Methods in Monte Carlo Computation, Astrophysical Data Analysis and Hypothesis Testing With Multiply-Imputed Data

Methods in Monte Carlo Computation, Astrophysical Data Analysis and Hypothesis Testing with Multiply-Imputed Data

A dissertation presented
by

Lazhi Wang

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
April 2015

Dissertation Advisor: Professor Xiao-Li Meng **Lazhi Wang**

Methods in Monte Carlo Computation, Astrophysical Data Analysis and Hypothesis Testing with Multiply-Imputed Data

Abstract

We present three topics in this thesis: the next generation warp bridge sampling, Bayesian methods for modeling source intensities, and large-sample hypothesis testing procedures in multiple imputation.

Bridge sampling is an effective Monte Carlo method to estimate the ratio of the normalizing constants of two densities. The Monte Carlo errors of the estimator are directly controlled by the overlap between the densities. In Chapter 1, we generalize the warp transformations in Meng and Schilling (2002), and introduce a class of stochastic transformation, called warp-U transformation, which aims at increasing the overlap of the densities of the transformed data without altering the normalizing constants. Warp-U transformation is determined by a Gaussian mixture distribution, which has reasonable amount of overlap with the density of unknown normalizing constant. We show warp-U transformation reduces the f-divergence of two densities, thus bridge sampling with warp-U transformed data has better statistical efficiency than that based on the original data. We then propose a computationally efficient method to find a Gaussian mixture distribution and investigate the performance of the corresponding warp-U bridge sampling. Finally, theoretical and simulation results are provided to shed light on how to choose the tuning parameters in the algorithm.

In Chapter 2, we propose a Bayesian hierarchical model to study the distribution of the X-ray intensities of stellar sources. One novelty of the model is its use of a zero-inflated gamma distribution for the source intensities to reflect the possibility of “dark” sources with practically zero luminosity. To quantify the evidence for “dark” sources, we develop a Bayesian hypothesis testing procedure based on the posterior predictive p-value. Statistical properties of the model and the test are investigated via simulation. Finally, we apply our method to a real dataset from Chandra.

Chapter 3 presents large-sample hypothesis testing procedures in multiple imputation, a common practice to handle missing data. Several procedures are classified, discussed, and compared in details. We also provide an improvement of a Wald-type procedure and investigate a practical issue of the likelihood-ratio based procedure.

Contents

Acknowledgments

I would like to thank my advisor Xiao-Li Meng for introducing me to the world of statistics and for guiding me through my entire graduate studies. His encouragement, enthusiasm, patience, and wisdom have not only pushed me to think deeply and critically, but also motivated me to be a better researcher, collaborator, and presenter. During these years, he has provided me with great inspiration, valuable feedback, and generous support for my research. In spite of his heavy duties as the Dean of Harvard Graduate School of Arts and Sciences, he has always been able to make time, even on weekends and the Christmas Day, to discuss and share his ideas with me. It has been my great privilege to work with him for nearly five years, and I owe him my utmost gratitude.

I would also like to take this opportunity to express my sincere thanks to Vinay Kashyap, David van Dyk, and Andreas Zezas for their invaluable contribution to my research in applying statistics to Astrophysical data. Their suggestions were extremely constructive and important for the development of the research and the writing in Chapter 2. I am also grateful to all the members of the International CHASC Astro-Statistics Collaboration for providing me with an intellectual stimulating environment and insightful comments .

I am very thankful to have Luke Bornn on my dissertation committee. The inspirational discussions in his seminars and classes opened my eyes to the fields of spacial statistic and Bayesian nonparametrics. I am incredibly fortunate to have worked with him on the project on structural health monitoring, during which I was greatly motivated by his energy, insights, and determination. I would like to extend my gratitude to Anthony Liu, who has put a great deal of effort into that project.

I would also like to take this moment to acknowledge all other members of the faculty in the Statistics Department, especially Joseph Blitzstein, Carl Morris, Jun Liu, Samuel Kou, Natesh Pillai, Donald Rubin, and Tirthankar Dasgupta, whose courses have shaped my understanding of statistics. I also thank the staff of the department, Betsey Cogswell, Alice Moses, James Matejek, and Maureen Stanton, for their constant help to make the journey smooth.

My experience would not have been as rich and enjoyable without the support and friendships of my fellow colleagues, David Jones, Peng Ding, Sobambo Sosina, Daniel Cervone, Yang Chen, Xufei Wang, Samuel Wong, Xiaojin Xu, Nathan Stein, Alexander Franks, Alex Blocker, Viviana Garcia-Horton, Jiannan Lu, among many others. Special thanks go to David Jones and Peng Ding for their help during the preparation for the qualifying examination and my entire graduate career.

Last but not least, I thank my family, especially my parents, for their unreserved patience, understanding, and love. They have always been the strength to lead me through adversity, and I dedicate my thesis to them.

To my family.

Chapter 1

Warp Bridge Sampling: the Next Generation

1.1 Motivations and Applications

MCMC methods enable us to simulate data from an unnormalized density without knowing the normalizing constant. However, in scientific and statistical studies, many problems are formulated as (ratios of) normalizing constants.

An example in physics and chemistry is the partition function, which describes the statistical properties of a system in thermodynamic equilibrium. It is the integral of the system density $q(\omega; T, v) = \exp\left(\frac{-H(\omega,v)}{kT}\right)$, where T is the temperature, k is the Boltzmann's constant, v is a vector of system characteristics, and $H(\omega, v)$ is the energy function. Because of the high dimentionality of the energy function, Monte Carlo methods are often used to estimate the integral (see, for example, Bennett, 1976; Ceperley, 1995; Voter and Doll, 1985).

Another example is the computation of the observed-data likelihood, $L(\Theta; Y_{\text{obs}})$, in the presence of massive missing data. More specifically, $L(\Theta; Y_{\text{obs}})$ can be formulated as the normalizing constant of the conditional distribution of Y_{mis} given (Y_{obs}, Θ), with the complete-data distribution as the unnormalized density, i.e.,

$$L(\Theta; Y_{\text{obs}}) \triangleq P(Y_{\text{obs}}|\Theta) = \int P(Y_{\text{mis}}, Y_{\text{obs}}|\Theta)\mathbf{u}(\mathrm{d}Y_{\text{mis}}).$$

An application lies in the genetic linkage analysis, where Θ represents the locations of disease genes relative to a set of markers, Y_{obs} is the vector of genotypes of markers for some members of a pedigree, and an example of the missing information is the ellele types inherited from the parents. For a large pedigree with many loci, direct calculation of the observed-data likelihood is often prohibitive. Fortunately, it is feasible to simulate Y_{mis} from the conditional distribution, $P(Y_{\text{mis}}|Y_{\text{obs}}, \Theta)$, and to evaluate $P(Y_{\text{obs}}, Y_{\text{mis}}|\Theta)$, so researchers often resort to Monte Carlo methods to estimate the observed-data likelihood.

In addition, Monte Carlo integration is often used to estimate Bayes factors for the purpose of model selection. Let Y be the data, fitted to two plausible models M_0 and M_1, parametrized by Θ_0 and Θ_1. The Bayes factor of the two models is defined as the ratio of the model likelihoods, $P(Y|M_0)$ and $P(Y|M_1)$, where

$$P(Y|M_i) = \int P(Y|\Theta_i, M_i)P(\Theta_i|M_i)\mathbf{u}(\mathrm{d}\Theta_i)$$

is the normalizing constant of the unnormalized density, $P(\Theta_i, Y|M_i)$, of Θ_i. In most applications, Monte Carlo draws of Θ_i from its posterior distribution, $P(\Theta_i|Y, M_i)$, are made for the purpose of statistical inference. So no additional sample is needed to estimate the Bayes factor via Monte Carlo methods.

Some good reviews of the applications of estimating the (ratios of) normalizing constants can be found in Meng and Wong (1996), Gelman and Meng (1998), Shao and Ibrahim (2000), and Tan (2013).

This chapter is organized as follows. First, we provide a brief review of bridge sampling and warp bridge sampling, highlighting the power of transformation in increasing the overlap of two densities and thus reducing the Monte Carlo errors in estimating the normalizing constants. Then, we introduce a general class of stochastic transformation that can warp two densities into having substantial overlap without altering their normalizing constants. Theoretical results and simulation studies are provided to demonstrate the potential of this class of transformation. In Section 1.4, we propose a computationally efficient method to find a specific transformation in the class and study the properties of the corresponding estimator. Finally, we compare both the computation costs and the statistical efficiencies of estimators with different tuning parameters, in the hope of providing some guidance for choosing these parameters.

1.2 Literature Review: Warp Bridge Sampling

Bridge sampling (Bennett, 1976; Meng and Wong, 1996) is an effective method to estimate the ratio of the normalizing constants of two unnormalized densities. The Monte Carlo errors of the estimator depend on the amount of overlap between the two densities. Warp bridge sampling aims to reduce the Monte Carlo errors by transforming the data so that the densities of the transformed data have more overlap.

To fix the idea, for $i = 1, 2$, let q_i be the two unnormalized densities with respect to a common measure $\mathbf{u}$, each with a normalizing constant c_i. We use p_i to denote the normalized density, i.e., $p_i(\omega) = c_i^{-1} q_i(\omega)$, for $\omega \in \Omega_i$, where Ω_i is the support of q_i. We are interested in estimating the ratio of the two normalizing constants, i.e., $r = c_1/c_2$ or $\lambda = \log(r)$, with the available draws, $\{w_{i,1}, w_{i,2}, \cdots, w_{i,n_i}\}$, from p_i.

1.2.1 Bridge Sampling

Bridge sampling relies on a simple fact that for any function, α, that is defined on $\Omega_1 \cap \Omega_2$ and satisfies

$$0 < \left| \int_{\Omega_1 \cap \Omega_2} \alpha(\omega) p_1(\omega) p_2(\omega) \mathbf{u}(\mathrm{d}\omega) \right| < \infty,$$

the following identity holds,

$$r = \frac{c_1}{c_2} = \frac{\mathrm{E}_2[q_1(\omega)\alpha(\omega)]}{\mathrm{E}_1[q_2(\omega)\alpha(\omega)]}, \tag{1.1}$$

where E_i represents the expectation with respect to the density p_i. The corresponding bridge sampling estimator of r is defined by replacing the expectations in (1.1) by the sampling averages, that is,

$$\widehat{r}_\alpha = \frac{n_2^{-1} \sum_{j=1}^{n_2} q_1(w_{2,j}) \alpha(w_{2,j})}{n_1^{-1} \sum_{j=1}^{n_1} q_2(w_{1,j}) \alpha(w_{1,j})}. \tag{1.2}$$

Different choices of α correspond to estimators with different statistical efficiencies, quantified by the asymptotic variance of $\widehat{\lambda}_\alpha = \log(\widehat{r}_\alpha)$, or equivalently, the asymptotic relative variance of $\widehat{r}_\alpha$, $\mathrm{E}(\widehat{r}_\alpha - r)^2/r^2$. Meng and Wong (1996) showed that asymptotically the variance of $\widehat{\lambda}_\alpha$ can be approximated by its first-order term $(n_1+n_2)^{-1}\mathcal{V}_\alpha(p_1,p_2)$, where

$$\mathcal{V}_\alpha(p_1,p_2) = \frac{\int p_1^* p_2^* (p_1^* + p_2^*)\alpha^2 \mathbf{u}(\mathrm{d}\omega)}{\left(\int p_1^* p_2^* \alpha \mathbf{u}(\mathrm{d}\omega)\right)^2} - \frac{1}{s_1} - \frac{1}{s_2}, \tag{1.3}$$

with $s_i = n_i/(n_1+n_2)$, and $p_i^* = s_i p_i$.

The importance sampling and the geometric bridge sampling are both special cases of bridge sampling, with $\alpha_{\text{imp}} \propto 1/q_2$ and $\alpha_{\text{geo}} \propto 1/\sqrt{q_1 q_2}$, respectively. Meng and Wong (1996) showed that the asymptotic variance of the geometric bridge sampling estimator, denoted as $\widehat{\lambda}_{\text{geo}} = \log(\widehat{r}_{\text{geo}})$, is

$$\mathrm{Var}\left(\widehat{\lambda}_{\text{geo}}\right) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left\{b\left[1 - H_{\mathrm{E}}^2(p_1,p_2)\right]^{-2} - 1\right\} + o\left(\frac{1}{n_1+n_2}\right), \tag{1.4}$$

where $b = \int_{\Omega_1\cap\Omega_2}\left[p_1^*(\omega) + p_2^*(\omega)\right]\mathbf{u}(\mathrm{d}\omega) \leqslant 1$, and $H_{\mathrm{E}}(p_1,p_2)$ is the Hellinger distance between p_1 and p_2, defined as

$$H_{\mathrm{E}}(p_1,p_2) = \left[\frac{1}{2}\int\left(\sqrt{p_1(\omega)} - \sqrt{p_2(\omega)}\right)^2 \mathbf{u}(\mathrm{d}\omega)\right]^{1/2}. \tag{1.5}$$

When all the draws are independent, Meng and Wong (1996) found that the optimal choice of α in terms of minimizing the asymptotic variance of $\widehat{\lambda}_\alpha$ is

$$\alpha_{\text{opt}}(\omega) \propto \frac{1}{s_1 q_1(\omega) + r s_2 q_2(\omega)}.$$

Since α_{opt} depends on the unknown quantity r, Meng and Wong (1996) proposed an iterative sequence, defined below, that converges to the optimal bridge sampling estimator, $\widehat{r}_{\text{opt}}$,

$$\widehat{r}^{(t+1)}_{\text{opt}} = \frac{\dfrac{1}{n_2}\displaystyle\sum_{j=1}^{n_2}\left[\frac{l_{2,j}}{s_1 l_{2,j} + s_2 \widehat{r}^{(t)}_{\text{opt}}}\right]}{\dfrac{1}{n_1}\displaystyle\sum_{j=1}^{n_1}\left[\frac{1}{s_1 l_{1,j} + s_2 \widehat{r}^{(t)}_{\text{opt}}}\right]}, \tag{1.6}$$

where $l_{i,j} = q_1(w_{i,j})/q_2(w_{i,j}),$ for $i = 1, 2,$ and $j = 1, 2, \cdots, n_i$. The sequence typically converges to $\widehat{r}_{\text{opt}}$ within 10 iterations in our simulation. We define the (sample-size adjusted) Harmonic distance between p_1 and p_2 as

$$H_{\text{A}}(p_1, p_2) = \left\{ \int_{\Omega_1 \cap \Omega_2} \left[p_1^*(\omega)^{-1} + p_2^*(\omega)^{-1} \right]^{-1} \mathbf{u}(\mathrm{d}\omega) \right\}^{-1} - s_1^{-1} - s_2^{-1}. \tag{1.7}$$

Then the asymptotic variance of $\widehat{\lambda}_{\text{opt}} = \log(\widehat{r}_{\text{opt}})$ is

$$\text{Var}(\widehat{\lambda}_{\text{opt}}) = \frac{1}{n_1 + n_2} H_{\text{A}}(p_1, p_2) + o\left(\frac{1}{n_1 + n_2}\right). \tag{1.8}$$

Kong et al. (2003) formulated the Monte Carlo estimation of integrals as a statistical inference problem, where the parameter to be estimated is the non-negative measure, $\mathbf{u}$, which we purposefully choose and pretend to be unknown. Let $\{w_{i,1}, \cdots, w_{i,n_i}\}$ be independent draws from the probability measure, $P_i(\mathrm{d}\omega) = c_i^{-1} q_i(\omega)\mathbf{u}(\mathrm{d}\omega)$, for

$i = 1, \cdots, k$. Then the maximum likelihood estimator (MLE) of $\mathbf{u}$ is a discrete measure defined on the set of all the simulated data, $\boldsymbol{\Omega}_k$, i.e.,

$$\widehat{\mathbf{u}}(\{\omega\}) = \frac{n\widehat{P}(\{\omega\})}{\sum_{s=1}^{k} n_s \widehat{c}_s^{-1} q_s(\omega)},$$

where n is the total sample size, $\widehat{P}$ is the discrete measure supported on $\boldsymbol{\Omega}_k$ with equal probability, and $\widehat{c}_i$ is the MLE of c_i, which satisfies

$$\widehat{c}_i = \int q_i(\omega)\widehat{\mathbf{u}}(\mathrm{d}\omega) = \sum_{l=1}^{k}\sum_{j=1}^{n_l} \frac{q_i(w_{l,j})}{\sum_{s=1}^{k} n_s \widehat{c}_s^{-1} q_s(w_{l,j})}. \tag{1.9}$$

In the special case of $k = 2$, the MLE of c_1/c_2 is equivalent to the optimal bridge sampling estimator, $\widehat{r}_{\text{opt}}$, in (1.6). When $k > 2$, Tan (2004) shows that the MLE of $\left(\frac{c_1}{c_k}, \cdots, \frac{c_{k-1}}{c_k}\right)$ defined in (1.9) has the minimal asymptotic variance-covariance matrix among the extended bridge sampling estimators.

1.2.2 Warp Bridge Sampling

According to (1.8), the only way to further reduce the asymptotic variance of $\widehat{\lambda}_{\text{opt}}$ without making additional draws from p_1 or p_2 is to reduce the Harmonic distance between the two densities. More specifically, we can apply a transformation, $\mathcal{F}_i$, to the original data such that the unnormalized density, $\widetilde{q}_i$, of the transformed data, $\widetilde{w}_{i,j}$, has the same normalizing constant as q_i, i.e.,

$$\widetilde{w}_{i,j} = \mathcal{F}_i(w_{i,j}) \overset{\text{iid}}{\sim} \widetilde{p}_i = \frac{1}{c_i}\widetilde{q}_i,$$

where $\widetilde{p}_i$ is the normalized density of $\widetilde{w}_{i,j}$. Then, the bridge sampling estimator with the transformed data also estimates r. If the transformations lead to

$$H_{\mathrm{A}}(\widetilde{p}_1, \widetilde{p}_2) < H_{\mathrm{A}}(p_1, p_2),$$

the optimal bridge sampling estimator based on $\{\widetilde{w}_{i,1}, \cdots, \widetilde{w}_{i,n_i}\}$ will have smaller asymptotic variance than that based on $\{w_{i,1}, \cdots, w_{i,n_i}\}$.

Warp transformation, proposed by Meng and Schilling (2002), is a class of transformations that aims at reducing the distance between the resulting densities while keeping the normalizing constants unchanged.

In the remainder of this chapter, we refer the warp-$\mathcal{X}$ bridge sampling to the bridge sampling with the data and the densities in (1.2) replaced with the warp-$\mathcal{X}$ transformed data, $\{\widetilde{w}_{i,1}^{(\mathcal{X})}, \cdots, \widetilde{w}_{i,n_i}^{(\mathcal{X})}\}$, and their corresponding densities, $\widetilde{q}_i^{(\mathcal{X})}$, where the superscript ${}^{(\mathcal{X})}$ represents the type of transformation. The corresponding estimator is denoted as $\widehat{\lambda}_{\alpha}^{(\mathcal{X})} = \log(\widehat{r}_{\alpha}^{(\mathcal{X})})$ for general choices of α, and $\widehat{\lambda}_{\mathrm{opt}}^{(\mathcal{X})} = \log\left(\widehat{r}_{\mathrm{opt}}^{(\mathcal{X})}\right)$ for the optimal bridge sampling. Let $\boldsymbol{\zeta}$ be the vector of parameters that characterizes the warp-$\mathcal{X}$ transformation. It is important to note that α is typically a functional of the two densities, as in the geometric bridge sampling and the optimal bridge sampling, so in the warp-$\mathcal{X}$ bridge sampling, α may also depend on $\boldsymbol{\zeta}$.

Warp-I transformation "moves" one density "closer" to the other density to reduce their distance. Let μ be a location parameter, e.g., the difference between the means or the modes of the two densities. The transformations applied to the data are

$$\widetilde{w}_{1,j}^{(\mathrm{I})} = w_{1,j} - \mu, \quad \widetilde{w}_{2,j}^{(\mathrm{I})} = w_{2,j},$$

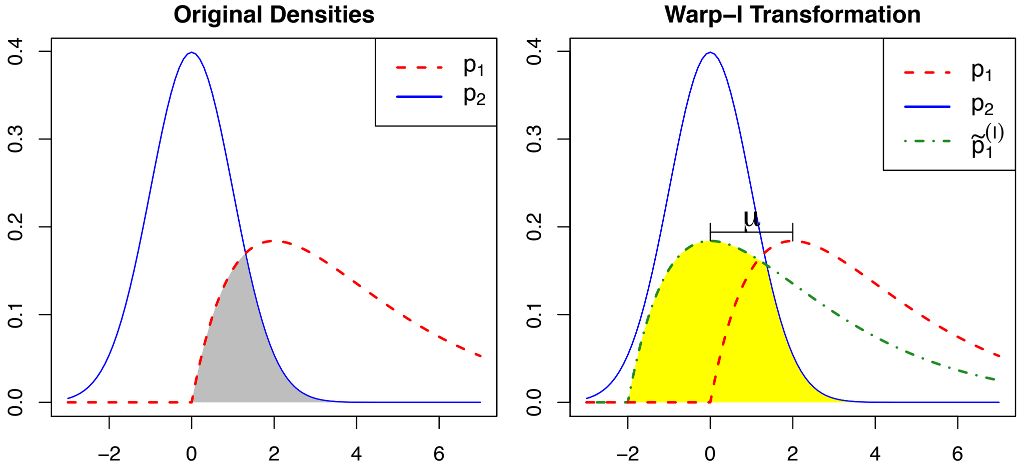

Figure 1.1: Graphical illustration of warp-I transformation. The dashed and the solid lines are the curves of p_1 and p_2. The dash-dot line is the density, $\widetilde{p}_1^{(\mathrm{I})}$, of the warp-I transformed data, obtained by moving p_1 to the left by μ units. The shaded areas are the overlap between two densities. After warp-I transformation, the overlap increases.

and the corresponding densities are

$$\widetilde{q}_1^{(\mathrm{I})}(w) = q_1(w+\mu), \quad \widetilde{q}_2^{(\mathrm{I})} = q_2.$$

The warp-I bridge sampling estimator of r is

$$\widehat{r}_\alpha^{(\mathrm{I})} = \frac{n_2^{-1}\sum_{j=1}^{n_2} \widetilde{q}_1^{(\mathrm{I})}(\widetilde{w}_{2,j}^{(\mathrm{I})})\alpha(\widetilde{w}_{2,j}^{(\mathrm{I})})}{n_1^{-1}\sum_{j=1}^{n_1} \widetilde{q}_2^{(\mathrm{I})}(\widetilde{w}_{1,j}^{(\mathrm{I})})\alpha(\widetilde{w}_{1,j}^{(\mathrm{I})})} = \frac{n_2^{-1}\sum_{j=1}^{n_2} q_1(w_{2,j}+\mu)\alpha(w_{2,j})}{n_1^{-1}\sum_{j=1}^{n_1} q_2(w_{1,j}-\mu)\alpha(w_{1,j}-\mu)}.$$

Figure 1.1 shows the densities before (left penal) and after (right penal) warp-I transformation, demonstrating the increase of the overlap.

Warp-II transformation matches both the center and the spread of the two densities to increase the amount of overlap. Let μ_i be a location parameter, $\mathcal{S}_i$ be a scaling

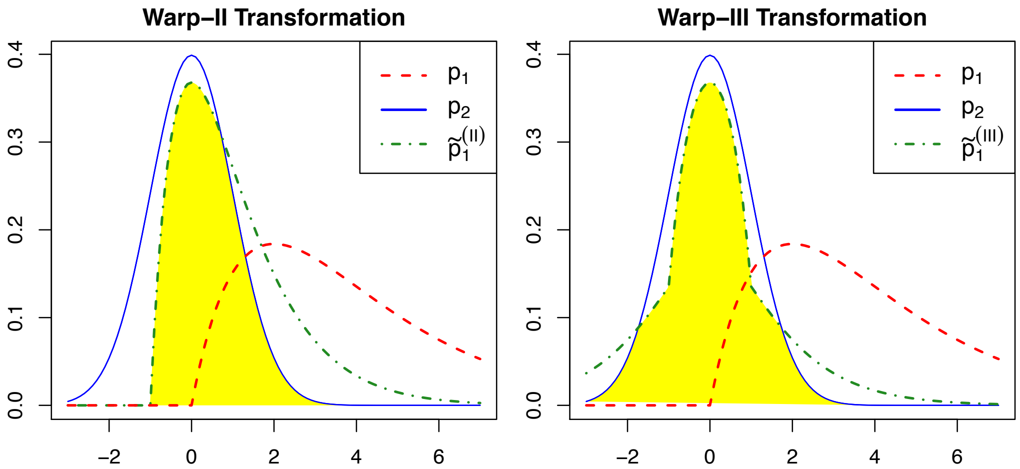

Figure 1.2: Graphical illustration of warp-II (left) and warp-III transformation (right). The dashed and the solid lines are the curves of p_1 and p_2. The dash-dot lines are $p_1^{(\mathrm{II})}$ (left) and $p_1^{(\mathrm{III})}$ (right), obtained by warp-II and warp-III transformation, respectively.

parameter, and $\boldsymbol{\zeta} = (\mu_1, \mu_2, \mathcal{S}_1, \mathcal{S}_2)$. The warp-II transformation applied to $w_{i,j}$ is

$$\widetilde{w}_{i,j}^{(\mathrm{II})} = \mathcal{S}_i^{-1}(w_{i,j} - \mu_i),$$

the unnormalized density of which is

$$\widetilde{q}_i^{(\mathrm{II})}(\omega) = |\mathcal{S}_i| q_i(\mathcal{S}_i \omega + \mu_i).$$

Then the corresponding warp-II bridge sampling estimator is

$$\widehat{r}_\alpha^{(\mathrm{II})} = \frac{n_2^{-1} \sum_{j=1}^{n_2} |\mathcal{S}_1| q_1(\mathcal{S}_1 \mathcal{S}_2^{-1}(w_{2,j} - \mu_2) + \mu_1) \alpha(\mathcal{S}_2^{-1}(w_{2,j} - \mu_2))}{n_1^{-1} \sum_{j=1}^{n_1} |\mathcal{S}_2| q_2(\mathcal{S}_2 \mathcal{S}_1^{-1}(w_{1,j} - \mu_1) + \mu_2) \alpha(\mathcal{S}_1^{-1}(w_{1,j} - \mu_1))}.$$

The dash-dot curve in Figure 1.2 (left) is an example of $\widetilde{p}_1^{(\mathrm{II})}$, which has more overlap with p_2 than p_1 or $\widetilde{p}_1^{(\mathrm{I})}$.

One common feature of warp-I and warp-II transformations is that the transformation $\mathcal{F}_i$ is a deterministic function that maps each $w_{i,j}$ to a unique value. Warp-III transformation, on the other hand, is a stochastic transformation that maps $w_{i,j}$ to different values with certain probabilities to induce symmetry. The warp-III transformation applied to $w_{i,j}$ is

$$\widetilde{w}_{i,j}^{(\mathrm{III})} = b_{i,j}\mathcal{S}_i^{-1}(w_{i,j} - \mu_i), \tag{1.10}$$

where $b_{i,j}$ takes -1 and 1 with equal probability. The unnormalized density of $\widetilde{w}_{i,j}^{(\mathrm{III})}$ is

$$\widetilde{q}_i^{(\mathrm{III})}(\omega) = \frac{|\mathcal{S}_i|}{2}\left[q_i\left(\mu_i - \mathcal{S}_i\omega\right) + q_i\left(\mu_i + \mathcal{S}_i\omega\right)\right]. \tag{1.11}$$

Figure 1.2 (right) shows an example of $\widetilde{p}_1^{(\mathrm{III})}$ (the dash-dot curve). More explanations and examples of warp I-III transformations can be found in Meng and Schilling (2002).

1.3 Warp-U Bridge Sampling

1.3.1 Setup of the Problem

The warp I-III transformations are very effective in increasing the overlap of the two densities and thus reducing the asymptotic variance of the bridge sampling estimator, especially when both densities are unimodal. When one or both densities have multiple modes, the power of these transformations is limited. In this section, we generalize warp-III transformation and introduce warp-U transformation, which is capable of “bringing” all the modes of a distribution together and forming a unimodal distribution (and hence the designation "warp-U").

We focus on the estimation of one normalizing constant and choose the other density to be a well-known density, ϕ, such as the standard normal distribution or t-distribution. The reasons are the following. First, the goal of the transformation is clear, that is, to transform the data so that the corresponding density will be "close" to ϕ. Second, the ratio of two normalizing constants can be obtained by estimating the two constants separately, which also bypasses the problem of different dimensionalities of the two densities in bridge sampling (Chen and Shao, 1997). Finally, if both datasets are transformed to have substantial overlap with a common distribution, ϕ, the resulting densities would also have substantial overlap with each other, thus the ratio can also be well estimated by the bridge sampling estimator with the two transformed datasets. We'll see an example of this in Section 1.5.4.

For clarification, we redefine the notations here. Let q be the unnormalized density of a continuous distribution, and $\{w_1, \cdots, w_n\}$ be n independent draws from $p = \frac{1}{c}q$, where c is the normalizing constant of q. To estimate c via bridge sampling, we choose the other distribution, ϕ, to be a well-known and normalized density, and make m independent draws, $\{z_1, \cdots, z_m\}$, from it. Warp-U transformation aims at reducing the divergence between ϕ and $\widetilde{p}$, the density of the warp-U transformed data $\widetilde{w}_j$, and thus reducing the asymptotic variance of the corresponding warp-U bridge sampling estimator $\widehat{\lambda}_{\alpha}^{(\mathrm{U})} = \log(\widehat{r}_{\alpha}^{(\mathrm{U})})$.

The density ϕ should be easy to evaluate, easy to simulate from, and typically symmetric and unimodal. Examples include the standard normal distribution, t-distribution and Laplace distribution. For concreteness, we assume ϕ to be the density of the standard normal distribution throughout the paper. The conclusions in this section also hold for any other choices of ϕ.

1.3.2 Intuition of Warp-U Transformation

Warp-U transformation is determined by a Gaussian mixture distribution, i.e.,

$$\phi_{\text{mix}}(\omega; \boldsymbol{\zeta}) = \sum_{k=1}^{K} \phi^{(k)}(\omega) = \sum_{k=1}^{K} \pi_k \left|\mathcal{S}_k\right|^{-1} \phi\left(\mathcal{S}_k^{-1}(\omega - \mu_k)\right), \tag{1.12}$$

where K is the number of components in ϕ_{mix}, $\phi^{(k)}$ represents the k-th component in ϕ_{mix} including its weight π_k, and $\boldsymbol{\zeta}$ is the vector of parameters, i.e., $\boldsymbol{\zeta} = (\pi_1, \cdots, \pi_K, \mu_1, \cdots, \mu_K, \mathcal{S}_1, \cdots, \mathcal{S}_K)$.

Alspach and Sorenson (1972) showed that the Gaussian sum approximation can converge uniformly to any piecewise continuous density function. So for a reasonable choice of K, we can find a ϕ_{mix} that has sufficient overlap with p. Section 1.4 discusses a computationally inexpensive method to find such a ϕ_{mix}, where the performance of the resulting warp-U bridge sampling is studied theoretically and via simulation. Before going into the details, we first assume ϕ_{mix} is known and fixed, and explain the intuition of the corresponding warp-U transformation. Figure 1.3 (left) shows the densities, p (red dashed line) and ϕ_{mix} (blue solid line), which have reasonable amount of overlap.

We explain, in Figure 1.4, how a Gaussian mixture distribution can be changed back to the standard normal distribution. The blue solid curve on the vertical plate in Figure 1.4(a) is ϕ_{mix}, which is decomposed into three components, $\phi^{(k)}$, for $k = 1, 2, 3$, corresponding to the three blue solid curves in Figure 1.4(b). Each component, $\phi^{(k)}$,

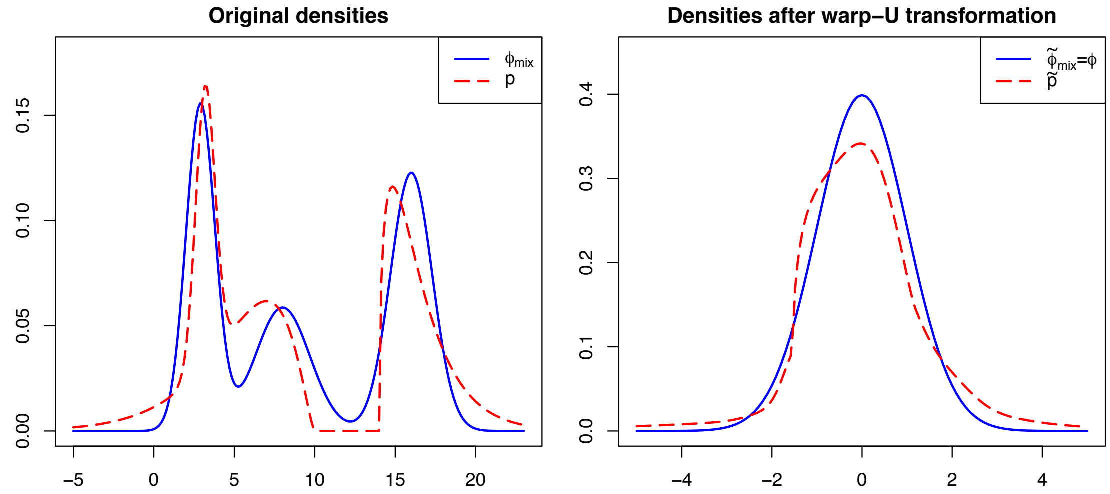

Figure 1.3: (Left) density p (dashed line) and a Gaussian mixture density ϕ_{mix} (solid line), which has substantial overlap with p; (Right) after warp-U transformation, ϕ_{mix} turns into the standard normal distribution (solid line) and p turns into $\widetilde{p}^{(\text{U})}$ (dashed line).

is moved by μ_k units to the origin and then rescaled by $\mathcal{S}_k^{-1}$, resulting in $\pi_k\phi$, as shown in Figure 1.4(d) (the blue solid curves). So after the transformation, the sum of the three components becomes the standard normal distribution.

From another prospective, if $X \sim \phi_{\text{mix}}$, then X can be represented stochastically, i.e.,

$$X = \mathcal{S}_\Theta Z + \mu_\Theta,$$

where $Z \sim \phi$, Θ is a discrete random variable with a probability mass function $P(\Theta = k) = \pi_k$ for $k = 1, 2, 3$, and Θ and Z are independent. Figure 1.4(b) shows the joint distribution of Θ and X, and their marginal distributions are on the two vertical plates. For $k \in \{1, 2, 3\}$, we define a deterministic function $\mathcal{F}_k(x; \boldsymbol{\zeta}) = \mathcal{S}_k^{-1}(x - \mu_k)$. Then, the random index Θ can induce a random transformation, $\mathcal{F}_\Theta(x; \boldsymbol{\zeta}) = \mathcal{S}_\Theta^{-1}(x - \mu_\Theta)$. By applying the random transformation to X, we obtain $\mathcal{F}_\Theta(X; \boldsymbol{\zeta}) = Z$, and

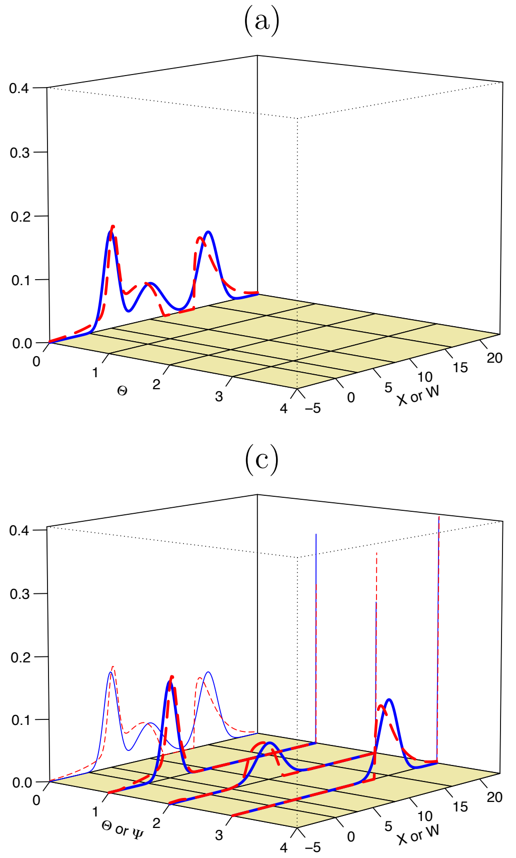

Figure 1.4: Illustration of warp-U transformation. (a) ϕ_{mix} (solid line) and p (dashed line); (b) the joint and marginal distributions of X and Θ (solid line); (c) the joint and marginal distributions of W and Ψ (dashed line); (d) the joint and marginal distributions of Θ and $\widetilde{X}$ (solid line) and those of Ψ and $\widetilde{W}$ (dashed line), where $\widetilde{X}$ and $\widetilde{W}$ are obtained via warp-U transformation.

thus we ϕ_{mix} into ϕ. In terms of data transformation, if (x_i, θ_i) is drawn from the joint distribution of (X, Θ), then $\widetilde{x}_i = \mathcal{S}_{\theta_i}^{-1}(x_i - \mu_{\theta_i})$ is a random draw from ϕ.

Now we describe how the warp-U transformation, determined by ϕ_{mix}, turns p into $\widetilde{p}$, the red dashed line in Figure 1.3. Let W be a random variable from p. To obtain the random transformation for W, we create a new random variable Ψ that serves the same purpose as Θ, i.e., to index the transformation. We define Ψ by assuming the conditional distribution of Ψ given W is,

$$\varpi(k|\omega) \triangleq P(\Psi = k|W = \omega) = \phi^{(k)}(\omega)/\phi_{\text{mix}}(\omega), \tag{1.13}$$

for $k = 1, \cdots, K$. As a result, p is also decomposed into K components, i.e., $p(\omega) = \sum_{k=1}^{K} p^{(k)}(\omega)$, where

$$p^{(k)}(\omega) = p(\omega, \Psi = k) = \phi^{(k)}(\omega)\frac{p(\omega)}{\phi_{\text{mix}}(\omega)}. \tag{1.14}$$

Figure 1.4(c) shows the joint distribution of (W, Ψ) (thick red dashed curves) and their marginal distributions (thin red dash curves in the two vertical plates).

The warp-U transformation applied to W is defined as

$$\widetilde{W} = \mathcal{F}_{\Psi}(W; \boldsymbol{\zeta}) = \mathcal{S}_{\Psi}^{-1}(W - \mu_{\Psi}) \sim \widetilde{p}. \tag{1.15}$$

To apply warp-U transformation to the data w_j, we first calculate the probability mass function $\varpi(\cdot|w_j)$ according to (1.13), then draw ψ_j from $\varpi(\cdot|w_j)$, and finally apply the deterministic transformation $\mathcal{F}_{\psi_j}(\cdot; \boldsymbol{\zeta})$ to w_j. Graphically, each $p^{(k)}$ in Figure 1.4(c) is re-centered and re-scaled, like their counterpart $\phi^{(k)}$. The red dashed lines in Figure 1.4(d) are the joint distribution of Ψ and the warp-U transformed variable, $\widetilde{W}$, the marginal distribution of which has a substantial overlap with ϕ.

Note that when $K = 1$, warp-U transformation degrades to warp-II transformation. When $K > 1$, Theorem 1 in Section 1.3.3 implies the divergence between $\widetilde{p}$ and ϕ is smaller than that between p and ϕ_{mix}, hence justifying warp-U transformation.

1.3.3 Key Theorem for Warp-U Transformation

In this section, we define the general form of warp-U transformation and show in theory the increase of overlap due to the transformation.

Let p and ϕ be the densities of any two continuous probability distributions, and $W \sim p$ and $Z \sim \phi$ be two random variables. Let π be probability distribution, and $\pi(\theta)\mathbf{u}(\mathrm{d}\theta)$ be its induced probability measure. More specifically, if π is a discrete distribution, $\mathbf{u}$ is a counting measure; if π is a continuous distribution, $\mathbf{u}$ is the Lebesgue measure. For any θ in the support of π, let $\mathcal{H}_\theta$ be a one-to-one and almost surely differentiable function indexed by θ, and $\mathcal{F}_\theta$ be its inverse function. Let Θ be a random variable from π, and it is independent of Z. We define $X = \mathcal{H}_\Theta(Z)$, which implies that the conditional distribution $X\big|\Theta = \theta$ is

$$\phi_{X|\Theta}(\omega|\theta) = \phi(\mathcal{F}_\theta(\omega)) \left|\mathcal{H}'_\theta(\mathcal{F}_\theta(\omega))\right|^{-1}, \tag{1.16}$$

and the density of X is

$$\phi_{\text{mix}}(\omega) = \int \phi_{X|\Theta}(\omega|\theta)\pi(\theta)\mathbf{u}(\mathrm{d}\theta) = \int \phi(\mathcal{F}_\theta(\omega)) \left|\mathcal{H}'_\theta(\mathcal{F}_\theta(\omega))\right|^{-1} \pi(\theta)\mathbf{u}(\mathrm{d}\theta). \tag{1.17}$$

We denote $\varpi(\cdot|\omega)$ to be the conditional distribution $\Theta|X = \omega$, which is given by

$$\varpi(\theta|\omega) = \frac{\phi_{X|\Theta}(\omega|\theta)\pi(\theta)}{\phi_{\text{mix}}(\omega)}. \tag{1.18}$$

We create a new random variable Ψ and its joint distribution with W by defining the conditional distribution to be $P(\Psi = \theta|W = \omega) = \varpi(\theta|\omega)$. Let $p_{\Psi,W}$ and $\phi_{\Theta,X}$ be the

$$\begin{tikzcd}
\boxed{\begin{array}{c} Z \sim \phi \\ \Theta \sim \pi \\ Z \perp \Theta \end{array}} \arrow[r] & \boxed{\begin{array}{c} X = \mathcal{H}_\Theta(Z) \sim \phi_{\text{mix}} \\ \Theta | X = \omega \sim \varpi(\cdot|\omega) \\ (\Theta, X) \sim \phi_{\Theta,X} \end{array}} \arrow[r] & \boxed{\widetilde{X} = \mathcal{F}_\Theta(X) \sim \phi} \\
 & \boxed{\begin{array}{c} W \sim p \\ \Psi | W = \omega \sim \varpi(\cdot|\omega) \\ (\Psi, W) \sim p_{\Psi,W} \end{array}} \arrow[r] & \boxed{\widetilde{W} = \mathcal{F}_\Psi(W) \sim \widetilde{p}}
\end{tikzcd}$$

Figure 1.5: Relationships of all the random variables and their densities in warp-U transformation.

joint distribution of (Ψ, W) and that of (Θ, X), respectively. Then,

$$p_{\Psi,W}(\theta, \omega) = \varpi(\theta|\omega)p(\omega) \tag{1.19}$$

$$\phi_{\Theta,X}(\theta, \omega) = \phi(\omega|\theta)\pi(\theta) = \varpi(\theta|\omega)\phi_{\text{mix}}(\omega), \tag{1.20}$$

and

$$\frac{p_{\Psi,W}(\theta, \omega)}{\phi_{\Theta,X}(\theta, \omega)} = \frac{p(\omega)}{\phi_{\text{mix}}(\omega)}. \tag{1.21}$$

The random transformation, $\mathcal{F}_\Psi$, applied to W,

$$\widetilde{W} = \mathcal{F}_\Psi(W) \sim \widetilde{p},$$

is called warp-U transformation. Figure 1.5 illustrates the relationships of these random variables and their densities.

Many measures can be used to quantify the discrepancy between two densities, p_1 and p_2. One of the most frequently used divergences is the f-divergence, which is defined as

$$\mathcal{D}_f(p_1, p_2) = \int p_2(\omega) f\left(\frac{p_1(\omega)}{p_2(\omega)}\right) \mathbf{u}(\mathrm{d}\omega),$$

where $f(\cdot)$ is a convex function and $f(1) = 0$. Examples of f-divergence include the squared Hellinger distance, the Kullback-Leibler divergence, and the Jeffreys divergence, the corresponding f functions being $f(x) = \left(\sqrt{x} - 1\right)^2/2$, $f(x) = -\log(x)$, and $f(x) = -(1 - x)\log(x)$, respectively. Statistical distances, such as the Hellinger distance (1.5), the Harmonic distance (1.7), and the L_p-distance ($\mathrm{p} > 0$), defined below, can also be used to assess the divergence of two densities,

$$L_\mathrm{p}(p_1, p_2) = \left[\int |p_1(\omega) - p_2(\omega)|^\mathrm{p} \mathrm{d}\omega\right]^{1/\mathrm{p}}. \tag{1.22}$$

The following theorem states warp-U transformation reduces the f-divergence between the two densities.

Theorem 1. Let $\mathcal{D}(p_1, p_2)$ be a monotonically increasing function of an f-divergence between two densities, p_1 and p_2. Then the following inequality holds,

$$\mathcal{D}(\widetilde{p}, \phi) \leqslant \mathcal{D}(p, \phi_\mathrm{mix}), \tag{1.23}$$

where $\widetilde{p}$ is the warped p with respect to ϕ, as defined by

$$\widetilde{p}(\omega) = \phi(\omega) \int \frac{p(\mathcal{H}_\theta(\omega))}{\phi_\mathrm{mix}(\mathcal{H}_\theta(\omega))} \pi(\theta) \mathbf{u}(\mathrm{d}\theta). \tag{1.24}$$

The equality in (1.23) holds if and only if Ψ and $\widetilde{W}$ are independent, e.g., when $K = 1$, or when $p = \phi_\mathrm{mix}$ almost surely.

