8-15-2017

Statistical Methods for Information Assessment and Data Compatibility with Applications

Daoyuan Shi
University of Connecticut - Storrs, daoyuan.shi@uconn.edu

Statistical Methods for Information Assessment and Data Compatibility with Applications

Daoyuan Shi, Ph.D.

University of Connecticut, 2017

Various modifications have been suggested in the past to extend the Shannon entropy to continuous random variables. We propose a new entropy called the fractional size adjusted entropy and later extend it to the generalized fractional size adjusted entropy. These two proposed entropies always exist and maintain non-negative values. The generalized fractional size adjusted entropy also includes many well-known entropies as its special cases, such as the Shannon entropy and the Rényi entropy. We apply our proposed entropies on various distributions and a phylogenetic example to demonstrate their good performances.

In addition, we propose a partition based measure to quantify the compatibility of two data sets using their respective posterior distributions. It is of great practical importance to compare and combine data from different studies in order to carry out appropriate and more powerful statistical inference. We further propose an information gain measure to quantify the information increase (or decrease) in combining two data sets. The compatibility measure and the information gain measure are well calibrated and efficient computational algorithms are provided for their calculations. We use a benchmark toxicology example, a six cities longitudinal health study and a melanoma clinical trials to illustrate how these measures are useful in combining current data with historical data and missing data.

Statistical Methods for Information Assessment and Data Compatibility with Applications

Daoyuan Shi

B.S., Nankai University, China, 2012

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2017

APPROVAL PAGE

Doctor of Philosophy Dissertation

Statistical Methods for Information Assessment and Data Compatibility with Applications

Presented by

Daoyuan Shi, B.S.

Major Advisor _____
Ming-Hui Chen

Major Advisor _____
Lynn Kuo

Associate Advisor _____
Paul O. Lewis

University of Connecticut

2017

To my family and friends

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deep gratitude to my major advisors Professor Ming-Hui Chen and Professor Lynn Kuo for their continuous support through all these years. They have supported me not only by providing the unique assistantship and collaboration opportunities, but also being incredibly wonderful academic and life mentors. Their wide and deep knowledge and detail oriented attitude inspire and guide me in my research and my life. I would like to thank them for leading me to the research area of statistics, partnering with me on the accumulation of advanced statistical knowledge, and showing me excellent examples of scholars and professionals.

I would like to thank my associate advisor Professor Paul O. Lewis for his great guidance and insightful advice, which are extremely important to me in my research. I benefit a lot from his knowledge in phylogenetics, his novel approaches and suggestions as well as his writing mentoring.

Special thanks to Pfizer Inc. for offering me the research fellow position and providing me the precious industry experience. I would like to especially thank Dr. Huaming Tan and Ronnie Wang for imparting me the pharmaceutical knowledge, and giving me advice on both professional and personal development.

My sincere thanks also go to Dr. Ian Stevenson from Department of Psychological Sciences. His knowledge in neuroscience and kindly support play important roles in my graduate study.

Many thanks are owed to all faculty members and my fellow graduate students. I would also like to thank Ms. Tracy Burke and Ms. Megan Petsa for their helpful administrative assistance.

Lastly, and most importantly, I want to dedicate this dissertation to my parents and my best friend Jordan, who always encourage and support me to pursue the life and career I adore. It is your unconditional loves and cares that lead me to conquer and to achieve.

This thesis is based upon work supported by the National Science Foundation under grant number DEB-1354146 to Paul O. Lewis, Ming-Hui Chen, Lynn Kuo, and Louise Lewis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Information Assessment

First introduced by Shannon (1948), entropy has been used as a key measurement for information. For a discrete random variable X with probability mass function (PMF) $f(X)$ supported at $\{x_1, x_2, ..., x_n\}$, the entropy is defined by

$$H(X) = E[-\log f(X)] = -\sum_i f(x_i) \log f(x_i). \tag{1.1}$$

It is widely used in mathematics, statistics, computer science, physics, neurobiology and electrical engineering. There is a rich literature on information measurement. Hartley (1928) defined a simple measure of uncertainty. If we pick a sample from a finite set A uniformly at random, then the information by Hartley function is $\log_b |A|$, where $|A|$ is the total number of possible choices of that sample. If $b = 2$, it coincides with the Shannon entropy in the case of discrete uniform distribution. Rényi (1961) defined another entropy as

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \big[\sum_i f^\alpha(x_i) \big], \tag{1.2}$$

where $\alpha \geq 0$ and $\alpha \neq 1$. Both Shannon entropy and Rényi entropy have good properties such as non-negativity, additivity and invariant under a one-to-one linear transformation for discrete random variables.

When Shannon entropy is extended to continuous random variables, it does not have the non-negativity anymore. The extended Shannon entropy for continuous random variables is called the differential entropy and is defined as

$$H(X) = E[-\log f(X)] = -\int f(x) \log f(x) dx, \tag{1.3}$$

where $f(X)$ is the probability density function (PDF). For example, the differential entropy for a continuous random variable with a uniform distribution $U(a, b)$ is $\log(b-a)$. When $b-a < 1$, the Shannon differential entropy is negative. A negative information measurement is undesirable. To fix this problem, Jaynes (1968) modified the differential entropy as

$$H(X) = -\int f(x) \log \frac{f(x)}{m(x)} dx,$$

where $m(x)$ was introduced as an invariant measure function. However Jaynes did not provide the exact form for this entropy. Awad and Alawneh (1987) used $m(x) = \sup f(x)$ but their version does not hold the additivity. Kittaneh et al. (2016) used $m(x) = E[f(x)]$. Their version has the additivity and is always non-negative by Jensen's inequality. But when $E[f(x)]$ does not exist, their entropy is not defined.

Rao et al. (2004) proposed another entropy called the cumulative residual entropy (CRE), defined as

$$\epsilon(X) = \int (1 - F(x)) \log(1 - F(x)) dx,$$

where $F(x)$ is the cumulative density function (CDF) of the random variable X. But the cumulative residual entropy lacks the additivity.

In order to adequately define the differential entropy for continuous random variables, we derive a new information measure called the fractional size adjusted entropy and then extend it to a much general version, the generalized fractional size adjusted entropy. It includes almost all the previous versions of entropy and maintain the non-negativity, the additivity and other good properties for both discrete and continuous random variables.

1.2 Data Compatibility

The need to compare and combine data across multiple studies in order to validate and extend results is widely recognized, and only increases as more data become available. For example, the ability to pool data and perform integrative data analysis is particularly important and timely in substance abuse and addiction science (Conway et al., 2014). In phylogenetics, measuring the amount of information in data and detecting conflict among data sets are important to systematists (Kluge and Farris, 1969; Lewis et al., 2016). In meta-analysis, it is important to be able to quantify the extent of heterogeneity among a collection of studies (Higgins and Thompson, 2002; Higgins, 2003).

