Faith, J., Branco, N., Bull, S., McLeod, R., Griffiths, A., Bitton, A., Greenberg, G., Lander, E., Siminovitch, K., and Hudson, T. (2001). Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nature Genetics* **29**, 223–228.

Rioux, J., Silverberg, M., Daly, M., Steinhart, A., McLeod, R., Griffiths, A., Green, T., Brettin, T., Stone, V., Bull, S., Bitton, A., Williams, C., Greenberg, G., Cohen, Z., Lander, E., Hudson, T., and Siminovitch, K. (2000). Genomewide search in canadian families with inflammatory bowel disease reveals two novel susceptibility loci. *American Journal of Human Genetics* **66**, 1863–1870.

Talmud, P., Hall, S., Holleran, S., Ramakrishnan, R., Ginsberg, H., and Humphries, S. (1998). Lpl promoter -93t/g transition influences fasting and postprandial plasma triglycerides response in african-americans and hispanics. *Journal of Lipid Research* **39**, 1189–1196.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528–550.

The International HapMap Consortium (2003). The international hapmap project. *Nature* **426**, 789–796.

Wang, T., Chen, W., Wang, T., Chen, C., Huang, L., and Ko, Y. (2004). Gene-gene synergistic effect on atopic asthma: tumour necrosis factor-alpha-308 and lymphotoxin-alpha-ncoi in taiwan's children. *Clinical & Experimental Allergy* **34**, 184–188.