We derive the expression of $\widetilde{p}$ in (1.24) below. Since $\widetilde{W} = \mathcal{F}_\Psi(W)$, the joint distribution of Ψ and $\widetilde{W}$ can be expressed as

$$\widetilde{p}_{\Psi,\widetilde{W}}(\theta,\omega) = p_{\Psi,W}(\theta,\mathcal{H}_\theta(\omega))|\mathcal{H}'_\theta(\omega)| = \phi(\omega)\frac{p(\mathcal{H}_\theta(\omega))}{\phi_{\text{mix}}(\mathcal{H}_\theta(\omega))}\pi(\theta), \tag{1.25}$$

where the second equation is obtained from (1.16), (1.18), and (1.19). The density of $\widetilde{W}$ in (1.24) is obtained by integrating out θ from (1.25). The proof of the inequality in (1.23) can be found in Appendix A.1.

The Hellinger distance, the Harmonic distance, the L_1 distance, and all the f-divergences satisfy the condition in Theorem 1, so the inequality in (1.23) holds for these definitions of divergence. However, the inequality does not necessarily hold for L_{p} distance when $\text{p} \neq 1$. As a simple counterexample, let $K = 1$ and $\phi_{\text{mix}}(\omega) = |\mathcal{S}|^{-1}\phi\left(\mathcal{S}^{-1}(\omega-\mu)\right)$, then $\widetilde{p}(\omega) = |\mathcal{S}|p(\mathcal{S}\omega+\mu)$. The L_{p} distance between $\widetilde{p}$ and ϕ is

$$L_{\text{p}}(\phi,\widetilde{p}) = \left(\int ||\mathcal{S}|p(\mathcal{S}\widetilde{\omega}+\mu) - \phi(\widetilde{\omega})|^{\text{p}}\,\mathrm{d}\widetilde{\omega}\right)^{1/\text{p}} = |\mathcal{S}|^{1-1/\text{p}}L_{\text{p}}(p,\phi_{\text{mix}}),$$

so the relationship between $L_{\text{p}}(\widetilde{p},\phi)$ and $L_{\text{p}}(p,\phi_{\text{mix}})$ depends on $|\mathcal{S}|$.

The transformation we discussed in Section 1.3.2 is a special case of warp-U transformation, where Θ is a discrete random variable with the point mass function $P(\Theta = k) = \pi_k$, and $\mathcal{H}_k(\omega) = \mathcal{S}_k\omega + \mu_k$.

1.3.4 Graphical Illustration of Theorem 1

In this section, we illustrate graphically how warp-U transformation increases the area of the overlapping region of two densities, defined as

$$O(p_1, p_2) = \int \min\{p_1(\omega), p_2(\omega)\} \mathrm{d}\omega = 1 - L_1(p_1, p_2)/2.$$

We also show theoretically the decrease of their L_1 distance.

Figure 1.6(a) shows a trimodal distribution p (dashed curve) and the Gaussian mixture with $K = 2$ components (solid line). As discussed in Section 1.3.2, p is decomposed into K components, denoted as $p^{(k)}$. Figure 1.6(b) shows $p^{(1)}$ (thin dashed line) and $\phi^{(1)}$ (thin solid line), as well as their overlapping region (shaded in red), and Figure 1.6(c) shows $p^{(2)}$, $\phi^{(2)}$, and their overlap (shade in yellow). The shaded region in Figure 1.6(a) is the overlap of p and ϕ_{mix}, which is exactly the sum of the shaded area in Figure 1.6(b) and that in Figure 1.6(c). This is because

$$\min\{\phi_{\text{mix}}, p\} = \phi_{\text{mix}} \min\{1, p/\phi_{\text{mix}}\} = \sum_{k=1}^{K} \phi^{(k)} \min\{1, p/\phi_{\text{mix}}\},$$

and by (1.14) or (1.21), we can replace p/ϕ_{mix} with $p^{(k)}/\phi^{(k)}$, and obtain

$$\min\{\phi_{\text{mix}}, p\} = \sum_{k=1}^{K} \min\{\phi^{(k)}, p^{(k)}\}.$$

More generally, for any function f,

$$\phi_{\text{mix}} f\left(\frac{p}{\phi_{\text{mix}}}\right) = \sum_{k=1}^{K} \phi^{(k)} f\left(\frac{p^{(k)}}{\phi^{(k)}}\right).$$

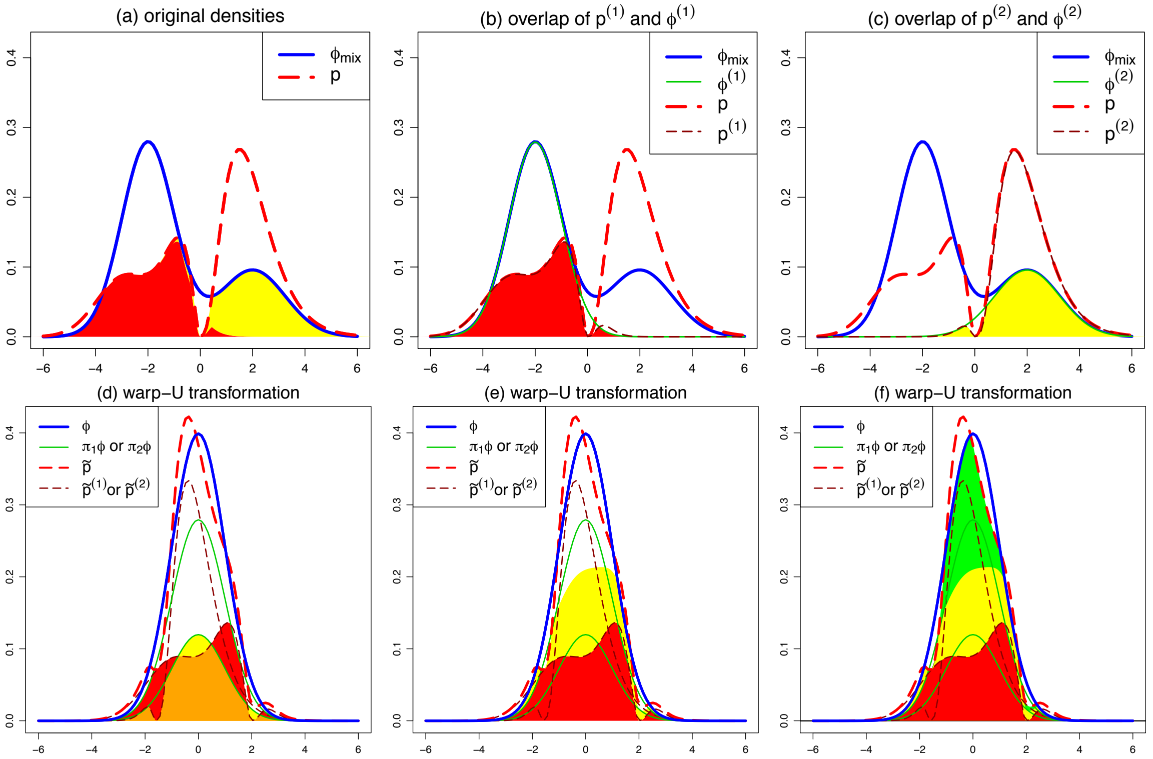

Figure 1.6: Graphical illustration of the increase in the area of the overlapping region after warp-U transformation. (a) p (dashed line) and ϕ_{mix} (solid line); (b) the first component of p, denoted as $p^{(1)}$ (thin dashed line), the first component of ϕ_{mix}, denoted as $\phi^{(1)}$ (thin solid line), and their overlap (shaded in red); (c) $p^{(2)}$, $\phi^{(2)}$, and their overlap (shaded in yellow); (d) the corresponding curves and shaded areas after warp-U transformation; (e) the yellow region is added on top of the red region; (f) the green area shows the additional overlap due to warp-U transformation.

For $k = 1, \cdots, K$, the warp-U transformation leads to the same relocating and rescaling of $\phi^{(k)}$ and $p^{(k)}$, resulting in

$$\widetilde{\phi}^{(k)}(\omega) = \pi_k \phi(\omega), \tag{1.26}$$

$$\widetilde{p}^{(k)}(\omega) = \pi_k \phi(\omega) \frac{p(\mathcal{S}_k \omega + \mu_k)}{\phi_{\text{mix}}(\mathcal{S}_k \omega + \mu_k)} = \widetilde{\phi}^{(k)}(\omega) \frac{p^{(k)}(\mathcal{S}_k \omega + \mu_k)}{\phi^{(k)}(\mathcal{S}_k \omega + \mu_k)}. \tag{1.27}$$

Figure 1.6(d) shows $\widetilde{p}^{(k)}$ (thin dashed lines), $\widetilde{\phi}^{(k)}$ (thin solid lines), and their overlap-

Table 1.1: The area of the overlapping region, the L_1 distance, the Hellinger distance, and the Harmonic distance between p and ϕ_{mix} and those between $\widetilde{p}$ and ϕ.

densities	Overlap	L_1 distance	Hellinger distance	Harmonic distance
(p, ϕ_{mix})	0.66	0.68	0.28	0.68
$(\widetilde{p}, \phi)$	0.92	0.16	0.08	0.05

ping regions (shaded in red and yellow), which remain the same as those in Figure 1.6(a). This is because by (1.27),

$$\int \widetilde{\phi}^{(k)}(\widetilde{\omega}) f\left(\frac{\widetilde{p}^{(k)}(\widetilde{\omega})}{\widetilde{\phi}^{(k)}(\widetilde{\omega})}\right) \mathrm{d}\widetilde{\omega} = \int \pi_k \phi(\widetilde{\omega}) f\left(\frac{p^{(k)}(\mathcal{S}_k\widetilde{\omega}+\mu_k)}{\phi^{(k)}(\mathcal{S}_k\widetilde{\omega}+\mu_k)}\right) \mathrm{d}\widetilde{\omega} = \int \phi^{(k)}(\omega) f\left(\frac{p^{(k)}(\omega)}{\phi^{(k)}(\omega)}\right) \mathrm{d}\omega,$$

where the second equation is obtained by replacing $\widetilde{\omega}$ with $\mathcal{S}_k^{-1}(\omega-\mu_k)$ and by the definition of $\phi^{(k)}(\omega)$.

Figure 1.6(e) combines the two shaded regions, which constitute only part of the total overlap of ϕ and $\widetilde{p}$. The additional overlap, shaded in green in Figure 1.6(f), is due to the concavity of $\min(\cdot)$, in other words, the reduction of the L_1 distance is due to the convexity of $f(\omega) = |\omega - 1|$. More specifically,

$$\sum_{k=1}^{K} \pi_k \phi(\omega) f\left(\frac{p(\mathcal{S}_k\omega+\mu_k)}{\phi(\mathcal{S}_k\omega+\mu_k)}\right) \geqslant \phi(\omega) f\left(\sum_{k=1}^{K} \pi_k \frac{p(\mathcal{S}_k\omega+\mu_k)}{\phi(\mathcal{S}_k\omega+\mu_k)}\right),$$

so by (1.27), we have

$$\int \sum_{k=1}^{K} \widetilde{\phi}^{(k)}(\omega) f\left(\frac{\widetilde{p}^{(k)}(\omega)}{\widetilde{\phi}^{(k)}(\omega)}\right) \mathrm{d}\omega \geqslant \int \phi(\omega) f\left(\frac{\widetilde{p}(\omega)}{\phi(\omega)}\right) \mathrm{d}\omega.$$

Graphically, the green region is from the overlap between $\widetilde{p}^{(k)}$ and the remainder of $\widetilde{\phi}^{(l)}$ that does not overlap with $\widetilde{p}^{(l)}$, for $k = 1, \cdots, K$, and $l \neq k$. Table 1.1 displays the overlap, the L_1 distance, the Hellinger distance, and the Harmonic distance between

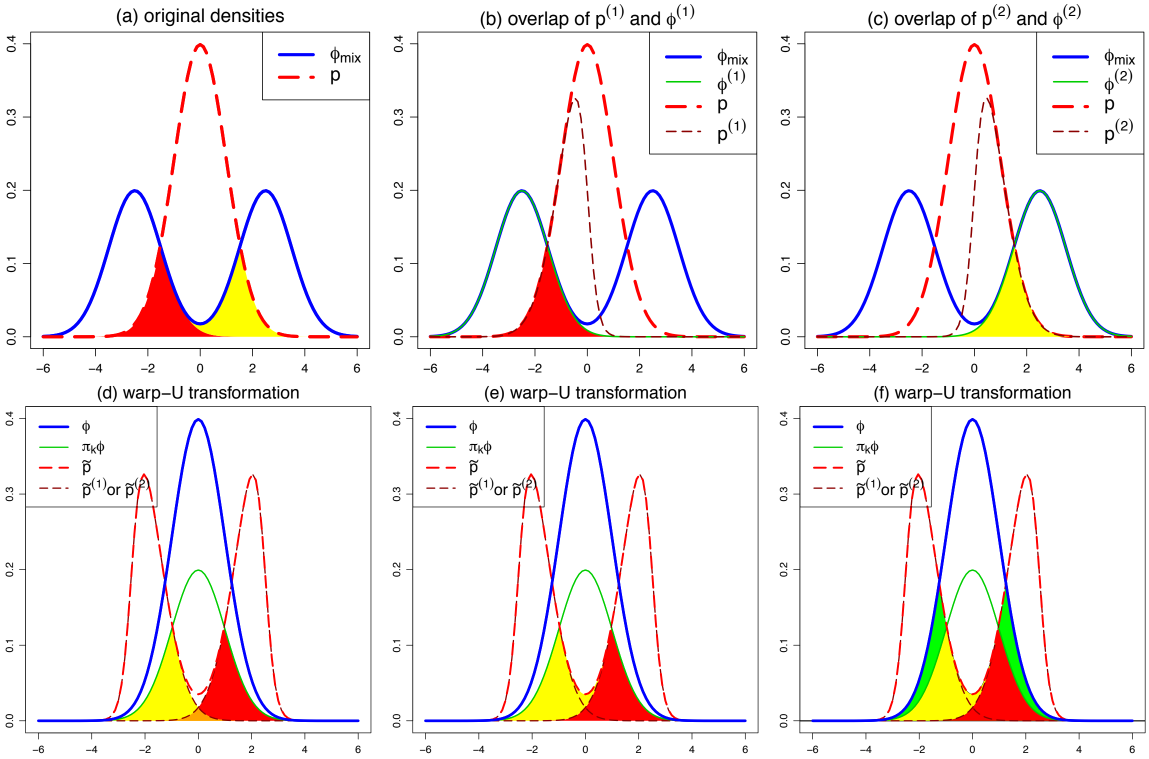

Figure 1.7: Graphical illustration of the increase of the overlap due to warp-U transformation, even the components $p^{(1)}$ and $p^{(2)}$ are moved further apart. See Figure 1.6 for more explanation.

p and ϕ_{mix} and those between $\widetilde{p}$ and ϕ. Consistent with Figure 1.6 and Theorem 1, the overlap increases and the distance decreases after the warp-U transformation.

In the example of Figure 1.6, due to the warp-U transformation, the two components of p are scaled and then moved to the origin, and the resulting density $\widetilde{p}$ is a single-modal distribution with more overlap with ϕ than that between p and ϕ_{mix}. Figure 1.7 illustrates that even if ϕ_{mix} does not match well with p and the corresponding warp-U transformation moves the components $p^{(k)}$ further apart, the inequality $O(\phi, \widetilde{p}) \geqslant O(\phi_{\text{mix}}, p)$ still holds. Figure 1.7(a) shows the one-modal density p (dashed line) and the bi-modal density ϕ_{mix} (solid line), which matches poorly with p. Figure

1.7(b) and (c) highlights $p^{(1)}$ and $p^{(2)}$ (thin dashed lines), which are moved further apart in the process of warp-U transformation, as shown in Figure 1.7(d). But warp-U transformation still results in additional overlap, highlighted in green in Figure 1.7(f).

1.3.5 Warp-U Bridge Sampling

According to Theorem 1, after the warp-U transformation defined in (1.15), the unnormalized density of $\{\widetilde{w}_1, \cdots, \widetilde{w}_n\}$ can be expressed as

$$\widetilde{q}(\omega; \boldsymbol{\zeta}) = \phi(\omega) \sum_{k=1}^{K} \frac{q(\mathcal{S}_k \omega + \mu_k)}{\phi_{\text{mix}}(\mathcal{S}_k \omega + \mu_k)} \pi_k, \tag{1.28}$$

where $\boldsymbol{\zeta}$ denotes the vector of parameters in ϕ_{mix}. By replacing q with cp, we get $\widetilde{q} = c\widetilde{p}$, meaning the normalizing constant of $\widetilde{q}$ is the same as that of q. As a result, we can estimate c with the bridge sampling estimator based on $\{z_1, \cdots, z_m\} \overset{\text{iid}}{\sim} \phi$ and $\{\widetilde{w}_1, \cdots, \widetilde{w}_n\} \overset{\text{iid}}{\sim} \widetilde{p}$, i.e.,

$$\widehat{c}_{\alpha}^{(\text{U})} = \widehat{r}_{\alpha}^{(\text{U})} = \frac{m^{-1} \sum_{j=1}^{m} \widetilde{q}(z_j; \boldsymbol{\zeta}) \alpha(z_j; \widetilde{p}, \phi)}{n^{-1} \sum_{j=1}^{n} \phi(\widetilde{w}_j) \alpha(\widetilde{w}_j; \widetilde{p}, \phi)}. \tag{1.29}$$

We emphasize that α is typically a functional of the two densities, e.g., the optimal choice of $\alpha(\cdot; \widetilde{p}, \phi)$ is proportional to $(s_1 \widetilde{p} + s_2 \phi)^{-1}$. Since ϕ_{mix} also has some overlap with p, the normalizing constant can also be estimated with the bridge sampling estimator based on $\{x_1, \cdots, x_m\} \overset{\text{iid}}{\sim} \phi_{\text{mix}}$ and $\{w_1, \cdots, w_n\} \overset{\text{iid}}{\sim} p$, i.e.,

$$\widehat{c}_{\alpha}^{(\text{mix})} = \widehat{r}_{\alpha}^{(\text{mix})} = \frac{m^{-1} \sum_{j=1}^{m} q(x_j) \alpha(x_j; p, \phi_{\text{mix}})}{n^{-1} \sum_{j=1}^{n} \phi_{\text{mix}}(w_j; \boldsymbol{\zeta}) \alpha(w_j; p, \phi_{\text{mix}})}. \tag{1.30}$$

Theorem 1 implies $D(\widetilde{p}, \phi) \leqslant D(p, \phi_{\text{mix}})$ for both the Harmonic distance and the Hellinger distance, so the asymptotic variance of $\widehat{\lambda}_{\alpha}^{(\mathrm{U})} = \log\left(\widehat{c}_{\alpha}^{(\mathrm{U})}\right)$ is smaller than that of $\widehat{\lambda}_{\alpha}^{(\mathrm{mix})} = \log\left(\widehat{c}_{\alpha}^{(\mathrm{mix})}\right)$ for both the geometric and the optimal bridge sampling.

We use a simulation to demonstrate the potential of warp-U bridge sampling by comparing it with other warp bridge sampling estimators. In this section, the vector of parameters, $\boldsymbol{\zeta}$, in the density, ϕ_{mix}, is fixed and independent of $\{w_1, \cdots, w_n\}$. For example, for fixed K, we can get a $\boldsymbol{\zeta}$ based on the expression of q, using methods such as iterative Laplace (Bornkamp, 2011) and fitting a Laplace approximation to each mode (Gelman et al., 2013, Chapter 12). The performance of warp-U bridge sampling where $\boldsymbol{\zeta}$ is estimated from draws from p is explored in Section 1.4.

The red dashed curve in Figure 1.8(a) is a tri-modal density q, the normalizing constant of which is to be estimated with $n = 1000$ i.i.d draws from it. An additional $m = 1000$ i.i.d draws are made from $\mathcal{N}(0, 1)$ to conduct bridge sampling. As shown in Figure 1.8(a), the two densities have very little overlap, and the Harmonic distance is 25.62. We apply the optimal bridge sampling algorithm in (1.6) to the $N = 10,000$ simulated replicate datasets, and obtain N vanilla optimal bridge sampling estimates of c with no transformation, denoted as $\widehat{c}_{\text{opt}}$. Figure 1.9(a) shows the histogram of $\widehat{\lambda}_{\text{opt}} - \lambda$, where $\widehat{\lambda}_{\text{opt}} = \log\left(\widehat{c}_{\text{opt}}\right)$. The root mean square error (RMSE) of $\widehat{\lambda}_{\text{opt}}$ is 0.109.

Figure 1.8(b) shows that after warp-I transformation, the overlap between the two densities increases and their Harmonic distance reduces to 4.47. The histogram of $\widehat{\lambda}_{\text{opt}}^{(\mathrm{I})} - \lambda$ is shown in Figure 1.9(b), and the RMSE reduces to 0.04. Warp-II and warp-III transformations reduce the Harmonic distance even further, as shown in Figure 1.8(c) and 1.8(d), and the RMSE of $\widehat{\lambda}_{\text{opt}}^{(\mathrm{II})}$ and $\widehat{\lambda}_{\text{opt}}^{(\mathrm{III})}$ are also reduced, see Figure 1.9(c)

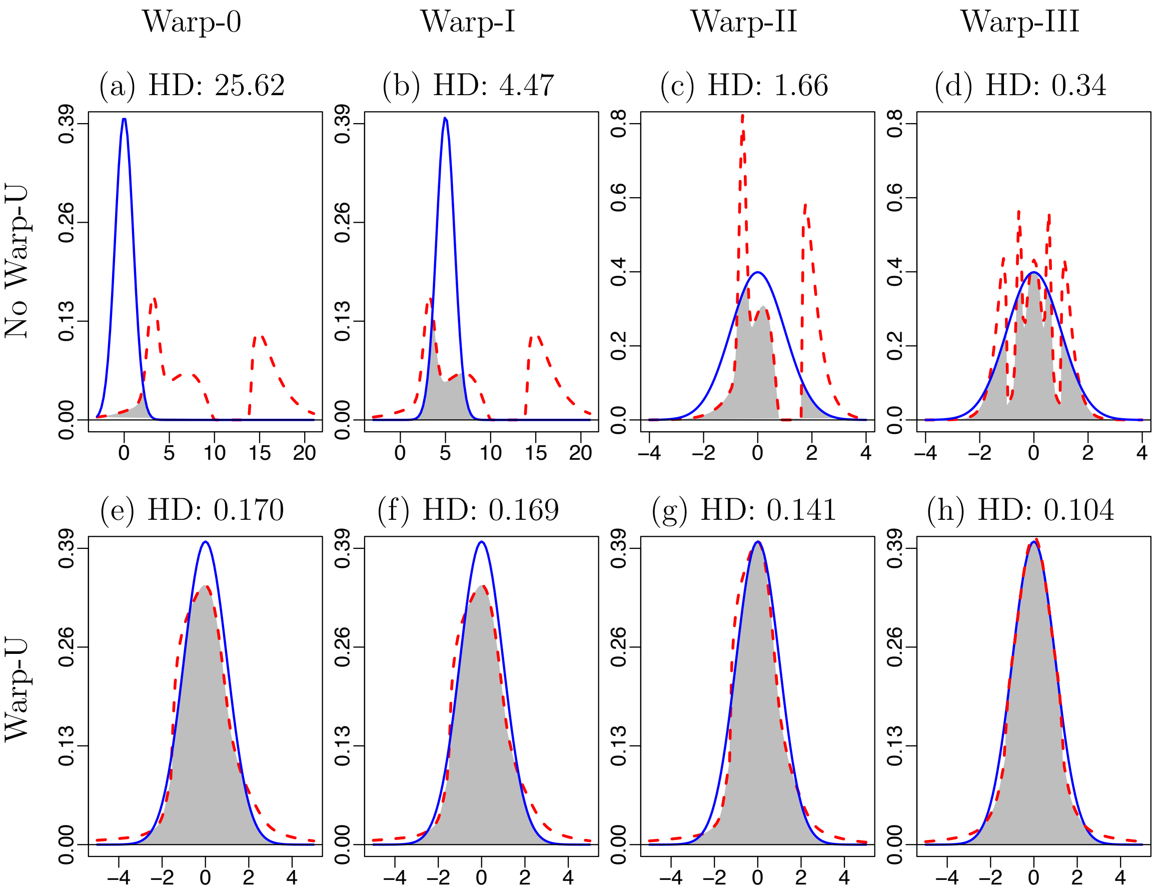

Figure 1.8: The two densities used in bridge sampling. Solid lines: the density of $\mathcal{N}(0,1)$. Dashed lines: (a) p, density of original data $\{w_1, \cdots, w_n\}$; (b) $\widetilde{p}^{(\mathrm{I})}$, density of warp-I transformed data; (c) $\widetilde{p}^{(\mathrm{II})}$; (d) $\widetilde{p}^{(\mathrm{III})}$; (e) $\widetilde{p}^{(\mathrm{U})}$; (f) $\widetilde{p}^{(\mathrm{U+I})}$, density after warp-U and then warp-I transformation; (g) $\widetilde{p}^{(\mathrm{U+II})}$; (h) $\widetilde{p}^{(\mathrm{U+III})}$.

and 1.9(d).

The Gaussian mixture distribution that specifies the warp-U transformation is shown in Figure 1.3 (left). The red dashed curve in Figure 1.8(e) is the density of the warp-U transformed data. The harmonic distance between $\widetilde{p}^{(\mathrm{U})}$ and ϕ reduces to 0.170, and the RMSE of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$ is 0.009. We apply warp-I, II, and III transformations to $\{\widetilde{w}^{(\mathrm{U})}_1, \cdots, \widetilde{w}^{(\mathrm{U})}_n\}$ and the overlap between the resulting densities and ϕ is reduced even further, see Figure 1.8(f-h). The Harmonic distance between ϕ and $\widetilde{p}^{(\mathrm{U+III})}$ is

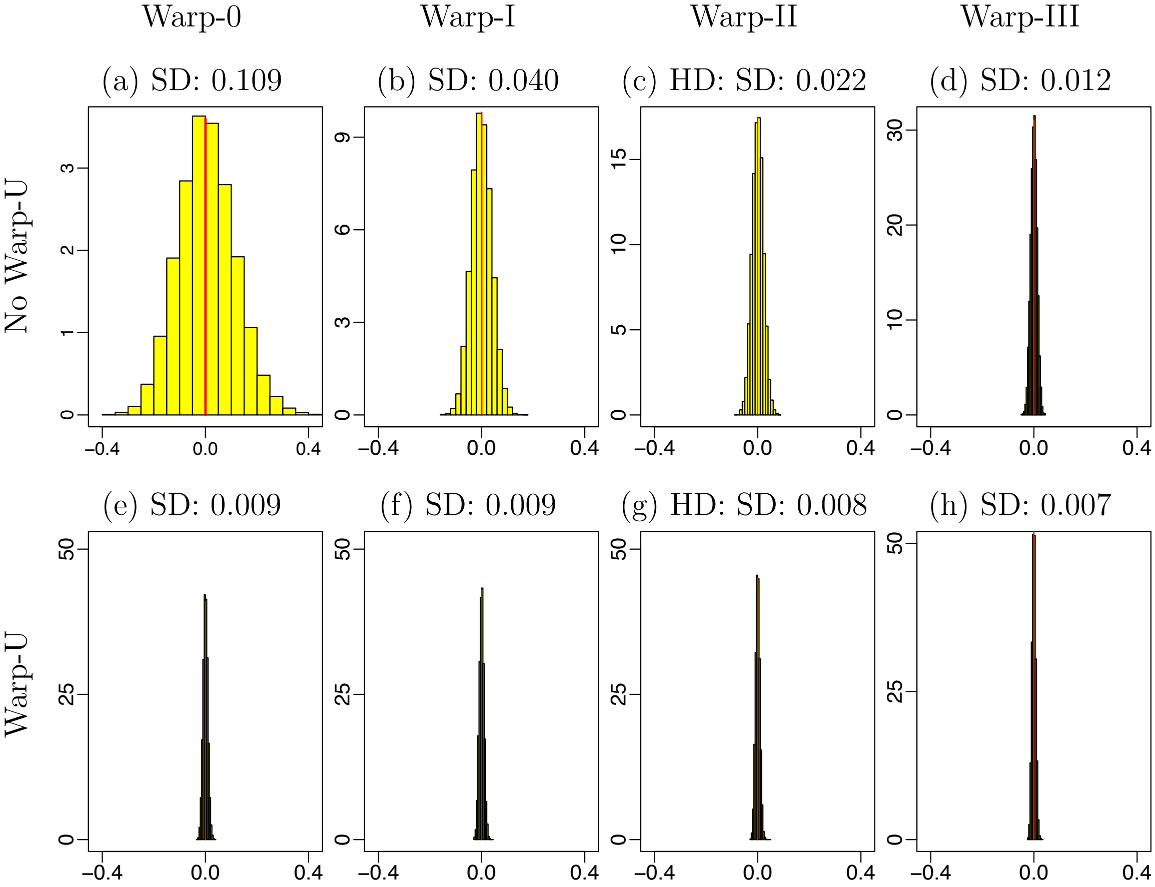

Figure 1.9: Histograms of $\widehat{\lambda}^{(\mathcal{X})}_{\rm opt} - \lambda$. (a) $\widehat{\lambda}_{\rm opt} - \lambda$, bridge sampling estimator with no transformation; (b) $\widehat{\lambda}^{(\rm I)}_{\rm opt} - \lambda$, warp-I bridge sampling; (c) $\widehat{\lambda}^{(\rm II)}_{\rm opt} - \lambda$; (d) $\widehat{\lambda}^{(\rm III)}_{\rm opt} - \lambda$; (e) $\widehat{\lambda}^{(\rm U)}_{\rm opt} - \lambda$; (f) $\widehat{\lambda}^{(\rm U+I)}_{\rm opt} - \lambda$; (g) $\widehat{\lambda}^{(\rm U+II)}_{\rm opt} - \lambda$; (h) $\widehat{\lambda}^{(\rm U+III)}_{\rm opt} - \lambda$.

the smallest and thus $\widehat{\lambda}^{(\rm U+III)}_{\rm opt}$ has the smallest RMSE. It is worth mentioning that compared with warp-U transformation, the additional distance reduction due to the additional warp transformation appears to be minor, as long as the Gaussian mixture distribution that determines the warp-U transformation has sufficient overlap with p.

We also compare the two optimal bridge sampling estimators defined in (1.29) and (1.30), denoted as $\widehat{c}^{(\rm U)}_{\rm opt}$ and $\widehat{c}^{(\rm mix)}_{\rm opt}$. Theorem 1 implies $H_{\rm A}(\widetilde{p}^{(\rm U)}, \phi) \leqslant H_{\rm A}(p, \phi_{\rm mix})$, so the asymptotic variance of $\widehat{\lambda}^{(\rm U)}_{\rm opt}$ is smaller than that of $\widehat{\lambda}^{(\rm mix)}_{\rm opt}$. The superiority of warp-U

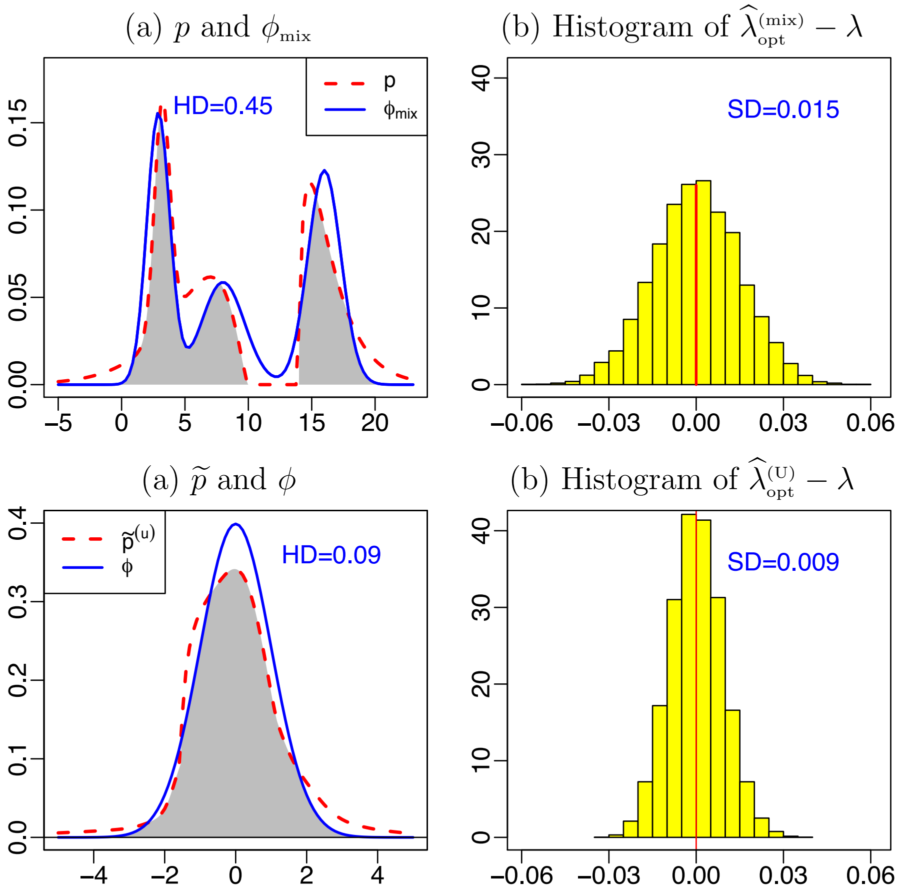

Figure 1.10: (a) Dashed line: p, solid line: ϕ_{mix}; (b) histogram of $\widehat{\lambda}^{(\text{mix})}_{\text{opt}} - \lambda$; (c) dashed line: $\widetilde{p}$, density after warp-U transformation, solid line: ϕ; (d) histogram of $\widehat{\lambda}^{(\text{U})}_{\text{opt}} - \lambda$.

transformation is confirmed by Figure 1.10.

1.4 Estimating Warp-U Transformation

In the previous sections, we did not discuss how to obtain a ϕ_{mix} with sufficient overlap with p, but instead assumed it is given. It is however the most crucial step in warp-U bridge sampling, since ϕ_{mix} completely specifies the warp-U transformation and determines the lower bound for the Monte Carlo errors of the corresponding warp-U bridge sampling.

In practice, a ϕ_{mix} should be obtained within a reasonable amount of time. In relatively low-dimensional ($\leqslant 10$) problems, we can obtain a ϕ_{mix} based on the expression of q (Bornkamp, 2011; Gelman et al., 2013). But in relatively high dimension, these methods can be extremely computationally expensive and are often unstable. In this section, we propose a simple model that can capture a reasonable amount of mass of p, and the computation costs are linear in the dimensionality.

1.4.1 A ϕ_{mix} with Diagonal Covariance Matrixes

Assume we have good data (e.g., i.i.d draws) in D dimensions from p that can represent the important regions of the density. We propose fitting the data to a mixture of normal distributions with diagonal covariance matrixes, that is,

$$\phi_{\text{mix}}(\omega; \boldsymbol{\zeta}) = (2\pi)^{-\frac{D}{2}} \sum_{k=1}^{K} \frac{\pi_k}{|\mathcal{S}_k|} \exp\left(-\frac{1}{2}(\omega - \mu_k)' \mathcal{S}_k^{-2} (\omega - \mu_k)\right), \tag{1.31}$$

where π_k is the weight of the normal distribution $\mathcal{N}(\mu_k, \mathcal{S}_k^2)$, $\mathcal{S}_k$ is a positive definite diagonal matrix,

$$\mathcal{S}_k = \begin{pmatrix} \sigma_{k,1} & 0 & \cdots & 0 \\ 0 & \sigma_{k,2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_{k,D} \end{pmatrix},$$

and $\boldsymbol{\zeta} = (\pi_1, \cdots, \pi_K, \mu_1, \cdots, \mu_K, \mathcal{S}_1, \cdots, \mathcal{S}_K)$.

The reasons for choosing the diagonal covariance matrix are as follows. First of all, without specifying the structures of the covariance matrixes, it is computationally very expensive to estimate these $D \times D$ covariance matrixes, especially in high dimensions; the computation costs are at least $O(D^2)$. Second, for full covariance matrixes, the estimates of the parameters may be unreliable, because there may not be enough degrees of freedom to estimate a total of $K(1 + D(D+3)/2)$ parameters. Finally, as shown in previous sections, it is not necessary (and computationally too expensive) to find a ϕ_{mix} that is almost identical to p in applying warp-U transformation; we only need a ϕ_{mix} that has sufficient overlap with p.

We propose estimating $\boldsymbol{\zeta}$ via the maximum likelihood method. It is important to note that the MLE of $\boldsymbol{\zeta}$ is not well-defined (Day, 1969; Kiefer and Wolfowitz, 1956), because the likelihood can go to infinity if we let μ_1 equal to a data point from p and $\mathcal{S}_1 \to 0$. Therefore, instead of maximizing the log-likelihood, we look for $\widetilde{\boldsymbol{\zeta}}$ that maximizes a penalized log-likelihood, proposed by Chen et al. (2008),

$$\widetilde{l}_n(\boldsymbol{\zeta}) = l_n(\boldsymbol{\zeta}) + \mathbf{p_n}(\boldsymbol{\zeta}),$$

where n is the sample size, l_n is the log-likelihood of the parameters under the model, and $\mathbf{p_n}$ is the penalty term that depends on the data. The penalty prevents any of the variances from approaching 0, and it converges to 0 as $n \to \infty$. Thus, asymptotically the penalized MLE (pMLE) of $\boldsymbol{\zeta}$ does maximize the original likelihood.

Chen and Tan (2009) suggest the penalty function

$$\mathbf{p_n}(\boldsymbol{\zeta}) = -a_n \sum_{k=1}^{K} \left\{ \text{trace}(\widehat{Q}_w^2 \Sigma_k^{-1}) + \log |\Sigma_k^{-1}| \right\},$$

where trace$(\cdot)$ represents the trace function, $a_n = 1/\sqrt{n}$, $\widehat{Q}_w^2$ is the empirical covariance matrix of the data and Σ_k is the covariance matrix of the Gaussian component. We change $\widehat{Q}_w$ to be diag$(\widehat{IQR}_1, \cdots, \widehat{IQR}_D)$, where $\widehat{IQR}_d$ is the inter-quantile range, i.e., the difference between the third quantile and the first quantile, of the data in the d-th dimension. We propose such $\widehat{Q}_w$ for two reasons. First, the empirical covariance matrix may be unreliable if p has very fat tails, such as Cauchy distribution. Second, since both $\widehat{Q}_w$ and $\Sigma_k = \mathcal{S}_k^2$ in our model are diagonal matrixes, the computation burden is lessened. The penalty $\mathbf{p_n}$ is then simplified to be

$$\mathbf{p_n}(\boldsymbol{\zeta}) = -\frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{d=1}^{D} \left\{ \widehat{IQR}_d^2 / \sigma_{k,d}^2 - \log(\sigma_{k,d}^2) \right\}.$$

The E-step and the M-step of the EM algorithm to search for the pMLE of $\boldsymbol{\zeta}$ are provided in Chen and Tan (2009). The EM algorithm is known to have some defects, especially when the original density p is sparsely scattered in a high-dimensional space. First, it is sensitive to the initial configuration; second, the algorithm is often trapped at local maxima due to the difficulty of passing through regions with very low likelihood. Fortunately, a good local maximizer of the penalized log-likelihood often suffices for the purpose of warp-U transformation.

For clarity, we discuss briefly what we do to obtain a good local maximizer via EM. The method we provide is likely not optimal, but it does yield a fairly reliable estimate of $\boldsymbol{\zeta}$ for warp-U transformation. To reduce the dependence of the final estimate of $\boldsymbol{\zeta}$ on the initial value, we apply EM to the same data repeatedly for M times, each time with a different starting point $\boldsymbol{\zeta}^{(0)}$. Based on simulation, it appears a small M usually suffices to provide a sound local maxima. The initial weights and covariance matrixes are set to be $1/K$ and $1.5\widehat{Q}_w^2$, respectively. For the first $M/2$ times, we randomly draw K points without replacement from the available data as the initial mean parameters. For the second $M/2$ times, along the dimension with the largest variance, we first divide the region where 95% of the data sit into K subregions so that each subregion contains approximately the same number of data points, and then sample one data point from each of the K subregions as the initial mean parameters. The stopping criterion we use for the EM algorithm is $|l_n^{(t)} - l_n^{(t-1)}| < 10^{-8} D |l_n^{(t)}|$. In our simulation, the EM algorithm mostly stops within 100 iterations. After obtaining M estimates of $\boldsymbol{\zeta}$, we choose the one with the largest likelihood as our parameters, $\widetilde{\boldsymbol{\zeta}}$, for warp-U bridge sampling.

1.4.2 Overcoming Adaptive Bias

For ease of reference, in this section and Section 1.4.3, we denote $\widetilde{\boldsymbol{\zeta}}_{\rm D}$ as the pMLE of $\boldsymbol{\zeta}$ estimated from the whole dataset, $\{w_1, \cdots, w_n\}$, from p, and $\widehat{\lambda}_{\rm D}^{(\rm U)} = \log\left(\widehat{c}_{\rm D}^{(\rm U)}\right)$ as the corresponding warp-U bridge sampling estimator. However, because $\widetilde{\boldsymbol{\zeta}}_{\rm D}$ is a function of the data, the distribution of the corresponding warp-U transformed data, $\{\widetilde{w}_1, \cdots, \widetilde{w}_n\}$, is no longer $\widetilde{p}(\cdot; \widetilde{\boldsymbol{\zeta}}_{\rm D})$, as expressed in (1.24) or (1.28). Consequently, additional bias is introduced to the estimator $\widehat{\lambda}_{\rm D}^{(\rm U)}$.

We use an example in 10 dimensions to demonstrate the additional bias in $\widehat{\lambda}_{\rm D}^{(\rm U)}$ and its impact on the RMSE of $\widehat{\lambda}_{\rm D}^{(\rm U)}$. We compare the performance of four warp-U bridge sampling estimators with the optimal choice of α, denoted as $\widehat{\lambda}_{\rm D,Diag}^{(\rm U)}$, $\widehat{\lambda}_{\rm D,Full}^{(\rm U)}$, $\widehat{\lambda}_{\rm I,Diag}^{(\rm U)}$, and $\widehat{\lambda}_{\rm I,Full}^{(\rm U)}$, where the first subscript specifies whether $\boldsymbol{\zeta}$ is estimated from the whole data set (D) or from other sources independent of the data (I), and the second subscript indicates whether the covariance matrixes are restricted to diagonal matrixes (Diag) or not (Full). The number of components in ϕ_{mix}, which determines the warp-U transformation, varies from 5 to 20. For each K and each type of covariance matrixes, we obtain a vector, $\widetilde{\boldsymbol{\zeta}}_{\text{I}}$, by maximizing the penalized log-likelihood based on a fixed dataset from p that is completely independent of the data used for bridge sampling. Note, in real applications, it is unlikely for us to have the luxury of a separate and large dataset just for the estimation of $\boldsymbol{\zeta}$. We only use $\widehat{\lambda}^{(\text{U})}_{\text{I,Diag}}$ and $\widehat{\lambda}^{(\text{U})}_{\text{I,Full}}$ here as the benchmark for comparison.

The density p is a mixture of 25 skewed t-distributions with degrees of freedom ranging from 1 to 4, and none of the covariance matrixes of these t-distributions are sparse. We simulate 10,000 replicate datasets, each of which contains 2,500 independent draws from p and 2,500 independent draws from $\mathcal{N}(0, I_{10})$.

Figure 1.11 shows the summary statistics of $\widehat{\lambda}^{(\text{U})}_{\mathcal{Y},\text{Diag}}$ (top row) and $\widehat{\lambda}^{(\text{U})}_{\mathcal{Y},\text{Full}}$ (bottom row), where the subscript "$\mathcal{Y}$" represents "D" or "I". The dotted and solid lines in the figure correspond to $\widehat{\lambda}^{(\text{U})}_{\text{D},\mathcal{Z}}$ and $\widehat{\lambda}^{(\text{U})}_{\text{I},\mathcal{Z}}$, respectively, where "$\mathcal{Z}$" represents "Diag" or "Full". The first column in Figure 1.11 shows the excessive bias of $\widehat{\lambda}^{(\text{U})}_{\text{D},\mathcal{Z}}$ compared with $\widehat{\lambda}^{(\text{U})}_{\text{I},\mathcal{Z}}$. Note, as K increases, the Gaussian mixture model, $\phi_{\text{mix}}(\cdot; \widetilde{\boldsymbol{\zeta}}_{\text{I}})$, fits to the fixed dataset better, but it does not necessarily result in smaller bias, thus we see the zigzag shape of the bias of $\widehat{\lambda}^{(\text{U})}_{\text{I},\mathcal{Z}}$. The second column shows the variance of $\widehat{\lambda}^{(\text{U})}_{\text{D},\mathcal{Z}}$ and that of $\widehat{\lambda}^{(\text{U})}_{\text{I},\mathcal{Z}}$ are quite similar. The variance decreases as K increases, because on average larger K corresponds to more overlap between p and the calibrated ϕ_{mix}, and thus more overlap between $\widetilde{p}$ and ϕ. In addition, $\widehat{\lambda}^{(\text{U})}_{\text{I},\mathcal{Z}}$ has slightly smaller variance than $\widehat{\lambda}^{(\text{U})}_{\text{D},\mathcal{Z}}$ for fixed K, because $\widetilde{\boldsymbol{\zeta}}_{\text{I}}$ is estimated from a much larger dataset than $\widetilde{\boldsymbol{\zeta}}_{\text{D}}$.