Unfortunately, there is no clear definition of "compatibility" with respect to comparing data sets. In statistics and information geometry, divergence is used to establish the distance of one probability distribution to another. The divergence is a weaker notion than the distance. Suppose P, Q, R are probability distributions. The distance $d(,)$ has four properties to hold:

- Non-negativity: $d(P, Q) \geq 0$.
- Identity of Indiscernible: $d(P, Q) = 0$ if and only if $P = Q$.
- Symmetry: $d(P, Q) = d(Q, P)$.

- Triangle Inequality: $d(P,R) \leq d(P,Q) + d(Q,R)$.

While for a divergence, it only needs to satisfy the first two properties.

There are many kinds of divergences. For example, the f-divergence (Kullback and Leibler, 1951; Ali and Silvey, 1966; Csiszár and Shields, 2004; Morimoto, 1963), Rényi's divergence (Rényi, 1961), and Bregman divergence (Bregman, 1967). The f-divergence is a family of divergences that are generated through functions $f(u)$, which is convex on $u > 0$ and $f(1) = 0$,

$$D_f(P||Q) = \int q(x) f\Big(\frac{p(x)}{q(x)}\Big) dx.$$

The Kullback-Leibler (KL) divergence is probably the most popular divergence within the f-divergence family. By taking $f(u) = u \log u$, the Kullback-Leibler divergence has the form of

$$D_{KL}(P||Q) = \int p(x) \log \frac{p(x)}{q(x)} dx.$$

Other divergences within the f-divergence family are:

Hellinger distance:

$$H^2(P,Q) = 2 \int \big(\sqrt{p(x)} - \sqrt{q(x)}\big)^2 dx,$$

$$\text{with } f(u) = (\sqrt{u} - 1)^2.$$

Jeffreys divergence:

$$D_J(P||Q) = \int (p(x) - q(x))(\log p(x) - \log q(x)) dx,$$

$$\text{with } f(u) = -\log u + u \log u.$$

Chernoff's α-divergence:

$$D^{\alpha}(P||Q) = \frac{4}{1-\alpha^2}(1 - \int p(x)^{\frac{1-\alpha}{2}} q(x)^{\frac{1+\alpha}{2}} dx),$$
$$\text{with } f(u) = \frac{4}{1-\alpha^2}(1 - u^{\frac{1+\alpha}{2}}).$$

Exponential divergence:

$$D_e(P||Q) = \int p(x)(\log p(x) - \log q(x))^2 dx,$$
$$\text{with } f(u) = (\log u)^2.$$

Kagan's divergence:

$$D_{\chi^2}(P||Q) = \frac{1}{2}\int \frac{(p(x) - q(x))^2}{p(x)} dx,$$
$$\text{with } f(u) = (1-u)^2.$$

(α, β)-product divergence:

$$D_{\alpha,\beta}(P||Q) = \frac{2}{(1-\alpha)(1-\beta)} \int (1 - \Big(\frac{q(x)}{p(x)}\Big)^{\frac{1-\alpha}{2}})(1 - \Big(\frac{q(x)}{p(x)}\Big)^{\frac{1-\beta}{2}})p(x)dx,$$
$$\text{with } f(u) = \frac{2}{(1-\alpha)(1-\beta)}(1 - u^{\frac{1-\alpha}{2}})(1 - u^{\frac{1-\beta}{2}}).$$

Besides the f-divergence, Rényi also defined a divergence for the discrete distributions as

$$D_{\alpha}(P||Q) = \frac{1}{\alpha - 1} \log \sum_i \frac{p(x_i)^{\alpha}}{q(x_i)^{\alpha-1}},$$

where $0 < \alpha < \infty$ and $\alpha \neq 1$.

Bregman divergence is another general divergence family with the form of

$$D_F(P||Q) = F(p(x)) - F(q(x)) - < \bigtriangledown F(q(x)), p(x) - q(x) >,$$

where F is a continuously-differentiable real-valued and strictly convex function defined on a closed convex set, $\bigtriangledown F(q(x)$ is the gradient function of $F(q(x)$, $<>$ means the inner product.

To measure and compare information, Shannon (1948) forms the foundation of information theory which is widely used (Cover and Thomas, 2006). In the Bayesian setting, Lindley (1956) defined the information provided by an experiment as the difference between the posterior differential entropy and the prior differential entropy. In meta-analysis, several measures such as Cochran's Q (Cochran, 1950) and I^2 (Higgins and Thompson, 2002; Higgins, 2003) have been used to quantify "heterogeneity," which describes the variety in effect across studies.

As discussed previously, for continuous distributions, the Shannon differential entropy may take negative values and is not calibrated. Since the Lindley information is based on the differential entropy, it has the same drawbacks. KL divergence is positive, but is difficult to calculate (but see Lefebvre et al. (2010)) and may not exist under certain situations. Moreover, KL divergence is not calibrated. Rényi divergence as well as other divergences in the f-divergence family and Bregman divergence have similar problems. Cochran's Q and I^2 focus on point estimates and fail to incorporate uncertainty of the point estimates.

We propose a new partition-based measure of compatibility that focuses on the posterior distributions of two data sets. We first partition one posterior distribution according to a specified probability vector. For example, for probability vector $\{1/4, 1/4, 1/4, 1/4\}$, the partition subset boundaries would be quartiles. We then use the same subset boundaries to determine the probability vector corresponding to the other posterior distribution.

The difference between the discrete Shannon entropies computed from the two probability vectors is the basis of our compatibility measure.

Our proposed measure always exists and takes a value between 0 and 1. In contrast to the Shannon differential entropy and the Lindley information measure, our proposed measure is always positive and well calibrated. Unlike Cochran's Q and I^2, the proposed measure compares whole distributions and therefore automatically takes the uncertainty into consideration. In addition, we extend the compatibility measure to a partial compatibility measure for the case in which the two data sets share some common parameters. Monte Carlo based computational algorithms are further developed for calculating both measures.

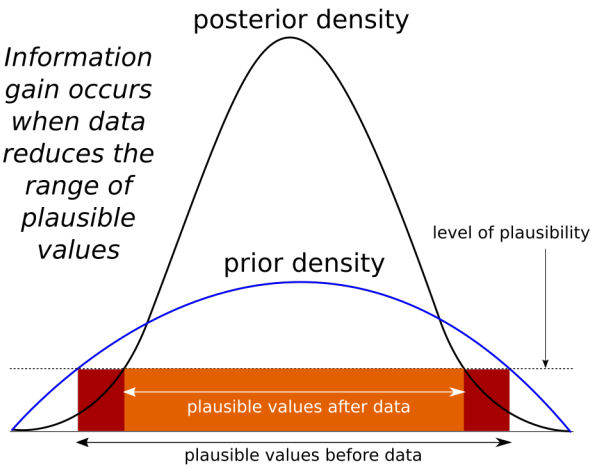
1.3 Information Gain

To quantify how much information is gained or lost by adding new data set B to historical data set A, we propose a new information gain measure. Information is *gained* if A+B has a similar location and a smaller variance than A alone. In this case, the new data B allows some possible outcomes previously considered to be plausible (on the basis of data A alone) to be effectively excluded from consideration. Information is *lost* if A+B has greater variance than A alone, or has a location shift, indicating that the effect of new data B has been to increase, not decrease, the number of plausible outcomes. Our compatibility measure can only determine whether the posterior distributions are similar, but our information gain measure allows us to distinguish information gain versus loss.