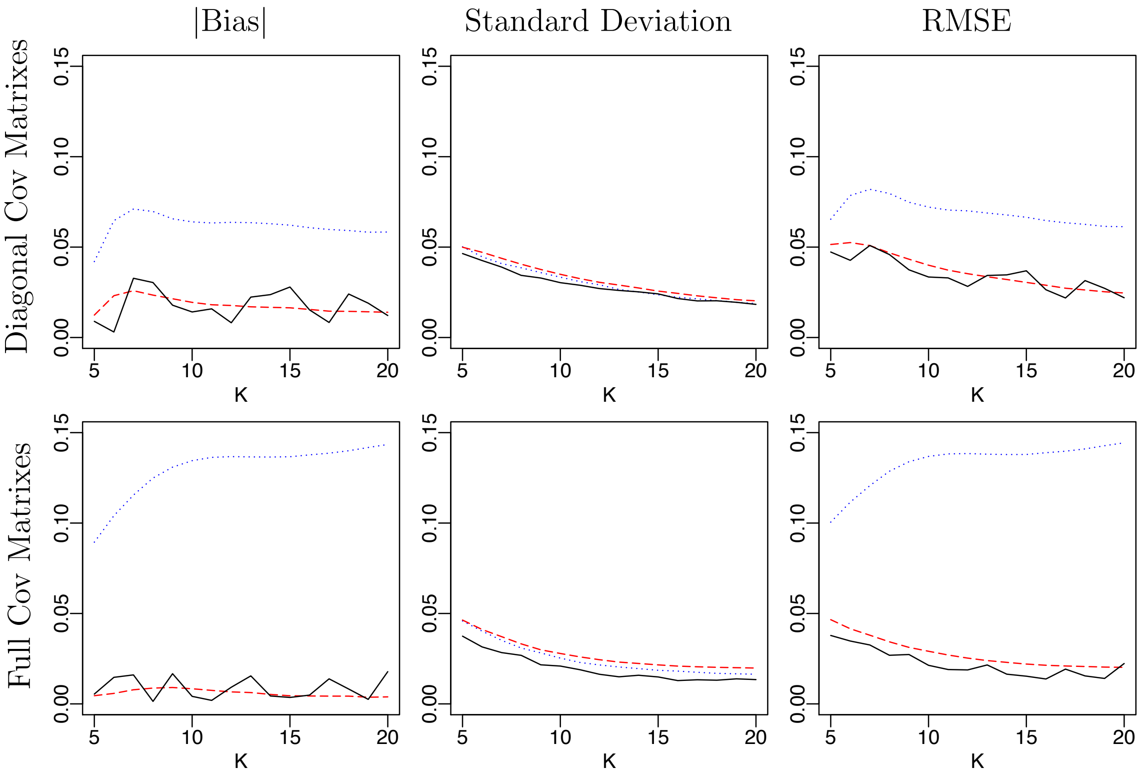

Figure 1.11: The three columns show the absolute value of the bias, the standard deviation, and the RMSE of (i) $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{D},\mathcal{Z}} = \log(\widehat{c}^{(\mathrm{U})}_{\mathrm{D},\mathcal{Z}})$ (dotted lines), the warp-U bridge sampling estimator specified by $\widetilde{\boldsymbol{\zeta}}_D$, which is estimated from $\{w_1, \cdots, w_n\}$, (ii) $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{I},\mathcal{Z}} = \log(\widehat{c}^{(\mathrm{U})}_{\mathrm{I},\mathcal{Z}})$ (solid lines), warp-U bridge sampling specified by $\widetilde{\boldsymbol{\zeta}}_{\mathrm{I}}$, which is independent of $\{w_1, \cdots, w_n\}$, and (iii) (dashed lines) the average of two warp-U bridge sampling estimators with half of data estimating $\boldsymbol{\zeta}$ and the other half for bridge sampling. The subscript "$\mathcal{Z}$" represents "Diag" (top row) or "Full" (bottom row) for the covariance matrixes in the Gaussian mixture model.

The last column in Figure 1.11 shows the RMSE of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{D},\mathcal{Z}}$ is dominated by the bias term, and is hence much larger than that of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{I},\mathcal{Z}}$.

Since the additional bias of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{D},\mathcal{Z}}$ is due to the dependence of $\widetilde{\boldsymbol{\zeta}}_{\mathrm{D}}$ and the data from p, an obvious solution to remove it is to use two independent subsets of draws from p to estimate $\boldsymbol{\zeta}$ and to do bridge sampling. However using partial data for bridge sampling will directly increase the asymptotic variance of the estimator. To remove the additional bias without substantially increasing the variance, two conditions should

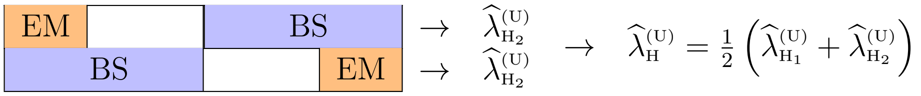
EM
BS
BS
EM
→ $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}_2}$
→ $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}_2}$
→ $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}} = \frac{1}{2}\left(\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}_1} + \widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}_2}\right)$

Figure 1.12: A proposed solution to remove the adaptive bias without increasing the variance of the warp-U bridge sampling estimator. Each of the two estimators $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}_1}$ and $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}_2}$ is obtained by using (part of) one half of the data for the estimation of $\boldsymbol{\zeta}$ and the other half for warp-U bridge sampling. The final estimator of λ is their average.

be satisfied: (i) independent subsets of draws should be used for estimating $\boldsymbol{\zeta}$ and in bridge sampling, and (ii) all the draws from p should be used at least once in bridge sampling.

We propose one solution that works relatively well in terms of both the statistical and computational efficiency. As visualized in Figure 1.12, we split the data into two halves, and obtain two separate bridge sampling estimators, denoted as $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}_1}$ and $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}_2}$. Each $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}_i}$ is obtained by using $L \leqslant n/2$ of one half of the data from p to estimate $\boldsymbol{\zeta}$ and the other half for the warp-U bridge sampling specified by the estimated $\boldsymbol{\zeta}$. The final estimator $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}}$ is the average of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}_1}$ and $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}_2}$. Empirical studies have shown the correlation of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}_1}$ and $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}_2}$ is very small (mostly < 0.06, see Figure 1.14), thus, the variance of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}}$ is nearly half of the variance of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H}_i}$. The dashed lines in Figure 1.11 show the bias, the standard deviation, and the RMSE of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H,Diag}}$ (top row) and $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H,Full}}$ (bottom row), which are very close to their corresponding benchmarks (solid lines).

1.4.3 Justification for Diagonal Covariance Matrixes

Using the simulation study in Section 1.4.2, we further justify the use of diagonal covariance matrixes in ϕ_{mix}. Figure 1.13 (left) shows the RMSE of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{I,Diag}}$, $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{I,Full}}$, $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H,Diag}}$, and $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H,Full}}$. On average, the RMSE of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{I,Diag}}$ (thin solid line) is 51% larger than that of

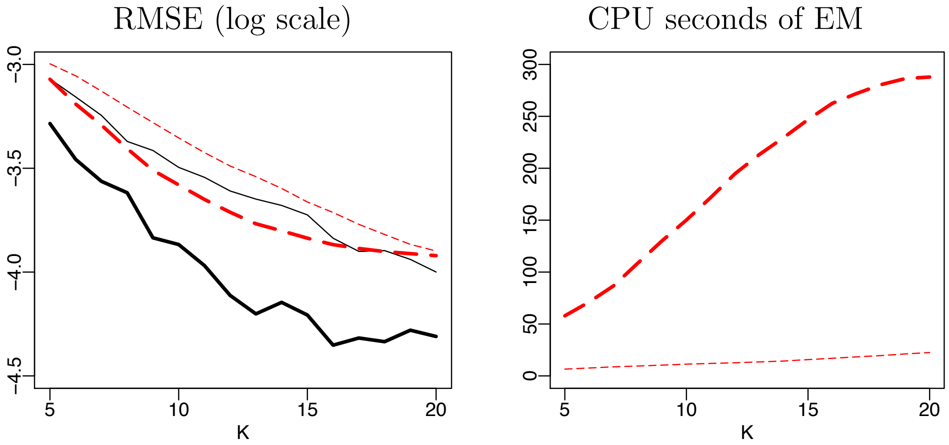

Figure 1.13: (Left) RMSE (on a logarithmic scale) of different estimators; (right) CPU seconds of each EM algorithm. (Solid lines) warp-U bridge sampling with $\widetilde{\boldsymbol{\zeta}}_{\mathrm{I}}$; (dashed lines) the average of the two warp-U bridge sampling estimators with half of data used for estimating $\boldsymbol{\zeta}$ and the other half for bridge sampling. (Thin lines) the covariance matrixes are diagonal matrixes; (thick lines) the covariance matrixes are not restricted to diagonal matrixes.

$\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{I,Full}}$ (thick solid line). However, the RMSE of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H,Diag}}$ (thin dashed line) is only 16.7% larger than that of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{H,Full}}$ (thick dashed line), and their difference diminishes as K increases. This is because when K is large, on one hand, an overfitting problem occurs for the full-covariance-matrix model due to the additional $KD(D-1)/2$ parameters in the model, and on the other hand, the diagonal-covariance-matrix model, being very different from p, continues to fit the data better and the resulting RMSE decreases at a stable rate.

Figure 1.13 (right) shows the CPU seconds for estimating $\boldsymbol{\zeta}$ via the EM algorithm. On average, it takes 12 times longer to obtain ϕ_{mix} with full covariance matrixes than ϕ_{mix} with diagonal covariance matrixes in this study, and the difference increases as the dimension increases. In addition, in the step of bridge sampling, evaluating ϕ_{mix} with full covariance matrixes is much more costly than ϕ_{mix} with diagonal covariance matrixes. Therefore, the small loss of statistical efficiency and the huge reduction of the computation costs justify the use of diagonal covariance matrixes. To reduce the RMSE, it is more effective to increase K than to use full covariance matrixes in the model.

In the subsequent sections, we only consider ϕ_{mix} to be the Gaussian mixture model with diagonal covariance matrixes. For simplicity, we drop the subscripts "H" and "Diag", and denote $\widehat{\lambda}^{(\mathcal{X})}_{\alpha,1}$ and $\widehat{\lambda}^{(\mathcal{X})}_{\alpha,2}$ as the estimators with half of data used for estimating $\boldsymbol{\zeta}$ and the other half in bridge sampling, where α specifies the functional used in bridge sampling. The combined estimator is $\widehat{\lambda}^{(\mathcal{X})}_{\alpha} = \dfrac{1}{2}\left(\widehat{\lambda}^{(\mathcal{X})}_{\alpha,1} + \widehat{\lambda}^{(\mathcal{X})}_{\alpha,2}\right)$.

1.4.4 Estimation of the Variance of $\widehat{\lambda}^{(\mathbf{U})}_{\alpha}$ and $\widehat{\lambda}^{(\mathbf{mix})}_{\alpha}$

In estimating λ, it is important to have some idea about the uncertainly associated with the point estimate. We analyze the variance of $\widehat{\lambda}^{(\mathcal{X})}_{\alpha,1}$ below. By symmetry, the results also applies to $\widehat{\lambda}^{(\mathcal{X})}_{\alpha,2}$. We do not intent to give rigorous proof for the variance here, rather, we use heuristic calculation to provide an estimate of the variance.

Let $\{w_1, \cdots, w_L\}$ be the i.i.d draws from p we use to estimate $\boldsymbol{\zeta}$, where $L \leqslant n/2$. We use $\widetilde{\boldsymbol{\zeta}}_L$ to denote the estimated vector, where the subscript "L" emphasizes the sample size. Specified by the estimate $\widetilde{\boldsymbol{\zeta}}_L$, we apply the corresponding warp-U transformation to the other half of data $\{w_{1+n/2}, \cdots, w_n\} \overset{iid}{\sim} p$. Let $\widehat{\lambda}^{(\text{U})}_{\alpha,1}$ be the bridge sampling estimator based on $\{z_{1+m/2}, \cdots, z_m\} \overset{iid}{\sim} \phi$ and the warp-U transformed data, $\{\widetilde{w}_{1+n/2}, \cdots, \widetilde{w}_n\} \overset{iid}{\sim} \widetilde{p}$. Then the conditional variance of $\widehat{\lambda}^{(\text{U})}_{\alpha,1}$ given $\widetilde{\boldsymbol{\zeta}}_L$ (Meng and Wong, 1996) is

$$\mathrm{Var}\left(\widehat{\lambda}^{(\mathrm{U})}_{\alpha,1}|\widetilde{\boldsymbol{\zeta}}_{L}\right) = \frac{2}{n+m}\mathcal{V}_{\alpha}\left(\widetilde{p},\phi\right) + o\left(\frac{1}{n+m}\right), \tag{1.32}$$

where $\mathcal{V}_{\alpha}\left(\widetilde{p},\phi\right)$ is defined in (1.3). By the law of total variance,

$$\mathrm{Var}\left(\widehat{\lambda}^{(\mathrm{U})}_{\alpha,1}\right) = \mathrm{E}_{L}\left[\mathrm{Var}\left(\widehat{\lambda}^{(\mathrm{U})}_{\alpha,1}|\widetilde{\boldsymbol{\zeta}}_{L}\right)\right] + \mathrm{Var}_{L}\left[\mathrm{E}\left(\widehat{\lambda}^{(\mathrm{U})}_{\alpha,1}|\widetilde{\boldsymbol{\zeta}}_{L}\right)\right],$$

where E_{L} and Var_{L} are taken over the sampling distribution of $\widetilde{\boldsymbol{\zeta}}_{L}$. Given $\widetilde{\boldsymbol{\zeta}}_{L}$, the asymptotic bias of $\widehat{\lambda}^{(\mathrm{U})}_{\alpha,1}$ is in the order of $(m+n)^{-1}$, so

$$\mathrm{Var}\left(\widehat{\lambda}^{(\mathrm{U})}_{\alpha,1}\right) = \frac{2}{n+m}E_{L}\left[\mathcal{V}_{\alpha}\left(\widetilde{p},\phi\right)\right] + o\left(\frac{1}{n+m}\right). \tag{1.33}$$

Figure 1.14 (left) shows the correlation between $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt},1}$ and $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt},2}$ for different values of K and m, based on 10,000 replications, within each of which $n = 10,000$ data are generated from p, as described in Section 1.4.2. The correlation between $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt},1}$ and $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt},2}$ is due to the fact that $L = 50K$ data points used in bridge sampling for one estimator are used for estimating $\boldsymbol{\zeta}$ for the other estimator, thus we observe the correlation increases with K. Figure 1.14 (left) shows the correlation is very small (< 0.06) even when $K = 50$, so

$$\mathrm{Var}\left(\widehat{\lambda}^{(\mathrm{U})}_{\alpha}\right) \approx \frac{1}{2}\mathrm{Var}\left(\widehat{\lambda}^{(\mathrm{U})}_{\alpha,i}\right) = \frac{1}{n+m}E_{L}\left[\mathcal{V}_{\alpha}\left(\widetilde{p},\phi\right)\right] + o\left(\frac{1}{n+m}\right).$$

For a given $\widetilde{\boldsymbol{\zeta}}_{L}$ estimated from $\{w_1,\cdots,w_L\}$, to estimate $\mathrm{Var}\left(\widehat{\lambda}^{(\mathrm{U})}_{\alpha,1}|\widetilde{\boldsymbol{\zeta}}_{L}\right)$, we divide $\{\widetilde{w}_{1+n/2},\cdots,\widetilde{w}_{n}\}$ and $\{z_{1+m/2},\cdots,z_{m}\}$ each into $S \geqslant 2$ non-overlapping subsets

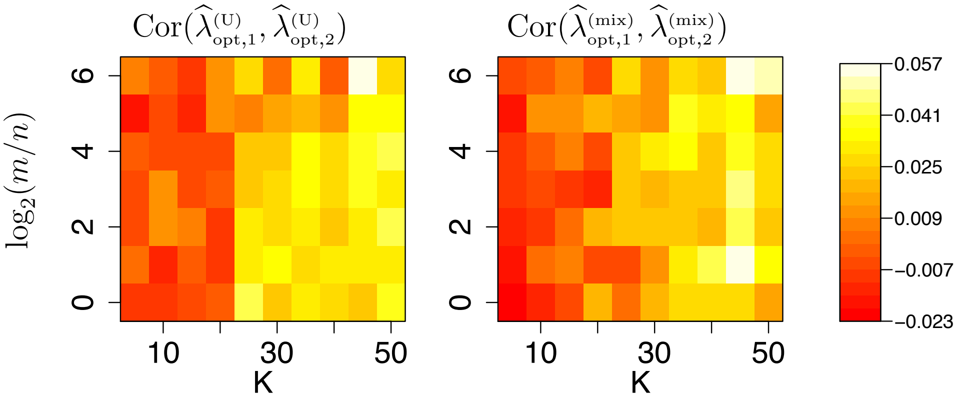

Figure 1.14: The correlation between $\widehat{\lambda}^{(\mathcal{X})}_{\mathrm{opt},1}$ and $\widehat{\lambda}^{(\mathcal{X})}_{\mathrm{opt},2}$ with different K and m/n.

of equal size, and obtain S separate estimators $\widehat{\lambda}^{(\mathrm{U})}_{\alpha,1,s}$, for $s = 1, \cdots, S$. The evaluations of $\widetilde{q}$ and ϕ at these data points are already done to obtain $\widehat{\lambda}^{(\mathrm{U})}_{\alpha,1}$, so little additional computation costs are required to compute $\widehat{\lambda}^{(\mathrm{U})}_{\alpha,1,s}$. The empirical variance of $\{\widehat{\lambda}^{(\mathrm{U})}_{\alpha,1,s}; s = 1, \cdots, S\}$, denoted as $\widehat{\nu}^{(\mathrm{U})}_{\alpha,1}$, estimates the variance of the bridge sampling estimator with $n/(2S)$ data from $\widetilde{p}$ and $m/(2S)$ data from ϕ. So according to the asymptotical expression of the bridge sampling estimator, $\mathcal{V}_\alpha(\widetilde{p}, \phi)$ in (1.3) can be estimated by $(n + m)\widehat{\nu}^{(\mathrm{U})}_{\alpha,1}/(2S)$. Similarly, for a given $\widetilde{\boldsymbol{\zeta}}_L$ estimated from L of the second half of the data from p, the corresponding $\mathcal{V}_\alpha(\widetilde{p}, \phi)$ can be estimated by $(n+m)\widehat{\nu}^{(\mathrm{U})}_{\alpha,2}/(2S)$, where $\widehat{\nu}^{(\mathrm{U})}_{\alpha,2}$ is the empirical variance of $\widehat{\lambda}^{(\mathrm{U})}_{\alpha,2,s}$. Finally, the asymptotic variance of the combined estimator $\widehat{\lambda}^{(\mathrm{U})}_{\alpha}$ is approximately $\dfrac{1}{n+m}E_L(\mathcal{V}_\alpha(\widetilde{p}, \phi))$, which can be estimated by

$$\widehat{\nu}^{(\mathrm{U})}_{\alpha} = \frac{1}{2}\frac{\widehat{\nu}^{(\mathrm{U})}_{\alpha,1} + \widehat{\nu}^{(\mathrm{U})}_{\alpha,2}}{2S} = \frac{1}{4S(S-1)}\sum_{i=1}^{2}\sum_{s=1}^{S}\left(\widehat{\lambda}^{(\mathrm{U})}_{\alpha,i,s} - \bar{\lambda}^{(\mathrm{U})}_{\alpha,i}\right)^2, \tag{1.34}$$

where $\bar{\lambda}^{(\mathrm{U})}_{\alpha,i} = \sum_{s=1}^{S}\widehat{\lambda}^{(\mathrm{U})}_{\alpha,i,s}/S$. There is a trade-off in choosing S, because small S

may cause inaccurate estimation of the variance by $\widehat{\nu}_{\alpha,1}^{(\mathrm{U})}$ and $\widehat{\nu}_{\alpha,2}^{(\mathrm{U})}$, whereas large S may break the asymptotic results we rely on to obtain (1.34), i.e., $\mathrm{Var}\left(\widehat{\lambda}_{\alpha,i,s}^{(\mathrm{U})} \middle| \widetilde{\boldsymbol{\zeta}}_L\right) \approx \frac{2S}{n+m}\mathcal{V}_\alpha(\widetilde{p}, \phi)$.

Similarly, for the bridge sampling estimator with $\{w_{1+n/2}, \cdots, w_n\} \overset{iid}{\sim} p$ and $\{x_{1+m/2}, \cdots, x_m\} \overset{iid}{\sim} \phi_{\mathrm{mix}}(\cdot; \widetilde{\boldsymbol{\zeta}}_L)$, we have

$$\mathrm{Var}\left(\widehat{\lambda}_{\alpha,1}^{(\mathrm{mix})}\right) = \frac{2}{n+m} E_L\left[\mathcal{V}_\alpha\left(p, \phi_{\mathrm{mix}}\right)\right] + o\left(\frac{1}{n+m}\right),$$

and the variance of the combined estimator $\widehat{\lambda}_{\alpha}^{(\mathrm{mix})} = (\widehat{\lambda}_{\alpha,1}^{(\mathrm{mix})} + \widehat{\lambda}_{\alpha,2}^{(\mathrm{mix})})/2$ is approximately half of $\mathrm{Var}\left(\widehat{\lambda}_{\alpha,1}^{(\mathrm{mix})}\right)$. Theorem 1 implies $\mathcal{V}_\alpha\left(p, \phi_{\mathrm{mix}}\right) \leqslant \mathcal{V}_\alpha\left(\widetilde{p}, \phi\right)$ for fixed $\widetilde{\boldsymbol{\zeta}}_L$ and for both the geometric and the optimal bridge sampling, so asymptotically we expect

$$\mathrm{Var}\left(\widehat{\lambda}_{\alpha}^{(\mathrm{U})}\right) \leqslant \mathrm{Var}\left(\widehat{\lambda}_{\alpha}^{(\mathrm{mix})}\right).$$

1.5 Computation Configurations

In the algorithm to obtain $\widehat{\lambda}_{\alpha}^{(\mathcal{X})}$, there are three tuning parameters:

- K: the number of components in the Gaussian mixture model $\phi_{\mathrm{mix}}(\cdot; \boldsymbol{\zeta})$,
- L: the number of data points from p to estimate $\boldsymbol{\zeta}$, $(L \leqslant n/2)$,
- m: the sample size of the dataset sampled from $N(0, I_D)$ or ϕ_{mix}.

In this section, we use simulation results and theoretical calculations to show how different choices of the tuning parameters affect the statistical efficiency and computation costs of $\widehat{\lambda}_{\mathrm{opt}}^{(\mathrm{U})} = \frac{1}{2}\left(\widehat{\lambda}_{\mathrm{opt},1}^{(\mathrm{U})} + \widehat{\lambda}_{\mathrm{opt},2}^{(\mathrm{U})}\right)$ and $\widehat{\lambda}_{\mathrm{opt}}^{(\mathrm{mix})} = \frac{1}{2}\left(\widehat{\lambda}_{\mathrm{opt},1}^{(\mathrm{mix})} + \widehat{\lambda}_{\mathrm{opt},2}^{(\mathrm{mix})}\right)$, in the hope of providing some guidance for choosing these parameters.

1.5.1 Computation Complexity and Parallel Computing

The computation costs for obtaining $\widehat{\lambda}_{\alpha}^{(\mathcal{X})}$ are determined by the tuning parameters. The first step of the algorithm is to estimate $\boldsymbol{\zeta}$ by applying the EM algorithm to L data points repeatedly for M times. This requires Mt_{EM} amount of time, where t_{EM} is the average time of executing EM algorithm once. If the number of iterations in the EM algorithm is fixed, then $t_{\text{EM}} \propto LK$. Since $\boldsymbol{\zeta}$ is estimated twice, the total execution time is $T_{\text{EM}} = 2Mt_{\text{EM}}$. Figure 1.15 (left) shows T_{EM} (we set $M = 2$) with different values of K for the 10-dimension example in Section 1.4.2. We set the sample size n to be 10,000, so we can have results with larger K. For each simulation configuration, we apply the entire algorithm to 10,000 replicate datasets from p. In Figure 1.15, we vary K from 5 to 250, and set $L = \min(50K, 5000)$. Consequently, as K increases, T_{EM} exhibit quadratic growth when $K \leqslant 100$, and linear growth when $K > 100$.

The second step is to apply warp-U transformation, specified by two different vectors of parameters, to the first and the second half of the n data points from p. For each data point w_i, the probability mass function $P(\Psi = k|w_i) = \frac{\pi_k|\mathcal{S}_k|^{-1}\phi\left(\mathcal{S}_k^{-1}(w_i-\mu_k)\right)}{\sum_{l=1}^{K}\pi_l|\mathcal{S}_l|^{-1}\phi\left(\mathcal{S}_l^{-1}(w_i-\mu_l)\right)}$ needs to be calculated, for $k = 1, \cdots, K$, to determine the probability of each linear transformation. So the execution time of warp-U transformation performed on w_i is $t_{\text{Tr}}^{(\text{U})} \approx K^2 t_{\phi}$, where t_{ϕ} is the amount of time for evaluating the density function of a normal distribution with diagonal covariance matrix. The computation costs of this step is $T_{\text{Tr}}^{(\text{U})} = t_{\text{Tr}}^{(\text{U})} n$.

The last step is to perform bridge sampling twice, each time using one half of

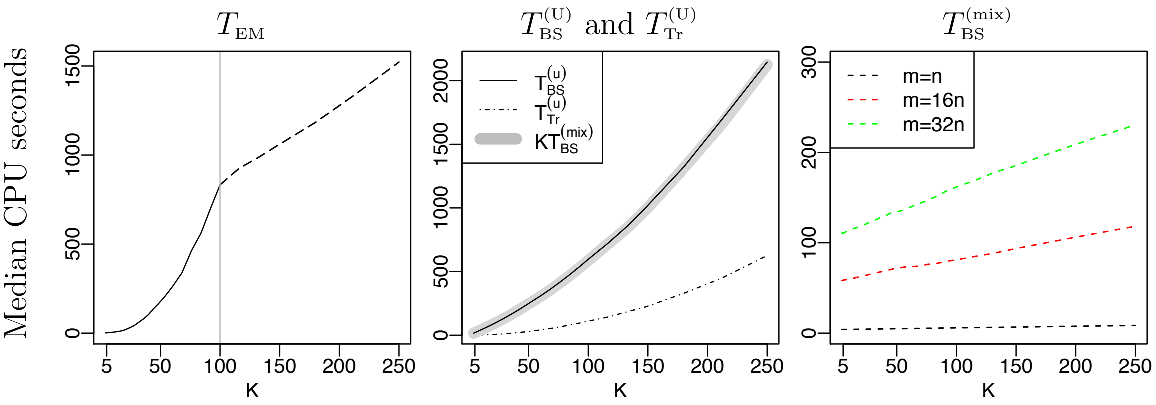

Figure 1.15: The time decomposition in the process of obtaining $\widehat{\lambda}_{\alpha}^{(\mathcal{X})}$. (Left) T_{EM}: time for estimating $\boldsymbol{\zeta}$ (twice); (middle) $T_{\text{BS}}^{(\text{U})}$ (solid line): time for the optimal bridge sampling with n warp-U transformed data and m data from ϕ; $T_{\text{Tr}}^{(\text{U})}$ (dash-dot line): time for the warp-U transformation of the n data points from p; $KT_{\text{BS}}^{(\text{mix})}$ (thick gray line); (right) $T_{\text{BS}}^{(\text{mix})}$: time for the optimal bridge sampling with n data from p and m data from ϕ_{mix}.

the transformed data and $m/2$ data points from ϕ, resulting in $\widehat{\lambda}_{\alpha,1}^{(\text{U})}$ and $\widehat{\lambda}_{\alpha,2}^{(\text{U})}$. The most expensive part in this step is the evaluation of $\widetilde{q}$, which takes $t_{\widetilde{q}} \approx t_{\phi}K^2 + t_qK$, where t_q and $t_{\widetilde{q}}$ represent the amount of time to evaluate q and $\widetilde{q}$, respectively. In the 10-dimension example, according to simulation, $t_{\widetilde{q}} \approx 250t_{\phi}$. The computation costs of bridge sampling also depend on the choice of α. For the optimal or the geometric bridge sampling, it takes $T_{\text{BS}}^{(\text{U})} = t_{\text{BS}}^{(\text{U})}(n+m)$ amount of time, where $t_{\text{BS}}^{(\text{U})} \approx t_{\phi} + t_{\widetilde{q}} \approx t_{\phi} + t_qK + t_{\phi}K^2$. Figure 1.15 (middle) shows $T_{\text{BS}}^{(\text{U})}$ (solid line) and $T_{\text{Tr}}^{(\text{U})}$ (dash-dot line) when $m = n = 10,000$.

In comparison, the last step, i.e., bridge sampling, is less expensive for $\widehat{\lambda}_{\alpha}^{(\text{mix})}$. For the optimal or geometric bridge sampling, $T_{\text{BS}}^{(\text{mix})} = t_{\text{BS}}^{(\text{mix})}(n+m)$, where $t_{\text{BS}}^{(\text{mix})} \approx t_q + t_{\phi}K$, which is approximately $1/K$ of $t_{\text{BS}}^{(\text{U})}$. Figure 1.15 (right) shows three lines of $T_{\text{BS}}^{(\text{mix})}$ with different values of m, growing linearly with K. For the same total sample size $n+m$, $T_{\text{BS}}^{(\text{U})} \approx KT_{\text{BS}}^{(\text{mix})}$, as illustrated in Figure 1.15 (middle), where the curve $KT_{\text{BS}}^{(\text{mix})}$ (thick gray line) is almost identical to $T_{\text{BS}}^{(\text{U})}$.

In summary, the computation costs to obtain $\widehat{\lambda}_{\alpha}^{(\mathrm{U})}$ and $\widehat{\lambda}_{\alpha}^{(\mathrm{mix})}$ are

$$T_{\alpha}^{(\mathrm{U})} = T_{\mathrm{EM}} + T_{\mathrm{Tr}}^{(\mathrm{U})} + T_{\mathrm{BS}}^{(\mathrm{U})} = 2Mt_{\mathrm{EM}} + nt_{\mathrm{Tr}}^{(\mathrm{U})} + (n+m)t_{\mathrm{BS}}^{(\mathrm{U})}, \tag{1.35}$$

$$T_{\alpha}^{(\mathrm{mix})} = T_{\mathrm{EM}} + T_{\mathrm{BS}}^{(\mathrm{mix})} = 2Mt_{\mathrm{EM}} + (n+m)t_{\mathrm{BS}}^{(\mathrm{mix})}, \tag{1.36}$$

where $t_{\mathrm{EM}} \propto LK$, both $t_{\mathrm{Tr}}^{(\mathrm{U})}$ and $t_{\mathrm{BS}}^{(\mathrm{U})}$ grow quadratically with K, and $t_{\mathrm{BS}}^{(\mathrm{mix})} \approx t_{\mathrm{BS}}^{(\mathrm{U})}/K$.

With the advance of computer technology, computation costs bear less and less importance than the statistical efficiency. The two estimators $\widehat{\lambda}_{\alpha,1}^{(\mathcal{X})}$ and $\widehat{\lambda}_{\alpha,2}^{(\mathcal{X})}$ can be computed simultaneous in different processors. Most of the computation in obtaining $\widehat{\lambda}_{\alpha,i}^{(\mathcal{X})}$ requires no communication, so the algorithm can be implemented in an embarrassingly parallel fashion, and the physical time can be reduced by simultaneously using multiple processors.

First, the $2M$ EM algorithms to estimate $\boldsymbol{\zeta}$ can be conducted independently with no interaction, we can distribute the $2M$ tasks among ρ processors. Theoretically, ignoring the possible overhead (time of thread creation/launching, data transformation, synchronization), the physical execution time of this step is $T_{\mathrm{EM}}^{\mathrm{P}} = \lceil \frac{2M}{\rho} \rceil t_{\mathrm{EM}}$, where $\lceil \cdot \rceil$ is the ceiling function. Besides, in the E-step of each iteration within the EM algorithm, the evaluations of the L data points at each of the K proposed normal distributions require no communication either, and thus can be speeded up further by parallel computing.

Second, the computation burden of the bridge sampling estimator lies in the evaluations of functions at $n+m$ data points, which also require no communication. With ρ processors, we can theoretically reduce the time to $T_{\mathrm{BS}}^{(\mathcal{X})\mathrm{P}} = \lceil \frac{n+m}{\rho} \rceil t_{\mathrm{BS}}^{(\mathcal{X})}$. The step of warp-U transformation requires $T_{\mathrm{Tr}}^{(\mathrm{U})\mathrm{P}} = \lceil \frac{n}{\rho} \rceil t_{\mathrm{Tr}}^{(\mathrm{U})}$. What is more, the evaluation of $\widetilde{q}$, which involves K^2 evaluations of ϕ and K evaluations of q, and the calculation of the transformation probability, which involves K^2 evaluations of ϕ, can be done in parallel, so both $t_{\widetilde{q}}$ and t_{Tr} can be reduced further with multiple processors.

Therefore, with ρ processors, even without paralleling the E-step in the EM algorithm and the evaluation of $\widetilde{q}$ or ϕ_{mix}, we can reduce the execution time of getting $\widehat{\lambda}_{\alpha}^{(\mathrm{U})}$ to

$$T_{\alpha}^{(\mathrm{U})\mathrm{P}} = T_{\mathrm{EM}}^{\mathrm{P}} + T_{\mathrm{Tr}}^{(\mathrm{U})\mathrm{P}} + T_{\mathrm{BS}}^{(\mathrm{U})\mathrm{P}} = \lceil \frac{2M}{\rho} \rceil t_{\mathrm{EM}} + \lceil \frac{n}{\rho} \rceil t_{\mathrm{Tr}}^{(\mathrm{U})} + \lceil \frac{n+m}{\rho} \rceil t_{\mathrm{BS}}^{(\mathrm{U})}. \tag{1.37}$$

Similarly, with ρ processors, the execution time of obtaining $\widehat{\lambda}_{\alpha}^{(\mathrm{mix})}$ is reduced to

$$T_{\alpha}^{(\mathrm{mix})\mathrm{P}} = T_{\mathrm{EM}}^{\mathrm{P}} + T_{\mathrm{BS}}^{(\mathrm{mix})\mathrm{P}} = \lceil \frac{2M}{\rho} \rceil t_{\mathrm{EM}} + \lceil \frac{n+m}{\rho} \rceil t_{\mathrm{BS}}^{(\mathrm{mix})}. \tag{1.38}$$

1.5.2 Choosing Tuning Parameters

Good statistical efficiencies often come with large computation costs, so when selecting tuning parameters and comparing $\widehat{\lambda}_{\mathrm{opt}}^{(\mathrm{U})}$ and $\widehat{\lambda}_{\mathrm{opt}}^{(\mathrm{mix})}$, both factors should be considered. We also compare the precision per CPU second (PpS), which accounts for both the statistical and computational efficiencies of the estimators,

$$\text{precision per CPU second} = PpS = \frac{\text{precision}}{\text{CPU seconds}} = \frac{1/\mathrm{Var}}{\text{CPU seconds}}.$$

In this section, we use simulation to compare estimators with different tuning parameters, K, L, and m. If not specified, $L = \min(50K, n/2)$ and $m = n$.

1.5.2.1 Impact of K on $\widehat{\lambda}^{(\mathcal{X})}_{\textbf{opt}}$

Figure 1.11 shows the variance and the MSE of $\widehat{\lambda}^{(\text{U})}_{\text{opt}}$ decrease as K increases. This can be explained by Figure 1.16. The dotted line is the average of the maximum log-likelihood $\bar{l}_{\text{fit}}$, defined as

$$\bar{l}_{\text{fit}} = \frac{1}{L}\sum_{i=1}^{L} \log\left(\phi_{\text{mix}}(w_i; \widetilde{\boldsymbol{\zeta}}_L)\right),$$

where $\{w_1, \cdots, w_L\}$ are used for estimating $\boldsymbol{\zeta}$ via EM algorithm. It measures how well the calibrated Gaussian mixture distribution fits to the L data points used for estimating $\boldsymbol{\zeta}$, so $\bar{l}_{\text{fit}}$ is an increasing function of K. The solid line in Figure 1.16 represents the average log-likelihood $\bar{l}^*$ based on the other half of the data and evaluated at $\widetilde{\boldsymbol{\zeta}}_L$, i.e.,

$$\bar{l}^* = \frac{2}{n}\sum_{i=n/2+1}^{n} \log\left(\phi_{\text{mix}}(w_i; \widetilde{\boldsymbol{\zeta}}_L)\right).$$

It indicates the divergence of p from the fitted $\phi_{\text{mix}}(\cdot; \widetilde{\boldsymbol{\zeta}}_L)$. For moderate K, on average, as the mixture model fits the L data points better, more mass of p will is captured by the calibrated $\phi_{\text{mix}}(\cdot; \widetilde{\boldsymbol{\zeta}}_L)$, and thus both $\bar{l}^*$ and the statistical efficiency of $\widehat{\lambda}^{(\text{U})}_{\alpha}$ increases as K increases.

However, for a large K, the Gaussian mixture model will overfit the $L \leqslant n/2$ data points from p. Figure 1.16 (right) shows $\bar{l}^*$ decreases slightly when K exceeds 100, indicating a slight increase of the divergence between p and $\phi_{\text{mix}}(\cdot; \widetilde{\boldsymbol{\zeta}}_L)$. Figure 1.17 shows the |bias|, standard deviation, and the RMSE (on a logarithmic scale) of $\widehat{\lambda}^{(\text{U})}_{\text{opt}}$ (solid lines) and $\widehat{\lambda}^{(\text{mix})}_{\text{opt}}$ (dashed lines), with K ranging from 5 to 250. When K exceeds 100, there is a slight increase in the variance and the RMSE of these estimators as K

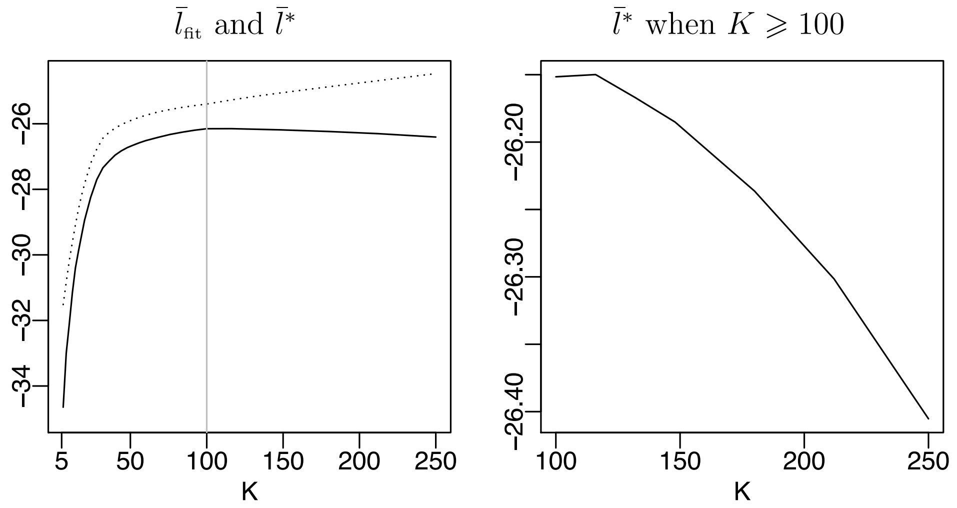

Figure 1.16: (Dotted line) $\bar{l}_{\text{fit}}$; (solid lines) $\bar{l}^*$. The gray vertical line marks the value of K, around which $\bar{l}^*$ changes from increasing to decreasing as K increases. The figure on the right-hand side shows $\bar{l}^*$ for K ranging from 100 to 250.

continues to increase.

Figure 1.18 (left) shows the computation costs of $\widehat{\lambda}^{(\text{U})}_{\text{opt}}$ (solid line) and $\widehat{\lambda}^{(\text{mix})}_{\text{opt}}$ (dashed line). When $K > 100$, $T^{(\text{U})}_{\text{opt}}$ exhibits a quadratic growth with K, whereas $T^{(\text{mix})}_{\text{opt}}$ grows linearly with K. Since there is no gain in statistical efficiency when increasing K beyond 100, the additional computation costs are completely wasted. Figure 1.18 (right) plots the PpS. The largest PpS is obtained when K is around $20 \sim 30$.

Based on our simulation, a rule of thumb in choosing K is $K \leqslant n/100$ in order to avoid overfitting and unnecessary computation costs. Unfortunately, we do not have a single rule to specify K, since there is a trade-off associated with the choice of K. On one hand, small K may result in insufficient overlap between ϕ_{mix} and p, which in turn may result in insufficient overlap between ϕ and $\widetilde{p}$. On the other hand, large K comes with expensive computation costs, and the rate of reduction of the variance decreases when K increases. Currently, we rely on users to specify a reasonable K to

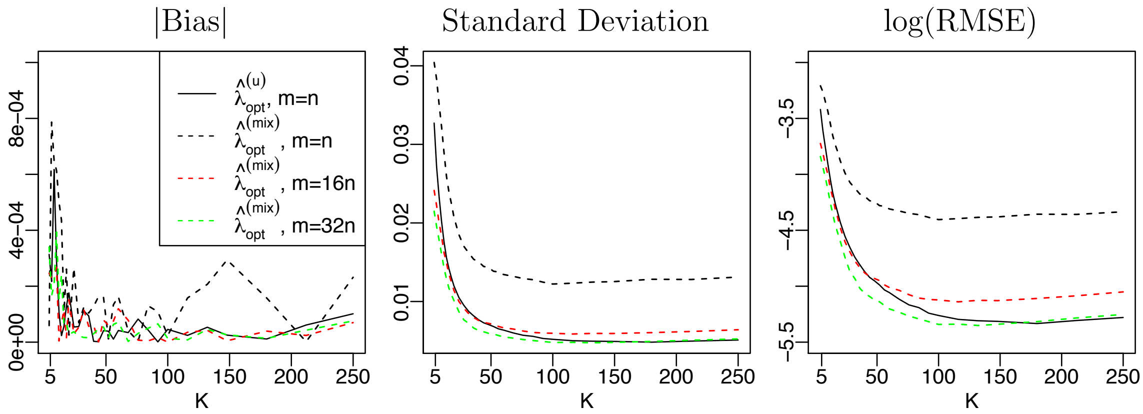

Figure 1.17: The three columns show |bias|, standard deviation and RMSE (on a logarithmic scale) of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$ (solid lines) and $\widehat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt}}$ (dashed lines). Different colors correspond to different values of m in the estimators. Black: $m = n$; red: $m = 16n$; green: $m = 32n$.

balance the statistically efficiency and the associated computation costs.

1.5.2.2 Impact of L on $\widehat{\lambda}^{(\mathcal{X})}_{\mathbf{opt}}$

Other factors being fixed, on average, larger L results in more overlap between p and ϕ_{mix}, hence more overlap between $\widetilde{p}$ and ϕ, and better statistical efficiency of $\widehat{\lambda}^{(\mathrm{U})}_{\alpha}$. If we are not concerned about the computation costs, we should use the whole half dataset to estimate $\boldsymbol{\zeta}$ and the other half in bridge sampling, for each $\widehat{\lambda}^{(\mathrm{U})}_{\alpha,i}$.