In many applied research fields, the investigators have access to previous studies measuring the same response and covariates as the current study. For example, in many cancer and AIDS clinical trails, current studies and previous studies often use similar treatments.

The results from previous studies can be treated as historical data (Ibrahim and Chen, 2000). Is historical data always compatible with the current data? Should we always use the historical data? Depending on the degree of compatibility, different decisions on how much historical data to use should be made. High compatibility allows us to use all of the historical data, whereas partial compatibility argues for using only some aspects of the historical data and low compatibility suggests not using the historical data at all.

Figure 1.1: Graphical Depiction of Information Gain



This is similar to the borrowing idea mentioned in Ibrahim et al. (2015). The measures we proposed provide a two-step procedure to those questions. First, check whether the historical data and current data are compatible with the compatibility measure. If they are, then evaluate how much historical information to borrow with the power prior approach to achieve maximal information gain.

1.4 Dissertation Outline

The rest of the thesis is organized as follows. In Chapter 2, we will introduce the fractional size adjusted entropy and the generalized fractional size adjusted entropy. We

will show their performances under different distributions and apply them to a phylogentic problem. In Chapter 3, we will present a detailed development of the data compatibility, including entropy, partition, compatibility measure, computational algorithm, and partial compatibility concept. Simulations and a toxicity example will be presented to illustrate how to apply this compatibility measure. A new information gain measure is proposed in Chapter 4 in order to quantity how much gain in terms of information by combining two data sets. Examples with clinical trials and missing data are included to show the potential applications of the information gain measure. Extensions and future research directions are given in Chapter 5 and the proofs of theorems are given in Appendix A.

Chapter 2

Information Assessment

2.1 Fractional Size Adjusted Entropy

2.1.1 Definitions and Properties

Definition 2.1.1. Let X be a random variable with PDF $f(x)$ if it is continuous and PMF $f(x)$ if it is discrete. Then the fractional size adjusted entropy of X is defined as followed

$$FSAE(X,k) = -E\Big[\log \frac{f^k(x)}{E[f^k(x)]}\Big],$$

where k can be any value such that $FSAE(X,k)$ exists. We define $0 \log 0 = 0$.

Followings are some properties of the fractional size adjusted entropy.

Property 2.1.2. The fractional size adjusted entropy is non-negative.

Property 2.1.3. The fractional size adjusted entropy is additive.

Property 2.1.4. The fractional size adjusted entropy is invariant under one-to-one linear transformation.

The detailed proofs are given in **Appendix A**.

We can easily extend the fractional size adjusted entropy for the joint distribution and the conditional distribution.

Definition 2.1.5. The fractional size adjusted joint entropy $FSAE(X, Y, k)$ of random variables (X, Y) with a joint distribution $f(x, y)$ is defined as

$$FSAE(X, Y, k) = -E_{X,Y}\Big[\log \frac{f^k(x, y)}{E_{X,Y}[f^k(x, y)]}\Big].$$

Definition 2.1.6. If $(X, Y) \sim f(x, y)$, and $f(x|y)$ is the probability density function of X given Y, then the conditional fractional size adjusted entropy $FSAE(X|Y, k)$ is defined as

$$FSAE(X|Y, k) = E_Y[FSAE(X|Y = y, k)] = E_Y\Big[- E_X\Big[\log \frac{f^k(x|y)}{E_X[f^k(x|y)]}\Big]\Big].$$

The proofs of the non-negativity, the additivity (in X for the conditional distribution) and the invariant under one-to-one linear transformation (in X for the conditional distribution) of the fractional size adjusted entropy for the joint distribution and the conditional distribution are similar to the proofs of **Property 2.1.2**, **Property 2.1.3** and **Property 2.1.4**.

2.1.2 Fractional Size Adjusted Entropy for Certain Distributions

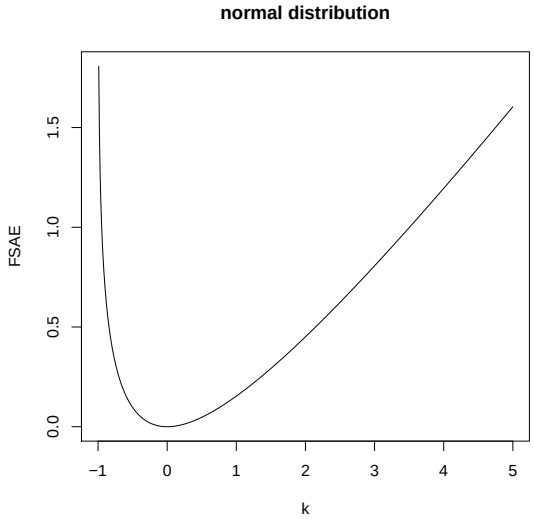
Example 2.1.1. Univariate Normal Distribution:

$$\begin{aligned} X &\sim N(\mu, \sigma^2), \\ f(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{(x-\mu)^2}{2\sigma^2}], \\ FSAE(X, k) &= \frac{k}{2} - \frac{1}{2}\log(k+1), k > -1. \end{aligned}$$

Table 2.1: FSAE for Univariate Normal Distribution

k	FSAE
-0.5	0.10
0.5	0.05
1	0.15
2	0.45
3	0.81
5	1.60

Figure 2.1: FSAE for Univariate Normal Distribution



We can see this entropy is a function only about k. So given k, the fractional size adjusted entropy for any univariate normal distribution is just a constant and is free of μ and σ^2.

Example 2.1.2. Bivariate Normal Distribution:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\Big\{ -\frac{1}{2(1-\rho^2)}\Big[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\Big]\Big\},$$

$$E\log f(x,y) = \log(2\pi) + 1 + \log\left(\sigma_X\sigma_Y\sqrt{1-\rho^2}\right),$$

$$\log Ef^k(x,y) = -k\log(2\pi) - k\log(\sigma_X\sigma_Y\sqrt{1-\rho^2})) - \log(k+1),$$

$$FSAE(X,Y,k) = k - \log(k+1), k > -1.$$

Actually the fractional size adjusted entropy for the bivariate normal distribution is twice as the fractional size adjusted entropy for the univariate normal distribution. For the multivariate normal distribution, we have the following remark.

Remark 2.1.7. The fractional size adjusted entropy of a p-dimensional vector with a multivariate normal distribution is

$$FSAE(\boldsymbol{X}, k) = \frac{p}{2}[k - \log(k+1)].$$

Example 2.1.3. Gamma Distribution:

$$X \sim gamma(\alpha, \beta),$$

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}}, \alpha > 0, \beta > 0,$$

$$FSAE(X, k) = k\alpha - \log\Gamma(\alpha) + \log\Gamma(k\alpha + \alpha - k) + k(1-\alpha)\psi(\alpha) - (\alpha k + \alpha - k)\log(1+k),$$

$$k > -1 \text{ when } \alpha \geq 1,$$

$$-1 < k < \frac{\alpha}{1-\alpha} \text{ when } 0 < \alpha < 1,$$

where $\psi(x)$ is called the digamma function and is defined as

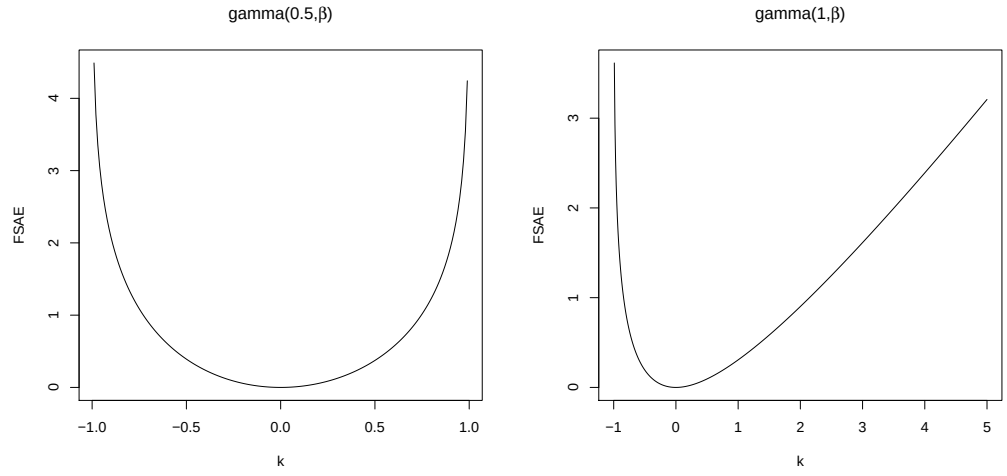
$$\psi(x) = \frac{d}{dx}\log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}.$$

From the formula, we can observe that the fractional size adjusted entropy of gamma(α, β) does not involve the scale parameter β.

Table 2.2: FSAE for Gamma Distributions

k	gamma(0.5,β)	gamma(1,β)	gamma(2,β)
-0.5	0.39	0.19	0.13
0.5	0.37	0.09	0.06
1	NA	0.31	0.19
2	NA	0.90	0.55

Figure 2.2: FSAE for Gamma Distributions



Example 2.1.4. Beta Distribution:

$$X \sim beta(\alpha, \beta),$$

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \alpha > 0, \beta > 0,$$

$$FSAE(X,k) = \log B((1+k)\alpha - k, (1+k)\beta - k) - \log B(\alpha,\beta)$$

$$- (\alpha-1)k\psi(\alpha) - (\beta-1)k\psi(\beta) + (\alpha+\beta-2)k\psi(\alpha+\beta),$$

$$k \text{ satisfies } (\alpha-1)k > -\alpha, (\beta-1)k > -\beta,$$

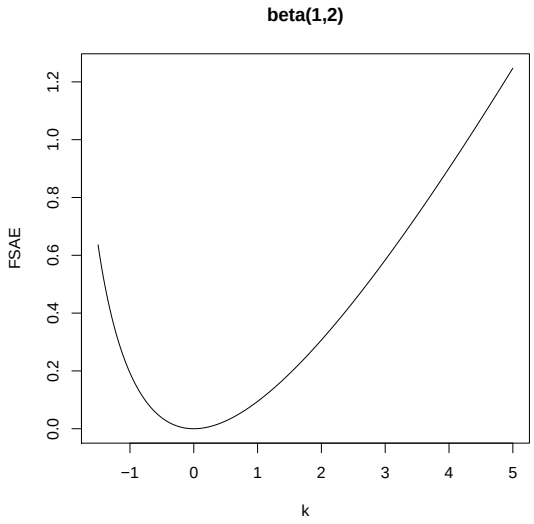
where $B(,)$ is the beta function with the form of

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx, \alpha > 0, \beta > 0.$$

Table 2.3: FSAE for Beta Distributions

k	beta(1,2)	beta(1,3)	beta(2,2)	beta(2,3)
-0.5	0.038	0.072	0.024	0.038
0.5	0.027	0.046	0.016	0.024
1	0.095	0.156	0.057	0.081
2	0.307	0.486	0.183	0.252

Figure 2.3: FSAE for Beta Distribution



Remark 2.1.8. Now consider a special case of the beta distribution. Suppose $X \sim beta(\frac{1}{2}, \frac{1}{2})$ with the PDF

$$f(x) = \frac{1}{\pi x^{1/2}(1-x)^{1/2}}.$$

First the Shannon differential entropy for a random variable X from a beta distribution $beta(\alpha, \beta)$ is given as

$$H(X) = \log B(\alpha, \beta) - (\alpha - 1)\psi(\alpha) - (\beta - 1)\psi(\beta) + (\alpha + \beta - 2)\psi(\alpha + \beta).$$

Plugging in $\alpha = \beta = \frac{1}{2}$, we have $H(X) = -0.2416$. A negative information is undesirable.

Next we will calculate the average entropy by Kittaneh et al. (2016) for this beta random variable. In order to calculate the average entropy, we first need to calculate:

$$\begin{aligned} E[f(X)] &= \int_0^1 \frac{1}{\pi^2 x(1-x)} dx \\ &= \int_0^1 \frac{1}{\pi^2 x} dx + \int_0^1 \frac{1}{\pi^2 (1-x)} dx \\ &= \frac{1}{\pi^2}(\log x|_0^1 + \log(1-x)|_1^0). \end{aligned}$$

$E[f(X)]$ does not exist, so the average entropy does not exist for this case.

Finally the fraction size adjusted entropy for this beta random variable

$$E[f^k(X)] = \int_0^1 \frac{1}{\pi^{k+1} x^{(k+1)/2}(1-x)^{(k+1)/2}} dx,$$
$$FSAE(X,k) = \log B(\frac{1-k}{2}, \frac{1-k}{2}) - 1.145 - 1.386k.$$

As long as $k < 1$ the fractional size adjusted entropy exists and it is non-negative.

From this special beta distribution example we show that, for certain distributions, the Shannon differential entropy is negative and the average entropy does not exist. But with a proper k value, the fraction size adjusted entropy always exists and is non-negative.

Example 2.1.5. Exponential Family:

A general form for the exponential family is given as followed

$$f(x|\theta) = h(x)\exp(\eta(\theta)T(x) - A(\theta)),$$
$$\int f(x|\theta)dx = 1,$$
$$\int h(x)\exp(\eta(\theta)T(x))dx = \exp[A(\theta)],$$

where $A(\theta)$ is the normalizing constant.