Chen et al. (2008) showed that if the data are from a mixture of K normal distributions with parameters $\boldsymbol{\zeta}_0$ and if the penalty term satisfies certain conditions, the pMLE $\widetilde{\boldsymbol{\zeta}}_{L}$ is consistence; that is, $\widetilde{\boldsymbol{\zeta}}_{L} \to \boldsymbol{\zeta}_0$ almost surely as $L \to \infty$. Further, Chen and Tan (2009) showed under these conditions, the central limit theorem holds for $\widetilde{\boldsymbol{\zeta}}_{L}$. So if p is a mixture of K normal distributions exactly as specified in (1.31), then as $L \to \infty$, $\phi_{\text{mix}}(\cdot; \widetilde{\boldsymbol{\zeta}}_{L})$ will converge to p, and thus the discrepancy between ϕ and $\widetilde{p}$ will diminish to zero. Under more likely circumstances where p is not in the family of (1.31), for fixed K, the distance between p and $\phi_{\text{mix}}(\cdot; \widetilde{\boldsymbol{\zeta}}_{L})$ remains positive and not

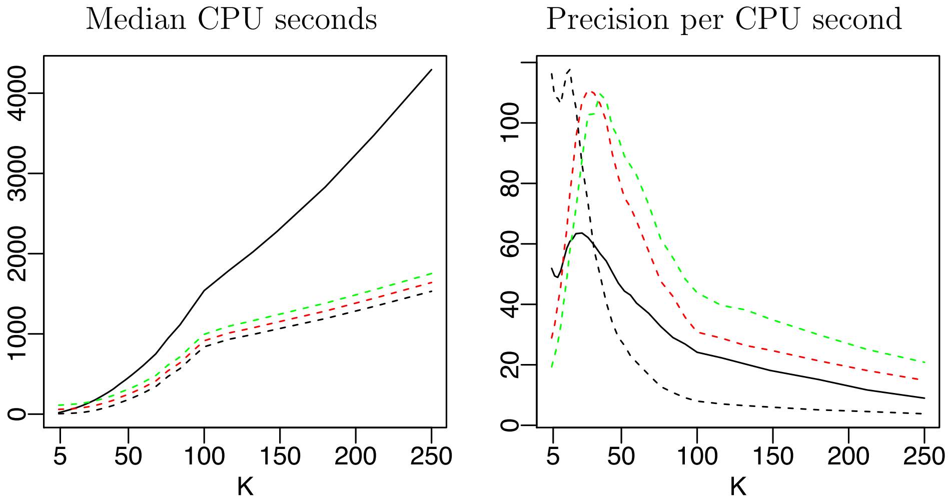

Figure 1.18: The total computation costs $T_{\text{opt}}^{(\mathcal{X})}$ (left), and the precision per CPU second (right) of the optimal bridge sampling estimators $\widehat{\lambda}_{\text{opt}}^{(\text{U})}$ (solid lines, $m = n$) and $\widehat{\lambda}_{\text{opt}}^{(\text{mix})}$ (dashed lines) with $m = n$ (black), $16n$ (red), and $32n$ (green).

negligible, even with $L \to \infty$, see Appendix A.2 for the theoretical calculations.

Figure 1.19 shows the impact of L on the performance of the algorithm to obtain $\widehat{\lambda}_{\text{opt}}^{(\mathcal{X})}$. The size of the sample to estimate $\boldsymbol{\zeta}$ should be linearly dependent on K, so we compare estimators with different values of L/K. In terms of the computation costs, for fixed K, L only affect T_{EM}, so both T_{EM} and $T_{\text{opt}}^{(\mathcal{X})}$ grows linearly as L, as shown in Figure 1.19 (left and middle), where K is set to be 5 (black lines), 25 (red lines), and 50 (green lines). On the side of statistical properties, for a fixed K, having more data to estimate each parameter on average results in more overlap between p and $\phi_{\text{mix}}(\cdot; \widetilde{\boldsymbol{\zeta}}_L)$, whereas small L/K may cause an overfitting problem and thus large divergence between p and $\phi_{\text{mix}}(\cdot; \widetilde{\boldsymbol{\zeta}}_L)$. Figure 1.19 (right) shows the RMSE decreases as L/K increases, but when $L/K > 50$, the reduction rate becomes very small. Figure 1.24 shows the similar impact of L/K on the RMSE of $\widehat{\lambda}_{\text{opt}}^{(\mathcal{X})}$ in a different example (50 dimensions). So we recommend setting L to be around $50K$.

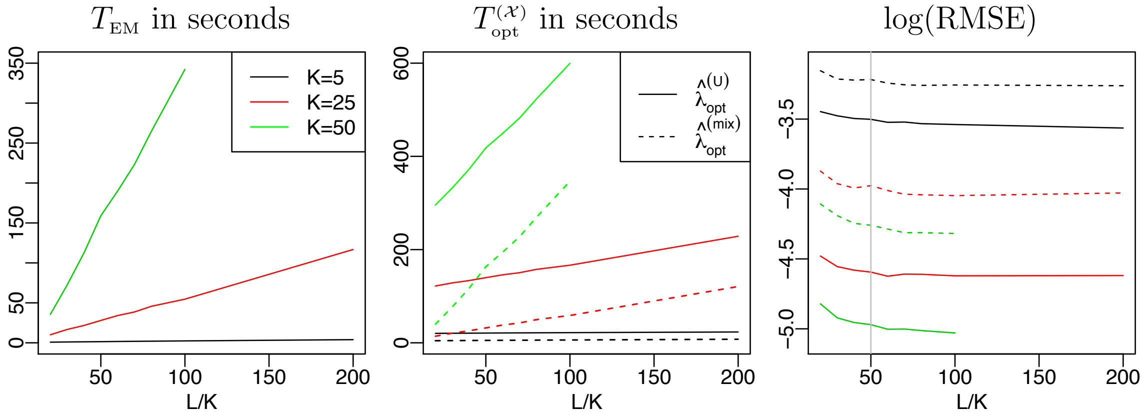

Figure 1.19: The impact of L on T_{EM} (left), $T^{(\mathcal{X})}_{\mathrm{opt}}$ (middle), and the RMSE of $\widehat{\lambda}^{(\mathcal{X})}_{\mathrm{opt}}$ (right). Black lines: $K = 5$; red lines: $K = 25$; green lines: $K = 50$. When $K = 50$, we can only take L/K up to 100, because $L \leqslant n/2$.

1.5.2.3 Impact of m and A Comparison of $\widehat{\lambda}^{(\mathrm{mix})}_{\alpha}$ and $\widehat{\lambda}^{(\mathrm{U})}_{\alpha}$

Similar to K and L/K, larger m improves the precision of $\widehat{\lambda}^{(\mathcal{X})}_{\alpha}$ but increases the computation costs. Figure 1.20 (left) shows the total computation costs of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$ (solid lines) and $\widehat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt}}$ (dashed lines) grow linearly as m increases from n to $64n$. Figure 1.20 (middle) shows the standard deviation of $\widehat{\lambda}^{(\mathcal{X})}_{\alpha}$ is inversely related to m, and the reduction rate decreases as m increases.

Consistent with our theoretical results, Figure 1.17, 1.19, and 1.20 all show with the same tuning parameters (K, L, m), the variance of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$ is smaller than that of $\widehat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt}}$, but $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$ is computationally much more costly than $\widehat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt}}$. Figure 1.17 (middle) shows difference between the variances of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$ and $\widehat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt}}$ increases as K increases. A possible explanation is the following. In the process of warp-U transformation, the overlap of $\widetilde{p}^{(\mathrm{k})}$ and $\widetilde{\phi}^{(\mathrm{k})}$ remains the same as that of $p^{(\mathrm{k})}$ and $\phi^{(\mathrm{k})}$, and the additional overlap comes from the rematching of $\widetilde{p}^{(\mathrm{k})}$ with the remainder of $\widetilde{\phi}^{(\mathrm{j})}$ (for $j \neq k$) that does not overlap with $\widetilde{p}^{(\mathrm{k})}$. The total number of possible rematches is $K(K-1)/2$,

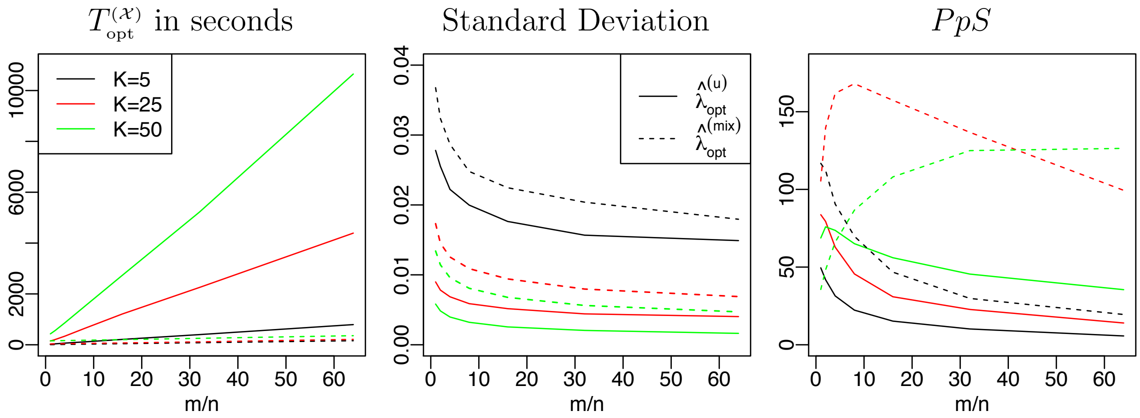

Figure 1.20: The total amount of time $T^{(\mathcal{X})}_{\text{opt}}$ (left), the standard deviation (middle), and the *PpS* of $\widehat{\lambda}^{(\text{U})}_{\text{opt}}$ (solid lines) and $\widehat{\lambda}^{(\text{mix})}_{\text{opt}}$ (dashed lines) with different values of m for $K = 5$ (black), 25 (red), and 50 (green).

so as K increases, it is more likely to form more additional overlap.

The advantage of $\widehat{\lambda}^{(\text{mix})}_{\alpha}$ over $\widehat{\lambda}^{(\text{U})}_{\alpha}$ is the inexpensive computation costs, $T^{(\text{mix})}_{\text{BS}}$, compared with $T^{(\text{U})}_{\text{BS}}$ and T_{EM}. When T_{EM} dominates $T^{(\text{mix})}_{\text{BS}}$, which is often true with large K (see Figure 1.15), we can increase m for the estimator $\widehat{\lambda}^{(\text{mix})}_{\alpha}$ to improve its statistical efficiency without significantly increasing the overall computation time. For easy reference, we use $\widehat{\lambda}^{(\mathcal{X})}_{\alpha}(m)$ to denote the estimator $\widehat{\lambda}^{(\mathcal{X})}_{\alpha}$ with a specific configuration of m. Figure 1.18 (left) show that for large K, the difference among the computation costs of $\widehat{\lambda}^{(\text{mix})}_{\text{opt}}$ when $m = n$ (black dashed line), $16n$ (red dashed line), and $32n$ (green dashed line), is negligible compared with T_{EM}. The variance of $\widehat{\lambda}^{(\text{mix})}_{\text{opt}}$, however, drops substantially when m increases from n to $32n$ in Figure 1.17 (middle). In fact, $\widehat{\lambda}^{(\text{mix})}_{\text{opt}}(16n)$ and $\widehat{\lambda}^{(\text{mix})}_{\text{opt}}(32n)$ are comparable with $\widehat{\lambda}^{(\text{U})}_{\text{opt}}(n)$ in terms of statistical efficiency, but $\widehat{\lambda}^{(\text{U})}_{\text{opt}}(n)$ is much more costly. Consequently, $\widehat{\lambda}^{(\text{mix})}_{\text{opt}}(16n)$ and $\widehat{\lambda}^{(\text{mix})}_{\text{opt}}(32n)$ have larger *PpS* than $\widehat{\lambda}^{(\text{U})}_{\alpha}(n)$ for moderate and large K, see Figure 1.18 (right).

Figure 1.20 shows the statistical efficiency of $\widehat{\lambda}^{(\text{U})}_{\text{opt}}$ can also be improved by increasing m, but the additional computation costs are significant. So in most cases, the

PpS of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$ decreases as m increases. It is, however, important to acknowledge that for a fixed sample from p of size n, the best statistical efficiency achieved by $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$ is better than that by $\widehat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt}}$, and the expensive computation costs of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$ can easily be overcome by parallel computing.

To sum up, it is sufficient to use $L = 50K$ data points from p to estimate $\boldsymbol{\zeta}$. When $L > 50K$, the additional computation costs may not be reflected in the improvement of the statistical efficiency. When $L < 50K$, overfitting may occur, resulting in more divergence between p and $\phi_{\mathrm{mix}}(\cdot; \widetilde{\boldsymbol{\zeta}}_L)$. The variance of $\widehat{\lambda}^{(\mathcal{X})}_{\mathrm{opt}}$ can be effectively reduced by increasing K and/or m up to some points. The rate of reduction in variance is different for K and m. When K is small, increasing K reduces the variance faster than increasing m; when K is large, increasing m is more beneficial to reduce the variance. For the estimator $\widehat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt}}$, having a large m, e.g., $m = 10n$, is recommended thanks to the inexpensive computation costs $T^{(\mathrm{mix})}_{\mathrm{BS}}$.

1.5.3 An Example in 50 Dimensions

In this section, we use an example in 50 dimensions to show our proposed algorithm works in high dimensions, and to further support the comparisons we made in Section 1.5.2. Here, p is a mixture of 30 distributions, including Gaussian distributions, t-distributions, Cauchy distributions, and multivariate distributions with gamma and/or exponential marginal distributions and with Gaussian copulas. Evaluating p at a point is about 700 times more costly than evaluating ϕ. The contour plots of p in Figure 1.21 show the density has very long tails and is quite skewed in some directions. The simulation results are based on 10,000 replications, and in each

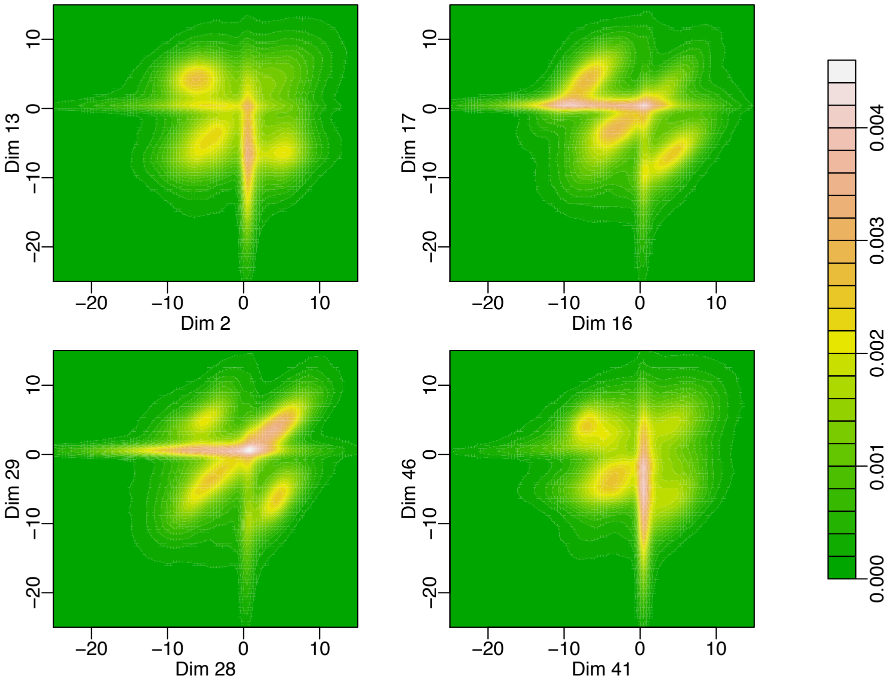

Figure 1.21: *The contours of the density p projected to different pairs of dimensions.*

replication, $n = 10,000$ data points from p are generated.

Figure 1.22 displays the computation costs of each of the steps in obtaining $\widehat{\lambda}_{\alpha}^{(\mathcal{X})}$. Figure 1.23 shows the total computation costs, the RMSE, and the PpS of $\widehat{\lambda}_{\text{opt}}^{(\mathcal{X})}$. As in the 10-dimension example, the RMSE decreases as K increases up to 100, and when $K > 100$, the mixture model overfits the data, resulting in a slight increase in the RMSE. On average, the RMSE of $\widehat{\lambda}_{\text{opt}}^{(\text{U})}$ is 60% of RMSE($\widehat{\lambda}_{\text{opt}}^{(\text{mix})}$), but the computation costs of $\widehat{\lambda}_{\text{opt}}^{(\text{U})}$ are 4.7 times of $T_{\text{opt}}^{(\text{mix})}$. In terms of the PpS, $\widehat{\lambda}_{\text{opt}}^{(\text{mix})}$ is superior to $\widehat{\lambda}_{\text{opt}}^{(\text{U})}$ for any value of K. The comparison of the examples in 10D and 50D indicates that the superiority of $\widehat{\lambda}_{\text{opt}}^{(\text{mix})}$ over $\widehat{\lambda}_{\text{opt}}^{(\text{U})}$ is more evident for larger t_q/t_ϕ. In addition, for large K, when we increase m from n (black lines) to $16n$ (red) and $32n$ (green), the total computation costs of $\widehat{\lambda}_{\text{opt}}^{(\text{mix})}$ only increases by a small fraction, but the gain in

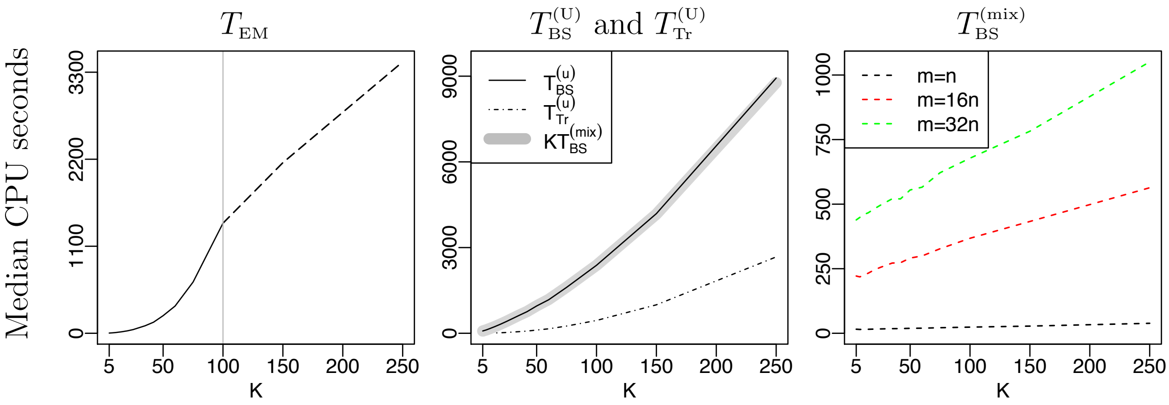

Figure 1.22: The time decomposition in the process of obtaining $\widehat{\lambda}^{(\mathcal{X})}_{\alpha}$ (50 dimension). See the caption in Figure 1.15.

statistical efficiency is substantial.

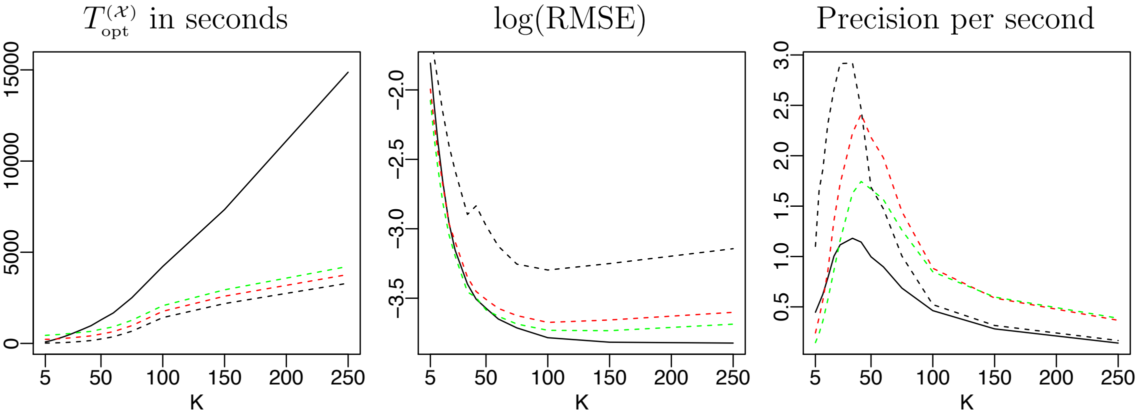

Figure 1.23: The total computation costs $T^{(\mathcal{X})}_{\rm opt}$ (left), the RMSE (middle) on a logarithmic scale, and the *PpS* (right) of $\widehat{\lambda}^{\rm (U)}_{\rm opt}$ (solid lines, $m = n$) and $\widehat{\lambda}^{\rm (mix)}_{\rm opt}$ (dashed lines) with $m = n$ (black), $16n$ (red), and $32n$ (green).

Figure 1.24 shows the impact of increasing L/K on $T_{\rm EM}$ (left), $T^{(\mathcal{X})}_{\rm opt}$ (middle), and the log(RMSE) (right) of estimators with $K = 5$ (black lines), 25 (red), and 50 (green). Consistent with the results in Figure 1.19, as L/K increases up to 50, the statistical efficiencies of the estimators improve considerably, but as we continue to increase L/K, the slope of the curves of log(RMSE) become very gradual. So this

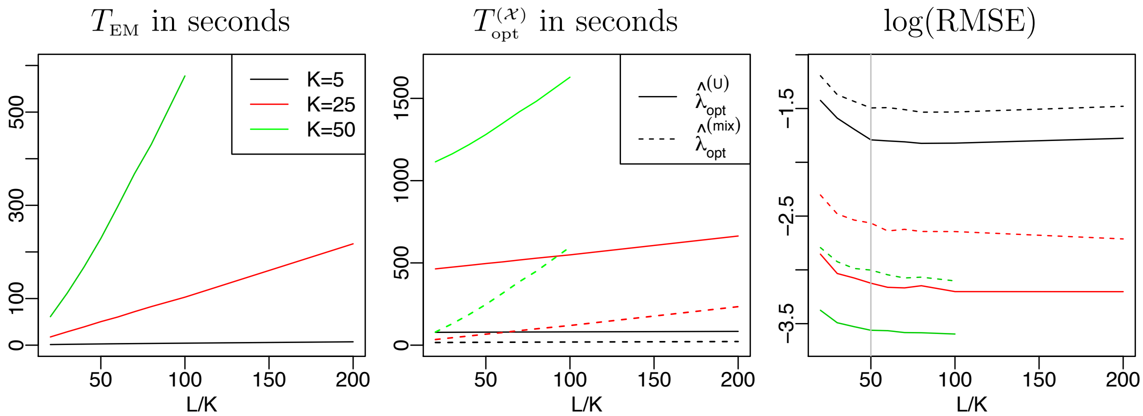

Figure 1.24: The impact of L on T_{EM} (left), $T_{\text{opt}}^{(\mathcal{X})}$ (middle), and the RMSE of $\widehat{\lambda}_{\text{opt}}^{(\mathcal{X})}$ (right) in the 50-dimension example. Black lines: $K = 5$; red: $K = 25$; green: $K = 50$. When $K = 50$, we can only take L/K up to 100, because $L \leqslant n/2$.

example also supports the choice of $L = 50K$.

1.5.4 Estimating c_1/c_2

So far, we have mainly focused on estimating one normalizing constant. Suppose we have $\{w_{i,1}, \cdots, w_{i,n_i}\} \overset{iid}{\sim} p_i = q_i/c_i$ for $i = 1, 2$, then the ratio of the two normalizing constants c_1 and c_2 can be obtained with the following three procedures:

1. Estimate λ_1 and λ_2 separately via warp-U bridge sampling. We denote the two estimators to be $\widehat{\lambda}_{\alpha,\text{I}}^{(\text{U})}$ and $\widehat{\lambda}_{\alpha,\text{II}}^{(\text{U})}$, so $\lambda = \log(c_1/c_2)$ is estimated by $\widehat{\lambda}_{\alpha,\text{I-II}}^{(\text{U})} = \widehat{\lambda}_{\alpha,\text{I}}^{(\text{U})} - \widehat{\lambda}_{\alpha,\text{II}}^{(\text{U})}$.

2. Estimate λ_1 and λ_2 separately by the algorithm of $\widehat{\lambda}_{\alpha}^{(\text{mix})}$, denoted as $\widehat{\lambda}_{\alpha,\text{I}}^{(\text{mix})}$ and $\widehat{\lambda}_{\alpha,\text{II}}^{(\text{mix})}$, and the corresponding estimator of λ is $\widehat{\lambda}_{\alpha,\text{I-II}}^{(\text{mix})} = \widehat{\lambda}_{\alpha,\text{I}}^{(\text{mix})} - \widehat{\lambda}_{\alpha,\text{II}}^{(\text{mix})}$.

3. Estimate the ratio directly by applying bridge sampling to the two sets of warp-U transformed data. More specifically, we divide the data $\{w_{i,1}, \cdots, w_{i,n_i}\}$ into two halves, estimate a Gaussian mixture distribution, $\phi_{\text{mix},i}$, from the $L_i(\leqslant n/2)$

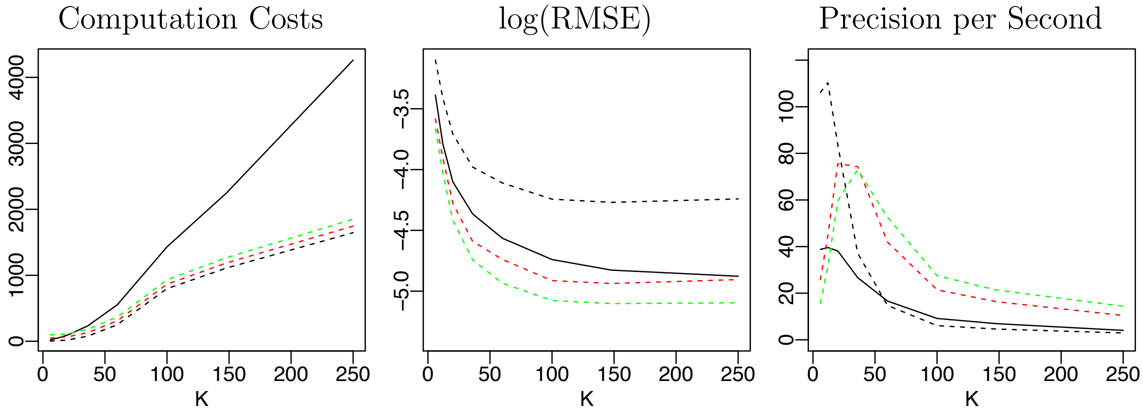

Figure 1.25: $T^{(\mathcal{X})}_{\text{opt}}$ (left), log(RMSE) (middle), and PpS (right) of $\widehat{\lambda}^{(\text{U})}_{\text{opt,II}}$ (solid lines, $m = n$), and $\widehat{\lambda}^{(\text{mix})}_{\text{opt,II}}$ (dashed lines) with $m = n$ (black), $16n$ (red), and $32n$ (green).

of the first half of the data, and apply the corresponding warp-U transformation to the second half of the data. Given the calibrated $\phi_{\text{mix},i}$, both the transformed datasets, $\{\widetilde{w}_{i,j}; j = 1 + \frac{n_i}{2}, \cdots, n_i\}$ for $i = 1, 2$, have substantial overlap with the common density, ϕ, so we expect they overlap with each other substantially. Therefore, we can apply bridge sampling to $\{\widetilde{w}_{i,j}; j = 1 + \frac{n_i}{2}, \cdots, n_i\} \overset{iid}{\sim} \widetilde{q}_i$ for $i = 1, 2$, and obtain one estimate of λ, denoted as $\widehat{\lambda}^{(\text{U})*}_{\alpha,1}$. Reversing the roles of the two halves of the datasets, we obtain a different estimate, $\widehat{\lambda}^{(\text{U})*}_{\alpha,2}$. By symmetry, the final estimator is $\widehat{\lambda}^{(\text{U})*}_{\alpha} = \frac{1}{2}\left(\widehat{\lambda}^{(\text{U})*}_{\alpha,1} + \widehat{\lambda}^{(\text{U})*}_{\alpha,2}\right)$.

In our simulation to compare the three estimators of λ, p_1 is the same as p in the 10-dimension example, and p_2 is a mixture of 20 skewed t-distributions, which is more spread out than p_1 and has heavier correlation among different dimensions. The results are based on 5,000 replications, and $n_1 = n_2 = 10,000$ data points are simulated from p_1 and p_2 in each replication.

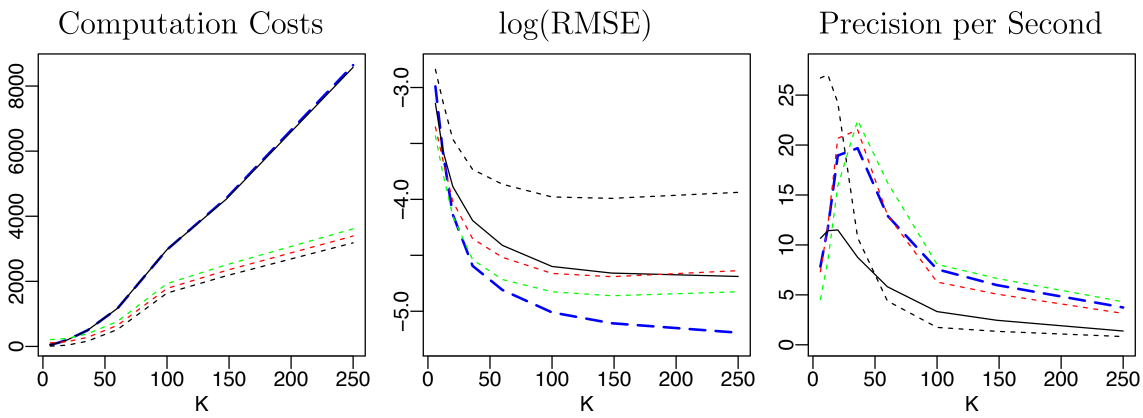

Figure 1.26: Total computation costs (left), log(RMSE) (middle), and *PpS* (right) of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt,I\text{-}II}}$ (solid lines, $m = n$), $\widehat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt,I\text{-}II}}$ (dashed lines) with $m = n$ (black), $16n$ (red), and $32n$ (green), and $\widehat{\lambda}^{(\mathrm{U})*}_{\mathrm{opt}}$ (long dashed lines).

Figure 1.17 and 1.18 show the statistical and computation efficiency, as well as the *PpS* of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt,I}}$ (solid lines), and $\widehat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt,I}}$ (dashed lines) with $m = n$ (black), $16n$ (red), and $32n$ (green). Figure 1.25 shows these summary statistics of $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt,II}}$ and $\widehat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt,II}}$. Figure 1.26 shows the summary statistics of the final estimators $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt,I\text{-}II}}$, $\widehat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt,I\text{-}II}}$, and $\widehat{\lambda}^{(\mathrm{U})*}_{\mathrm{opt}}$ (blue long dashed lines). The computation costs of $\widehat{\lambda}^{(\mathrm{U})*}_{\mathrm{opt}}$ is almost identical to $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt,I\text{-}II}}$, because $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt,I\text{-}II}}$ involves an additional evaluation of the normal density $2m$ times, which is negligible. Figure 1.26 (middle) shows $\mathrm{Var}\left(\widehat{\lambda}^{(\mathcal{X})}_{\mathrm{opt,I\text{-}II}}\right) = \mathrm{Var}\left(\widehat{\lambda}^{(\mathcal{X})}_{\mathrm{opt,I}}\right) + \mathrm{Var}\left(\widehat{\lambda}^{(\mathcal{X})}_{\mathrm{opt,II}}\right)$, which is consistent with the fact $\widehat{\lambda}^{(\mathcal{X})}_{\mathrm{opt,I}}$ and $\widehat{\lambda}^{(\mathcal{X})}_{\mathrm{opt,II}}$ are independent. Interestingly, $\widehat{\lambda}^{(\mathrm{U})*}_{\mathrm{opt}}$ has a much better statistical efficiency than $\widehat{\lambda}^{(\mathrm{U})}_{\mathrm{opt,I\text{-}II}}$ or $\widehat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt,I\text{-}II}}$, and the reduction of the RMSE is less affected by the overfitting issue than other estimators as K increases to 250. Figure 1.26 (right) shows the *PpS* of $\widehat{\lambda}^{(\mathrm{U})*}_{\mathrm{opt}}$ is comparable with that of $\widehat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt,I\text{-}II}}$ when $m = 16n$ or $32n$. So $\widehat{\lambda}^{(\mathrm{U})*}_{\mathrm{opt}}$ has the advantage of having the lowest RMSE and the largest *PpS*.

1.6 Challenges and Opportunities

This chapter generalizes the warp-I, II, and III transformation proposed by Meng and Schilling (2002), and introduces a class of stochastic transformation that aims at reducing the f-divergence of two densities without changing their normalizing constants. Asymptotically, the bridge sampling estimator with the warp-U transformed data has smaller variance than that based on the original data.

Warp-U transformation is determined by a mixture distribution, ϕ_{mix}, that has reasonable amount of overlap with the density p, the normalizing constant of which is of interest. We suggest using the penalized EM algorithm proposed by Chen et al. (2008) to fit a mixture of normal distributions with diagonal covariance matrixes to the data from p. This method is computationally inexpensive, scales linearly with the dimension, and can capture a large mass of p for a reasonably chosen number of components in ϕ_{mix}. Adaptive bias is introduced if the estimated parameters in ϕ_{mix} and the data used in bridge sampling are dependent. We propose one solution that removes the bias without incurring additional variance. More specifically, the data are divided into two halves, and we obtain two estimators by using part of one half of the data to estimate $\boldsymbol{\zeta}$ and the other half in bridge sampling. The two resulting estimators have very small correlation, so the statistical efficiency of the combined estimator, i.e., their average, is as good as if the parameters in ϕ_{mix} are estimated with other resources.

The selections of the three tuning parameters (K, L, m) are discussed in details with theoretical and simulation results. In addition, we compare the statistical and computational efficiencies of the optimal warp-U bridge sampling estimator $\widehat{\lambda}_{\text{opt}}^{(\text{U})}$ and

$\widehat{\lambda}_{\text{opt}}^{(\text{mix})}$ (the optimal bridge sampling estimator with data from p and ϕ_{mix}). For a fixed ϕ_{mix} and sample sizes (n, m), asymptotically, $\widehat{\lambda}_{\text{opt}}^{(\text{U})}$ has better statistical efficiency than $\widehat{\lambda}_{\text{opt}}^{(\text{mix})}$. However, $\widehat{\lambda}_{\text{opt}}^{(\text{U})}$ is computationally more expensive than $\widehat{\lambda}_{\text{opt}}^{(\text{mix})}$, especially for large K. Thus, if computation costs are of concern, $\widehat{\lambda}_{\text{opt}}^{(\text{mix})}$ with a large sample from ϕ_{mix} is often a better choice than $\widehat{\lambda}_{\text{opt}}^{(\text{U})}$ with a small sample from ϕ. Another advantage of warp-U transformation is that we can apply bridge sampling to the two transformed datasets to estimate the ratio of two normalizing constants directly.

Like any research, we also face many challenges which require further exploration.

First of all, Theorem 1 implies that for any continuous density ϕ and p, the f-divergence between (ϕ_{mix}, p) is larger than that between $(\phi, \widetilde{p})$, the densities due to the warp-U transformation. We have only explored the Gaussian distribution as the base density ϕ, so one future direction is to investigate other base densities. For heavy-tail problems, using t-distribution may be more effective in capturing the mass of p, requiring fewer components and thus smaller computation costs. If the support of p is within a bounded region, a base density with bounded support may be more appropriate. Nonetheless, we suggest using the diagonal covariance matrix for the components in ϕ_{mix} and using our strategy in Figure 1.12 to remove the adaptive bias.

Second, the possibility of using parallel computation techniques to speed up computation is discussed in this chapter. The step of bridge sampling with the warp-U transformed data is embarrassingly parallelizable, since no communication of the results is needed. With many computation resources available, the bottleneck may lies in the estimation of $\boldsymbol{\zeta}$ via the EM algorithm, which is not easily to parallelize. So a parallel version of EM or clustering methods, such as k-means algorithm, may be used to reduce the computation costs even further.

Third, we assume the number of components in ϕ_{mix} is pre-specified by the user and have not provided much guidance on how to choose K, except that K should be $\leqslant n/100$ to avoid overfitting. Both the computation costs and the statistical efficiency of the warp-U bridge sampling estimator increase with K. So a reasonable criterion for choosing K should take into account both factors, for example, a weighted average of the statistical efficiency and computation costs, where the weights represent the relative importance of the two factors. Whereas we can get a good estimate of the computation costs as a function of K, it is difficult to estimate how the MSE of the estimator changes with K. Some widely-used criteria for choosing the optimal K, such as AIC and BIC, deal with the trade-off between the goodness of fit of the model and its complexity, and may be used for our purpose. But keep in mind these criteria may not directly represent the statistical and computational efficiency of the warp-U bridge sampling estimator. The method proposed by Lee et al. (2006), which estimates the optimal K by dynamically adding the components one by one based on incremental k-means until some criteria are met, may be a possible solution to our problem.

Another interesting area of research is the connection of the estimator $\widehat{c}_{\text{opt}}^{(\text{U})}$ with the likelihood method proposed by Kong et al. (2003). The estimator $\widehat{c}_{\text{opt}}^{(\text{mix})}$ is essentially a special case of the likelihood method. More specifically, let $q_0 = q = cp$ be the unnormalized density, q_i be the pdf of $\mathcal{N}(\mu_i, \Sigma_i)$, for $i = 1, \cdots, K$, $\phi_{\text{mix}} = \sum_{i=1}^{K} \pi_i q_i$, n_i be the number of draws from q_i, $n = n_0$, and $m = \sum_{i=1}^{K} n_i$, then, when $m \to \infty$

and $n_i/m \to \pi_i$, the likelihood estimator defined below is equivalent to $\widehat{c}^{(\mathrm{mix})}_{\mathrm{opt}}$,

$$\widehat{c} = \sum_{i=0}^{K} \sum_{j=1}^{n_i} \frac{q(w_{i,j})}{n\widehat{c}^{-1} q(w_{i,j}) + m\left(\sum_{k=1}^{K} \frac{n_k}{m} q_k(w_{i,j})\right)}. \tag{1.39}$$

There are many other questions we have not explored. Theorem 1 states $\mathcal{D}(\widetilde{p}, \phi) \leqslant \mathcal{D}(p, \phi_{\mathrm{mix}})$ for any increasing function of the f-divergence. However, we have little knowledge about the amount of reduction in the divergence, and when the superiority of $\widehat{\lambda}^{(\mathrm{U})}_{\alpha}$ to $\widehat{\lambda}^{(\mathrm{mix})}_{\alpha}$ is more evident in terms of the statistical efficiency. In addition, we have not investigated how the dependence of the data would affect our algorithm. We would also like to explore more applications of the warp-U transformation.

Chapter 2

Bayesian Methods for Modeling Source Intensities

2.1 Introduction

One of the goals of source detection is to obtain the luminosity function, which specifies the distribution of source intensities in a population. The luminosity function of a population in X-rays can be estimated using the Chandra X-ray data. The Chandra X-ray Observatory consists of high-resolution count-based detectors, which record the arrival time, the 2D sky coordinates and the energy of each of the X-ray photon arrives at the detectors.

We use independently constructed catalogues from optical, radio, or previous X-ray surveys to locate the position of each source in a population. Centering at each source location in the detector, we use the point spread function (PSF) to determine a circular aperture, also referred to as the source region, so that $\sim 90\%$ of the source photons are expected to fall within the region. In this paper, we develop a novel method to simultaneously modeling the distribution of source intensities of a population, based on the 2D binned counts in the source regions. Thus, we bypass the traditional problem of detection entirely, and directly determine the quantity of astrophysical interest, the population luminosity function. Currently, this is among the first principled methods to use the independently constructed catalogues in an X-ray source detection algorithm.

Several complications are associated with the data. First, the data are contaminated with background counts, which are recorded events that are not originated from the source of interest. To quantify the background counts, we also collect the 2D binned count in a vast region with no sources. A common practice to obtain source counts is to directly subtract the estimated background counts from the observed counts in source regions. This ad-hoc method, however, often leads to negative source counts. To overcome this problem, we model each observed count as the sum of two independent Poisson random variables, the expected values of which are determined by the background rate and the source intensity, respectively.

Second, the photon counts in some source regions are very low, and could be completely from the background. We call a source X-ray dark, if its X-ray source intensity is zero. We are particularly interested in whether such X-ray dark sources exist as a discernible subpopulation of the population. To represent this possibility, we model the distribution of source intensities as a mixture of a gamma distribution for sources with non-zero intensities, and a zero-inflated component (a positive probability at zero) for sources that are intrinsically non-emitting. Note that in most flux-truncated analyses, power-law models (akin to a Pareto distribution) are fit to the data, which by design ignore the downward slopes in the source brightness distribution at low luminosities. The zero-inflated component models a subpopulation of sources that are not just weak, but completely dark, and is an entirely new construct hitherto never considered in astronomical problems. Including such a population in the modeling provides powerful new avenues to investigate, e.g., areas of the H-R diagram where sources have never been reliably detected, such as dA stars, or K and M giants and supergiants. The individual source intensity is constrained by both the observed counts at the source locations and the expected distribution, which acts as a "smoothing" constraint even for weak sources. This obviates the need to determine upper limits to undetected sources, since for every catalog object a full posterior probability distribution of its intensity is obtained and used in the construction of the luminosity function.

Third, there are a number of overlapping source regions in the data, especially at locations far away from the center of the field. We bypass the problem by counting the photons in each of the segments formed by these overlapping regions. We are able to obtain the proportion of photons from each source that are expected to arrive at that segment. This information allows us to principally model the source intensities based on photon counts in the segments.

Finally, the background rate may have an increasing trend as the projected angle on the sky from the center of the field increases. Instead of assuming a constant background rate across the field, we assume it is piecewise constant, i.e., the whole field is divide into several non-overlapping regions and the background within each region is homogeneous.

The remainder of the chapter is organized into three sections. In Section 2.2, we develop the Bayesian method for modeling the X-ray luminosity function. A simulation study, with simulated data mimicking typical stellar clusters, is conducted to demonstrate the performance of our model. In Section 2.3, we propose a likelihood-ratio based Bayesian hypothesis testing procedure, where the posterior predictive p-value is computed to quantify the evidence against the null hypothesis of no X-ray dark sources. The actual levels and powers of the test are examined under a variety of simulation configurations. In Section 2.4, we apply our model and the hypothesis testing procedure to two subsets of the Chandra/HRC-I observation of the stellar open cluster, NGC 2516.

2.2 Modeling the Luminosity Function

2.2.1 Statistical Model

In this section, we introduce the hierarchical model that we use to describe and fit the luminosity function. The model accounts for background contamination and the possibility of dark sources. We take a Bayesian perspective to model fittings, employing non-informative prior distributions for the parameters characterizing the luminosity function because external information is not available for these parameters.

2.2.1.1 Basic Bayesian Hierarchical Model

Suppose we observe a field with n sources for $\mathcal{T}$ seconds, and that for each source we count Y_i photons in its source region. Source region i is a circular aperture centered at the putative location of source i, and its radius is determined by the point spread function (PSF) so that $\sim 90\%$ of the source photons are expected to fall within the region. Let a_i (pixels) denote the area of source region i. We assume the number of putative sources and their locations are determined with other instruments. Jones et al. (2014), for example, investigate a Bayesian method for simultaneously inferring the number of sources, their locations, and their expected photon counts. For simplicity, here we assume there is no overlap among the n source regions; this assumption is relaxed in Section 2.2.1.2. Due to background contamination, Y_i can be written as the sum of photon counts from the source, $\mathcal{S}_i$, and from the background, $\mathcal{B}_i$, i.e.,

$$Y_i = \mathcal{S}_i + \mathcal{B}_i, \tag{2.1}$$

where $\mathcal{S}_i$ and $\mathcal{B}_i$ are independent. We emphasize that we only observe Y_i, and not $\mathcal{S}_i$ or $\mathcal{B}_i$.

In addition to the counts in the individual source regions, we observe a pure background count, X, from a presumably source-free region of area A pixels in the observed field. The exposure time for the pure background observation is also $\mathcal{T}$ seconds. In this section, we assume constant background rate across the field; an extension of the model that allows for different background rates appears in Section 2.2.1.2.