Then the fractional size adjusted entropy for the exponential family is

$$\begin{aligned} FSAE(X) &= kH(X) + \log \int f^{1+k}(x|\theta)dx \\ &= -k\int h(x)\exp[\eta(\theta)T(x) - A(\theta)]\log h(x)dx - k\int h(x)\exp[\eta(\theta)T(x) - A(\theta))(\eta(\theta)T(x)]dx \\ &+ k\log A(\theta) + \log \int h^{1+k}(x)\exp[(1+k)\eta(\theta)T(x) - (1+k)A(\theta)]dx \\ &= -kE\log h(x) - k\eta(\theta)ET(x) - \log A(\theta) + \log \int h^{1+k}(x)\exp[(1+k)\eta(\theta)T(x)]dx. \end{aligned}$$

2.2 Generalized Fractional Size Adjusted Entropy

2.2.1 Definitions and Properties

We can see from previous normal examples that the fractional size adjusted entropy is free of σ^2. In other words, given k, all the normal distributions have the same fractional size adjusted entropy. If we're only interested in the shape of the distribution, that may be a good property. But most of the time, we still want to consider their variations. Since the fractional size adjusted entropy doesn't involve scale parameter because it's invariant under one-to-one linear transformation, we introduce a more general form of the fractional size adjusted entropy.

Definition 2.2.1. For a random variable X with PDF or PMF $f(x)$. We define the generalized fractional size adjusted entropy of X as

$$GFSAE(X, k_1, k_2) = -E[\log \frac{f^{k_1}(x)}{Ef^{k_2}(x)}].$$

In order to make sure that the generalized fractional size adjusted entropy is non-negative, we have to choose proper values of k_1 and k_2. One way to ensure the non-negativity is to make

$$Ef^{k_1}(x) \leq Ef^{k_2}(x). \tag{2.1}$$

Then by Jensen's inequality, the generalized fractional size adjusted entropy is guaranteed to be non-negative.

Remark 2.2.2. By setting $k_1 = k_2$, the generalized fractional size adjusted entropy reduces to the fractional size adjusted entropy. And by setting $k_1 = \alpha - 1, k_2 = 0$, the generalized fractional size adjusted entropy is proportional to the Rényi entropy defined in (1.2).

Remark 2.2.3. The relationship between the generalized fractional size adjusted entropy and the fractional size adjusted entropy is

$$GFSAE(X, k_1, k_2) = FSAE(X, k_2) + (k_1 - k_2)H(X),$$

where $H(X)$ is the Shannon differential entropy.

We could also extend the generalized fractional size adjusted entropy for the joint distribution and the conditional distribution.

Definition 2.2.4. The generalized fractional size adjusted joint entropy $GFSAE(X, Y, k_1, k_2)$ of random variables (X, Y) with a joint distribution $f(x, y)$ is given by

$$GFSAE(X, Y, k_1, k_2) = -E[\log \frac{f^{k_1}(X, Y)}{Ef^{k_2}(X, Y)}].$$

Definition 2.2.5. If $(X, Y) \sim f(x, y)$, and $f(x|y)$ is the probability density function of X given Y, then the conditional generalized fractional size adjusted entropy $GFSAE(X|Y, k_1, k_2)$ is defined as

$$GFSAE(X|Y, k_1, k_2) = E_Y[GFSAE(X|Y = y, k_1, k_2)].$$

2.2.2 Generalized Fractional Size Adjusted Entropy for Certain Distributions

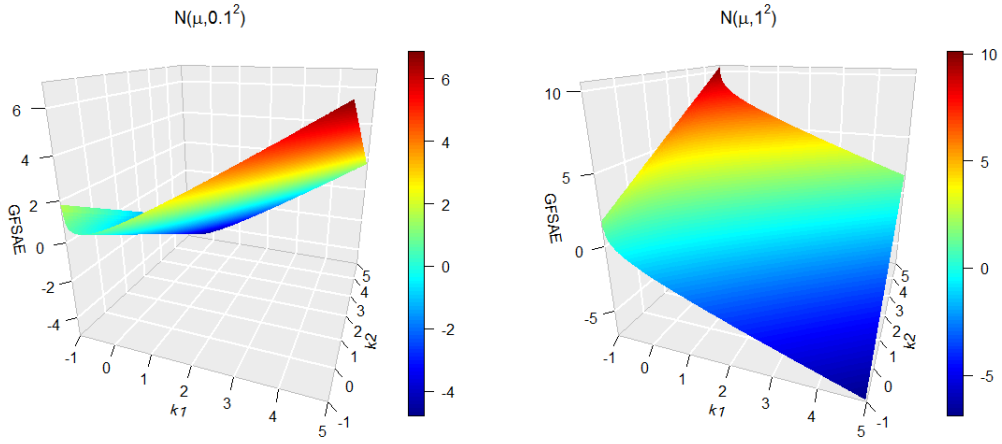
Example 2.2.1. Normal Distribution:

$$\begin{aligned} X &\sim N(\mu, \sigma^2), \\ f(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{(x-\mu)^2}{2\sigma^2}\Big], \\ Ef^k(x) &= \frac{1}{(2\pi\sigma^2)^{k/2}(1+k)^{1/2}}, \\ GFSAE(X, k_1, k_2) &= \frac{k_1 - k_2}{2}\log(2\pi\sigma^2) + \frac{k_1}{2} - \frac{1}{2}\log(1 + k_2), k_2 > -1. \end{aligned}$$

Table 2.4: GFSAE for Univariate Normal Distributions

k_1	k_2	$N(\mu, 0.1^2)$	$N(\mu, 1^2)$	$N(\mu, 10^2)$
1	1	0.15	0.15	0.15
1	2	1.33	-0.97	-3.27
1	3	2.57	-2.03	-6.64
2	1	-0.73	1.57	3.87
2	3	1.69	-0.61	-2.91
3	1	-1.61	2.99	7.60
3	2	-0.43	1.87	4.17

Figure 2.4: GFSAE for Univariate Normal Distributions



With the generalized fractional size adjusted entropy, different σ values have different patterns. From Figure 2.4 we can see there are two patterns. One is when $\sigma < 0.4$, the generalized fractional size adjusted entropy increases as k_1 increases and k_2 decreases. The other pattern is when $\sigma \geq 0.4$, the generalized fractional size adjusted entropy decreases as k_1 increases and k_2 decreases. We can see from the formula that, as long as $\sigma < 0.4$, we have the first pattern because $\log(2\pi\sigma^2)$ will be smaller than 0. As long as $\sigma \geq 0.4$, we will have the second pattern.

Example 2.2.2. Gamma Distribution:

$$X \sim gamma(\alpha, \beta),$$

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}}, \alpha > 0, \beta > 0,$$

$$Ef^k(x) = \frac{\Gamma(\alpha k + \alpha - k)(\frac{\beta}{1+k})^{\alpha k+\alpha-k}}{\Gamma^{1+k}(\alpha)\beta^{\alpha+\alpha k}},$$

$$GFSAE(X, k_1, k_2) = k_1\alpha + (k_1 - k_2)\log\beta + (k_1 - k_2 - 1)\log\Gamma(\alpha) + \log\Gamma(\alpha k_2 + \alpha - k_2)$$

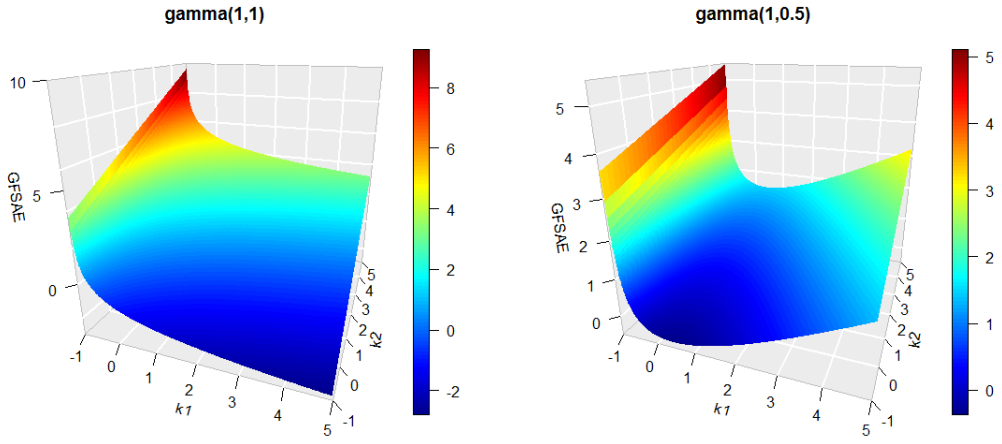
$$+ k_1(1-\alpha)\psi(\alpha) - (\alpha k_2 + \alpha - k_2)\log(1 + k_2),$$

$$k_2 > -1, \alpha k_2 + \alpha - k_2 > 0.$$

Table 2.5: GFSAE for Gamma Distributions

k_1	k_2	gamma(1,1)	gamma(1,0.5)	gamma(1,2)	gamma(0.5,1)	gamma(2,1)
1	1	0.31	0.31	0.31	NA	0.19
1	2	-0.10	0.59	-0.79	0.19	-1.03
1	3	-0.39	1	-1.77	NA	-2.18
2	1	1.31	0.61	2	NA	1.77

Figure 2.5: GFSAE for Gamma Distributions



Similarly, the generalized fractional size adjusted entropies differ with different values of β in gamma distributions when α is fixed. Figure 2.5 shows two patterns of the

generalized fractional size adjusted entropy for the gamma distribution with different combinations of k_1, k_2 values.

Example 2.2.3. Beta Distribution:

$$
\begin{aligned}
X &\sim beta(\alpha, \beta), \\
f(x) &= \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \alpha > 0, \beta > 0, \\
Ef^k(x) &= \frac{B(k\alpha - k + \alpha, k\beta - k + \beta)}{B(\alpha,\beta)^{1+k}}, \\
GFSAE(X, k_1, k_2) &= \log B(k_2\alpha - k_2 + \alpha, k_2\beta - k_2 + \beta) + (k_1 - 1 - k_2)\log B(\alpha,\beta) \\
&\quad - k_1(\alpha-1)\psi(\alpha) - k_1(\beta-1)\psi(\beta) + k_1(\alpha+\beta-2)\psi(\alpha+\beta), \\
&\alpha k_2 + \alpha - k_2 > 0, \beta k_2 + \beta - k_2 > 0.
\end{aligned}
$$

Table 2.6: GFSAE for Beta Distributions

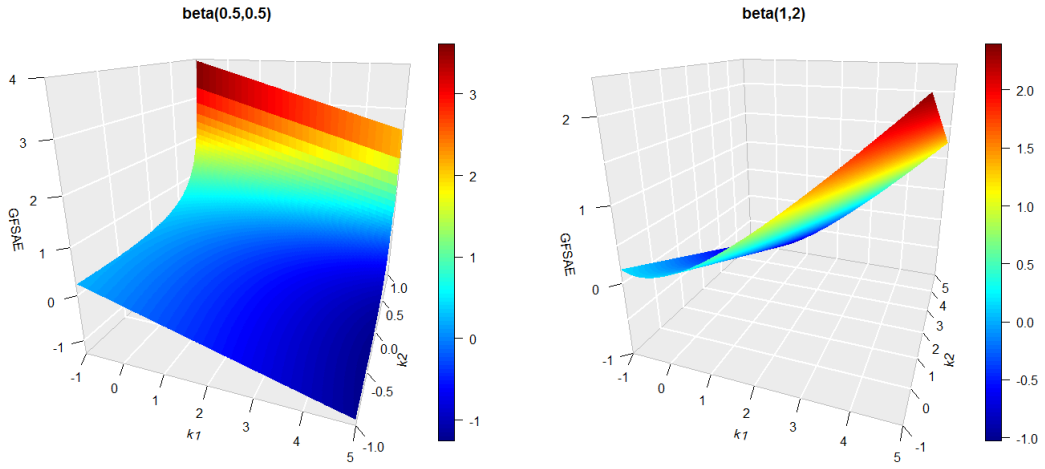
k_1	k_2	beta(1,2)	beta(1,3)	beta(2,3)
1	1	0.09	0.16	0.08
1	2	0.5	0.92	0.49
1	3	0.97	1.77	0.94
2	1	-0.10	-0.28	-0.15
2	3	0.78	1.33	0.70

Figure 2.6 again shows two patterns. One is when at least one of α, β is smaller than 1, the other is when both α, β are lager than 1. We will see more examples about the beta distribution in **Section 2.4**.

2.3 Applying to the Bayesian Framework

When we apply the fractional size adjusted entropy measure to the Bayesian framework, a remarkable property is that the normalizing constant is not necessary. As it is known to all, the normalizing constant is often analytically intractable due to the complex

Figure 2.6: GFSAE for Beta Distribution



posterior structure in the Bayesian analysis. Suppose for an unknown parameter θ, we have a prior $\pi(\theta)$ and the likelihood function given data D is $f(D|\theta)$. Then the posterior distribution for θ is given by

$$\pi(\theta|D) = \frac{q(\theta)}{c} = \frac{\pi(\theta)f(D|\theta)}{\int \pi(\theta)f(D|\theta)d\theta}.$$

Here c is the normalizing constant and $q(\theta)$ is the posterior kernel. Under this setting, we will show that the calculation of the fractional size adjusted entropy only needs the posterior kernel instead of the posterior density.

$$\begin{aligned} FSAE(\theta, k) &= -E[\log \frac{\pi^k(\theta|D)}{E[\pi^k(\theta|D)]}] \\ &= -E[\log \pi^k(\theta|D)] + \log E[\pi^k(\theta|D)] \\ &= -kE[\log \pi(\theta|D)] + \log E[\pi^k(\theta|D)] \\ &= -kE[\log q(\theta)] + k\log c + \log E[q^k(\theta)] - k\log c \\ &= -kE[\log q(\theta)] + \log E[q^k(\theta)]. \end{aligned}$$

However to calculate the generalized fractional size adjusted entropy, we have to involve the normalizing constant when $k_1 \neq k_2$. There are many literatures discussing how to

calculate the normalizing constant. Several Monte Carlo methods include the importance sampling by Geweke (1989), the harmonic mean from Newton and Raftery (1994), the generalized harmonic mean from Gelfand and Dey (1994), the serial approaches from Chib (1995) and Chib and Jeliazkov (2001), the inflated density ratio method from Petris and Tardella (2003), the thermodynamic integration from Lartillot and Philippe (2006), the constrained estimator with the highest posterior density region from Robert and Wraith (2009), the stepping stone sampling from Xie et al. (2010) and the partition weighted kernel estimator from Wang et al. (2017). These methods vary in using Monte Carlo samples or kernels in the integration. Here we just briefly provide the calculating formula for the partition weighted estimator method in Wang et al. (2017).