The arrival of photons at the detector can be modeled as a Poisson process, that is,

$\mathcal{S}_i, \mathcal{B}_i$, and X are independent Poisson random variables. In addition to the exposure time and source intensity, λ_i (count/s/cm^2), the expected count from source i in its source region is affected by two known constants: (i) the proportion, r_i, of photons from the source that are expected to fall in the source region as determined by the PSF, and (ii) the telescope effective area, e_i (cm^2), at the source location, which characterizes the efficiency of the telescope, i.e.,

$$\mathcal{S}_i \big| \lambda_i \overset{\text{indep}}{\sim} \text{Poisson}(r_i e_i \lambda_i \mathcal{T}). \tag{2.2}$$

Similarly, with a background rate of ξ (count/s/pixel), the photon count in the background region is modeled

$$X \big| \xi \sim \text{Poisson}(A \xi \mathcal{T}), \tag{2.3}$$

and the background count in source region i is modeled

$$\mathcal{B}_i \big| \xi \overset{\text{indep}}{\sim} \text{Poisson}(a_i \xi \mathcal{T}). \tag{2.4}$$

The background count rates are not adjusted for the effective area because (i) the effective area adjusts only for photon counts, whereas background events include both X-ray photons and charged particles, and (ii) the background rate, ξ, is quantified in terms of the observed count. Finally, the observed count in source region i, $Y_i = \mathcal{S}_i + \mathcal{B}_i$, also follows a Poisson distribution,

$$Y_i \big| (\lambda_i, \xi) \overset{\text{indep}}{\sim} \text{Poisson}\left((a_i \xi + r_i e_i \lambda_i) \mathcal{T}\right). \tag{2.5}$$

We take a Bayesian statistical perspective which involves the computation of the posterior distribution of the unknown parameters, $(\xi, \boldsymbol{\lambda})$ with $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_n)$, given the observed data, $\boldsymbol{D} = (Y_1, \cdots, Y_n, X)$. To do this, we must specify the likelihood function and the prior distribution; a brief introduction to Bayesian data analysis and parameter fitting is given in Appendices A.3 and A.4. Equations (2.3) and (2.5) together define the distribution of $\boldsymbol{D}$, and thus can be used to specify the likelihood function of $(\xi, \boldsymbol{\lambda})$,

$$L(\xi, \boldsymbol{\lambda}|\boldsymbol{D}) = \exp\left(-A\mathcal{T}\xi\right)\frac{(A\mathcal{T}\xi)^X}{X!}\prod_{i=1}^{n}\exp[-(a_i\xi + r_ie_i\lambda_i)\mathcal{T}]\frac{[(a_i\xi + r_ie_i\lambda_i)\mathcal{T}]^{Y_i}}{Y_i!}. \tag{2.6}$$

As for prior distributions, we first specify the prior distribution on ξ as [1]Gamma$[\mu_0, \theta_0]$. This choice simplifies computation because we can derive a closed-form posterior distribution for ξ, given all other model parameters and data. Since the density function of gamma distribution can take various shapes, it is a flexible model for the prior distribution, see Figure 2.1 (left panel) for examples. If prior information is available for the background rate, μ_0 and θ_0 can be chosen accordingly. Otherwise, a nearly flat prior distribution with large variance can be used to reflect prior ignorance. Practically speaking, because the background count X is collected over a large region (A pixels) and over a long period of time ($\mathcal{T}$ seconds), it is quite informative for ξ, and thus a weakly informative prior distribution has little impact on the posterior

[1] For the conventional parametrization of a gamma distribution, Gamma(α, β), the mean and variance are $\mu = \alpha/\beta$ and $\theta = \alpha/\beta^2$, respectively. To simplify interpretation, we instead use the mean-variance parametrization, Gamma$[\mu, \theta]$; the square brackets are used to distinguish the mean-variance parametrization, i.e., the Gamma$[\mu, \theta]$ distribution has mean μ and variance θ, and is equivalent to Gamma(α, β).

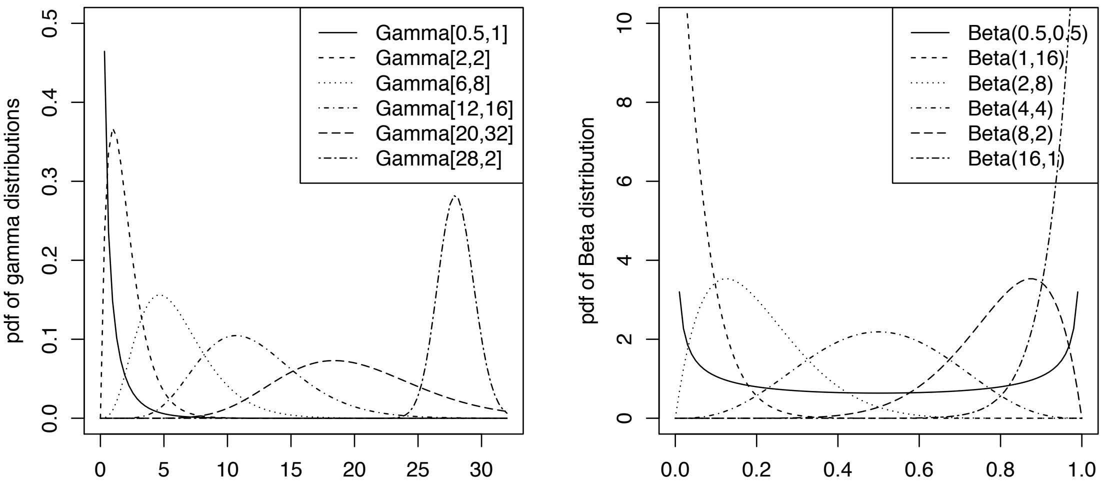

Figure 2.1: (Left) Examples of gamma distributions; (Right) Examples of Beta distributions. The notation used to specify these distributions appears in footnotes 1 and 3.

distribution of ξ. In our numerical studies, we set

$$A\mathcal{T}\xi \sim \text{Gamma}\left[10^6, 10^{18}\right], \text{ or equivalently, } \xi \sim \text{Gamma}\left[\frac{10^6}{A\mathcal{T}}, \frac{10^{18}}{(A\mathcal{T})^2}\right], \tag{2.7}$$

which is nearly a flat prior distribution for $A\mathcal{T}\xi$.

To capture the possibility of X-ray dark sources, we model the population prior distribution of source intensities as a zero-inflated gamma distribution, i.e., a mixture of a gamma distribution and the δ function at zero.[2] The gamma component of this prior distribution describes the intensities of sources that are not X-ray dark, whereas the δ function represents the X-ray dark sources in the field, i.e., those having $\lambda_i = 0$. We model $\{\lambda_1, \cdots, \lambda_n\}$ as independent random variables from this population prior

[2] The δ function is a discrete distribution. If $X \sim \delta_0$, then $P(X = 0) = 1$.

distribution, i.e., independently,

$$\lambda_i \big| \mu, \theta, \pi_d \begin{cases} = 0 & \text{with probability } \pi_d, \\ \sim \text{Gamma}[\mu, \theta] & \text{with probability } 1 - \pi_d, \end{cases} \tag{2.8}$$

where π_d is the proportion of dark sources. The prior distribution of $\boldsymbol{\lambda}$ can be written

$$P(\boldsymbol{\lambda} \big| \mu, \theta, \pi_d) = \prod_{i=1}^{n} \left(\pi_d \delta_0(\lambda_i) + (1 - \pi_d) \frac{(\mu/\theta)^{(\mu^2/\theta)}}{\Gamma(\mu^2/\theta)} \lambda_i^{(\mu^2/\theta)-1} e^{-(\mu/\theta)\lambda_i} \right), \tag{2.9}$$

where $\delta_0(\lambda_i) = 1$ if $\lambda_i = 0$, and $\delta_0(\lambda_i) = 0$ if $\lambda_i \neq 0$.

Because we observe multiple sources, we can fit the parameters, μ, θ, and π_d of the population prior distribution on $\boldsymbol{\lambda}$. To do so, we must also set prior distributions on μ, θ, and π_d, and study their posterior distribution. There is an important difference between the prior distributions on ξ and on $\boldsymbol{\lambda}$. A priori, ξ is assumed to follow Gamma$[\mu_0, \theta_0]$, where μ_0 and θ_0 are constants of our choice, whereas the distribution of $\boldsymbol{\lambda}$ is modeled in a hierarchical fashion, i.e., we can leverage the replicates $\{\lambda_1, \cdots, \lambda_n\}$ to fit the population prior distribution and set fixed prior distributions on the parameters describing this population distribution.

We assume μ, θ, and π_d are a priori independent. A natural prior distribution for π_d is a Beta distribution,[3] the support of which is the unit interval. Examples of Beta distributions are shown in Figure 2.1 (right panel). Prior knowledge as to likely values of the proportion of the dark sources can be incorporated into this prior

[3] The probability density function of Beta(α, β) is $P(x) = (\Gamma(\alpha)\Gamma(\beta))^{-1}\Gamma(\alpha + \beta)x^{\alpha-1}(1 - x)^{\beta-1}$, where $0 < x < 1$. The mean and variance of the distribution are $\alpha/(\alpha + \beta)$ and $\alpha\beta/\left[(\alpha + \beta + 1)(\alpha + \beta)^2\right]$, respectively.

distribution. Otherwise, we can assume π_d follows Beta$(1,1)$, which is a uniform distribution between zero and one,

$$\pi_d \sim \text{Beta}(1,1) = \text{Uniform}(0,1). \tag{2.10}$$

Setting a prior distribution on (μ, θ) is more subtle. It may be tempting to specify a flat prior distribution on (μ, θ), because these parameters can in principle take on any positive values. A flat prior distribution on $(0, \infty)$, however, is not a proper distribution (since it cannot be normalized) and leads to technical difficulties in this case, see Appendix A.6. Thus, we must specify a prior distribution that can be normalized. Nonetheless, we would like the prior distribution of (μ, θ) to reflect ignorance since we have no prior knowledge about their values. To accomplish this, we specify a relatively flat and heavy-tailed prior distribution with the aim of ensuring that the posterior distribution is driven by $\boldsymbol{D}$, rather than the prior distribution.

We start by identifying a rough range of $\boldsymbol{\lambda}$ using background subtraction. Assuming a constant background rate, we expect that approximately, $\widehat{\mathcal{B}}_i = Xa_i/A$ photons in source region i are due to the background and $\widehat{\mathcal{S}}_i = Y_i - \widehat{\mathcal{B}}_i$ photons are due to source i. Thus, $\widehat{\lambda}_i = \widehat{\mathcal{S}}_i/(r_i e_i \mathcal{T})$ is a rough estimate of λ_i. Figure 2.2 shows histograms of Y_i and $\widehat{\lambda}_i$ for the non-overlapping sources within 6 arcmin from the center of the field in the Chandra/HRC-I observation of the open cluster NGC 2516, in which 44.6% of the $\widehat{\lambda}_i$ are negative. We emphasize that $\widehat{\lambda}_i$ is a poor estimator of λ_i and we do not espouse its use as such. We only use it as a very rough guide in setting a prior distribution. The empirical mean and variance of the positive $\widehat{\lambda}_i$ are 1.2×10^{-6} and 8.7×10^{-12}, respectively. Based on these calculations, we use two independent

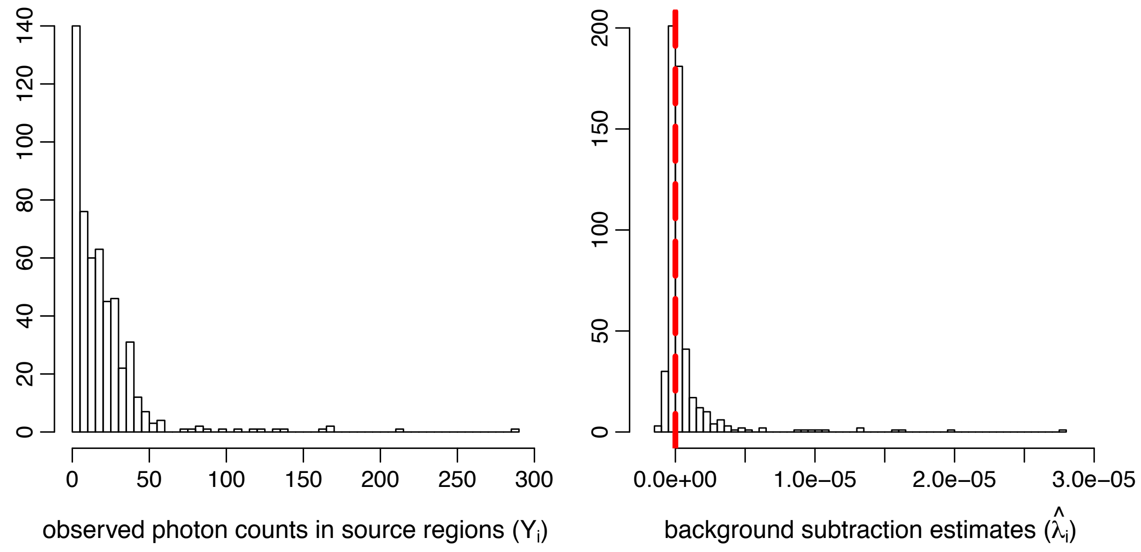

Figure 2.2: (Left) Histogram of the observed photon counts Y_i of non-overlapping sources located within 6 arcmin from the center of the field in the Chandra/HRC-I observation of the open cluster NGC 2516; (right) histogram of the empirical estimates of λ_i obtained by subtracting the estimated background counts from the observed counts.

zero-truncated Cauchy distributions,[4] $\text{tCauchy}\left(\eta_0, \gamma_0\right)$, as the prior distributions for μ and θ. The Cauchy distribution is an appropriate choice for a weakly informative prior distribution (Gelman et al., 2006, 2008) due to its heavy tails. We truncate the prior distribution to be positive because both μ and θ must be greater than zero. Figure 2.3 (left panel) shows several examples of normalized zero-truncated Cauchy distributions. The mode, η_0, of the prior distribution on μ is set to match the mean of the positive $\widehat{\lambda}_i$, and the scale parameter, γ_0, is set to be $100\eta_0$, so that the density of $\text{tCauchy}(\eta_0, \gamma_0)$ only drops by a half within $101\eta_0$ units of the mode. Using the

[4] The probability density of $\text{Cauchy}(\eta, \gamma)$ is $P(x) = \gamma / \left[\pi\left(\gamma^2 + (x-\eta)^2\right)\right]$ for $-\infty < x < \infty$. The zero-truncated Cauchy distribution, denoted as $\text{tCauchy}(\eta, \gamma)$, is proportional to the positive part of $\text{Cauchy}(\eta, \gamma)$. The density of $\text{tCauchy}(\eta, \gamma)$ is $P(x) = \gamma / \left[C\left(\gamma^2 + (x-\eta)^2\right)\right]$ for $x > 0$, where $C = \pi/2 + \arctan(\eta/\gamma)$ is the normalizing constant that ensures $\int_0^\infty P(x)\mathrm{d}x = 1$. The mode of the density is $\min(\eta,\, 0)$. Larger value of the scale parameter, γ, results in to a heavier right-tail.

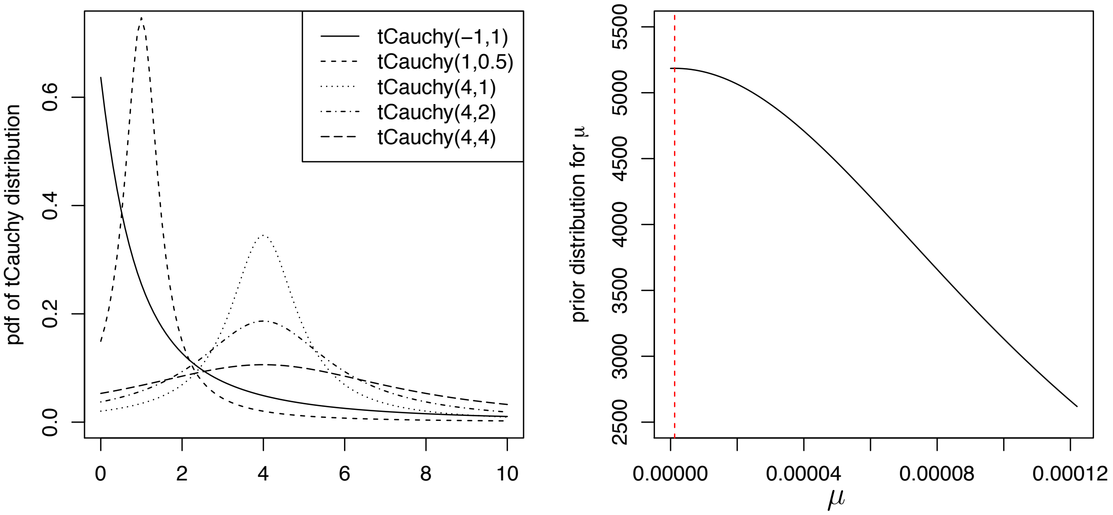

Figure 2.3: (Left) Examples of normalized tCauchy distributions; (Right) density function of the prior distribution for μ, with the mode of the distribution marked by the red vertical line.

same method, we obtain the prior distribution for θ. So

$$\mu \sim \text{tCauchy}(1.2 \times 10^{-6}, 1.2 \times 10^{-4}), \;\; \theta \sim \text{tCauchy}(8.7 \times 10^{-12}, 8.7 \times 10^{-10}). \quad (2.11)$$

Figure 2.3 (right panel) plots this prior distribution of μ, which is fairly flat near the mode and has a heavy right tail.

In all, the unknown parameters that we aim to estimate include μ, θ, π_d, ξ, and $\boldsymbol{\lambda}$. By Bayes' Theorem, their joint posterior distribution is

$$P(\mu, \theta, \pi_d, \xi, \boldsymbol{\lambda} \big| \boldsymbol{D}) \propto L(\xi, \boldsymbol{\lambda} | \boldsymbol{D}) P(\xi) P(\boldsymbol{\lambda} \big| \mu, \theta, \pi_d) P(\pi_d) P(\mu, \theta), \quad (2.12)$$

where the terms on the right-hand side of (2.12) are given by (2.6), (2.7), (2.9), (2.10), and (2.11), respectively.

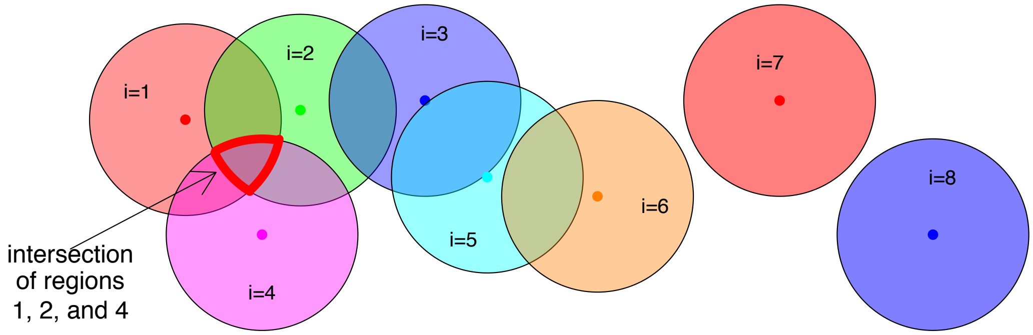

Figure 2.4: An example of eight source regions in the detector. Source regions 7 and 8 do not overlap with other source regions. Source regions 1-6 overlap and form a total of 13 segments. The highlighted segment is the intersection of source regions 1, 2, and 4.

2.2.1.2 Model Extension

When two or more sources are spatially close and their respective PSF-defined source regions overlap, photons observed in one source region are a mixture of photons from multiple sources and from the background. In this section, we extend the basic model of Section 2.2.1.1 to handle overlapping sources.

Instead of modeling the photon count in each source region, we model the count, Y_s, in each segment defined by either a single non-overlapping source region or the intersection of multiple source regions. The subscript, s, denotes the set of sources whose regions overlap and form the segment. For example in Figure 2.4, the highlighted segment is defined by the intersection of source regions 1, 2, and 4, so $s = \{1, 2, 4\}$, and $Y_{\{1,2,4\}}$ is the photon count in the highlighted segment. Each Y_s consists of a mixture of photons from the sources in s and from the background, so

$$Y_s = \sum_{i \in s} \mathcal{S}_{s,i} + \mathcal{B}_s, \tag{2.13}$$

where $\mathcal{S}_{s,i}$ is the photon count from source i detected in segment s, and $\mathcal{B}_s$ is the background count in segment s, with $\mathcal{B}_s$ and $\mathcal{S}_{s,i}$ for $i \in s$ assumed independent. We emphasize only Y_s, and not $\mathcal{S}_{s,i}$ or $\mathcal{B}_s$ is observed.

As in the basic model, the expected count in each segment is computed as a function of several constants: (i) the area, a_s (pixels), of the segment, (ii) the expected proportion, $r_{s,i}$, of photons from source $i \in s$ that are recorded in segment s, and (iii) the effective area, e_s (cm^2), of the segment. In particular, given the source intensity λ_i, $\mathcal{S}_{s,i}$ is modeled as

$$\mathcal{S}_{s,i} \big| \lambda_i \overset{\text{indep}}{\sim} \text{Poisson}(r_{s,i} e_s \lambda_i \mathcal{T}). \tag{2.14}$$

In the basic model of Section 2.2.1.1, we assume the background rate is constant across the field. In reality, this assumption is unrealistic because we observe an increase in the observed background rate as the projected angle (in arcmin) on the sky from the center of the field increases from 0 to 16, see Table 2.1. Here, we extend the model to allow for piecewise homogeneous background. More specifically, we divide the field into K fixed regions and assume a constant background rate, denoted as ξ_k, in each region. A priori, we assume $\xi_1, \cdots, \xi_K$ are independently distributed as Gamma$\left[\mu_0, \theta_0\right]$, as in (2.7).

Let X_k be the observed background count and A_k (pixels) be the area of the pure background in region k, for $k = 1, \cdots, K$. As in (2.3), given $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_K)$, $\boldsymbol{X} = (X_1, \cdots, X_K)$ can be modeled as

$$X_k \big| \xi_k \overset{\text{indep}}{\sim} \text{Poisson}(A_k \xi_k \mathcal{T}). \tag{2.15}$$

Table 2.1: Background counts and average background counts per pixel in different regions in the Chandra/HRC-I observation of the open cluster NGC 2516.

Projected Angle	Count (count)	Area (pixels)	Average count per pixel
0-6	219962	22029408	0.0100
6-8	146332	14093856	0.0104
8-16	285300	26448800	0.0108

If segment s is in region k, $\mathcal{B}_s$ is modeled as

$$\mathcal{B}_s \big| \xi_k \overset{\text{indep}}{\sim} \text{Poisson}(a_s \xi_k \mathcal{T}), \tag{2.16}$$

and thus

$$Y_s \big| \xi_k, \boldsymbol{\lambda} \overset{\text{indep}}{\sim} \text{Poisson}\left(\left(a_s \xi_k + \sum_{i \in s} r_{s,i} e_s \lambda_i \right) \mathcal{T} \right). \tag{2.17}$$

We use an MCMC sampler to fit the model, and to obtain parameter estimates and error bars, see Appendix A.4 for details of the model fitting algorithm.

2.2.2 Simulation Study

We use a series of simulation configurations to evaluate the statistical properties of estimators based on our model. We make three simplifications in our simulation: (i) there are no overlapping sources, (ii) the background rate is constant, and (iii) the three constants (a_i, e_i, r_i) associated with each source are the same for all sources, and thus we remove the subscript i and use (a, e, r) to denote these parameters.

We choose the simulation parameters to mimic the Chandra observation of the open cluster NGC 2516. The following parameters are fixed in the simulation, (i) the exposure time $\mathcal{T} = 5 \times 10^4$ seconds, (ii) the area of the pure background region

Table 2.2: Parameters in the simulation and their corresponding values in the model.

parameters in simulation	λ_i^* (count)	μ^* (count)	θ^* (count^2)	ξ^* (count)
corresponding values to model	$re\mathcal{T}\lambda_i$	$re\mathcal{T}\mu$	$(re\mathcal{T})^2\theta$	$a\xi\mathcal{T}$

$A = 2.5 \times 10^7$ pixels, (iii) the background rate $\xi = 2 \times 10^{-7}$ count/s/pixel, and (iv) the number of sources $n = 1000$. Each replicate dataset is simulated according to the following steps:

1. Simulate the background count, X, from Poisson$(A\xi\mathcal{T})$, where $A\xi\mathcal{T} = 2.5 \times 10^5$.
2. Simulate the expected source counts, $\{\lambda_1^*, \cdots, \lambda_n^*\}$, independently from a zero-inflated gamma distribution, i.e., each $\lambda_i^* = 0$ with probability π_d, and $\lambda_i^* \sim$ Gamma $[\mu^*, \theta^*]$ with probability $1 - \pi_d$.
3. For $i = 1, \cdots, n$, simulate the background count $\mathcal{B}_i \overset{\text{indep}}{\sim}$ Poisson(ξ^*) and the source count $\mathcal{S}_i \overset{\text{indep}}{\sim}$ Poisson(λ_i^*), and thus the observed photon count is $Y_i = \mathcal{B}_i + \mathcal{S}_i$.

Note, the parameters that are marked by the superscript * are the scaled versions of those used in our model, see Table 2.2.

The simulation depends on the parameters μ^*, θ^*, π_d, and ξ^*. We fix $\mu^* = 15$ and consider 100 simulation configurations that we generate using a full factorial design, i.e., crossing two values of ξ^*, five values of θ^*, and ten values of π_d, in particular, $\xi^* \in \{15, 30\}$, $\theta^* \in \{50, 100, 300, 500, 1000\}$, and $\pi_d \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. Under each configuration, $m = 300$ replicate datasets are generated and fit via MCMC. We compute the $(1 - \rho)$ highest posterior density (HPD, see Appendix A.3) intervals with $\rho \in (0, 1)$ for μ^*, θ^*, and π_d using each of the 100×300 replicate datasets.

The actual coverage rate of the $(1 - \rho)$ HPD interval for each parameter under each configuration is estimated by the proportion of the 300 replicate $(1 - \rho)$ intervals that contain the true parameter value under that configuration. Ideally, the coverage rate of a $(1 - \rho)$ interval is exactly $(1 - \rho)$. The top row of Figure 2.2.2 illustrates one replicate of the counts in source regions, simulated with $\xi^* = 30, \pi_d = 0.5$, and $\theta^* = 100, 500$, and 1000, from left to right. As the variance, θ^*, of the population distribution increases, with the mean fixed at $\mu^* = 15$, more of the λ_i become concentrated near 300. The bottom row of Figure 2.2.2 shows the estimated coverage rates of the HPD intervals for $\mu^*, \theta^*,$ and π_d. The horizontal coordinate is the nominal coverage, and the vertical coordinate is the estimated actual (observed) coverage. Because the observed coverages lie near the 45° line, we see that the actual coverages of the HPD intervals for the three parameters are close to their nominal rates. The one exception occurs when $\theta^* = 1000$ and nominal rates are large. In that case, the coverage rates for the three parameters appear to be slightly higher than the nominal rates.

Table 2.3 shows the summaries of the posterior mode estimates and HPD intervals of π_d based on the 100×300 replicate datasets. Since the background rate and the exposure time are fixed in our simulation, larger ξ^* corresponds to larger source regions. In each cell of Table 2.3, the three summaries from top to bottom are (i) the estimated actual coverage rate of the 95% HPD interval for π_d, (ii) the average length of these 95% HPD intervals, and (iii) the root mean-squared error (rMSE) of

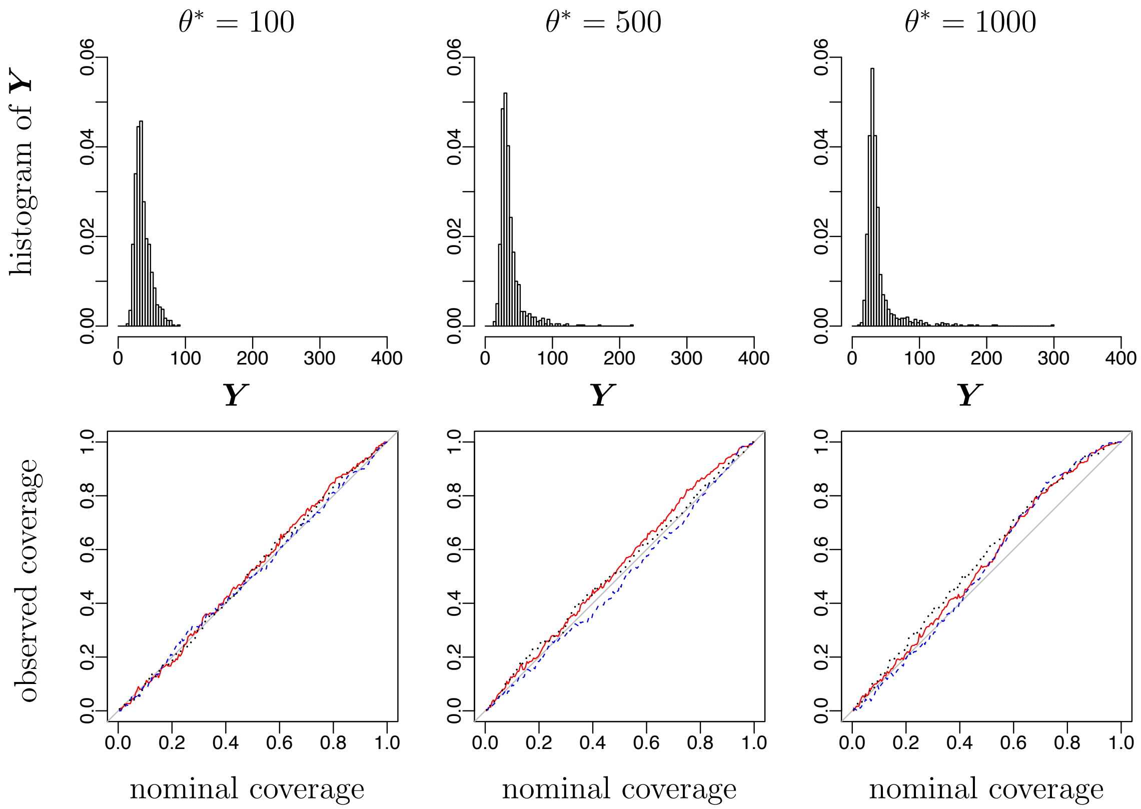

Figure 2.5: (Top row) histograms of $\boldsymbol{Y}$ generated under three simulation settings with $\xi^* = 30$, $\pi_d = 0.5$ and $\theta^* = 100, 500$ and 1000 from left to right; (bottom row) the coverage rates of the HPD intervals for μ^* (blue dashed lines), θ^* (black dotted lines) and π_d (red solid lines).

the point estimates of π_d, estimated by

$$\text{rMSE}(\widehat{\pi}_d) \approx \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}\left(\widehat{\pi}_d^{(j)} - \pi_d\right)^2},$$

where $\widehat{\pi}_d^{(j)}$ is the posterior mode estimate of π_d based on the j-th replicate dataset simulated under a particular configuration.

Coverage rates of most of the 95% HPD intervals in Table 2.3 are around 95% when $\pi_d > 0$, confirming that these HPD intervals exhibit approximately their nominal

Table 2.3: The coverage rate of 95% HPD interval, the average length of the interval, and the rMSE of the point estimates of π_d.

ξ^*	θ^*	π_d									
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
15	50	–	94.0%	94.7%	96.0%	96.3%	96.0%	97.7%	94.7%	96.0%	97.0%
		0.02	0.07	0.08	0.09	0.09	0.1	0.1	0.09	0.09	0.1
		0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	100	–	96.3%	94.0%	95.7%	95.0%	96.3%	96.7%	95.7%	97.0%	99.3%
		0.04	0.09	0.11	0.12	0.12	0.13	0.13	0.13	0.13	0.18
		0.01	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02
	300	–	97.0%	94.0%	94.3%	93.3%	95.7%	95.0%	99.3%	99.0%	98.7%
		0.11	0.18	0.24	0.27	0.3	0.3	0.32	0.34	0.36	0.38
		0.04	0.06	0.07	0.06	0.07	0.06	0.06	0.05	0.04	0.03
	500	–	95.7%	95.0%	92.7%	96.0%	94.3%	95.0%	97.3%	97.0%	98.7%
		0.17	0.25	0.32	0.36	0.41	0.44	0.46	0.48	0.47	0.41
		0.07	0.09	0.1	0.11	0.09	0.1	0.09	0.07	0.06	0.03
	1000	–	95.3%	97.0%	98.0%	97.7%	97.0%	98.3%	97.7%	99.3%	98.3%
		0.32	0.38	0.44	0.5	0.54	0.57	0.61	0.61	0.6	0.49
		0.17	0.16	0.16	0.16	0.15	0.13	0.11	0.09	0.07	0.04
30	50	–	92.3%	95.0%	92.0%	92.3%	94.7%	94.3%	95.3%	96.7%	98.0%
		0.03	0.09	0.11	0.12	0.13	0.13	0.13	0.13	0.14	0.2
		0.01	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
	100	–	93.3%	94.7%	94.0%	95.3%	95.3%	96.7%	97.3%	97.7%	98.7%
		0.06	0.12	0.15	0.16	0.17	0.17	0.18	0.19	0.21	0.29
		0.02	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.03
	300	–	97.7%	97.0%	91.0%	91.7%	96.3%	93.3%	96.3%	98.7%	99.0%
		0.13	0.2	0.28	0.32	0.35	0.38	0.4	0.43	0.46	0.42
		0.05	0.06	0.07	0.09	0.08	0.08	0.08	0.07	0.05	0.03
	500	–	96.0%	97.3%	97.0%	96.3%	95.7%	95.3%	97.3%	98.7%	99.3%
		0.21	0.28	0.35	0.41	0.46	0.49	0.52	0.52	0.53	0.43
		0.09	0.1	0.11	0.11	0.12	0.11	0.1	0.08	0.06	0.04
	1000	–	95.3%	96.7%	95.3%	94.3%	98.7%	98.3%	99.0%	99.0%	99.0%
		0.35	0.42	0.48	0.53	0.57	0.61	0.64	0.63	0.59	0.48
		0.19	0.18	0.18	0.18	0.17	0.14	0.13	0.11	0.08	0.05

If the true coverage rate is 95%, the standard deviation of the estimated coverage rate is 1.3%. So coverage rates between 92.4% and 97.6% are statistically indistinguishable from 95% at 2σ level.

coverage rates. When π_d is large, however, the intervals have slightly higher coverages than the nominal rates. We do not include the coverage rates when $\pi_d = 0$, because none of the intervals contain the boundary of the distribution.

Both the rMSE and the average length of the intervals are indicators of the uncertainty associated with the estimates of π_d. Holding other parameters constant, both the average length of the 95% HPD intervals and the rMSE increase as θ^* increases. The explanation is as follows. Figure 2.2.2 shows that for fixed mean, μ^*, the gamma distribution with larger variance, θ^*, is more concentrated at 0. It is hence more difficult for the model to distinguish the X-ray dark sources ($\lambda_i^* = 0$) from the dim sources with small but positive intensities. The rMSE and the average length of the intervals also increase as ξ^* or the area of source regions increases, because the noise (i.e., count from the background) becomes increasingly overwhelming whereas the signal (i.e., count from the source) remains unchanged. Overall, the rMSE's of the point estimates are all reasonably small, meaning that the point estimates of π_d are reasonably close to the true values.

2.3 Testing for X-ray Dark Sources

2.3.1 Hypothesis Testing

We are particularly interested in whether there are any X-ray dark sources in the pppulation. In our model, the proportion of X-ray dark sources is π_d, so we can address the existence of dark sources via a statistical hypothesis test, where the null and alternative hypotheses are

$$H_0 : \pi_d = 0 \quad \text{and} \quad H_1 : \pi_d > 0. \tag{2.18}$$

The null hypothesis claims no X-ray dark sources in the pppulation.

In classical hypothesis testing, the significance level α, which controls the probability of a false positive, is pre-specified, typically at 1% or 5%. A test statistic, T, is a summary of the data that captures the discrepancy of the data generated under H_0 and H_1. The corresponding p-value is the probability of sampling a value of the test statistic under H_0 at least as extreme as the observed value, $T_{\text{obs}} = T(\boldsymbol{D})$, that is,

$$\text{p-value} = P(T(\boldsymbol{D}_{\text{rep}}) > T_{\text{obs}} \big| H_0), \tag{2.19}$$

where $\boldsymbol{D}_{\text{rep}}$ is a replicate dataset sampled under H_0. For clarity, we assume larger values of T are more consistent with H_1. If the p-value is less than α, we say the null hypothesis is rejected; otherwise, we have insufficient evidence to reject H_0. To compute the p-value, the distribution of $T(\boldsymbol{D}_{\text{rep}})$ under H_0, also known as the reference (or null) distribution, is required. Unfortunately this is problematic if there are unknown parameters under H_0 because in this case we cannot directly sample D_{rep} under H_0 to derive the reference distribution needed to compute a p-value. This is the case in (2.18) because H_0 only states that there are no X-ray dark sources without specifying the distribution of the source intensities of the X-ray luminous sources, namely Gamma$[\mu, \theta]$.

Posterior predictive p-values (ppp-values) were designed by Rubin et al. (1984) to address this difficulty, and extended by Meng (1994) and Gelman et al. (1996). A ppp-value is often used to assess goodness of fit of a posited model (Gelman et al., 1996). Assume the observed data, $\boldsymbol{D}$, is fit to a model parametrized by a vector of unknown parameters Θ. The ppp-value is defined by averaging the classical p-value in (2.19) over the posterior distribution of Θ, that is,

$$\text{ppp-value} = P(T(\boldsymbol{D}_{\text{rep}}) > T_{\text{obs}} \big| \boldsymbol{D}) = \int P(T(\boldsymbol{D}_{\text{rep}}) > T_{\text{obs}} \big| \Theta) P(\Theta \big| \boldsymbol{D}) \mathrm{d}\Theta. \quad (2.20)$$

To test the existence of X-ray dark sources, we calculate the ppp-value under H_0 (i.e., the model with $\pi_d = 0$). A ppp-value can be interpreted much like a classical p-value. A small value means that data generated under the posterior predictive distribution

$$P(\boldsymbol{D}_{\text{rep}} | \boldsymbol{D}) = \int P(\boldsymbol{D}_{\text{rep}} | \Theta) P(\Theta | \boldsymbol{D}) \mathrm{d}\Theta,$$

is unlikely to have given rise to the observed test statistic, $T_{\text{obs}} = T(\boldsymbol{D})$, hence the null model should be rejected and the alternative model preferred. We use a Monte Carlo simulation to approximate the ppp-value. This proceeds as follows:

1. Obtain posterior draws $\left\{ (\mu^{(l)}, \theta^{(l)}, \boldsymbol{\xi}^{(l)}); l = 1, \cdots, M \right\}$ from the posterior distribution of $(\mu, \theta, \boldsymbol{\xi})$ under the null model. This is done using the MCMC sampler described in Appendix A.4, except we fix $\pi_d = 0$.

2. For $l = 1, \cdots, M$, simulate a dataset, $\boldsymbol{D}^{(l)}_{\text{rep}}$, under H_0, as described in (2.8), (2.15), and (2.17), where the parameters are $(\pi_d, \mu, \theta, \boldsymbol{\xi}) = (0, \mu^{(l)}, \theta^{(l)}, \boldsymbol{\xi}^{(l)})$.

3. Calculate the test statistic $T^{(l)}_{\text{rep}} = T(\boldsymbol{D}^{(l)}_{\text{rep}})$, for $l = 1, \cdots, M$.

4. Estimate the ppp-value by the proportion of the $T^{(l)}_{\text{rep}}$ that are greater than T_{obs}, that is,

$$\text{ppp-value} \approx \frac{1}{M} \sum_{l=1}^{M} I\left(T^{(l)}_{\text{rep}} > T_{\text{obs}}\right), \quad (2.21)$$

where the indicator function $I(c)$ equals to 1 if the statement c is true, and otherwise is 0.

In practice, we simplify the procedure by fixing $\boldsymbol{X}_{\text{rep}}^{(l)}$ at the observed values, $\boldsymbol{X}$, because we are primarily concerned about the source counts. In addition, K is typically much smaller than n, and the combined source segments within region k is only a small fraction of A_k, in other words, $\boldsymbol{X}$ contain substantial data for estimating $\boldsymbol{\xi}$. For example, in the Chandra observation of the open cluster NGC 2516, within 6 arcmin from the center of the field, the combined source segments is only 4.3% of the background region. As a result, the posterior distribution of $\boldsymbol{\xi}$ is largely determined by the observed background counts $\boldsymbol{X}$ and is concentrated around $X_k/(A_k\mathcal{T})$. Consequently, this simplification has little impact on the final ppp-value.

An important difference between the classical p-value and the ppp-value lies in their null distributions. Under H_0, the classical p-value in (2.19) is uniformly distributed in the unit interval, implying the probability of observing a p-value less than α when H_0 is true, and thus falsely rejecting H_0 is α. However, the distribution of the ppp-value under H_0 may differ from uniform, so α may not be the actual false positive rate. Fortunately, Section 2.3.3 shows via simulation that the null distributions of the ppp-values in our model under different configurations are very close to a uniform distribution when $\alpha < 0.5$. As a result, just like the classical hypothesis test, the false positive rate of our procedure is approximately equal to the nominal significance level, α.

2.3.2 Test Statistic

The discussion in Section 2.3.1 presupposes a suitable test statistic that can be used to distinguish H_0 and H_1. A general choice, and one that we employ, is the likelihood ratio (LR) statistic,

$$LR(\boldsymbol{D}_{\text{rep}}) = \frac{\sup_{\mu,\theta,\pi_d,\boldsymbol{\xi}} L_1(\mu,\theta,\pi_d,\boldsymbol{\xi};\boldsymbol{D}_{\text{rep}})}{\sup_{\mu,\theta,\boldsymbol{\xi}} L_0(\mu,\theta,\boldsymbol{\xi};\boldsymbol{D}_{\text{rep}})}, \tag{2.22}$$

where L_0 and L_1 are the likelihood functions under the null and the alternative models, i.e.,

$$\begin{aligned} L_1(\mu,\theta,\pi_d,\boldsymbol{\xi};\boldsymbol{D}_{\text{rep}}) = P(\boldsymbol{D}_{\text{rep}}\big|\mu,\theta,\pi_d,\boldsymbol{\xi}) &= \int P(\boldsymbol{D}_{\text{rep}},\boldsymbol{\lambda}\big|\mu,\theta,\pi_d,\boldsymbol{\xi})\mathrm{d}\boldsymbol{\lambda} \\ &= \int P(\boldsymbol{Y}_{\text{rep}}|\boldsymbol{\lambda},\boldsymbol{\xi})P(\boldsymbol{\lambda}|\mu,\theta,\pi_d)\mathrm{d}\boldsymbol{\lambda}, \end{aligned} \tag{2.23}$$

with the probabilities in the second line defined in (2.17) and (2.9), respectively, and

$$L_0(\mu,\theta,\boldsymbol{\xi};\boldsymbol{D}_{\text{rep}}) = P(\boldsymbol{D}_{\text{rep}}\big|\mu,\theta,\pi_d=0,\boldsymbol{\xi}) = L_1(\mu,\theta,\pi_d=0,\boldsymbol{\xi};\boldsymbol{D}_{\text{rep}});$$

see Freeman et al. (1999) and Protassov et al. (2002) for other applications of the likelihood ratio statistic in Astronomy. Note that the reason for integrating out $\boldsymbol{\lambda}$ in (2.23) is because $\boldsymbol{\lambda}$ are random effects in the model and the integrated likelihood incorporates the associated uncertainty. The function $P(\boldsymbol{D}_{\text{rep}},\boldsymbol{\lambda}\big|\mu,\theta,\pi_d,\boldsymbol{\xi})$ is integrable because it can be written as

$$P(\boldsymbol{D}_{\text{rep}},\boldsymbol{\lambda}\big|\mu,\theta,\pi_d,\boldsymbol{\xi}) = P(\boldsymbol{\lambda}\big|\boldsymbol{D}_{\text{rep}},\mu,\theta,\pi_d,\boldsymbol{\xi})P(\boldsymbol{D}_{\text{rep}}\big|\mu,\theta,\pi_d,\boldsymbol{\xi}),$$

where the posterior distribution of $\boldsymbol{\lambda}$ given $(\boldsymbol{D}_{\text{rep}}, \mu, \theta, \pi_d, \boldsymbol{\xi})$ is integrable because the prior distribution for $\boldsymbol{\lambda}$ is proper. However, the integration is intractable, so we make two modifications to simplify the computation of the likelihood function.

In the first modification, we replace each ξ_k with an estimate based on the background counts, $\widehat{\xi}_k = X_k/(A_k\mathcal{T})$, for $k = 1, \cdots, K$. This substitution allows us to avoid maximizing out $\boldsymbol{\xi}$ in computing LR. We expect this simplification has little effect on LR, because $\boldsymbol{X}$ contains substantial data for estimating $\boldsymbol{\xi}$, and thus the values of $\boldsymbol{\xi}$ that maximize the likelihood functions L_0 and L_1 should be very close to $\widehat{\boldsymbol{\xi}} = (\widehat{\lambda}_1, \cdots, \widehat{\lambda}_K)$.

Second, we only consider sources whose regions do not overlap with other source regions when computing LR. The resulting LR is a legitimate test statistic because (i) it is a function of the observed data, and (ii) the source intensities do not determine whether the source regions overlap or not, meaning we can view the sources without overlap as a random subset of all sources. The downside of this simplification is the reduction of the sample size, which leads to the decrease of the statistical power of the test. It offers a substantial computational advantage, however, in that it allows us to express the integral over $\boldsymbol{\lambda}$ in (2.23) as the product of univariate integrals, each of which can be obtained analytically. We emphasize that we use all the data to fit the null model and obtain the posterior draws of (μ, θ).

With these two simplifications, the simplified likelihood is

$$\widetilde{L}_1(\mu, \theta, \pi_d, \widehat{\boldsymbol{\xi}}; \boldsymbol{D}_{\text{rep}}) = \prod_{i \in \mathbb{S}} \int P(Y_{\text{rep},i} | \lambda_i, \widehat{\xi}_k) P(\lambda_i | \mu, \theta, \pi_d) \mathrm{d}\lambda_i, \tag{2.24}$$

where $\mathbb{S}$ is the set of sources with no overlap, $Y_{\text{rep},i} | \lambda_i, \widehat{\xi} \sim \text{Poisson}(r_i e_i \lambda_i \mathcal{T} + a_i \widehat{\xi} \mathcal{T})$

if source region i is in background region k. The likelihood function under H_0, $\widetilde{L}_0(\mu, \theta, \widehat{\boldsymbol{\xi}}; \boldsymbol{D}_{\text{rep}})$, equals to $\widetilde{L}_1(\mu, \theta, \pi_d = 0, \widehat{\boldsymbol{\xi}}; \boldsymbol{D}_{\text{rep}})$. Thus, our final test statistic is LR, as given in (2.22), but with L_0 and L_1 replaced by $\widetilde{L}_0$ and $\widetilde{L}_1$, respectively.