Following the previous settings, θ is the unknown parameter, $q(\theta)$ is the posterior kernel function and c is the normalizing constant. Let $\{A_1, \ldots, A_K\}$ forms a partition of the working parameter space Ω, where $K > 0$ is an integer, $\omega_1, \ldots, \omega_K$ are the weights assigned to these K regions, respectively. Then the partition weighted estimator is given by

$$\frac{1}{\hat{c}} = \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{\omega_k}{q(\theta_t)}1\{\theta_t \in A_k\}}{\sum_{k=1}^{K}\omega_k V(A_k)},$$

where T is the total number of MCMC samples, $V(A_k)$ is the volume of the kth subset in the partition with $V(A_k) = \int_\Omega 1\{\theta \in A_k\}d\theta$.

There are two mild assumptions to use this estimator.

Assumption 1: The volume of each region is finite $V(A_k) < \infty$ for $k = 1, \ldots, K$.

Assumption 2: $q(\theta)$ is positive and continuous on $\bar{A}_k$, which is the closure of A_k for $k = 1, \ldots, K$.

To choose the weights ω_k, Wang et al. (2017) provided an optimal value

$$\omega_{k,opt} = \frac{V(A_k)}{\alpha_k\big[\sum_{k=1}^{K}\frac{V^2(A_k)}{\alpha_k}\big]},$$
$$\text{where } \alpha_k = E\Big[\frac{1\{\theta \in A_k\}}{q^2(\theta)}\Big].$$

While in practice when $q(\theta)$ is roughly constant over A_k, simply choosing $\omega_k = q(\theta_k^*)$ where $\theta_k^* \in A_k$ yields a pretty good result. Then the partition weighted estimator reduces to

$$\frac{1}{\hat{c}} = \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{q(\theta_k^*)}{q(\theta_t)}1\{\theta_t \in A_k\}}{\sum_{k=1}^{K} q(\theta_k^*)V(A_k)}.$$

How to construct A_k to make $q(\theta)$ roughly constant over A_k is provided in detail in Wang et al. (2017) and we just skip those explanations.

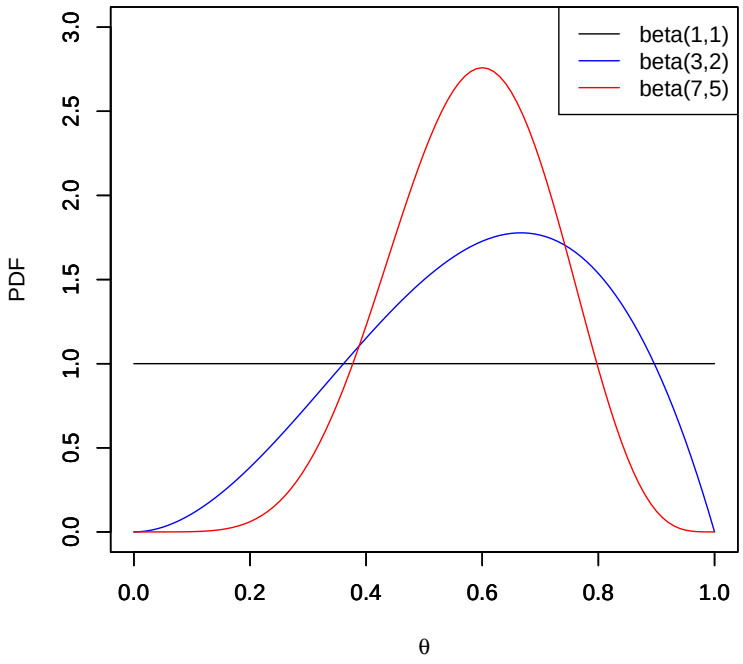
2.4 Coin Flipping Example

Imagine flipping a coin $n = 3$ times and getting $y = 2$ heads. Assuming a flat $beta(1, 1)$ prior on θ, which is the probability of heads on any given flip, this would yield a $beta(y + 1, n - y + 1) = beta(3, 2)$ posterior distribution for θ. Now imagine flipping the coin 7 more times and getting 4 more heads. Assuming the posterior from the first result as the new prior yielding a new $beta(7, 5)$ posterior for θ with the updated result.

Figure 2.7 shows the densities of the original prior $beta(1, 1)$, the first posterior $beta(3, 2)$ and the second posterior $beta(7, 5)$.

We would like to continue this coin flipping process. With 10, 40, 90 more flips, we get the updated posterior distributions $beta(12, 10), beta(28, 24), beta(52, 50)$ respectively. The results are simulated with fixing $\theta = 0.5$. Now we would like to use different measures to get an idea of how much information of θ in each flipping results (posterior distributions) we can get.

Figure 2.7: Three Beta Densities



Form Table 2.7, we can see all the measures return 0 for the original prior $beta(1,1)$. After that, the Shannon differential entropy (Entropy) always returns negative values. The fractional size adjusted entropies (FSAE) with different k values and the generalized fractional size adjusted entropies (GFSAE) with different k_1, k_2 values are always positive and increasing as the sample size increasing. As we have more samples, we should be able to gain more information. Since the results are simulated from the same pattern, the additional gain with each updated result should be small. Next we will calculate the difference in entropies between each result.

In Table 2.8, DD means the difference of densities. It's a naive measure defined as the sum of the area that are not shared by two distributions. From the results in Table 2.8, we can see KL divergence doesn't have the additivity. Lindley entropy has a different pattern

Table 2.7: Different Measures for Coin Flipping Results

	beta(1,1)	beta(3,2)	beta(7,5)	beta(12,10)	beta(28,24)	beta(52,50)
Entropy	0	-0.235	-0.580	-0.848	-1.263	-1.592
FSAE(1)	0	0.081	0.122	0.136	0.146	0.150
FSAE(2)	0	0.252	0.366	0.405	0.431	0.441
FSAE(5)	0	0.961	1.338	1.459	1.544	1.573
GFSAE(1,2)	0	0.486	0.946	1.253	1.694	2.033
GFSAE(1,3)	0	0.935	1.823	2.425	3.300	3.974
GFSAE(1,5)	0	1.901	3.656	4.852	6.595	7.941
GFSAE(2,3)	0	0.700	1.243	1.577	2.037	2.382
GFSAE(2,4)	0	1.175	2.150	2.781	3.674	4.355
GFSAE(2,10)	0	4.257	7.853	10.270	13.771	16.469