The integration in (2.24) can be obtained analytically, see Appendix A.5 for details. Since $\widetilde{L}_0$ and $\widetilde{L}_1$ are functions of only a few parameters, their supremes required for LR in (2.22) can be obtained via numerical methods, such as Newton-Raphson method and Nelder-Mead method (Nelder and Mead, 1965; Gerald et al., 1989), which are implemented in many computer programs, including R and Python.

2.3.3 Simulation Study

As mentioned in Section 2.3.1, a general challenge of using ppp-values is their non-uniform distribution under H_0. This means that ppp-values may be somewhat less likely to reject H_0 than classical p-values. Thus, there will be fewer false positives, but also somewhat less statistical power, i.e., ppp-values may be somewhat less able to detect a subpppulation of X-ray dark sources. In this section, we explore the distribution of the ppp-values via simulation. In addition, we examine the statistical power of our hypothesis testing procedure using the 100 × 300 replicate datasets described in Section 2.2.2.

Figure 2.6 shows histograms of the ppp-values and a Q-Q plot that compares the uniform distribution with that of the ppp-values under H_0. Figure 2.6 fixes (μ^*, ξ^*, π_d) at $(15, 30, 0)$ and varies $\theta^* = 100, 500$, and 1000, from left to right. Since we fix $\pi_d = 0$, these simulations all correspond to H_0. In this case, the distribution of the ppp-values under H_0 is quite close to the uniform distribution when $ppp \leqslant 0.5$,

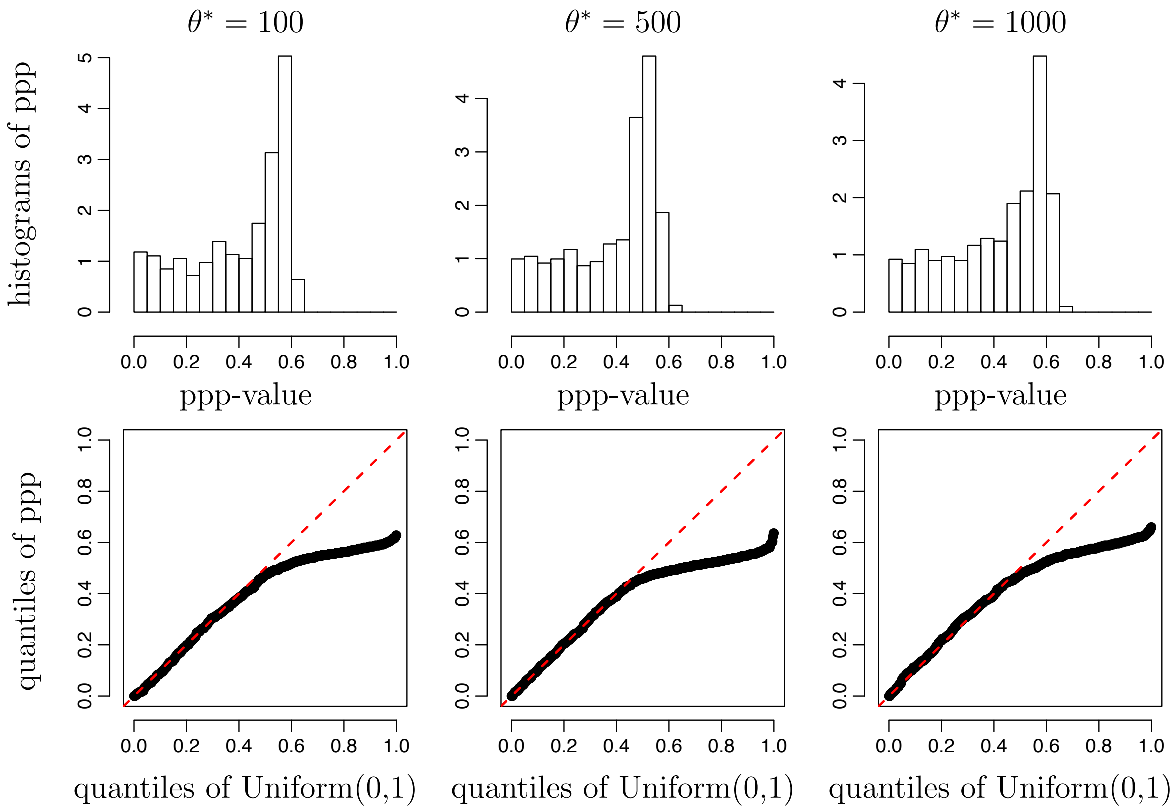

Figure 2.6: (Top row) histograms of the ppp-values under H_0, with the simulation parameters being $\xi^* = 30, \pi_d = 0, \mu^* = 15$, and $\theta^* = 100, 500$, and 1000 (from left to right). (Bottom row) the Q-Q plots comparing the uniform distribution with that of the ppp-values under H_0.

implying the probability of false positive is approximately equal to the significance level α for any reasonable value of α.

We apply our hypothesis testing procedure to each of the 100×300 datasets used in Section 2.2.2, including those simulated under H_1 with $\pi_d > 0$. For each simulation configuration, we compute the ppp-value, denoted as $ppp^{(j)}$, for each of the $m = 300$ replicate datasets, and estimate the probability of rejecting H_0 (i.e., rejection rate) by the proportion of the $\{ppp^{(j)}; j = 1, \cdots, m\}$ that are less than the significance level,

Table 2.4: The rejection rates of our hypothesis testing procedure.

ξ^*	θ^*	π_d									
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
15	50	4.7%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	100	5.0%	98.7%	100%	100%	100%	100%	100%	100%	100%	96.7%
	300	4.7%	41.3%	79.0%	90.0%	92.0%	95.0%	94.7%	89.0%	76.0%	45.3%
	500	4.7%	21.7%	40.7%	48.0%	58.7%	63.0%	60.7%	58.7%	46.3%	26.7%
	1000	6.7%	9.0%	17.3%	21.0%	23.3%	29.7%	22.7%	19.0%	15.0%	10.3%
30	50	4.3%	99.3%	100%	100%	100%	100%	100%	100%	100%	96.0%
	100	4.0%	83.7%	98.7%	100%	100%	100%	100%	99.7%	99.0%	81.0%
	300	5.0%	21.0%	52.7%	70.0%	83.3%	81.7%	79.7%	75.7%	55.0%	36.3%
	500	5.3%	15.3%	31.7%	38.0%	43.0%	48.7%	47.0%	44.3%	27.3%	18.3%
	1000	4.7%	9.3%	11.7%	19.0%	20.7%	21.7%	21.7%	18.3%	15.7%	11.3%

α, that is,

$$\text{rejection rate} \approx \frac{1}{m}\sum_{j=1}^{m} I\left(ppp^{(j)} < \alpha\right).$$

The statistical power of the test is the probability that it rejects H_0 when H_1 is true. We expect the power to increase with π_d. The rejection rate represents the power of the test when the data are generated under H_1, and the probability of false positive under H_0. Ideally, the power is large and the probability of false positive equals the significance level.

Table 2.4 shows the rejection rates at $\alpha = 5\%$ significance level for data generated under the 100 simulation configurations. The rejection rates corresponding to $\pi_d = 0$, i.e., the probabilities of false positive, are all around 5%. All other columns in Table 2.4 show the powers of the test. Other parameters being constant, we find (i) larger ξ^* or larger source regions lead to lower power, because the quality of the data (i.e., signal-to-noise ratio) is degraded, and (ii) the power decreases with θ^*, because with large θ^* there are more dim sources in the data, making the null and the alternative models less distinguishable.

The impact of π_d on the power is more complicated. For small π_d, the power increases as π_d increases, because there are more source regions with small counts that can not be explained by the null model. However, for large π_d, increasing π_d leads to power loss. This can be explained by Figure 2.3.3, which shows histograms of the posterior mode estimates of μ^* (left) and θ^* (right) fit under the null model, based on each of the 3×300 datasets simulated with $\xi^* = 30, \mu^* = 15, \theta^* = 300$, and $\pi_d = 0.1$ (black lines), 0.5 (red lines) and 0.9 (green lines). As π_d increases, the posterior mode estimates of μ^* and θ^* fit under the null model move towards zero in order to accommodate the large proportion of dark sources. That is, when there are many dark sources, when H_0 is fit, the distribution of the luminous sources shifts to accommodate them. Thus, for large π_d, as π_d increases, data simulated from the fitted null model and data from the true model become less distinguishable, leading to the decrease in power. It is important to note that the impact of π_d on the power of the test is sensitive to the choice of the parameter distribution used for the non-dark source intensities; in other words, if we were to model the distribution of non-dark source intensities with a different distribution (e.g., log-normal distribution), the influence of π_d on the power may be different. In addition, we assume no overlapping sources in the simulation, and thus use all the data in computing the likelihood functions L_0 and L_1 in (2.22). In the presence of densely overlapping sources, the power is expected to be lower due to the reduction of effective sample size.

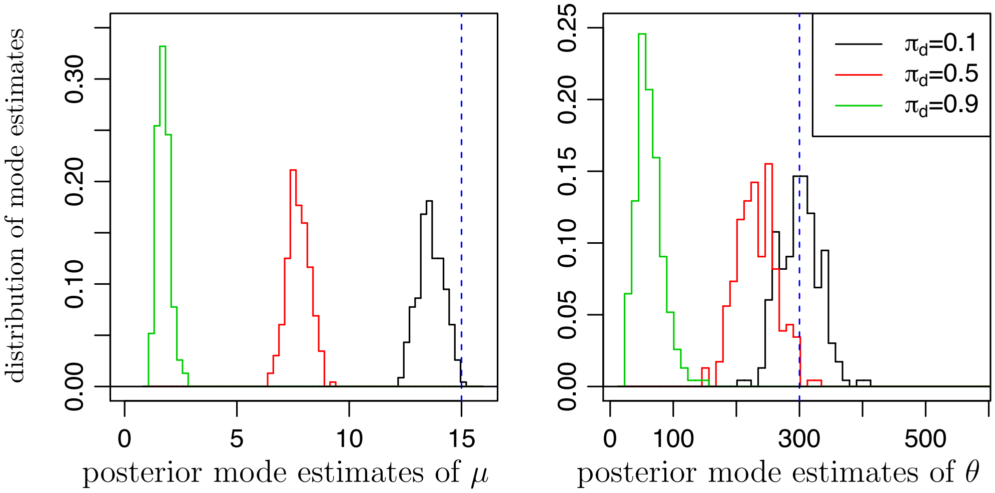

Figure 2.7: Histograms of the posterior mode estimates of μ^* (left) and θ^* (right) fit under H_0 (no dark sources). The 3×300 datasets are generated from the alternative model with $\xi^* = 30, \mu^* = 15, \theta^* = 300$, and $\pi_d = 0.1$ (black lines), 0.5 (red lines) and 0.9 (green lines). The vertical dashed lines mark the true values of μ^* and θ^*. When the proportion of dark sources grows, the distribution under H_0 for the luminous sources adjusts to accommodate the dark sources.

2.4 Application

We apply the model and the hypothesis testing procedure to two subsets of the Chandra/HRC-I observation of the open cluster NGC 2516, with exposure time of $\mathcal{T} = 4.9 \times 10^4$ seconds. The first subset of data consist of the 649 sources within 6 arcmin from the center of the field (CF), where the background rate is assumed constant. The average source regions is ~ 1400 pixels, and the pure background region is $A_1 = 2.2 \times 10^7$ pixels. Out of the 649 source regions, 525 have no overlap with other source regions. In addition to the 649 sources, the second dataset also includes the 520 sources between 6 and 8 arcmin from the CF. For ease of reference, we call the region within 6 arcmin and between 6-8 arcmin from the CF region 1 and region 2, respectively. Region 2 is also assumed to have a spatially uniform background,

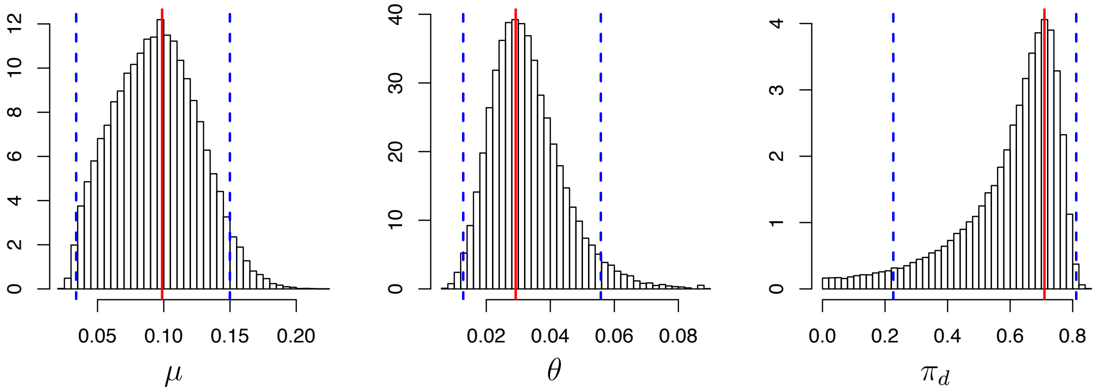

Figure 2.8: Histograms of the posterior distributions of μ (left), θ (middle) and π_d (right) given the first dataset. The red solid lines are the posterior mode estimators and the blue dash lines show the lower and upper bounds of the 95% HPD intervals of the parameters.

but with a background rate different from region 1. In region 2, the average source regions is around 6900 pixels, the background region is $A_2 = 1.4 \times 10^7$ pixels, and 227 out of the 520 source regions do not overlap with other source regions.

For the first dataset, the posterior distributions of μ, θ and π_d are shown in Figure 2.8, with the posterior mode estimates and the 95% HPD intervals of these parameters marked by the solid and dashed vertical lines, respectively. The mode of the posterior distribution of π_d is 0.71, and its 95% HPD interval is $(0.23, 0.81)$, suggesting that a large proportion of sources in the optical catalog are X-ray dark. While the large spread in the posterior distribution of π_d allows it to be zero, the more likely scenario is that there does exist a separate X-ray dim/dark population among the X-ray bright stars.

Figure 2.4 (left) shows the population distribution of source intensities on the scale of natural logarithm. Each gray line is a zero-inflated gamma distribution with the parameters, (μ, θ, π_d), drawn from their joint posterior distribution. The red line corresponds to the zero-inflated gamma distribution with (μ, θ, π_d) set to their

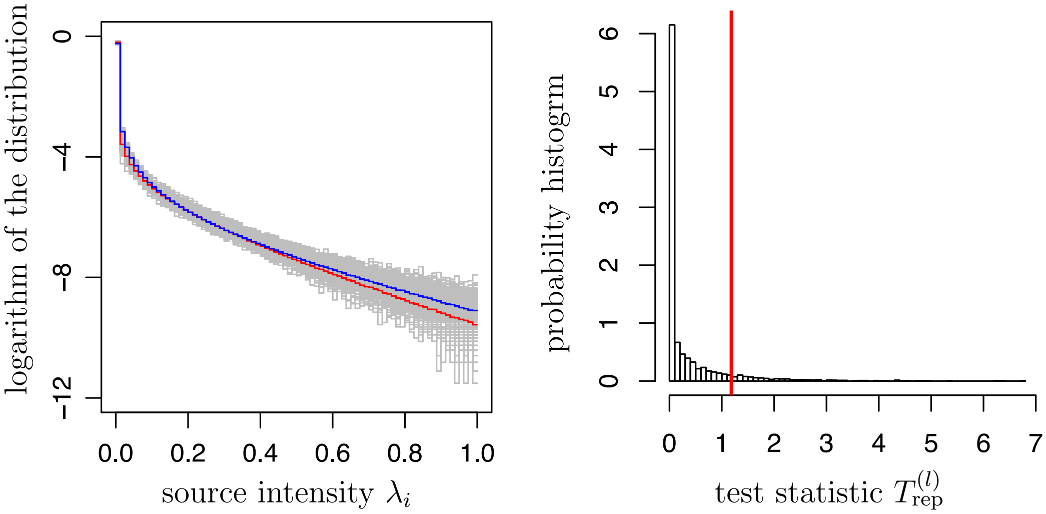

Figure 2.9: (Left) The fitted distribution of source intensities on a logarithmic scale (base e). The red line is the zero-inflated gamma distribution with the parameters set to their posterior mode estimates. The blue line is the fitted gamma distribution for the source intensities under the null model. Each gray line is a zero-inflated gamma distribution with parameters, (μ, θ, π_d), drawn from their joint posterior distribution. (Right) The distribution of the test statistics $\{T^{(l)}_{\rm rep}; l = 1, \cdots, M\}$ simulated from the null model fit to the data within 6-arcmin from the center of the field.

posterior mode estimates. For comparison, we plot the gamma distribution (blue line) with the mean and variance parameters being their posterior mode estimates fit under the null model. The blue and red lines are quite similar, and they both imply a large proportion of dim/dark sources in the population. We apply the hypothesis testing procedure described in Section 2.3 to the data, and obtain the distribution of the simulated test statistic, $T^{(l)}_{\rm rep} = \log(LR(\boldsymbol{D}^{(l)}_{\rm rep}))$, for $l = 1, \cdots, M$, as shown in Figure 2.4 (right). The observed test statistic $T_{\rm obs} = \log(LR(\boldsymbol{D})) = 1.181$ and the ppp-value = 8.9%.

When we include data in region 2, the posterior distributions of the parameters changes in two ways. First, with the combined dataset, the 95% HPD interval of π_d is $(0.12, 0.78)$, which is wider than that based on dataset 1. The reason is that while in region 1, the average of source regions is 1400 pixels, the average increases to 3847 pixels in the combined dataset. As discussed in Section 2.2.2, holding other factors constant, the uncertainty associated with the estimates of π_d increases as the source regions (or ξ^* in the simulation) increase. Even though dataset 2 consists of more data, the impact of increasing source regions overwhelms the impact of increasing sample size, and the information from data in region 1 is diluted by the noisy data in region 2, resulting in a wider interval for π_d. The point estimates and the bounds of the 95% HPD interval estimates of the parameters μ, θ, and π_d also decrease slightly with the added data in region 2. As the posterior distribution of π_d shifts toward zero, more sources whose observed counts are small relative to the background are accommodated as dim (small but positive source intensities) rather than dark sources. This leads to a decrease in the estimates of both the mean and the variance of the non-zero source intensities. For dataset 2, the observed test statistic is 0.363 and the estimated ppp-value is 23.2%. So there is insufficient evidence to conclude the existence of dark sources.

2.5 Summary

We have developed a Bayesian hierarchical model to investigate the population distribution of source intensities. The main innovations are the introduction of X-ray dark sources as a subpopulation and modeling the distribution of source intensities as a zero-inflated gamma distribution. In addition, we extend the model to allow for overlapping sources and piecewise homogeneous background. The simulation shows under a variety of simulation configurations, the model produces point estimates with reasonably small errors and interval estimates with good coverage rates.

We have also proposed a Bayesian hypothesis testing procedure based on our model. We selected the likelihood ratio as the test statistic and provided a detailed description of how a posterior predictive p-value can be computed. The simulation shows the probability of false positive of the test is approximately equal to the pre-specified significance level. We have also examined the statistical power of the test via simulation under various configurations, and analyzed the impact of different factors on the power.

Finally, we apply the Bayesian model and the hypothesis testing procedure to two subsets of the Chandra/HRC-I observation of the open cluster, NGC 2516. The first subset are data within 6 arcmin from the center of the field, where the source regions are relatively small and most of the source regions do not overlap. The second subset includes all the data within 8 arcmin from the center of the field. The added sources are further from the center of the field, and have larger regions and a greater proportion of overlapping sources. For these two datasets, the posterior distribution of π_d suggests a large proportion of sources could be X-ray dark, although the result comes with large uncertainty.

Chapter 3

Large-Sample Hypothesis Testing with Multiply-Imputed Data

3.1 Introduction

Missing data is a common occurrence in research studies in healthcare science, sociology, political science, etc. For example, the AIDS surveillance data of the US Centers for Disease Control (Tu et al., 1993; Barnard and Meng, 1999) suffer from severe incompleteness, including a non-negligible fraction of unreported deaths and censored time of the reported deaths. The New York City School Choice Scholarship Program (Barnard et al., 2002, 2003; Krueger and Zhu, 2004), although carefully designed and implemented, is pervaded by missing data, e.g., family background, children's pre-test and post-test scores.

In the frequentist framework, methods such as the EM algorithm are effective in handling missing data; in Bayesian analysis, missing data are treated as random variables, and the posterior distribution given the observed data is used to make inferences. Unfortunately, these techniques are typically difficult to implement, especially in large-sample survey data with complex structures.

Imputation handles missing data by replacing them with imputed values and thus allows the standard complete-data analysis to be applied to the completed data. In addition, for public-use data, data producers can use their expert knowledge to make informed and sensible imputation. Inconsistency in the analysis among users can be alleviated by sharing the same imputed data to all users. The downside of single imputation, however, is that we ignore the uncertainty associated with the imputed data, and thus underestimate the variance of the estimates. The numerical example in Li et al. (1991b) illustrates that the actual levels of a large-sample Wald test based on a single imputation are much higher than the nominal levels.

Multiple imputation, proposed by Rubin (1978, 1987), rectifies the problem by imputing the missing data several times. The standard complete-data analysis is then applied to each of the completed datasets, and the resulting inferences are then combined via some simple combining rules to form the final repeated-imputation inference. Multiple imputation inherits the advantages of simple imputation, meanwhile, the final combined inference properly accounts for the uncertainty due to missing data. Theoretical and applied justification for the use of multiple imputation includes Rubin and Schenker (1986), Rubin (1987), Schenker and Welsh (1988), Rubin and Schenker (1991), and Schenker et al. (1993).

In this chapter, we review, compare, and modify some current large-sample hypothesis testing procedures based on multiply-imputed data. We first present the notations and background that are necessary for understanding the procedures, which are classified according to (i) what is available to form the final inference, e.g., the complete-data moments estimates (point and variance-covariance estimates) or the test statistics, and (ii) whether the derivation is based on the assumption of equal fraction of missing information. In Section 3.3, we describe and compare procedures proposed by Li et al. (1991b) and Xie (2011) based on the moment estimates. We also provide a modification to a procedure by Xie (2011) to make it behave well under all circumstances. In Section 3.4, we discuss and compare procedures by Li et al. (1991a), Meng and Rubin (1992), and Xie (2011), when the moment estimates are not available.

3.2 Notations and Background

3.2.1 Hypothesis Testing without Imputation

Let $X = \{x_1, \cdots, x_n\}$ be the complete data with the density $f(X|\psi)$, parametrized by a vector of parameters $\psi \in \mathcal{R}^h$. We are interested in testing the null hypothesis, $H_0 : \theta = \theta_0$, against the alternative hypothesis, $H_a : \theta \neq \theta_0$, where $\theta = \theta(\psi) \in \mathcal{R}^k$ is a vector function of the model parameters ψ.

Let $\widehat{\theta} = \widehat{\theta}(X)$ be the maximum likelihood estimate (MLE) of θ, and $U = U(X)$ be the associated variance-covariance matrix. When n is large, asymptotically we have

$$U^{-1/2}(\widehat{\theta} - \theta_t)\big|\theta = \theta_t \sim N(0, I_k),$$

where θ_t is the true value of θ. With lower-order variability, U is approximately equal to the true variance-covariance matrix, $U_t = \mathrm{Var}(\widehat{\theta}|\theta = \theta_t)$, that is, $(U|\theta = \theta_t) \approx U_t$. Under the null hypothesis, the test statistic

$$D = (\widehat{\theta} - \theta_0)^t U^{-1} (\widehat{\theta} - \theta_0)/k,$$

which is proportional to the Wald χ^2 statistic, is asymptotically distributed as χ^2_k/k. So the p-value is computed as $P = \Pr(\chi^2_k/k > D)$.

With the presence of missing data, however, the analysis becomes more complicated. We denote X_{obs} and X_{mis} as the observed and missing data, and $X = (X_{\text{obs}}, X_{\text{mis}})$ as the complete data. Based on the observed-data likelihood, we can obtain the MLE of θ, denoted as $\widehat{\theta}_{\text{obs}} = \widehat{\theta}_{\text{obs}}(X_{\text{obs}})$, and the associated variance-covariance matrix, $T = T(X_{\text{obs}})$. Asymptotically,

$$T^{-1}(\widehat{\theta}_{\text{obs}} - \theta_t)\big|\theta = \theta_t \sim N(0, I_k), \tag{3.1}$$

and $(T|\theta = \theta_t) \approx T_t$, where $T_t = \mathrm{Var}(\widehat{\theta}_{\text{obs}}|\theta = \theta_t)$. The asymptotically optimal test is based on the observed Wald statistic, i.e.,

$$D_{\text{obs}} = (\widehat{\theta}_{\text{obs}} - \theta_0)^t T^{-1} (\widehat{\theta}_{\text{obs}} - \theta_0)/k,$$

and the p-value is $P_{\text{obs}} = \Pr(\chi^2_k/k > D_{\text{obs}})$.

Note that the consequence of missing data is the loss of information. More specifically, information from the observed and the complete data can be quantified by T_t^{-1} and U_t^{-1}, respectively. Thus, the loss of information due to missing data is $U_t^{-1} - T_t^{-1}$, and the increase in variance is $B_t = T_t - U_t$. The ratios of missing to observed information can be represented by the eigenvalues, denoted as $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_k)$, of the matrix $(U_t^{-1} - T_t^{-1})T_t = U_t^{-1}B_t$. So the vector of the ratios of complete to observed information is $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_k)$, where $\xi_i = 1 + \lambda_i$.

3.2.2 Multiple Imputation

Although there are methods such as the EM algorithm to compute the observed estimates, i.e., $\widehat{\theta}_{\text{obs}}$ and T, they are often complicated to implement. Multiple imputation is an easy-to-implement and yet principled method to handle missing data. It involves two separate steps: imputation and analysis. In the imputation step, missing data are generated m times, resulting in m completed datasets

$$X_*^{(l)} = (X_{\text{obs}}, X_{\text{mis}}^{(l)}), \text{ for } l = 1, \cdots, m.$$

In the analysis step, the complete-data procedure is applied to each of the m completed datasets, producing m point estimates, $\widehat{\theta}_{*l} = \widehat{\theta}(X_*^{(l)})$, and variance-covariance estimates, $U_{*l} = U(X_*^{(l)})$. We denote $\mathcal{S}_{\text{m}}$ as the set of moment estimates, i.e.,

$$\mathcal{S}_{\text{m}} = \{(\widehat{\theta}_{*l}, U_{*l}); l = 1, \cdots, m\}.$$

These estimates are combined to produce the multiple imputation point estimate

$$\bar{\theta}_m = \frac{1}{m} \sum_{l=1}^{m} \widehat{\theta}_{*l},$$

with the associated variance-covariance estimate

$$T_m = \bar{U}_m + (1 + m^{-1})B_m,$$

where $\bar{U}_m$ quantifies the within-imputation variance,

$$\bar{U}_m = \frac{1}{m}\sum_{l=1}^{m} U_{*l},$$

and B_m measures the between-imputation variance,

$$B_m = \frac{1}{m-1}\sum_{l=1}^{m}(\widehat{\theta}_{*l} - \bar{\theta}_m)(\widehat{\theta}_{*l} - \bar{\theta}_m)^t.$$

The justification of the variance estimate, T_m, is the following. Assuming the imputation is conducted properly (Rubin, 1987, Chapter 4), for example, from the posterior predictive distribution, $P(X_{\text{mis}}|X_{\text{obs}})$, then,

$$\widehat{\theta}_{*l}|X_{\text{obs}}, \theta = \theta_t \overset{iid}{\sim} \mathcal{N}(\widehat{\theta}_{\text{obs}}, B_t), \tag{3.2}$$

$$(m-1)B_t^{-1/2}B_mB_t^{-1/2}|X_{\text{obs}}, \theta = \theta_t \sim \text{Wishart}_k(I_k, m-1). \tag{3.3}$$

By (3.2) and the distribution of $\widehat{\theta}_{\text{obs}}$ in (3.1), we have

$$\bar{\theta}_m|\theta = \theta_t \sim \mathcal{N}(\theta_t, U_t + (1 + m^{-1})B_t). \tag{3.4}$$

Replacing (U_t, B_t) by $(\bar{U}_m, B_m)$, we obtain T_m as an estimate of the variance of $\bar{\theta}_m$.

3.2.3 Hypothesis Testing with Multiply-Imputed Data

If we have the luxury of obtaining a large number of completed datasets, B_m would be an unbiased and accurate estimate of B_t, and T_m an accurate estimate of $\mathrm{Var}(\bar{\theta}_m | \theta = \theta_t)$. Then, we can compute the p-value by referring the test statistic,

$$\tilde{\tilde{D}}_m = (\bar{\theta}_m - \theta_0)^t T_m^{-1} (\bar{\theta}_m - \theta_0)/k, \tag{3.5}$$

to the distribution, χ^2_k/k. Unfortunately, m is typically $3 \sim 10$, so there are not enough degrees of freedom to estimate the $k \times k$ matrix, B_t, and (3.5) hardly produces satisfactory results. Instead, a commonly used test statistic in practice is

$$D_m = \frac{(\bar{\theta}_m - \theta_0)^t \bar{U}_m^{-1} (\bar{\theta}_m - \theta_0)}{k(1 + \bar{r}_m)}, \tag{3.6}$$

where $\bar{r}_m = (1 + m^{-1})\mathrm{tr}(B_m \bar{U}_m^{-1})/k$. The rationale for D_m is the following. With the assumption of equal fraction of missing information (EFMI), i.e., $U_t = \bar{\lambda} B_t$, the variance of $\bar{\theta}_m$ is $(1 + \bar{\gamma}_m)U_t$, where $\bar{\gamma}_m \equiv (1 + m^{-1})\bar{\lambda}$ can be estimated by $\bar{r}_m$ with $k(m-1)$ degrees of freedom. Then (3.6) is obtained by replacing T_m in (3.5) with $(1 + \bar{r}_m)\bar{U}_m$.

The exact distribution of D_m under the null hypothesis is intractable, but it can be well approximated by an F distribution. The reference distribution in Li et al. (1991b) is derived under the assumption of EFMI, and the actual levels of the resulting test are around the nominal levels when the assumption is not severely violated. Xie (2011) proposes another F distribution to approximate the distribution of D_m without such assumption. The resulting levels are closer than those from Li et al. (1991b) to the nominal levels when both $k(m-1)$ and the coefficient of variation of $\boldsymbol{\xi}$, defined below, are large,

$$C_{\xi}^{2} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\xi_i - \bar{\xi}}{\bar{\xi}}\right)^2 = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\lambda_i - \bar{\lambda}}{1+\bar{\lambda}}\right)^2,$$

where $\bar{\xi}$ and $\bar{\lambda}$ are the averages of $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}$, respectively. In section 3.3, we describe the two F distributions in Li et al. (1991b) and Xie (2011), and propose a modification to the procedure in Xie (2011) by accounting for the additional variability due to not knowing $\boldsymbol{\lambda}$. The modified procedure does not assume EFMI and outperforms both methods, in terms of the actual level, in most situations.

When the dimension of the vector of parameters is large, the complete-data procedure may not provide the $k \times k$ matrix U_{*l}, but rather the Wald χ^2 statistic,

$$d_{*l} = (\widehat{\theta}_{*l} - \theta_0)^t U_{*l}^{-1} (\widehat{\theta}_{*l} - \theta_0). \tag{3.7}$$

Let $\mathcal{S}_{\mathrm{d}} = \{d_{*l}; l = 1, \cdots, m\}$ be the set of Wald χ^2 statistic from each of the m complete-data inferences. Without $\mathcal{S}_{\mathrm{m}}$, the complication in forming a final repeated-imputation inference is that $\bar{r}_m$ can not be computed with $\mathcal{S}_{\mathrm{m}}$, and thus D_m is unknown. Currently, there are three reasonably behaved methods to estimate $\bar{r}_m$. Li et al. (1991a) proposed estimating $\bar{r}_m$ based entirely on $\mathcal{S}_{\mathrm{d}}$. Xie (2011) developed an estimator by using an additional Wald χ^2 statistic, denoted as d_{full}, from the complete-data inference applied to the combined data, $X_{\text{full}} = (X_{*1}, \cdots, X_{*\mathrm{m}})$. Meng and Rubin (1992) exploited the asymptotic equivalence between the Wald χ^2 statistic and the log likelihood ratio, and obtained a procedure based on (i) the MLE of ψ under both the null and alternative hypotheses, and (ii) the computer code for calculating the

Table 3.1: Classification of large sample hypothesis testing procedures.

Available Information		Assuming $B_t = \bar{\lambda} U_t$	Not Assuming $B_t = \bar{\lambda} U_t$
$\mathcal{S}_{\mathrm{m}}$		Li et al. (1991b)	Xie (2011) and its modification
No	$\mathcal{S}_{\mathrm{d}}$	Li et al. (1991a)	
	$\mathcal{S}_{\mathrm{d}}$, d_{full}	Xie (2011)	
$\mathcal{S}_{\mathrm{m}}$	MLE, LLR	Meng and Rubin (1992)	

complete-data log likelihood ratio. The resulting test statistic in each of the three procedures is then referred to an F distribution, derived under the assumption of EFMI. We describe and compare these procedures in Section 3.4.

Table 3.1 classifies the above-mentioned large-sample hypothesis testing procedures according to what information is available and whether the assumption of EFMI holds. When $\mathcal{S}_{\mathrm{m}}$ is not available, there are currently no satisfactory methods without assuming EFMI.

3.3 Hypothesis Testing Based on $\mathcal{S}_{\mathrm{m}}$

With $\mathcal{S}_{\mathrm{m}}$, the multiple-imputation test statistic D_m in (3.6) can be computed. Approximating the distribution of D_m under H_0 by F distributions with different degrees of freedom results in a set of hypothesis testing procedures.

3.3.1 Sampling Distribution of D_m under H_0

Since D_m is invariant under nonsingular linear transformation, without loss of generality, we set $\theta_0 = 0, U_t = I_k$ and $B_t = \operatorname{diag}(\lambda_1, \cdots, \lambda_k)$. With proper imputation,

$U_{*l} \approx U_t$, so we also assume $\bar{U}_m = I_k$. The test statistic is then simplified as

$$D_m = \frac{\sum_{i=1}^{k} \bar{\theta}_{m,i}^2}{k(1+\bar{r}_m)}, \tag{3.8}$$

where

$$\bar{r}_m = \left(1 + \frac{1}{m}\right) \frac{\sum_{i=1}^{k} \sum_{l=1}^{m} (\theta_{*l,i} - \bar{\theta}_{m,i})^2}{k(m-1)},$$

and the subscript i represents the i-th component of the vector. From (3.4), under the null hypothesis, $\bar{\theta}_{m,i}|\theta = 0 \overset{ind}{\sim} \mathcal{N}(0, 1+\gamma_i)$, where $\gamma_i = (1+m^{-1})\lambda_i$. So the numerator in (3.8) is distributed as a linear combination of k independent χ_1^2 random variables. In addition, $\sum_{l=1}^{m}(\theta_{*l,i} - \bar{\theta}_{m,i})^2|\theta = \theta_t \overset{ind}{\sim} \lambda_i \chi_{m-1}^2$, so $\bar{r}_m$ follows a linear combination of k independent χ_{m-1}^2 random variables. Because of the independence of $\bar{\theta}_m$ and $\{\theta_{*l} - \bar{\theta}_m; l = 1, \cdots, m\}$, the exact distribution of D_m under H_0 is

$$D_m|\theta = \theta_0 \sim \frac{\sum_{i=1}^{k}(1+\gamma_i)\chi_{1,i}^2}{\sum_{j=1}^{k}\left(1 + \gamma_i \frac{\chi_{m-1,j}^2}{m-1}\right)} \triangleq \mathcal{Y}_m, \tag{3.9}$$

where $\chi_{d,i}^2$ are independent χ_d^2 random variables for $i \in \{1, \cdots, k\}$ and $d \in \{1, m-1\}$.

The distribution of $\mathcal{Y}_m$ depends on the unknown parameters, $\boldsymbol{\gamma} = \{\gamma_1, \cdots, \gamma_k\}$, so an approximated and known distribution is needed to conduct the hypothesis test. Since each γ_i can only be estimated with $m-1$ degrees of freedom, the procedure obtained by replacing γ_i in $\mathcal{Y}_m$ with its estimate typically has unsatisfactory results. In the following sections, we describe the procedures proposed by Li et al. (1991b) and Xie (2011), and discuss a modification to the estimated distribution in Xie (2011). The levels and powers of these procedures are compared theoretically and in simulation.

3.3.2 Approximating $\mathcal{Y}_m$ Assuming $B_t = \bar{\lambda} U_t$: Li et al. (1991b)

With the assumption of EFMI, every γ_i equals to their average, $\bar{\gamma}_m$, so $\mathcal{Y}_m$ simplifies to

$$\mathcal{Y}_m^{\text{LRR}} = \frac{\chi_k^2/k}{(1+\bar{\gamma}_m \chi_v^2/v)/(1+\bar{\gamma}_m)},$$

where $v = k(m-1)$. The distribution of $\mathcal{Y}_m^{\text{LRR}}$ can be further approximated by an F distribution with degrees of freedom k and w, where so far the best choice of w is proposed by Li et al. (1991b),

$$w(\bar{\gamma}_m) = \begin{cases} 4 + (v-4)(1+(1-2/v)/\bar{\gamma}_m)^2, & \text{if } v > 4, \\ (m-1)(k+1)(1+1/\bar{\gamma}_m)^2/2, & \text{otherwise,} \end{cases} \tag{3.10}$$

which is obtained by matching the mean and variance of $\mathcal{Y}_m^{\text{LRR}}$ with a scaled $F_{k,w}$ distribution. Then, the distribution $F_{k,\widehat{w}_m}$, with $\widehat{w}_m = w(\bar{r}_m)$, is used as the reference distribution of D_m in computing the p-value, $P_{\text{LRR}} = \Pr(F_{k,\widehat{w}_m} > D_m)$.

3.3.3 Approximating $\mathcal{Y}_m$ without Assuming $B_t = \bar{\lambda} U_t$: Xie (2011) with Modification

The approximation of $\mathcal{Y}_m$ in Xie (2011) is achieved by ignoring the variability in the denominator of $\mathcal{Y}_m$, or equivalently, assuming $m \to \infty$. Then $\mathcal{Y}_m$ is simplified to be a weighted sum of independent χ_1^2 random variables,

$$\mathcal{Y}_m^{\text{X}} = \frac{1}{k} \sum_{i=1}^{k} \frac{1+\gamma_i}{1+\bar{\gamma}_m} \chi_{1,i}^2. \tag{3.11}$$

Xie (2011) proposed to approximate $\mathcal{Y}_m^{\mathrm{x}}$ further by a gamma random variable using method of moments, i.e., $\mathcal{Y}_m^{\mathrm{x}} \dot{\sim} \text{Gamma}(\rho_m/2, \rho_m/2) \sim \chi^2_{\rho_m}/\rho_m \sim F_{\rho_m,\infty}$, where

$$\rho_m = k(1 + C_{\xi,m})^{-1}, \tag{3.12}$$

with

$$C_{\xi,m} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\gamma_i - \bar{\gamma}_m}{1+\bar{\gamma}_m}\right)^2 \to C_\xi, \text{ as } m \to \infty. \tag{3.13}$$

Compared with the distribution, $F_{k,w(\bar{\gamma}_m)}$, in Li et al. (1991b), we notice that (i) for the numerator degrees of freedom, $\rho_m \leqslant k$ (the equality holds when $B_t = \bar{\lambda}U_t$), and (ii) for the denominator degrees of freedom, $w(\bar{\gamma}_m) < \infty$.

How closely $\mathcal{Y}_m^{\mathrm{x}}$ approximates $\mathcal{Y}_m$ can be examined by their ratio, i.e.,

$$\mathcal{Z} = \frac{\mathcal{Y}_m^{\mathrm{x}}}{\mathcal{Y}_m} = \frac{\sum_{j=1}^{k}(1+\gamma_i\frac{\chi^2_{m-1,j}}{m-1})}{k(1+\bar{\gamma}_m)}.$$

The mean of $\mathcal{Z}$ is 1, and its variance is $2/\beta_m$, where

$$\beta_m = k(m-1)\left[\left(\frac{\bar{\gamma}_m}{1+\bar{\gamma}_m}\right)^2 + C_{\xi,m}\right]^{-1}. \tag{3.14}$$

When the variance of $\mathcal{Z}$ is large, $\mathcal{Y}_m^{\mathrm{x}}$ ignores too much variability in $\mathcal{Y}_m$. To fix this problem, we propose approximating $\mathcal{Z}$ by the distribution $\chi^2_{\beta_m}/\beta_m$, which matches the first two moments of $\mathcal{Z}$.

In practice, $C_{\xi,m}$, ρ_m, and β_m can be estimated by replacing $\bar{\gamma}_m$ and $\overline{\gamma_m^2} = \sum_{i=1}^{m}\gamma_i^2/m$ with their estimates $\bar{r}_m$ and $\overline{r_m^2} = \sum_{i=1}^{m} r_i^2/m$, in (3.13), (3.12), and (3.14). We use $\widehat{C}_{\xi,m}$, $\widehat{\rho}_m$, and $\widehat{\beta}_m$ to denote these estimates. According to simulation, we find when $\widehat{\beta}_m \leqslant 20(m-1)$ or $k \leqslant 4$, using $F_{\widehat{\rho}_m, \widehat{\beta}_m}$ as the reference distribution generally leads to levels closer to the nominal levels than using $F_{\widehat{\rho}_m, \infty}$. So we propose referring D_m to the distribution $F_{\widehat{\rho}_m, \widehat{\eta}_m}$, where

$$\widehat{\eta}_m = \begin{cases} \widehat{\beta}_m, & \text{if } \widehat{\beta}_m \leqslant 20(m-1) \text{ or } k \leqslant 4, \\ +\infty, & \text{otherwise,} \end{cases} \tag{3.15}$$

and the corresponding p-value is $P_{\mathrm{x},\eta} = \Pr(F_{\widehat{\rho}_m, \widehat{\eta}_m} > D_m)$.

3.3.4 Comparison

The procedures we described above use the same test statistic, D_m, which is then referred to an F distribution to compute the p-value. In this section, we compare the levels and powers of these procedures when $m \to \infty$ and when m is finite. For ease of reference, we give names to the few tests we compare in this section and list the corresponding test statistics and the reference distributions in Table 3.2. In addition, we refer $\mathcal{T}_{\text{obs}}$ to the observed test with the test statistic D_{obs} referred to the distribution χ^2_k/k.

3.3.4.1 Levels Comparison as $m \to \infty$

Let D_∞ be the limit of D_m as $m \to \infty$, then, $D_\infty = k^{-1} \sum_{i=1}^{k} \widehat{\theta}^2_{\text{obs},i}/\bar{\xi}$. The distribution of D_∞ under H_0 is $\mathcal{Y}_\infty$, the limit of $\mathcal{Y}_m$ as $m \to \infty$, i.e.,

$$\mathcal{Y}_m \to \mathcal{Y}_\infty \triangleq \frac{1}{k} \sum_{i=1}^{k} \frac{\xi_i}{\bar{\xi}} \chi^2_{1,i}.$$

Table 3.2: Summary of Tests For Comparison.

Finite m				$m \to \infty$			
Test name	Stat	Exact ref.	Test ref.	Test name	Stat	Exact ref.	Test ref.
$\mathcal{T}^m_{\text{LRR}}$	D_m	$\mathcal{Y}_m$	$F_{k,\widehat{w}_m}$	$\mathcal{T}^\infty_{\text{LRR}}$	D_∞	$\mathcal{Y}_\infty$	χ^2_k/k
$\mathcal{T}^m_{\text{X}}$			$F_{\widehat{\rho}_m,\infty}$	$\mathcal{T}^\infty_{\text{X}}$			$\chi^2_{\rho_\infty}/\rho_\infty$
$\mathcal{T}^m_{\text{X},\beta}$			$F_{\widehat{\rho}_m,\widehat{\beta}_m}$				
$\mathcal{T}^m_{\text{X},\eta}$			$F_{\widehat{\rho}_m,\widehat{\eta}_m}$				
$\mathcal{T}^m_{\text{Exact}}$			$\mathcal{Y}_m$	$\mathcal{T}^\infty_{\text{Exact}}$			$\mathcal{Y}_\infty$

Ref. is short for reference distribution; Test ref. is the distribution used in the test as the reference; Exact ref. is the exact distribution of the test statistic under H_0.

For the approximated reference distributions, as $m \to \infty$,

$$F_{k,\widehat{w}_m} \to F_{k,\infty} \sim \chi^2_k/k, \tag{3.16}$$

$$F_{\widehat{\rho}_m,\star} \to F_{\rho_\infty,\infty} \sim \chi^2_{\rho_\infty}/\rho_\infty, \tag{3.17}$$

where $\rho_\infty = k(1+C^2_\xi)^{-1} \leqslant k$.

The actual level of an α-nominal-level test, $\mathcal{T}$, based on D_∞ can be computed by $\Pr(\mathcal{Y}_\infty > Q_{\mathcal{T}}(\alpha))$, where $Q_{\mathcal{T}}(\alpha)$ is the $1-\alpha$ percentile of the reference distribution used in test $\mathcal{T}$. Figure 3.1 shows the actual levels of the 5%-nominal-level tests, $\mathcal{T}^\infty_{\text{LRR}}$ (black lines) and $\mathcal{T}^\infty_{\text{X}}$ (red lines), from a full factorial experiment with three factors,

1. The dimension of θ: $k = 5$ (solid lines), and 50 (dashed lines);

2. The average ratios of complete to observed information: $\bar{\xi} = 1.2$, 1.5, and 2;

3. The coefficient of variation of $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_k)$: $C_\xi = 0, 0.05, 0.10, \cdots, 0.50$.

For each $(k, \bar{\xi}, C_\xi)$, we first generate the vector $\boldsymbol{\xi}$ with the pre-specified mean and coefficient of variation. Then, we simulate 1 million independent draws from $\mathcal{Y}_\infty$, and estimate the actual level of each test by the proportion of the sample greater than the

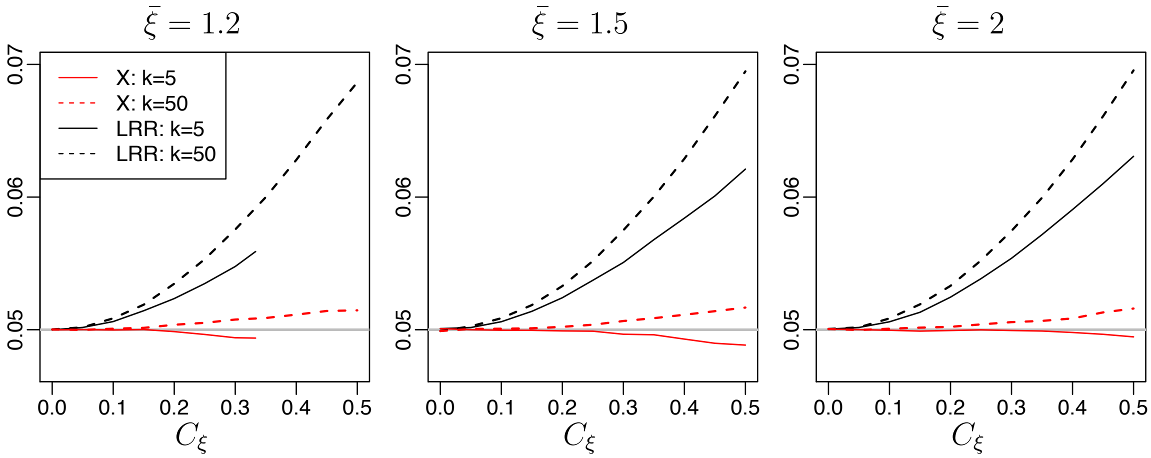

Figure 3.1: Actual levels of the 5%-nominal-level tests, with the test statistic D_∞ referred to the distributions χ^2_k/k (black lines) and $\chi^2_{\rho_\infty}/\rho_\infty$ (red lines), under different settings. The solid and dashed lines correspond to the levels when $k = 5$ and 50, respectively. The nominal level, 5%, is marked by the gray lines.

$1 - \alpha$ percentile of the corresponding reference distribution. Note that when $k = 5$ and $\bar{\xi} = 1.2$, the largest value of C_ξ is 0.33, thus in Figure 3.1 (left), the solid lines are plotted up to $C_\xi = 0.33$.

When the variation of $\boldsymbol{\xi}$ is zero, $\chi^2_{\rho_\infty}/\rho_\infty \sim \chi^2_k/k$ is the exact distribution of D_∞ under H_0, so the actual levels of $\mathcal{T}^\infty_{\text{LRR}}$ and $\mathcal{T}^\infty_{\text{X}}$ equal to the nominal levels. When $C_\xi > 0$, the first two central moments of $\chi^2_{\rho_\infty}/\rho_\infty$ match exactly with those of $\mathcal{Y}_\infty$, so the red lines in Figure 3.1 are close to the nominal level, 5%. The distribution χ^2_k/k, however, only matches with $\mathcal{Y}_\infty$ in expectation, and its variance is smaller than that of $\mathcal{Y}_\infty$, more specifically,

$$\text{Var}\left(\mathcal{Y}_\infty\right) = \left(1 + C_\xi^2\right) \text{Var}\left(\chi^2_k/k\right).$$

As a result, the actual levels of test $\mathcal{T}^\infty_{\text{LRR}}$ are larger than the nominal levels, and they increase as C_ξ increases, as the black lines in Figure 3.1 show.

3.3.4.2 Powers Comparison as $m \to \infty$

Under the alternative hypothesis that $\theta = \theta_t$, the exact distribution of D_∞ is

$$\mathcal{Y}_{t,\infty} \sim \frac{1}{k\bar{\xi}} \sum_{i=1}^{k} \left(\theta_{t,i} + \sqrt{\xi_i} Z_i\right)^2 \sim \frac{1}{k} \sum_{i=1}^{k} \frac{\xi_i}{\bar{\xi}} \chi^2_{1,i}(a_i),$$

where $Z_i \overset{iid}{\sim} \mathcal{N}(0,1)$, and $\{\chi^2_{1,i}(a_i); i = 1, \cdots, k\}$ are independent noncentral χ^2 random variables with degrees of freedom 1 and noncentrality parameters $a_i = \theta^2_{t,i}/\xi_i$. The power of an α-nominal-level test, $\mathcal{T}$, based on D_∞, is then $\Pr(\mathcal{Y}_{t,\infty} > Q_{\mathcal{T}}(\alpha))$. For comparison, we also investigate the power of test $\mathcal{T}_{\text{obs}}$ based on D_{obs}. Given $\theta = \theta_t$, D_{obs} follows the scaled non-central χ^2 distribution, $\chi^2_k(\nu)/k$, where $\nu = \sum_{i=1}^k \theta^2_{t,i}/\xi_i$. So the power of the α-level test is $\Pr(\chi^2_k(\nu)/k > Q_k(\alpha))$, where $Q_k(\alpha)$ is the $1-\alpha$ percentile of χ^2_k/k.

In Figure 3.2, we show the powers of four 5%-nominal-level tests: (i) $\mathcal{T}^\infty_{\text{LRR}}$ (black solid lines), (ii) $\mathcal{T}^\infty_{\text{X}}$ (red solid lines), (iii) $\mathcal{T}^\infty_{\text{Exact}}$ (blue dashed lines), and (iv) $\mathcal{T}_{\text{obs}}$ (green dashed lines). We fix $\Delta \triangleq \sum_{i=1}^k \theta^2_{t,i}/k = 0.5$, and simulate 30 replicates of $(\boldsymbol{\xi}, \theta_t)$ for each $(k, \bar{\xi}, C_\xi)$. The power of each test is then estimated by the proportion of the 1 million draws from the distribution of the test statistic, specified by $\boldsymbol{\xi}$ and θ_t, that are greater than the 95% percentile of the reference distribution. The plotted value for each $(k, \bar{\xi}, C_\xi)$ is the average of the corresponding 30 estimated powers.

When $C_\xi = 0$, the four tests are equivalent, so the powers are the same. When $C_\xi > 0$, we find (i) compared with test $\mathcal{T}_{\text{obs}}$, tests based on D_∞ suffers from power loss due to the ignorance of the variability in $\boldsymbol{\xi}$, and the loss is more severe for larger C_ξ; (ii) tests $\mathcal{T}^\infty_{\text{X}}$ and $\mathcal{T}^\infty_{\text{Exact}}$ have almost identical powers, because their reference

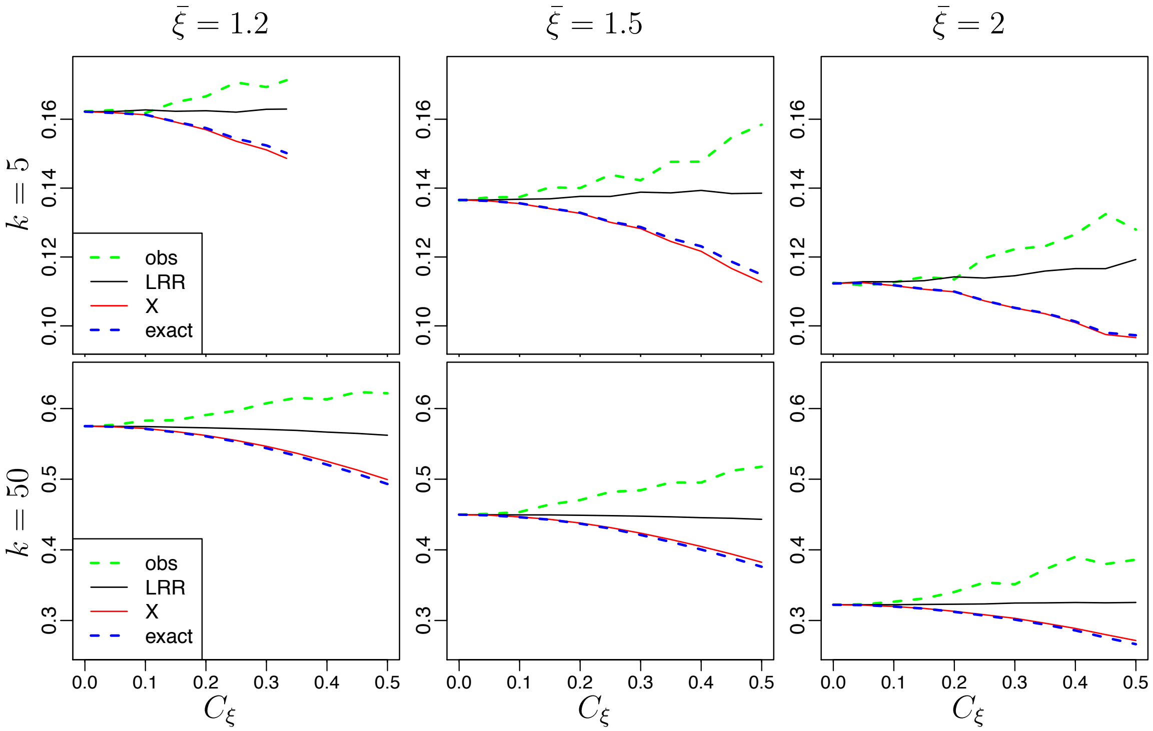

Figure 3.2: Powers of four 5%-nominal-level tests: (i) (green dashed lines) test $\mathcal{T}_{\text{obs}}$, with the test statistic D_{obs} referred to χ^2_k/k, (ii) (blue dashed lines) test $\mathcal{T}^{\infty}_{\text{Exact}}$, with the test statistic D_∞ referred to $\mathcal{Y}_\infty$, (iii) (black solid lines) test $\mathcal{T}^{\infty}_{\text{LRR}}$, with D_∞ referred to χ^2_k/k, and (iv) (red solid lines) test $\mathcal{T}^{\infty}_{\text{X}}$, with D_∞ referred to $\chi^2_{\rho_\infty}/\rho_\infty$. The top and bottom rows show the results when $k = 5$ and 50, respectively.

distributions, $\chi^2_{\rho_\infty}/\rho_\infty$ and $\mathcal{Y}_\infty$, have the same mean and variance; and (iii) test $\mathcal{T}^{\infty}_{\text{LRR}}$, which has higher levels than tests $\mathcal{T}^{\infty}_{\text{X}}$ and $\mathcal{T}^{\infty}_{\text{Exact}}$, also has larger powers, due to the smaller variance of its reference distribution, χ^2_k/k.

In addition, other factors being fixed, the powers of tests $\mathcal{T}^{\infty}_{\text{X}}$ and $\mathcal{T}^{\infty}_{\text{Exact}}$ decrease as C_ξ increases. This is because the null and alternative distributions of D_∞ are more distinguishable when C_ξ is smaller. More specifically, for fixed Δ and $\bar{\xi}$, the difference in expectation between the distributions of D_∞ under H_a and H_0 is fixed, i.e.,

$$\mathrm{E}(\mathcal{Y}_{t,\infty}) - \mathrm{E}(\mathcal{Y}_\infty) = \Delta/\bar{\xi};$$

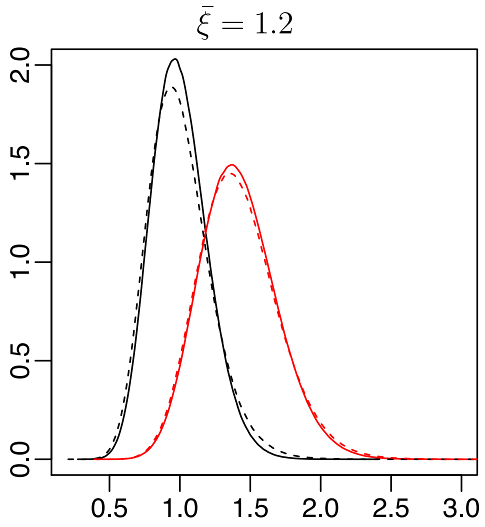

Figure 3.3: Distributions of the test statistis, D_∞, under the null (distribution $\mathcal{Y}_\infty$, black lines) and the alternative (distribution $\mathcal{Y}_{t,\infty}$, red lines) hypotheses. Solid lines correspond to results when $C_\xi = 0$, and dashed lines correspond to $C_\xi = 0.5$.

however, $\mathrm{Var}(\mathcal{Y}_\infty)$ increases more than $\mathrm{Var}(\mathcal{Y}_{t,\infty})$ as C_ξ increases, see below,

$$\mathrm{Var}(\mathcal{Y}_{t,\infty}) = \mathrm{Var}(\mathcal{Y}_\infty)\left(1 + \frac{1}{(1+C_\xi^2)k}\sum_{i=1}^{k}\frac{\theta_{t,i}^2\xi_i}{\bar{\xi}^2}\right).$$

This is also confirmed in Figure 3.3, which shows the distribution of $\mathcal{Y}_\infty$ (black lines) becomes more spread out than $\mathcal{Y}_{t,\infty}$ (red lines) when C_ξ increases from zero (solid lines) to 0.5 (dashed lines). On the contrary, the reference distribution of test $\mathcal{T}_{\text{LRR}}^{\infty}$ is invariant to C_ξ, whereas the distribution of D_∞ under H_a gets slightly more spread out and separate from the reference distribution, so its power increases sightly with C_ξ. Figure 3.2 also shows the powers of all the tests decrease as $\bar{\xi}$ increases. This is because of the reduction in expectation between the distributions of the test statistics under H_0 and H_a. Finally, larger k corresponds to smaller variability in the test statistics under both H_0 and H_a, and thus larger power.

In summary, tests $\mathcal{T}_{\text{X}}^{\infty}$ and $\mathcal{T}_{\text{Exact}}^{\infty}$ have nearly identical behavior in terms of level and power, because their reference distributions have the same first two moments. Both the level and the power of test $\mathcal{T}_{\text{LRR}}^{\infty}$ are larger than $\mathcal{T}_{\text{Exact}}^{\infty}$ because its reference distribution underestimates the variance of the test statistic under H_0.

3.3.4.3 Levels Comparison with Finite m

In this section, we compare the actual levels of the tests based on D_m for finite m. We again conduct a full factorial experiment with the three factors $k \in \{2, 5, 10, 50\}$, $\bar{\xi} \in \{1.2, 1.5, 2\}$, and $C_\xi \in \{0, 0.05, \cdots, 0.5\}$. The simulation is conducted according to the following steps. For each $(k, \bar{\xi}, C_\xi)$,

1. Generate $\boldsymbol{\xi}$ that satisfies (i) each component is $\geqslant 1$, and (ii) the mean and coefficient of variance of $\boldsymbol{\xi}$ equal to the pre-specified values. Let $\lambda_i = \xi_i - 1$, $B_t = \text{diag}(\lambda_1, \cdots, \lambda_k)$, $T_t = \text{diag}(\xi_1, \cdots, \xi_k)$, and $U_t = I_k$.
2. Generate θ_t so that $\Delta = \sum_{i=1}^{k} \theta_{t,i}^2 / k$ equals to the pre-specified value. In computing the levels, $\Delta = 0$ and $\theta_t = (0, \cdots, 0)$.
3. Simulate $\widehat{\theta}_{\text{obs}} \sim \mathcal{N}(\theta_t, T_t)$.
4. For $l = 1, \cdots, m$, simulate $\widehat{\theta}_{*l} \overset{iid}{\sim} \mathcal{N}(\widehat{\theta}_{\text{obs}}, B_t)$.
5. Compute D_m as defined in (3.8), and the reference distribution, $\mathcal{F}$, of the test.
6. Compute the p-value by $\Pr(\mathcal{F} > D_m)$.
7. Repeat Step 3-6 for 10^6 times and compute the level or power of an α-nominal-level test by the proportion of the computed p-values that are less than α.

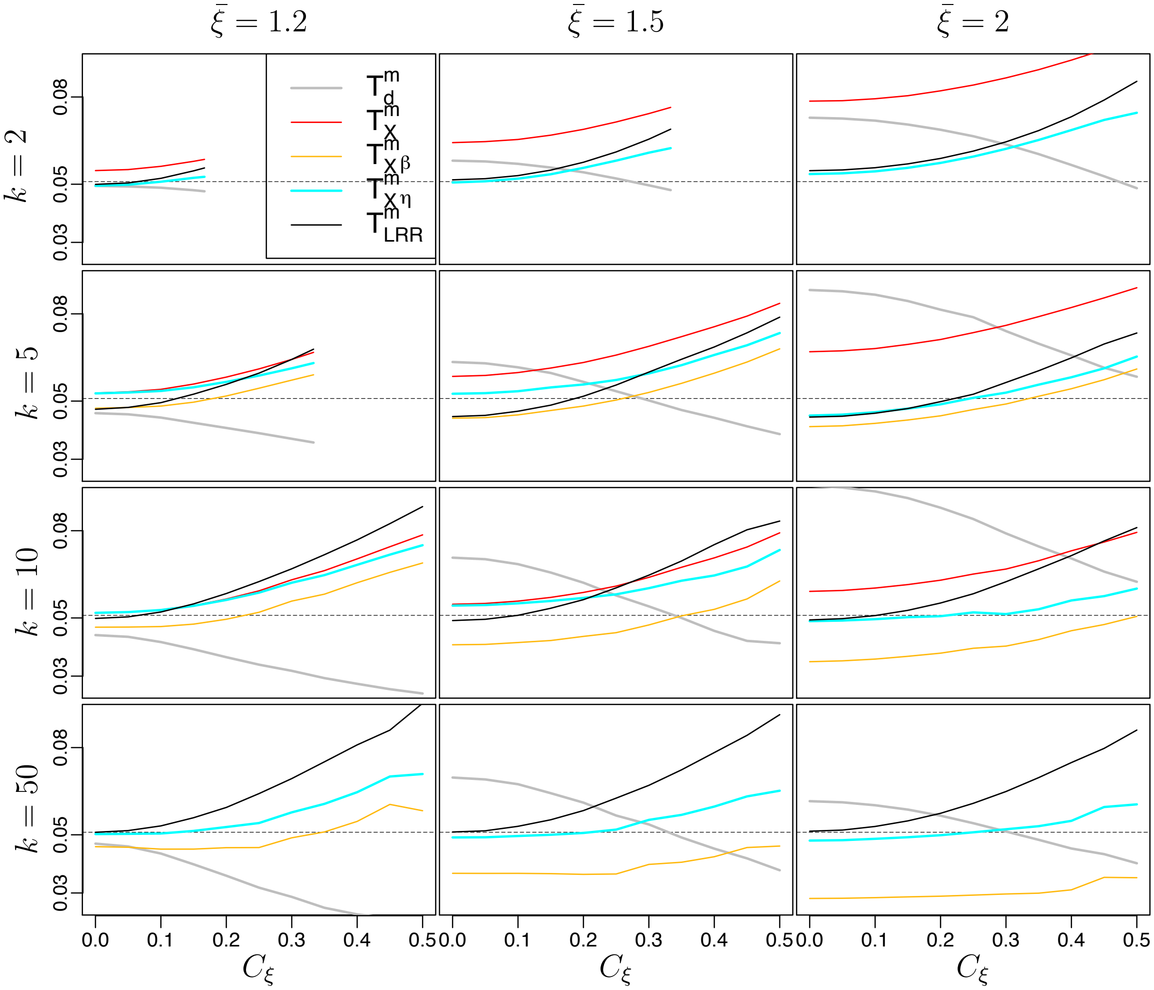

Figure 3.4: Actual levels of the 5%-nominal-level tests, $\mathcal{T}_{\text{LRR}}^m$ (black lines), $\mathcal{T}_{\text{X}}^m$ (red lines), $\mathcal{T}_{\text{X},\beta}^m$ (orange lines), and $\mathcal{T}_{\text{X},\eta}^m$ (cyan lines), with $m = 3$, under different settings. When $k = 2$, the cyan and orange lines overlap; when $k = 50$, the cyan and red lines overlap.

Figure 3.4 shows the actual levels of the 5%-nominal-level tests, $\mathcal{T}_{\text{LRR}}^m$ (black lines), $\mathcal{T}_{\text{X}}^m$ (red lines), $\mathcal{T}_{\text{X},\beta}^m$ (orange lines), and $\mathcal{T}_{\text{X},\eta}^m$ (cyan lines), with $m = 3$. Figure 3.5 shows the results when m increases to 10. Note, the maximum of C_ξ can be less than 0.5 for some pairs of $(k, \bar{\xi})$, so there are no results with large C_ξ in these cases. The levels of test $\mathcal{T}_{\text{LRR}}^m$, for $m = 3, 10$, and ∞, are close to 5% when $C_\xi < 0.1$, but they become substantially larger than 5% when C_ξ is large. Test $\mathcal{T}_{\text{X}}^m$ produces levels considerably

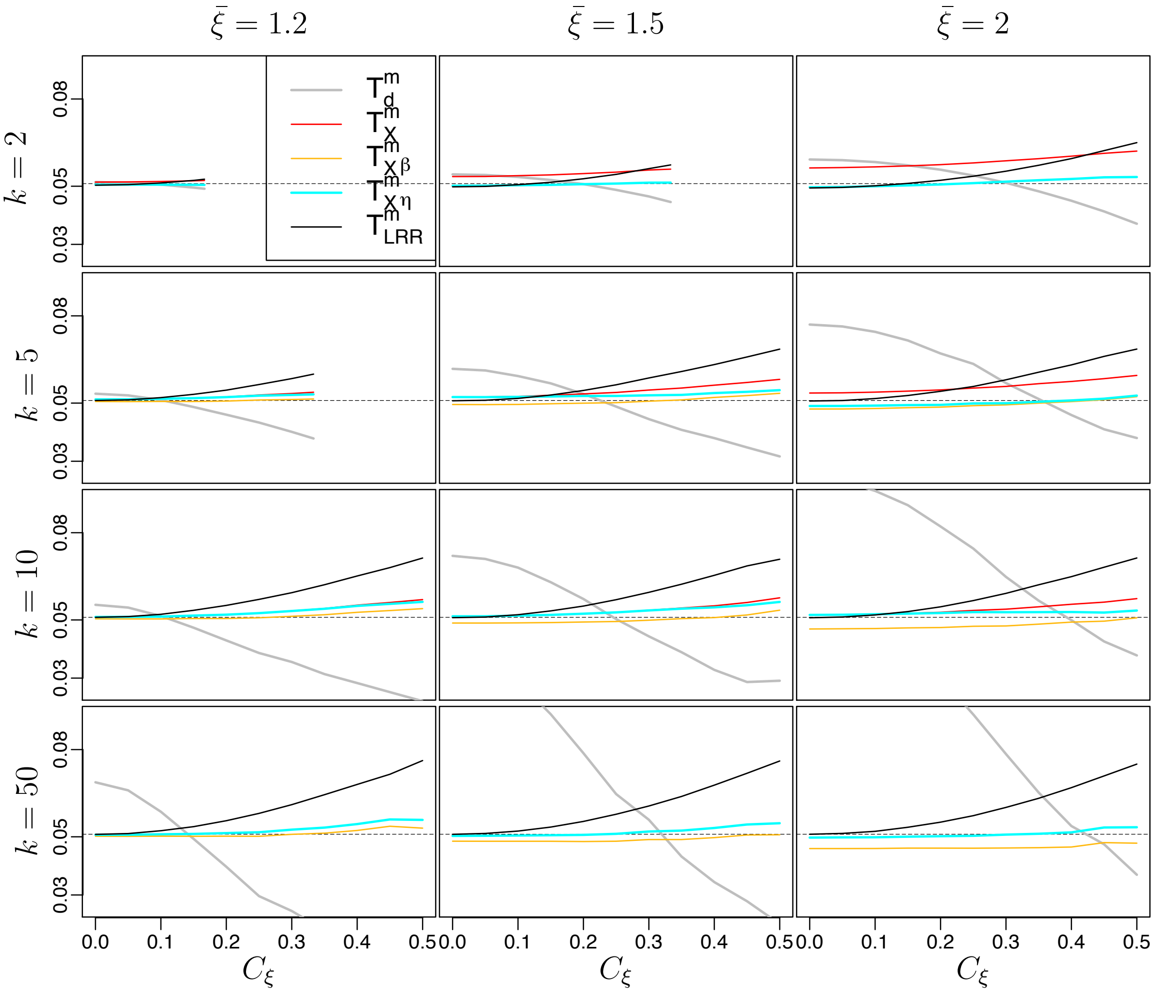

Figure 3.5: See the caption of Figure 3.4. Here $m = 10$.

larger than 5% when k is small, whereas test $\mathcal{T}_{\mathrm{x},\beta}^m$ has levels much smaller than 5% when k is large. The levels of test $\mathcal{T}_{\mathrm{x},\eta}^m$ are the closest to 5% among all the four tests under nearly all the situations. In addition, the levels of tests $\mathcal{T}_{\mathrm{x},\star}^m$ (namely, $\mathcal{T}_{\mathrm{x}}^m$, $\mathcal{T}_{\mathrm{x},\beta}^m$, and $\mathcal{T}_{\mathrm{x},\eta}^m$) get much closer to 5% when m increases to 10.

3.3.4.4 Powers Comparison with Finite m

Tests $\mathcal{T}^m_{\text{LRR}}$ and $\mathcal{T}^m_{\text{X},\star}$ use the same test statistic, D_m, but different approximations to its distribution under H_0. As a result, other factors being fixed, their actual levels and powers are different. To make fair comparison of the powers of these tests, we compare their receiver operating characteristic (ROC) curves. The horizontal axis of a ROC curve is the actual level of an α-nominal-level test, for $\alpha \in (0, 1)$, and the vertical axis is the corresponding power. A test with larger area under the curve is considered to have better performance.

Figure 3.6 shows the ROC curves of several tests: (i) $\mathcal{T}_{\text{obs}}$ (green dashed lines), (ii) $\mathcal{T}^m_{\text{LRR}}$ and $\mathcal{T}^m_{\text{X},\star}$ with $m = 3$ (black solid lines), (iii) $\mathcal{T}^m_{\text{LRR}}$ and $\mathcal{T}^m_{\text{X},\star}$ with $m = 10$ (red dotted lines), and (iv) $\mathcal{T}^\infty_{\text{LRR}}$ and $\mathcal{T}^\infty_{\text{X},\star}$, which have nearly identical ROC curve as test $\mathcal{T}^\infty_{\text{Exact}}$ (blue dot-dash lines). Other factors being fixed, the ROC curves of tests $\mathcal{T}^m_{\text{LRR}}$ and $\mathcal{T}^m_{\text{X},\star}$ nearly overlap. This implies their performances are almost identical, and thus tests with higher true positive rates also have higher false positive rates. Figure 3.6 also shows increasing m from 3 to 10 results in a notable improvement of the performances of tests $\mathcal{T}^m_{\text{LRR}}$ and $\mathcal{T}^m_{\text{X},\star}$. We also find test $\mathcal{T}_{\text{obs}}$ is more superior than tests based on D_m, when k, C_ξ, and/or $\bar{\xi}$ are large, because more information in the observed data is ignored by D_m.

To conclusion, among the four tests, $\mathcal{T}^m_{\text{LRR}}$, $\mathcal{T}^m_{\text{X}}$, $\mathcal{T}^m_{\text{X},\beta}$, and $\mathcal{T}^m_{\text{X},\eta}$, the levels of test $\mathcal{T}^m_{\text{X},\eta}$ are closest to the nominal levels under most situations. Even though test $\mathcal{T}^m_{\text{LRR}}$ appears to have higher powers than tests $\mathcal{T}^m_{\text{X},\star}$, it also has the highest levels. Their ROC curves are nearly identical, so we recommend referring the test statistic D_m to distribution $F_{\widehat{\rho}_m,\widehat{\eta}_m}$ to compute the p-value. In addition, increasing m from 3 to 10

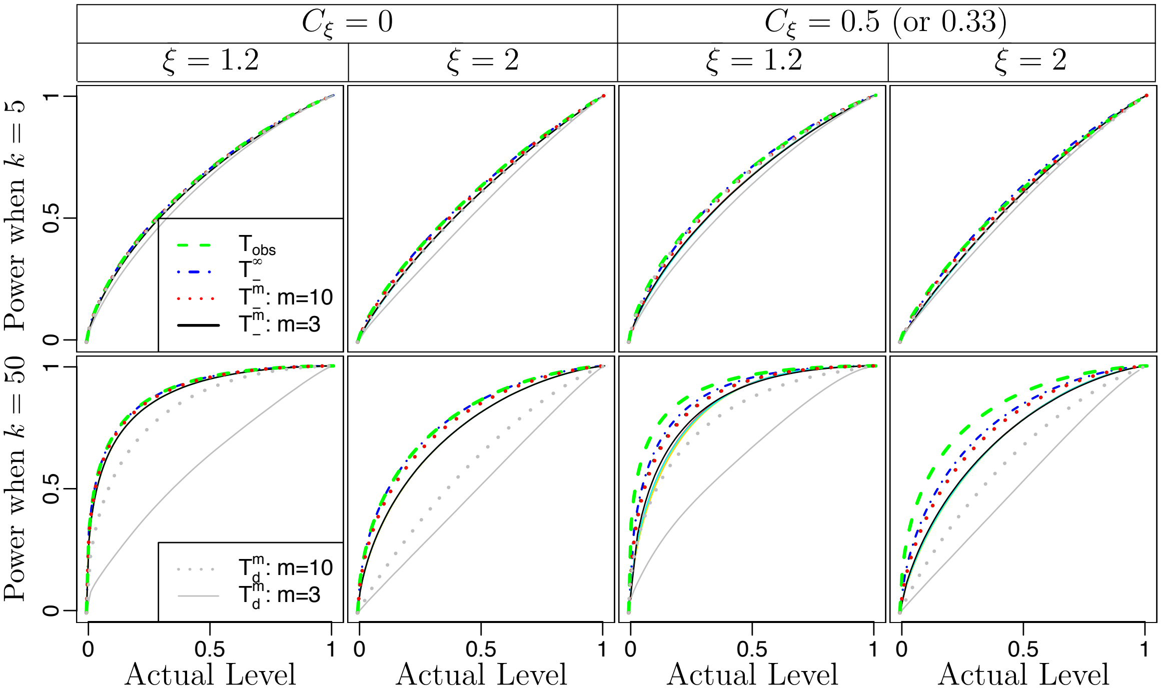

Figure 3.6: ROC curves of tests (i) $\mathcal{T}_{\rm obs}$ (green dashed lines), (ii) $\mathcal{T}^{m}_{\rm LRR}$ and $\mathcal{T}^{m}_{\rm X,\star}$ with $m = 3$ (black solid lines), (iii) $\mathcal{T}^{m}_{\rm LRR}$ and $\mathcal{T}^{m}_{\rm X,\star}$ with $m = 10$ (red dotted lines), (iv) $\mathcal{T}^{\infty}_{\rm LRR}$, $\mathcal{T}^{\infty}_{\rm X,\star}$, and $\mathcal{T}^{\infty}_{\rm Exact}$ (blue dot-dash lines), and (v) $\mathcal{T}^{m}_{\rm d}$, the test based on $\mathcal{S}_{\rm d}$ by Li et al. (1991a) (see Section 3.4.1), with $m = 3$ (gray solid lines) and $m = 10$ (gray dotted lines). The horizontal coordinate of a ROC curve is the actual level of an α-nominal-leve test, with $\alpha \in (0, 1)$, and the vertical coordinate is the corresponding powers. Note, when $k = 5$ and $\bar{\xi} = 1.2$, the largest possible value of C_ξ is 0.33, so we plot the ROC curves correspond to $C_\xi = 0.33$.

greatly improves both the actual levels and the ROC curves of these tests.

3.4 Hypothesis Testing without $\mathcal{S}_{\rm m}$

When $\mathcal{S}_{\rm m}$ is not available, the combined test statistic D_m can not be computed, so we need to estimate D_m, as well as its corresponding sampling distribution. Fortunately, D_m can be approximated by a function of $\bar{r}_m$ and the average, denoted as

$\bar{d}_m$, of the Wald χ^2 statistics defined in (3.7), that is,

$$D_m \approx \widetilde{D}_m \triangleq \left(\frac{\bar{d}_m}{k} - \frac{m-1}{m+1} \bar{r}_m \right) / (1 + \bar{r}_m). \tag{3.18}$$

So when $\bar{d}_m$ is available, an estimate of $\bar{r}_m$ corresponds to an estimate of D_m, and thus a testing procedure.

3.4.1 Estimating $\bar{r}_m$ Based on $\mathcal{S}_{\mathrm{d}}$: Li et al. (1991a)

So far, the best estimate of $\bar{r}_m$ based purely on the set, $\mathcal{S}_{\mathrm{d}}$, of the Ward χ^2 statistics was proposed by Li et al. (1991a),

$$\widehat{r}_d = \left(1 + \frac{1}{m}\right) \left(\frac{1}{m-1} \sum_{l=1}^{m} \left(\sqrt{d_{*l}} - \overline{\sqrt{d}} \right)^2 \right),$$

where $\overline{\sqrt{d}}$ is the mean of $\{\sqrt{d_{*l}}; l = 1, \cdots, m\}$. Let $\widehat{D}_d$ be the corresponding test statistic with $\bar{r}_m$ replaced by $\widehat{r}_d$ in (3.18). Li et al. (1991a) suggest using $F_{k,\widehat{w}_d}$ as the reference distribution, where $\widehat{w}_d = (m-1)(1+\widehat{r}_d)^2/k^{3/m}$. The factor $k^{-3/m}$ adjusts the loss of degrees of freedom, because we only have the m scalar χ^2 statistics, rather than the $k(m-1)$ degrees of freedom from the m pairs of $k \times 1$ point estimate and $k \times k$ variance-covariance estimate to compute $\bar{r}_m$. For ease of reference, we call this test procedure $\mathcal{T}_{\mathrm{d}}^m$.

The gray lines in Figure 3.4 and 3.5 show the actual levels of 5%-nominal-level test $\mathcal{T}_{\mathrm{d}}^m$ where $m = 3$ and 10, respectively. It appears increasing m does not improve the levels. As Li et al. (1991a) pointed out, the factor $k^{-3/m}$ was chosen to make the test perform especially well when $m = 3$ in terms of the actual levels. Figure 3.6

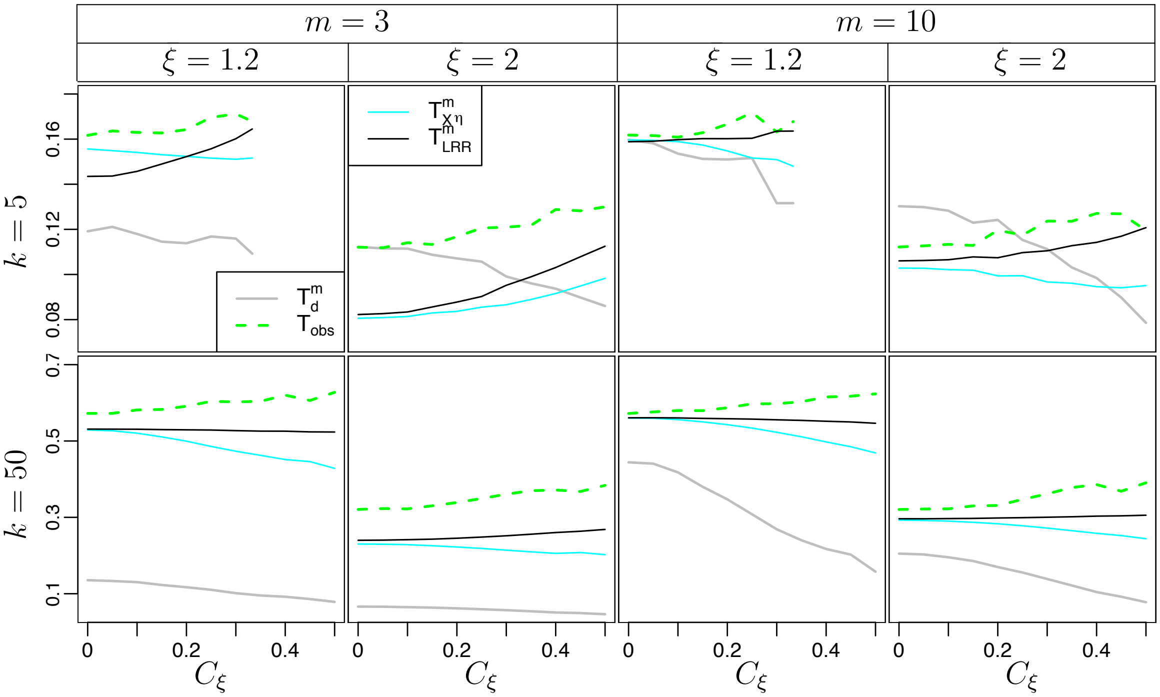

Figure 3.7: Powers of 5%-nominal-level tests, (i) $\mathcal{T}_{\text{obs}}$ (green dashed lines), (ii) $\mathcal{T}^m_{\text{X},\eta}$ (cyan solid lines), (iii) $\mathcal{T}^m_{\text{LRR}}$ (black solid lines), and (iv) $\mathcal{T}^m_{\text{d}}$ (gray solid lines) .

shows the ROC curves of the test when $m = 3$ (solid gray lines) and 10 (dashed gray lines). When k is large, the test based purely on $\mathcal{S}_{\text{d}}$ has much worse performance than tests based on $\mathcal{S}_{\text{m}}$, because the set $\mathcal{S}_{\text{d}}$ contains much less information than $\mathcal{S}_{\text{m}}$. Figure 3.7 compares the powers of the 5%-nominal-level test $\mathcal{T}^m_{\text{d}}$ (gray solid lines) with tests $\mathcal{T}_{\text{obs}}$ (green dashed lines), $\mathcal{T}^m_{\text{X},\eta}$ (cyan solid lines), and $\mathcal{T}^m_{\text{LRR}}$ (black solid lines). Consistent with Figure 3.6, compared with procedures based on $\mathcal{S}_{\text{m}}$, we find (i) $\mathcal{T}^m_{\text{d}}$ is associated with a severe loss of power, especially when k is large; and (ii) increasing m significantly improves the power of $\mathcal{T}^m_{\text{d}}$.

3.4.2 Estimating $\bar{r}_m$ Based on Likelihood Ratio: Meng and Rubin (1992)

3.4.2.1 Procedure

Meng and Rubin (1992) proposed using a complete-data log likelihood ratio to estimate $\bar{r}_m$. The procedure is based on the asymptotic equivalence of the Wald χ^2 statistic and the log likelihood ratio statistic, defined as

$$d'(\widehat{\psi}^0, \widehat{\psi}|X) \triangleq 2 \log \frac{f(X|\widehat{\psi})}{f(X|\widehat{\psi}^0)}, \tag{3.19}$$

where $\widehat{\psi}^0 = \widehat{\psi}^0(X)$ and $\widehat{\psi} = \widehat{\psi}(X)$ are the estimates of the model parameters, ψ, that maximize the complete-data likelihood, $f(X|\psi)$, under the null and alternative hypotheses. The heuristic derivation of their equivalence is the following. Without loss of generality, we can assume θ is a sub-vector of ψ, i.e., $\psi = (\theta, \vartheta)$, and the asymptotic variance-covariance matrix of the complete-data MLE, $\widehat{\psi} = (\widehat{\theta}, \widehat{\vartheta})$, is $\text{diag}(U, W)$. Then, asymptotically the complete-data likelihood is

$$f(X|\psi) \propto \exp\left[-\frac{1}{2}(\widehat{\theta} - \theta)^t U^{-1}(\widehat{\theta} - \theta) - \frac{1}{2}(\widehat{\vartheta} - \vartheta)^t W^{-1}(\widehat{\vartheta} - \vartheta)\right]. \tag{3.20}$$

The MLE under the null hypothesis is $\widehat{\psi}^0 = (0, \widehat{\vartheta})$. The equivalence of $d'(\widehat{\psi}^0, \widehat{\psi}|X)$ and $\widehat{\theta}^t U^{-1}\widehat{\theta}$ is obtained by replacing $\widehat{\psi}$ with $(\widehat{\theta}, \widehat{\vartheta})$, and $\widehat{\psi}^0$ with $(0, \widehat{\vartheta})$.

Let $d(\bar{\theta}_m, \bar{U}_m) = \bar{\theta}_m^t \bar{U}_m^{-1} \bar{\theta}_m$, then $D_m = d(\bar{\theta}_m, \bar{U}_m)/k(1 + \bar{r}_m)$. Compare this expression of D_m with (3.18), we can obtain an approximation of $\bar{r}_m$, i.e.,

$$\bar{r}_m \approx \frac{m+1}{k(m-1)} \left[\bar{d}_m - d(\bar{\theta}_m, \bar{U}_m)\right]. \tag{3.21}$$

Meng and Rubin (1992) proposed estimating $d(\bar{\theta}_m, \bar{U}_m)$ using (i) the MLE of ψ based on each of the completed dataset X_{*l} under both the null and alternative hypotheses, denoted as $\widehat{\psi}^0_{*l} = \widehat{\psi}^0(X_{*l})$ and $\widehat{\psi}_{*l} = \widehat{\psi}(X_{*l})$, and (ii) the code for computing the complete-data likelihood ratio, defined in (3.19), as a function of the point estimates. Let $\bar{\psi}^0_m = m^{-1}\sum_{l=1}^m \widehat{\psi}^0_{*l}$ and $\bar{\psi}_m = m^{-1}\sum_{l=1}^m \widehat{\psi}_{*l}$ be the averages of the m point estimates under H_0 and H_a. Asymptotically,

$$d(\bar{\theta}_m, \bar{U}_m) \approx \bar{d}_{\mathrm{L}} \triangleq \frac{1}{m}\sum_{l=1}^m d'(\bar{\psi}^0_m, \bar{\psi}_m | X_{*l}),$$

which can be obtained by replacing $\bar{\psi}^0_m$ in (3.19) with $(0, \bar{\vartheta}_m)$, and $\bar{\psi}_m$ with $(\bar{\theta}_m, \bar{\vartheta}_m)$. So we can estimate $\bar{r}_m$ by

$$\widehat{r}_{\mathrm{L}} = \frac{m+1}{k(m-1)}(\bar{d}_m - \bar{d}_{\mathrm{L}}) \approx \frac{m+1}{k(m-1)}(\bar{d}'_m - \bar{d}_{\mathrm{L}}),$$

where $\bar{d}'_m \triangleq m^{-1}\sum_{l=1}^m d'(\widehat{\psi}^0_{*l}, \widehat{\psi}_{*l} | X_{*l})$ is asymptotically equal to $\bar{d}_m$. The resulting test statistic is

$$D_{\mathrm{L}} = \frac{\bar{d}_{\mathrm{L}}}{k(1+\widehat{r}_{\mathrm{L}})}, \tag{3.22}$$

which is referred to distribution $F_{k,w(\widehat{r}_{\mathrm{L}})}$, where $w(\widehat{r}_{\mathrm{L}})$ is obtained by replacing $\bar{\gamma}_m$ in (3.10) by $\widehat{r}_{\mathrm{L}}$. Note, since this procedure assumes the normality of the MLE in the derivation, it is advised to use parameters, the MLE of which are close to a normal distribution.

3.4.2.2 A Practical Issue: Negative $\widehat{r}_{\mathrm{L}}$

The procedure is especially useful when the dimension of ψ is large, and the log likelihood ratio is easy to compute. Examples include testing a special structure in a contingency table, testing the significance of certain explanatory variables in generalized linear models, etc. When the sample size is large, and consequently the MLE $\widehat{\psi}$ and $\widehat{\theta}$ approximately follow normal distributions, the resulting test statistic and the p-value are essentially the same as those in Li et al. (1991b). However, although rare with a large sample, $\widehat{r}_{\mathrm{L}}$ may be negative in some cases in practice.

Here, we discuss conditions for $\widehat{r}_{\mathrm{L}}$ to be non-negative in exponential families. Let $T(X)$ be the sufficient statistic, and $\zeta = \zeta(\psi)$ be the natural vector of parameters. Then the complete-data likelihood can be written as

$$f(X|\psi) = h(X) \exp\left(\zeta^t T(X) - A(\psi)\right),$$

and the log likelihood ratio is

$$\log \frac{f(X|\psi)}{f(X|\psi^0)} = [\zeta(\psi) - \zeta(\psi^0)]^t T(X) - [A(\psi) - A(\psi^0)].$$

If the MLE $\widehat{\psi}$ is a linear function of $T(X)$, without loss of generality, $\widehat{\psi}(X) = T(X)$, then $\bar{d}_{\mathrm{L}}$ and $\bar{d}'_m$ can be written as functions of $\{(\widehat{\psi}^0_{*l}, \widehat{\psi}_{*l}); l = 1, \cdots, m\}$. More specifically, $\bar{d}_{\mathrm{L}} = 2\Lambda(\bar{\psi}^0_m, \bar{\psi}_m)$, and $\bar{d}'_m = 2m^{-1}\sum_{l=1}^m \Lambda(\widehat{\psi}^0_{*l}, \widehat{\psi}_{*l})$, where Λ is defined as

$$\Lambda(\psi^0, \psi) = [\zeta(\psi) - \zeta(\psi^0)]^t\psi - [A(\psi) - A(\psi^0)].$$

In addition to the linearity of $T(X)$ and $\widehat{\psi}$, if Λ is a convex function, then we have

$$\bar{d}'_m = \frac{2}{m}\sum_{l=1}^m \Lambda(\widehat{\psi}^0_{*l}, \widehat{\psi}_{*l}) \geqslant 2\Lambda(\bar{\psi}^0_m, \bar{\psi}_m) = \bar{d}_{\mathrm{L}},$$

and thus $\widehat{r}_{\mathrm{L}}$ is guaranteed to be nonnegative.

The contingency table example in Meng and Rubin (1992) satisfies these two conditions, and thus $\widehat{r}_{\mathrm{L}} \geqslant 0$. Let π be the array of probabilities for the table. We are interested in testing a specific structure in π, such as conditional independence, against the saturated model, i.e., no structure in π. Let $\widehat{\pi}^0$ and $\widehat{\pi}$ be the MLE of π under the null and alternative models, c be the index of a cell in the table, x_c be the count in that cell, and $n = \sum_c x_c$ be the total counts in the table. Then, the log likelihood ratio is

$$d'(\widehat{\pi}^0, \widehat{\pi}|X) = 2\sum_c x_c \left[\log(\widehat{\pi}_c) - \log(\widehat{\pi}^0_c)\right],$$

under the constrain that $\sum_c \widehat{\pi}^0_c = \sum_c \widehat{\pi}_c = 1$. Under the saturated model, $\widehat{\pi}_c = x_c/n$. So the sufficient statistic of the model and the MLE under the alternative model have a linear relationship, as long as n is fixed in all the completed datasets. The function

$$\Lambda(\pi^0, \pi) = n\sum_c \pi_c(\log(\pi_c) - \log(\pi^0_c))$$

under the constrains that $\sum_c \pi_c^0 = \sum_c \pi_c = 1$, $\pi_c \geqslant 0$, and $\pi_c^0 \geqslant 0$, has a non-negative definite hessian matrix, so it is a convex function, so in this case $\widehat{r}_{\mathrm{L}} \geqslant 0$.

3.4.3 Estimating $\widehat{D}_m$ based on $\mathcal{S}_{\mathbf{d}}$ and d_{full}: Xie (2011)

Xie (2011) proposed another method to estimate $d(\bar{\theta}_m, \bar{U}_m)$, given we have $\mathcal{S}_{\mathrm{d}}$, the m completed datasets, and the computer code to calculate the Wald χ^2 statistic. Treating the m completed datasets as independent data, we can apply the computer code to the combined dataset, $X_{\mathrm{full}} = (X_{*1}, \cdots, X_{*\mathrm{m}})$, and obtain the Wald χ^2 statistic, d_{full}. Asymptotically,

$$d_{\mathrm{full}} = (\widehat{\theta}_{\mathrm{full}} - \theta_0)^t U_{\mathrm{full}}^{-1} (\widehat{\theta}_{\mathrm{full}} - \theta_0),$$

where $\widehat{\theta}_{\mathrm{full}} = \widehat{\theta}(X_{\mathrm{full}}) \approx \bar{\theta}_m$, and $U_{\mathrm{full}} = U(X_{\mathrm{full}}) \approx m^{-1}\bar{U}_m$, so $d(\bar{\theta}_m, \bar{U}_m) \approx m^{-1} d_{\mathrm{full}}$. Let $\widehat{r}_{\mathrm{full}}$ and D_{full} be the estimates of $\bar{r}_m$ and D_m, obtained by replacing $d(\bar{\theta}_m, \bar{U}_m)$ in (3.21) with $m^{-1} d_{\mathrm{full}}$, and $\widehat{r}_{\mathrm{L}}$ in (3.22) by $\widehat{r}_{\mathrm{full}}$, respectively. Then, the p-value is computed by referring D_{full} to distribution $F_{k,w(\widehat{r}_{\mathrm{full}})}$, where $w(\cdot)$ is defined in (3.10).

3.4.4 Comparison

When the standard complete-data analysis does not produce the variance-covariance matrix, U_{*l}, the test statistic D_m in (3.6) can not be computed. Fortunately, D_m can be approximated by a function of $\bar{d}_m$ and $\bar{r}_m$. We typically can obtain the Wald χ^2 statistic, d_{*l}, from each of the m complete-data inferences. So an estimate of $\bar{r}_m$ leads to an estimate of D_m, which is referred to an $F_{k,\widehat{w}}$ distribution, derived under the assumption of EFMI. Currently, we do not have an estimate of $\sum_{i=1}^{k} r_i^2$ to allow for the use of $F_{\widehat{\rho}_m,\widehat{\eta}_m}$.

The procedures by Meng and Rubin (1992) and Xie (2011) have interesting connections. First, they are asymptotically equivalent to $\mathcal{T}_{\text{LRR}}^m$, so they are asymptotically superior to the procedure by Li et al. (1991a). But keep in mind that they require resources other than $\mathcal{S}_{\text{d}}$, and involve more computation. Second, the two procedures are identical in certain situations. Xie (2011) treated the m datasets as independent data, so in an exponential family, the likelihood of X_{full} is

$$f(X_{\text{full}}|\psi) = \prod_{l=1}^{m} f(X_{*l}|\psi) = \left(\prod_{l=1}^{m} h(X_{*l})\right) \exp\left[m\left(\zeta^t \frac{T(X_{\text{full}})}{m} - A(\psi)\right)\right], \qquad (3.23)$$

where $T(X_{\text{full}}) = \sum_{l=1}^{m} T(X_{*l})$ is the sufficient statistic. If the MLE of ψ under H_0 and H_a are linear functions of the sufficient statistic, then $\widehat{\psi}_{\text{full}}^0 = \bar{\psi}_m^0$, $\widehat{\psi}_{\text{full}} = \bar{\psi}_m$, and

$$d'(\widehat{\psi}_{\text{full}}^0, \widehat{\psi}_{\text{full}}|X_{\text{full}}) = 2\log\left(\frac{f(X_{\text{full}}|\widehat{\psi}_{\text{full}})}{f(X_{\text{full}}|\widehat{\psi}_{\text{full}}^0)}\right) = 2\sum_{l=1}^{m} \log\left(\frac{f(X_{*l}|\bar{\psi}_m)}{f(X_{*l}|\bar{\psi}_m^0)}\right) = m\bar{d}_{\text{L}},$$

where $\widehat{\psi}_{\text{full}}^0$ and $\widehat{\psi}_{\text{full}}$ are the estimates of ψ that maximize (3.23) under H_0 and H_a, respectively. In addition to the linearity condition, if the complete-data analysis substitute the log likelihood ratio for the Wald χ^2 statistic, then $d_{\text{full}} = d'(\widehat{\psi}_{\text{full}}^0, \widehat{\psi}_{\text{full}}|X_{\text{full}}) = m\bar{d}_{\text{L}}$, and consequently, the procedures by Meng and Rubin (1992) and Xie (2011) lead to the exact same results.

The procedure by Xie (2011) has several advantages over the likelihood-ratio based procedures by Meng and Rubin (1992). First, the computer code for computing the Wald χ^2 statistic is usually readily available, whereas the code for computing likelihood ratio may require additional effort to achieve. Second, the Wald χ^2 statistic is invariant to the transformation of parameters, but $\bar{d}_{\mathrm{L}}$ can be sensitive to the transformation. Third, the MLE $\widehat{\psi}^0_{*l}$ and $\widehat{\psi}_{*l}$ are not needed. However, the computation of the procedure by Xie (2011) can be burdensome, since the complete-data analysis needs to be applied to the gigantic dataset X_{full}.

3.5 Concluding Remarks

In this chapter, we have studied and compared procedures for forming the repeated-imputation inference based on multiply-imputed data. Due to the lack of degrees of freedom to estimate individual eigenvalue of $B_t U_t^{-1}$, the test statistic D_m in (3.6) is used in practice, and it is referred to an F distribution to compute the final p-value. The procedures can be classified according to whether or not the moments estimates, i.e., point estimate and the associated variance-covariance estimate, from each of the m completed dataset are available.

When the set of moments estimates are available, the statistic D_m can be computed, so different reference distributions result in different procedures. Li et al. (1991b) derived the distribution $F_{k,w(\bar{r}_m)}$ under the assumption of equal fraction of missing information (EFMI). So when this assumption approximately holds, the procedure works well in terms of the actual levels and powers, but the actual levels can be considerably higher than the nominal levels when the assumption is severely violated. Xie (2011) developed the distribution $F_{\widehat{\rho}_m,\infty}$ without assuming EFMI, and are especially suited for high dimension, large fractions of missing information, and large variability of the fractions. We propose a modification to the procedure by Xie

(2011) to account for the additional uncertainty in the denominator of D_m, and suggest criteria for deciding the denominator degrees of freedom for the F distribution.

We have conducted a thorough comparison of the actual levels and powers of these procedures under different conditions and for finite m and when $m \to \infty$. When $m \to \infty$, the reference proposed by Xie (2011), with or without our modification, have the same first two moments with the true distribution of D_m under the null hypothesis, so its behavior are nearly identical to using the actual null distribution of D_m. For finite m, among all procedures, the actual levels of the modified procedure we propose are closest to the nominal levels, under nearly all the conditions. For fixed m, the ROC curves of these procedures are nearly identical, meaning higher powers (true positive rates) come with the price of higher actual levels (false positive rates). So in practice, we recommend using the modified distribution, $F_{\widehat{\rho}_m, \widehat{\eta}_m}$, as the reference distribution. Overall, these calibrated F distributions are reasonably close to the exact distribution of D_m under the null hypothesis, so the actual levels are quite close to the nominal levels, and the powers are quite close to these using the exact null distribution of D_m. When the assumption of EFMI does not hold, these procedures all lose some powers due to using D_m instead of D_{obs} as the test statistic.

When the moments estimates are not available, both the statistic D_m and its reference distribution need to be be estimated. The procedure by Li et al. (1991a) only uses the m scalar Wald χ^2 statistics to estimate these quantities, and thus suffers from a severe loss of power, compared with methods based on moments estimates. Meng and Rubin (1992) proposed to use log likelihood ratio and the MLE of the model parameters to estimate D_m, and Xie (2011) suggest obtaining the Wald χ^2 statistic from the complete-data inference applied to the combined data X_{full} to estimate D_m. Both procedures are asymptotic equivalent to the procedure by Li et al. (1991b).

We have also discussed a practical issue in the likelihood-ratio based procedure by Meng and Rubin (1992), i.e., the estimate of $\bar{r}_m$ can be negative. We propose a sufficient condition that can guarantee the non-negativeness of the estimate. A direction for future research is to investigate transformations that can be applied to the model parameters to ensure the estimate of $\bar{r}_m$ is non-negative.

The existing procedures for combining the Wald χ^2 statistics all implicitly assume EFMI, since the reference distributions are some modified versions of $F_{k,w(\bar{r}_m)}$, derived under that assumption. Another direction for research is to explore other resources that can be used to estimate the test statistic or its reference distribution, and to develop procedures without the assumption of EFMI.

Appendix A

Appendix

A.1 Proof of Theorem 1 in Chapter 1

In this appendix, we derive the inequality in Theorem 1. Due to the convexity of the function f, we have

$$\phi(\omega) f\left(\int \frac{p(\mathcal{H}_\theta(\omega))}{\phi_{\text{mix}}(\mathcal{H}_\theta(\omega))}\pi(\theta)\mathbf{u}(\mathrm{d}\theta)\right) \leqslant \int \phi(\omega) f\left(\frac{p(\mathcal{H}_\theta(\omega))}{\phi_{\text{mix}}(\mathcal{H}_\theta(\omega))}\right)\pi(\theta)\mathbf{u}(\mathrm{d}\theta). \quad \text{(A.1)}$$

By the expression of $\widetilde{p}$, the term on the left of the equal sign is equal to $\phi(\omega) f\left(\frac{\widetilde{p}(\omega)}{\phi(\omega)}\right)$. Let $D_f(\widetilde{p}, \phi)$ be the f-divergence between $\widetilde{p}$ and ϕ, and $D_f(p, \phi_{\text{mix}})$ be the f-divergence between p and ϕ_{mix}. Then,

$$D_f(\widetilde{p}, \phi) \leqslant \int \int \phi(\omega) f\left(\frac{p(\mathcal{H}_\theta(\omega))}{\phi_{\text{mix}}(\mathcal{H}_\theta(\omega))}\right)\pi(\theta)\mathbf{u}(\mathrm{d}\theta)\mathrm{d}\omega \triangleq D_f^*(p, \phi_{\text{mix}}). \quad \text{(A.2)}$$

Below we show $D_f^*(p, \phi_{\text{mix}}) = D_f(p, \phi_{\text{mix}})$, thus the inequality in Theorem 1 holds.

Since $\mathcal{H}_\theta$ and its inverse function, $\mathcal{F}_\theta$, are both monotonic and differentiable functions, we can substitute (ω, θ) with $(\mathcal{F}_\theta(\widetilde{\omega}), \theta)$ in $D_f^*(p, \phi_{\text{mix}})$, and obtain

$$D_f^*(p, \phi_{\text{mix}}) = \int \left[\int \phi(\mathcal{F}_\theta(\widetilde{\omega})) / \left| \mathcal{H}_\theta'(\mathcal{F}_\theta(\widetilde{\omega})) \right| \pi(\theta) \mathbf{u}(\mathrm{d}\theta) \right] f \left(\frac{p(\widetilde{\omega})}{\phi_{\text{mix}}(\widetilde{\omega})} \right) \mathrm{d}\widetilde{\omega}. \tag{A.3}$$

The expression in the square brackets is equal to $\phi_{\text{mix}}(\widetilde{\omega})$, so

$$D_f^*(p, \phi_{\text{mix}}) = \int \phi_{\text{mix}}(\widetilde{\omega}) f \left(\frac{p(\widetilde{\omega})}{\phi_{\text{mix}}(\widetilde{\omega})} \right) \mathrm{d}\widetilde{\omega} = D_f(p, \phi_{\text{mix}}). \tag{A.4}$$

A.2 Non-negligible Variance When $L \to \infty$

In this appendix, we show that if p is not in the family specified in (1.31), the term $\mathcal{V}_\alpha(\widetilde{p}, \phi)$ remains positive and non-negligible when $L \to \infty$.

For fixed α, p, and ϕ, we define a non-negative continuous function of $\boldsymbol{\zeta}$, $f_\alpha(\boldsymbol{\zeta}) = \mathcal{V}_\alpha(\widetilde{p}, \phi)$. Assume the sequence of random variable $\left\{ \widetilde{\boldsymbol{\zeta}}_L; L = 1, \cdots, \infty \right\}$ is defined in the probability space $(\Omega, \mathcal{F}, P_0)$. Then, $f_{\alpha,L} = f_\alpha(\widetilde{\boldsymbol{\zeta}}_L)$, for $L = 1, 2, \cdots$, is a sequence of non-negative random variables in $(\Omega, \mathcal{F}, P_0)$, and

$$E_L \left[\mathcal{V}_\alpha(\widetilde{p}, \phi) \right] = \int_\Omega f_{\alpha,L}(\omega) dP_0(\omega).$$

Let f_α^* be the inferior limit of the sequence $\{f_{\alpha,L}\}_{L=1}^\infty$, i.e., for any $\omega \in \Omega$,

$$f_\alpha^*(\omega) = \liminf_{L \to \infty} f_{\alpha,L}(\omega).$$

Then, $f^*_\alpha \geqslant 0$ almost surely. Fatou's lemma implies

$$\liminf_{L\to\infty} \int_\Omega f_{\alpha,L}(\omega) dP_0(\omega) \geqslant \int_\Omega f^*_\alpha(\omega) dP_0(\omega). \tag{A.5}$$

The integral of f^*_α is greater than zero, meaning that using a gigantic dataset to estimate $\boldsymbol{\zeta}$ will not remove the discrepancy between p and the calibrated ϕ_{mix}, if p is not a Gaussian mixture distribution exactly as specified in (1.31).

A.3 Bayesian Inference

In this appendix, we briefly introduce the basic concepts and computational methods needed for our Bayesian inference. Readers looking for a more complete introduction should read Gelman et al. (2014).

Assume the vector of observed data $\boldsymbol{D} = (x_1, x_2, \cdots, x_n)$ follow a probability distribution that is parametrized by a vector of parameters, Θ. We wish to use $\boldsymbol{D}$ to estimate Θ. The likelihood of Θ is defined as the probability of $\boldsymbol{D}$ given Θ, that is, $L(\Theta; \boldsymbol{D}) = P(\boldsymbol{D}\big|\Theta)$. Information on Θ that is available before $\boldsymbol{D}$ is observed may be summarized by the prior distribution of Θ, denoted as $P(\Theta)$. Bayes' Theorem allows us to calculate the posterior distribution of Θ, $P(\Theta\big|\boldsymbol{D})$,

$$P(\Theta\big|\boldsymbol{D}) = \frac{P(\Theta)P(\boldsymbol{D}\big|\Theta)}{P(\boldsymbol{D})}.$$

The posterior distribution combines the prior information with new information contained in $\boldsymbol{D}$, and is a complete summary of information as to likely values of Θ.

If no prior knowledge is available, a “non-informative” prior distribution can be used. For instance, a flat prior distribution from $-\infty$ to $+\infty$ is often used as a non-informative prior for the mean parameter of a distribution. This flat prior distribution is not a proper distribution, in the sense that it does not integrate to 1 or any finite positive value. In Bayesian inference, this type of improper prior distribution is legitimate so long as the posterior distribution is proper.

Summaries of the posterior distribution can be used as fitted values (e.g. the posterior mode, mean, or median) and error bars (e.g. a posterior probability interval) of the parameters. (Asymmetric) error bars can be computed by finding an interval of parameter values with a given posterior probability. The highest probability density (HPD) interval[1], for example, is the shortest interval with a given posterior probability. Figure A.1 shows the lower and upper bounds of the 95% HPD interval of a gamma distribution,[2] marked by two vertical dotted lines. The areas under the curve between the bounds is 0.95.

When analytical calculation of the Bayesian estimators is infeasible, we often resort to Monte Carlo methods to approximate them. For example, the posterior mean of a function, f, of Θ can be approximated by

$$\mathrm{E}\left(f(\Theta)\big|\boldsymbol{D}\right) = \int f(\Theta) P\left(\Theta\big|\boldsymbol{D}\right) \mathrm{d}\Theta \approx \frac{1}{m}\sum_{i=1}^{m} f(\Theta^{(i)}),$$

where $\Theta^{(1)}, \cdots, \Theta^{(m)}$ is a sample of size m from $P(\Theta\big|\boldsymbol{D})$. The $1-\alpha$ ET interval is

[1] The $1-\alpha$ HPD interval of a continuous random variable can be obtained by moving down a horizontal line from the top of the probability density function until the area under the density curve between the intersections of the horizontal line and the curve is $1-\alpha$.

[2] The probability density function of Gamma(α, β) is $P\left(x\right) = \beta^{\alpha} e^{-\beta x} x^{\alpha - 1}/\Gamma(\alpha)$ for $x > 0$.

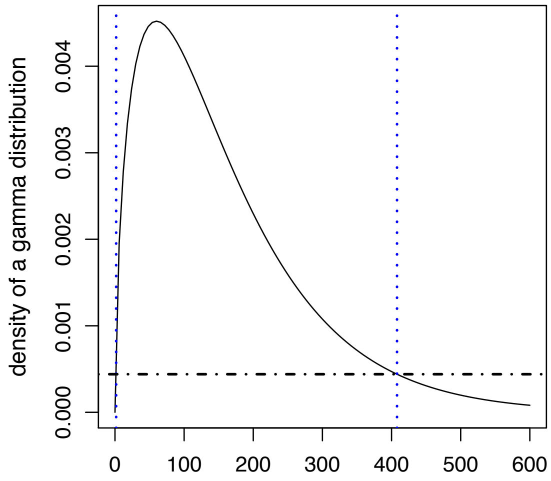

Figure A.1: The 95% HPD interval (the bounds are marked by the blue dotted lines) and the 95% ET interval (marked by the red dashed lines) of a gamma distribution.

estimated by the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of the sample, and the $1-\alpha$ HPD interval can be estimated by the shortest interval that contains a proportion of the sample equal to $1-\alpha$. See Park et al. (2006) for more details on how to approximate Bayesian point and interval estimators from a Monte Carlo sample.

Two popular algorithms to obtain posterior samples are Gibbs sampling (Geman and Geman, 1984) and Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). Both algorithms are Markov Chain Monte Carlo (MCMC) methods that provide a sequence of draws from a Markov Chain, the distribution of which converges to the posterior distribution as the length of the sequence increases. Gibbs sampling obtains observations from a joint probability distribution of random variables by sequentially drawing from the conditional distribution of each parameter while fixing the others at their current values. Metropolis-Hastings algorithm uses a proposal distribution and an accept-reject rule that again ensures the sequence converges to the posterior distribution. Some good reviews of the sampling algorithms can be found in the statistical literature, such as Chib and Greenberg (1995), Casella and George (1992) and Smith and Roberts (1993), or in the astrophysics literature, such as Van Dyk et al. (2001) and Xu et al. (2014).

A.4 Sampling Algorithm of the Bayesian Model

We provide a detailed description of the algorithm to make the MCMC draws from the posterior distribution of the parameters

$$P(\mu,\theta,\pi_d,\boldsymbol{\mathcal{B}},\boldsymbol{\lambda},\boldsymbol{\xi}\big|\boldsymbol{X},\boldsymbol{Y}) \propto P(\mu,\theta)P(\pi_d)P(\boldsymbol{\lambda}\big|\mu,\theta,\pi_d)P(\boldsymbol{\xi})P(\boldsymbol{\mathcal{B}}\big|\boldsymbol{\xi})P(\boldsymbol{X}\big|\boldsymbol{\xi})P(\boldsymbol{Y}\big|\boldsymbol{\lambda},\boldsymbol{\mathcal{B}}), \tag{A.6}$$

where $P(\mu,\theta)$ is defined in (2.11), $P(\pi_d)=1$, and $P(\boldsymbol{\lambda}\big|\mu,\theta,\pi_d)$ is defined in (2.9). Let $\mathcal{R}_k$ be the set of segments in region k, for $k=1,\cdots,K$. Let $\mathcal{R}$ be the union of all the segments in the data. Then,

$$\begin{aligned} P(\boldsymbol{\xi})P(\boldsymbol{\mathcal{B}}\big|\boldsymbol{\xi})P(\boldsymbol{X}\big|\boldsymbol{\xi}) &\propto \prod_{s\in\mathcal{R}} (a_s\mathcal{T})^{\mathcal{B}_s} / (\mathcal{B}_s!) \\ &\times \prod_{k=1}^{K} \exp\left[-\left(\beta_0 + A_k\mathcal{T} + \sum_{s\in\mathcal{R}_k} a_s\mathcal{T}\right)\xi_k + \left(\alpha_0 + X_k + \sum_{s\in\mathcal{R}_k}\mathcal{B}_s - 1\right)\log(\xi_k)\right] \end{aligned}$$

where $\alpha_0 = 10^{-6}$ and $\beta_0 = 1.1$ are constants in the prior distribution of ξ. The term $P(\boldsymbol{Y}\big|\boldsymbol{\lambda},\boldsymbol{\mathcal{B}})$ is

$$\prod_{s\in\mathcal{R}} \exp\left[-\sum_{i\in s} r_{s,i}e_s\lambda_i\mathcal{T} + (Y_s - \mathcal{B}_s)\log\left(\sum_{i\in s} r_{s,i}e_s\lambda_i\mathcal{T}\right) - \log((Y_s-\mathcal{B}_s)!)\right].$$

The Gibbs sampler makes draws from the joint posterior distribution by alternatively sampling a single or a group of parameters from the posterior distribution of these parameters while fixing other parameters at their current values. In the initial step, we pick an initial guess of $(\mu^{(0)}, \theta^{(0)}, \pi_d^{(0)}, \boldsymbol{\xi}^{(0)})$, and simulate $\boldsymbol{\lambda}^{(0)}$ according to the zero-inflated gamma distribution with parameters $(\mu^{(0)}, \theta^{(0)}, \pi_d^{(0)})$. In the $(t+1)$-th iteration, we alternatively make draws according to the following steps:

1. Draw $\pi_d^{(t+1)}$ from Beta$(n_d^{(t)} + 1, n - n_d^{(t)} + 1)$, where $n_d^{(t)}$ is the number of X-ray dark sources (i.e., number of $\lambda_i^{(t)} = 0$) in the t-th iteration.

2. For $k = 1, \cdots, K$, for $s \in \mathcal{R}_k$, draw $\mathcal{B}_s^{(t+1)}$ from the binomial distribution Binomial $\left(Y_s, \varpi_s^{(t)}\right)$, where

$$\varpi_s^{(t)} = \frac{a_s \xi_k^{(t)}}{a_s \xi_k^{(t)} + \sum_{i \in s} r_{s,i} e_s \lambda_i^{(t)}}.$$

3. For $k = 1, \cdots, K$, draw $\xi_k^{(t+1)}$ from Gamma $\left(\alpha_k^{(t+1)}, \beta_k\right)$, where [3]

$$\alpha_k^{(t+1)} = \alpha_0 + X_k + \sum_{s \in \mathcal{R}_k} \mathcal{B}_s^{(t+1)}, \quad \beta_k = \beta_0 + A_k \mathcal{T} + \sum_{s \in \mathcal{R}_k} a_s \mathcal{T}.$$

4. Draw $(\mu^{(t+1)}, \theta^{(t+1)})$. For simplicity, in the expression below, $\alpha = \mu^2/\theta$ and $\beta = \mu/\theta$. Other parameters being fixed, the posterior distribution of μ, θ is

[3] To make it easier to describe the algorithm, we use the conventional parametrization of the gamma distribution, i.e., Gamma(α, β), which is equivalent to Gamma $[\mu, \theta]$, where $\mu = \alpha/\beta$ and $\theta = \alpha/\beta^2$.

proportional to

$$Q^{(t)}(\mu,\theta) = P\left(\mu,\theta\right)\left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^{n-n_d^{(t)}} \exp\left(-\sum_{i=1}^{n}\lambda_i^{(t)} + (\alpha-1)\sum_{\lambda_i^{(t)}>0}\log\left(\lambda_i^{(t)}\right)\right),$$

The Metropolis-Hastings algorithm is used to draw $(\mu^{(t+1)},\theta^{(t+1)})$ jointly from the above distribution. We use a normal distribution $\mathcal{N}(\tau^{(t)},\Sigma^{(t)})$ as the proposal density, where $\tau^{(t)}$ is the mode of $Q^{(t)}$ and $\Sigma^{(t)}$ is the inverse of its Hessian matrix at $\tau^{(t)}$. A new draw (μ',θ') from $\mathcal{N}(\tau^{(t)},\Sigma^{(t)})$ is proposed. We accept (μ',θ') as $\left(\mu^{(t+1)},\theta^{(t+1)}\right)$ with probability γ, where

$$\gamma = \min\left(1, \frac{Q^{(t)}(\mu^{(t)},\theta^{(t)})}{Q^{(t)}(\mu',\theta')}\frac{\phi^{(t)}(\mu^{(t)},\theta^{(t)})}{\phi^{(t)}(\mu',\theta')}\right), \tag{A.7}$$

where $\phi^{(t)}$ is the density of $\mathcal{N}(\tau^{(t)},\Sigma^{(t)})$. With probability $1-\gamma$, $\left(\mu^{(t+1)},\theta^{(t+1)}\right)$ is set to be the current value $\left(\mu^{(t)},\theta^{(t)}\right)$. In general, the proposal density $\phi^{(t)}$ approximates the target density $Q^{(t)}$ quite well, and thus the acceptance ratio is quite high. For example, the acceptance rate of the algorithm applied to the real data in Section 2.4 is 0.7.

5. Let $\mathcal{C}_i$ be the set of segments that constitute source region i, i.e., $\mathcal{C}_i = \bigcup_{i\in s}\{s\}$. Let $\mathcal{S}_i^{(t+1)} = \sum_{s\in\mathcal{C}_i}\left(Y_s - \mathcal{B}_s^{(t+1)}\right)$, $\widetilde{\alpha}^{(t+1)} = (\mu^{(t+1)})^2/\theta^{(t+1)}$, and $\widetilde{\beta}^{(t+1)} = \mu^{(t+1)}/\theta^{(t+1)}$. For $i = 1,\cdots,n$,
   - If $\mathcal{S}_i^{(t+1)} > 0$, we generate $\lambda_i^{(t+1)}$ from Gamma $\left(\widetilde{\alpha}_i^{(t+1)},\widetilde{\beta}_i^{(t+1)}\right)$, where $\widetilde{\alpha}_i^{(t+1)} = \widetilde{\alpha}^{(t+1)} + \mathcal{S}_i^{(t+1)}$ and $\widetilde{\beta}_i^{(t+1)} = \widetilde{\beta}^{(t+1)} + \sum_{s\in\mathcal{C}_i} r_{s,i}e_s\mathcal{T}$.
   - If $\mathcal{S}_i^{(t+1)} = 0$, we draw $\lambda_i^{(t+1)}$ from the zero-inflated gamma distribution,

i.e.,

$$\lambda_i^{(t+1)} \begin{cases} = 0 & \text{with probability } \widetilde{\pi}_*^{(t+1)}, \\ \sim \text{Gamma}\left(\widetilde{\alpha}_i^{(t+1)}, \widetilde{\beta}_i^{(t+1)}\right) & \text{with probability } 1 - \widetilde{\pi}_*^{(t+1)}, \end{cases} \tag{A.8}$$

where

$$\widetilde{\pi}_*^{(t+1)} = \frac{\pi_d^{(t+1)}}{\pi_d^{(t+1)} + (1 - \pi_d^{(t+1)}) \left(\beta^{(t+1)}/\widetilde{\beta}^{(t+1)}\right)^{\alpha^{(t+1)}}}.$$

The 5 steps are repeated to generate the Monte Carlo sequence that converges to the joint posterior distribution. We usually discard the first T_0 draws because they may not follow the right distribution, especially if the initial values of the parameters are far from the center of the distribution. This step is called burn-in. Our sampling algorithm is very fast, and we set T_0 to be 20000.

Trace plots of the Monte Carlo draws of μ, θ, and π_d from the posterior distribution appears to be convergent. The leg-1 autocorrelation of these parameters are 0.82, 0.51, and 0.95, and we made a total of 150,000 draws after the initial burn-in.

A.5 Computation of the Test Statistic

Here, we provide the analytical expression of the simplified likelihood function $\widetilde{L}_1$, which is defined in (2.24). Under the simplified alternative model, for $i \in \mathbb{S}$, if source i is in region k, then $Y_i^{rep}|\lambda_i \overset{\text{indep}}{\sim} \text{Poisson}\left(r_i e_i \lambda_i \mathcal{T} + \widehat{\xi}_i^*\right)$, where $\widehat{\xi}_i^* = a_i \widehat{\xi}_k \mathcal{T}$. A priori,

λ_i independently follows the zero-inflated gamma distribution. So the likelihood is

$$\widetilde{L}_1\left(\mu, \theta, \pi_d, \widehat{\boldsymbol{\xi}}; \boldsymbol{D}^{rep}\right) = \prod_{i \in \mathbb{S}} P(Y_i^{rep} | \mu, \theta, \pi_d, \widehat{\boldsymbol{\xi}}),$$

where

$$\begin{aligned} P(Y_i^{rep} | \mu, \theta, \pi_d, \widehat{\boldsymbol{\xi}}) &= \int P(Y_i^{rep} | \lambda_i, \widehat{\boldsymbol{\xi}}) P(\lambda_i | \mu, \theta, \pi_d) \mathrm{d}\lambda_i \\ = \mathbb{C}_i &\left[\pi_d + (1 - \pi_d) \sum_{j=0}^{Y_i^{rep}} \begin{pmatrix} Y_i^{rep} \\ \\ j \end{pmatrix} (\mathbb{W}_i)^j \, \frac{\Gamma(\alpha + j)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(r_i e_i \boldsymbol{\mathcal{T}} + \beta)^{\alpha + j}} \right], \end{aligned}$$

with $\mathbb{W}_i = r_i e_i \boldsymbol{\mathcal{T}} / \widehat{\xi}_i^*$ and $\mathbb{C}_i = \exp\left[-\widehat{\xi}_i^* + Y_i^{rep} \log\left(\widehat{\xi}_i^*\right)\right] / (Y_i^{rep}!)$. Note that in the above expression, $\alpha = \mu^2/\theta$, and $\beta = \mu/\theta$. The likelihood function $\widetilde{L}_0\left(\mu, \theta; \boldsymbol{D}^{rep}\right) = \prod_{i \in \mathbb{S}} P(Y_i^{rep} | \mu, \theta, \pi_d = 0)$.

Although the expressions of $\widetilde{L}_0$ and $\widetilde{L}_1$ are very complex, they are functions of two and three parameters. In our simulation, we use the function “optimize” implemented in R programming to find the supremes of $\log\left(\widetilde{L}_0\left(\mu, \theta; \boldsymbol{D}^{rep}\right)\right)$ and $\log\left(\widetilde{L}_1\left(\mu, \theta, \pi_d; \boldsymbol{D}^{rep}\right)\right)$. The reason for taking the logarithm of both functions is to reduce computer-precision related errors. The “optimize” function in R programming works reasonably well. When occasionally it produces a negative test statistic (i.e. logarithm of the likelihood ratio), we set the test statistic to be 0.

A.6 Reason for Using A Proper Prior Distribution for (μ, θ)

Any Bayesian model must have a proper posterior distribution, meaning the posterior distribution has to be integrable. We show in this appendix that the prior distribution of (μ, θ) in our model has to be proper (i.e., integrable) to ensure the integrability of the posterior distribution. For simplicity, we assume (i) the dataset only has non-overlapping sources, and (ii) the background is homogenous.

We first compute the posterior distribution of $(\mu, \theta, \pi_d, \boldsymbol{\mathcal{B}})$ by integrating out ξ and λ from the posterior distribution in (A.6),

$$P(\mu, \theta, \pi_d, \boldsymbol{\mathcal{B}} \big| \boldsymbol{D}) = \int P(\mu, \theta, \pi_d, \boldsymbol{\mathcal{B}}, \boldsymbol{\lambda}, \xi \big| \boldsymbol{D}) \mathrm{d}\xi \mathrm{d}\boldsymbol{\lambda} = f(\boldsymbol{\mathcal{B}}; \boldsymbol{D}) g(\mu, \theta, \pi_d, \boldsymbol{\mathcal{B}}; \boldsymbol{D}),$$

where

$$f(\boldsymbol{\mathcal{B}}; \boldsymbol{D}) \propto \frac{\Gamma(\alpha_0 + X + \sum \mathcal{B}_i)}{(\beta_0 + A\mathcal{T} + \sum a_i \mathcal{T})^{\alpha_0 + X + \sum \mathcal{B}_i}} \prod_{i=1}^{n} \frac{a_i^{\mathcal{B}_i} \left(r_i e_i\right)^{Y_i - \mathcal{B}_i}}{\mathcal{B}_i!(Y_i - \mathcal{B}_i)!},$$

and

$$\begin{aligned} g(\mu, \theta, \pi_d, &\boldsymbol{\mathcal{B}}; \boldsymbol{D}) = P(\mu, \theta) \\ &\times \prod_{i=1}^{n} \left(\pi_d I(Y_i = \mathcal{B}_i) + (1 - \pi_d) \frac{(\mu/\theta)^{(\mu^2/\theta)}}{(\mu/\theta + e_i)^{(\mu^2/\theta + Y_i - \mathcal{B}_i)}} \frac{\Gamma(\mu^2/\theta + Y_i - \mathcal{B}_i)}{\Gamma(\mu^2/\theta)} \right). \end{aligned}$$

Since $0 \leqslant \mathcal{B}_i \leqslant Y_i$, $f(\boldsymbol{\mathcal{B}}; \boldsymbol{D})$ can be bounded by two positive constants $\mathbb{C}_1$ and $\mathbb{C}_2$, that is

$$0 < \mathbb{C}_1 \leqslant f(\boldsymbol{\mathcal{B}}; \boldsymbol{D}) \leqslant \mathbb{C}_2 < +\infty.$$

Therefore, the integrability of $P(\mu, \theta, \pi_d, \boldsymbol{\mathcal{B}} \big| \boldsymbol{D})$ depends completely on the integrability of $g(\mu, \theta, \pi_d, \boldsymbol{\mathcal{B}}; \boldsymbol{D})$. The integral of the positive function $g(\mu, \theta, \pi_d, \boldsymbol{\mathcal{B}}; \boldsymbol{D})$ is

$$\sum_{b_1=0}^{Y_1} \cdots \sum_{b_n=0}^{Y_n} \left[\int g\left(\mu, \theta, \pi_d, \mathcal{B} = (b_1, \cdots, b_n); \boldsymbol{D}\right) \mathrm{d}\mu \mathrm{d}\theta \mathrm{d}\pi_d \right],$$

which is great than the single term that corresponds to $\mathcal{B} = (Y_1, \cdots, Y_n)$, which is

$$\int g\left(\mu, \theta, \pi_d, \mathcal{B} = (Y_1, \cdots, Y_n); \boldsymbol{D}\right) \mathrm{d}\mu \mathrm{d}\theta \mathrm{d}\pi_d = \int P(\mu, \theta) \mathrm{d}\mu \mathrm{d}\theta \int \pi_d^n \mathrm{d}\pi_d$$

If the prior on (μ, θ) is improper, i.e., $\int P(\mu, \theta) \mathrm{d}\mu \mathrm{d}\theta = +\infty$, then

$$\int g(\alpha, \beta, \pi_b, \boldsymbol{\mathcal{B}}; \boldsymbol{Y}) \mathrm{d}\mu \mathrm{d}\theta \mathrm{d}\pi_d \geqslant \int P(\mu, \theta) \mathrm{d}\mu \mathrm{d}\theta \int \pi_d^n \mathrm{d}\pi_d = +\infty,$$

which means the posterior distribution if improper. Therefore, a proper prior distribution for (μ, θ) is required to ensure the posterior distribution is proper.

Bibliography

Daniel L Alspach and Harold W Sorenson. Nonlinear bayesian estimation using gaussian sum approximations. *Automatic Control, IEEE Transactions on*, 17(4): 439–448, 1972.

John Barnard and Xiao-Li Meng. Applications of multiple imputation in medical studies: from aids to nhanes. *Statistical Methods in Medical Research*, 8(1):17–36, 1999.

John Barnard, Constantine Frangakis, Jennifer Hill, and Donald B Rubin. School choice in ny city: A bayesian analysis of an imperfect randomized experiment. In *Case studies in Bayesian statistics*, pages 3–97. Springer, 2002.

John Barnard, Constantine E Frangakis, Jennifer L Hill, and Donald B Rubin. Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in new york city. *Journal of the American Statistical Association*, 98(462):299–323, 2003.

Charles H Bennett. Efficient estimation of free energy differences from monte carlo data. *Journal of Computational Physics*, 22(2):245–268, 1976.

Björn Bornkamp. Approximating probability densities by iterated laplace approximations. *Journal of Computational and Graphical Statistics*, 20(3), 2011.

George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

David M Ceperley. Path integrals in the theory of condensed helium. *Reviews of Modern Physics*, 67(2):279, 1995.

Jiahua Chen and Xianming Tan. Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, 100(7):1367–1383, 2009.

Jiahua Chen, Xianming Tan, and Runchu Zhang. Inference for normal mixtures in mean and variance. *Statistica Sinica*, 18(2):443, 2008.

Ming-Hui Chen and Qi-Man Shao. Estimating ratios of normalizing constants for densities with different dimensions. *Statistica Sinica*, 7(3):607–630, 1997.

Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

Neil E Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56(3):463–474, 1969.

PE Freeman, C Graziani, DQ Lamb, TJ Loredo, EE Fenimore, T Murakami, and A Yoshida. Statistical analysis of spectral line candidates in gamma-ray burst grb 870303. *The Astrophysical Journal*, 524(2):753, 1999.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 6(4):733–760, 1996.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, pages 1360–1383, 2008.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

Curtis F Gerald, Patrick O Wheatley, and Fengshan Bai. *Applied numerical analysis*, volume 19903. Addison-Wesley New York, 1989.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

David E Jones, Vinay L Kashyap, and David A van Dyk. Disentangling overlapping astronomical sources using spatial and spectral information. *arXiv preprint arXiv:1411.7447*, 2014.

Jack Kiefer and Jacob Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.

A Kong, P McCullagh, X-L Meng, D Nicolae, and Z Tan. A theory of statistical models for monte carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):585–604, 2003.

Alan B Krueger and Pei Zhu. Another look at the new york city school voucher experiment. *American Behavioral Scientist*, 47(5):658–698, 2004.

Younjeong Lee, Ki Yong Lee, and Joohun Lee. The estimating optimal number of gaussian mixtures based on incremental k-means for speaker identification. *International Journal of Information Technology*, 12(7):13–21, 2006.

Kim-Hung Li, Xiao-Li Meng, Trivellore E Raghunathan, and Donald B Rubin. Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, 1(1):65–92, 1991a.

Kim-Hung Li, Trivellore E Raghunathan, and Donald B Rubin. Large-sample significance levels from multiply imputed data using moment-based statistics and an f reference distribution. *Journal of the American Statistical Association*, 86(416): 1065–1073, 1991b.

Xiao-Li Meng. Posterior predictive p-values. *The Annals of Statistics*, pages 1142–1160, 1994.

Xiao-Li Meng and Donald B Rubin. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79(1):103–111, 1992.

Xiao-Li Meng and Stephen Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586, 2002.

Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 6(4):831–860, 1996.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

Taeyoung Park, Vinay L Kashyap, Aneta Siemiginowska, David A Van Dyk, Andreas Zezas, Craig Heinke, and Bradford J Wargelin. Bayesian estimation of hardness ratios: Modeling and computations. *The Astrophysical Journal*, 652(1):610, 2006.

Rostislav Protassov, David A Van Dyk, Alanna Connors, Vinay L Kashyap, and Aneta Siemiginowska. Statistics, handle with care: Detecting multiple model components with the likelihood ratio test. *The Astrophysical Journal*, 571(1):545, 2002.

Donald B Rubin. Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American statistical association*, volume 1, pages 20–34. American Statistical Association, 1978.

Donald B Rubin. Multiple imputation for nonresponse in surveys (wiley series in probability and statistics). 1987.

Donald B Rubin and Nathaniel Schenker. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81(394):366–374, 1986.

Donald B Rubin and Nathaniel Schenker. Multiple imputation in health-are databases: An overview and some applications. *Statistics in medicine*, 10(4):585–598, 1991.

Donald B Rubin et al. Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984.

Nathaniel Schenker and AH Welsh. Asymptotic results for multiple imputation. *The Annals of Statistics*, pages 1550–1566, 1988.

Nathaniel Schenker, Donald J Treiman, and Lynn Weidman. Analyses of public use decennial census data with multiply imputed industry and occupation codes. *Applied Statistics*, pages 545–556, 1993.

Qi-Man Shao and Joseph G Ibrahim. *Monte Carlo methods in Bayesian computation.* Springer Series in Statistics, New York, 2000.

Adrian FM Smith and Gareth O Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–23, 1993.

Zhiqiang Tan. On a likelihood approach for monte carlo integration. *Journal of the American Statistical Association*, 99(468):1027–1036, 2004.

Zhiqiang Tan. Calibrated path sampling and stepwise bridge sampling. *Journal of Statistical Planning and Inference*, 143(4):675–690, 2013.

Xin Ming Tu, Xiao-Li Meng, and Marcello Pagano. Survival differences and trends in patients with aids in the united states. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 6(10):1150–1156, 1993.

David A Van Dyk, Alanna Connors, Vinay L Kashyap, and Aneta Siemiginowska. Analysis of energy spectra with low photon counts via bayesian posterior simulation. *The Astrophysical Journal*, 548(1):224, 2001.

Arthur F Voter and Jimmie D Doll. Dynamical corrections to transition state theory for multistate systems: Surface self-diffusion in the rare-event regime. *The Journal of chemical physics*, 82(1):80–92, 1985.

Xianchao Xie. *Two Tales of Frequentist Properties of Bayesianly Motivated Methods: Multiple Imputation and Shrinkage Estimation*. Harvard University, 2011.

Jin Xu, David A van Dyk, Vinay L Kashyap, Aneta Siemiginowska, Alanna Connors, Jeremy Drake, Xiao-Li Meng, Pete Ratzlaff, and Yaming Yu. A fully bayesian method for jointly fitting instrumental calibration and x-ray spectral models. *The Astrophysical Journal*, 794(2):97, 2014.