Table 2.8: Distance Between Distributions

	beta(3,2) from beta(1,1)	beta(7,5) from beta(3,2)	beta(7,5) from beta(1,1)
KL	0.515	0.323	2.25
Lindley	0.235	0.345	0.580
DD	0.543	0.455	0.906
FSAE(1)	0.081	0.041	0.122
FSAE(2)	0.252	0.114	0.366
FSAE(5)	0.961	0.377	1.338
FSAE(1,2)	0.486	0.460	0.946
FSAE(1,3)	0.935	0.888	1.823
FSAE(1,5)	1.901	1.755	3.656
FSAE(2,3)	0.700	0.543	1.243
FSAE(2,4)	1.175	0.975	2.150
FSAE(2,10)	4.257	3.596	7.853

than other measures. All other measures show that the additional gain from $beta(3,2)$ to $beta(7,5)$ is the smallest and the gain from $beta(1,1)$ to $beta(7,5)$ is the largest, while Lindley entropy shows the additional gain from $beta(3,2)$ to $beta(7,5)$ is in the middle of $beta(1,1)$ to $beta(7,5)$ and $beta(1,1)$ to $beta(3,2)$.

We will discuss more about the comparison of distributions in **Chapter 3** and information gain in **Chapter 4**.

2.5 Analysis of *Protosiphon botryoides* Data

In 1958, a spade full of soil was collected from a cornfield on the University of Connecticut campus in Storrs, Connecticut, USA. The original plan was to isolate sexually reproducing algae. In 2001 and 2008, Lewis and Trainor repeated the growth experiment using the original soil (Lewis and Trainor, 2012). The alga resulting from the treatment during 2001 was isolated into culture as strain FRT2000 and deposited as UTEX B 2969. Cells of *Protosiphon botryoides* isolate FRT2000 were observed. The nearest matches from other *Protosiphon* isolates were compiled separately into an 18S rDNA and *rbc*L alignment.

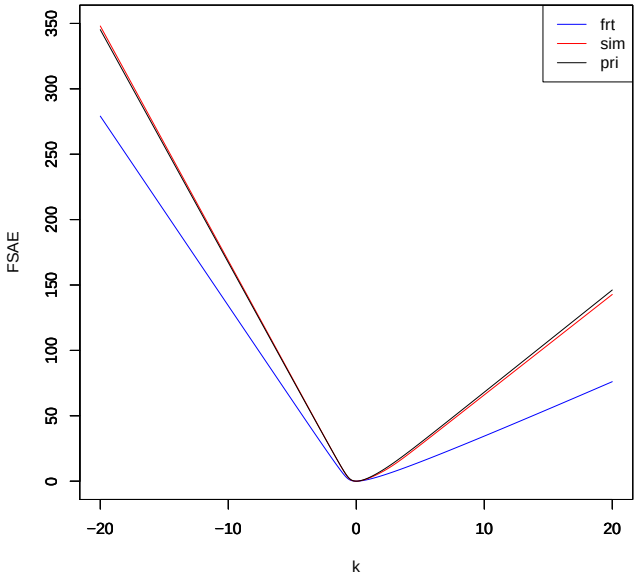
General time reversible (GTR) model with invariant sites rates is used for analysis. Three different data sets are generated. frt is generated based on *rbc*L, *sim* is a simulated data with 19995 out of 20000 sites are constant, *pri* is generated from the prior with each edge length prior is exponential distribution with mean 0.1. For each data set, log posterior kernel values are calculated. Based on these kernel values, Table 2.9 and Figure 2.8 show the fractional size adjusted entropies with different k values.

Table 2.9: FSAE for Three Datasets

k	frt	sim	pri
1	1.45	2.37	2.75
2	4.21	7.02	8.07
3	7.48	13.01	14.56
5	14.73	27.65	29.28
-1	5.44	8.02	8.20
-2	18.82	25.29	25.15
-3	32.94	43.10	42.74
-5	61.72	78.91	78.22

We can see that the *sim* and *pri* are very similar, while frt is different from those two for all k values. *sim* has very concentrated estimates for edge lengths with single tree

Figure 2.8: FSAE for Three Datasets from *Protosiphon botryoides* Data



topology. frt provides more variable estimates with multiple tree topologies. The results also indicate that the *rbc*L data has conflicted information with the prior assumption.

Table 2.10: GFSAE for Three Datasets

k_1	k_2	frt	sim	pri
1	2	45.66	35.45	19.80
5	6	60.00	63.73	48.68
9	10	75.92	94.42	79.78
13	12	1.19	52.90	71.93

For the generalized fractional size adjusted entropy, since k_1 and k_2 are not equal, the normalizing constant is necessary for the calculations. The normalizing constant for sim is -27798, for frt is -2771. Since pri is from the prior distribution, the normalizing constant is automatically 1. Table 2.10 shows the generalized fractional size adjusted entropy for three data sets with different combinations of k_1, k_2 values. The relative pattern of

the generalized fractional size adjusted entropies for these three data sets changes with different k_1, k_2 values. Future research is needed and we will discuss it in **Section 5.4**.

Chapter 3

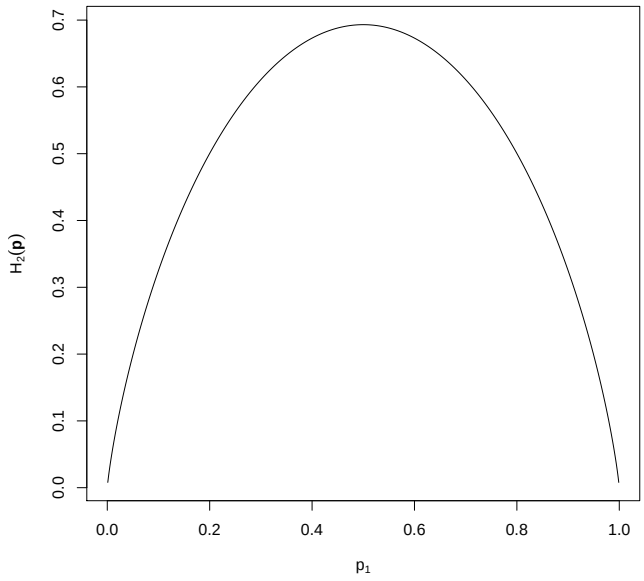
Data Compatibility

3.1 Entropy Function

The Shannon entropy (Shannon, 1948) is one of the most widely used measures of information. For a discrete random variable X with probability mass function $f(X)$ supported at $\{x_1, \ldots, x_K\}$, the Shannon entropy is defined in (1.1). Let $p_i = f(x_i)$ for $i = 1, \ldots, K$. Since $f(X)$ is a probability mass function, we have $\sum_{i=1}^{K} p_i = 1$. Define $\boldsymbol{p} = (p_1, \ldots, p_K)$ and $H_K(\boldsymbol{p}) = -\sum_{i=1}^{K} p_i \log(p_i)$. Obviously $H(X) = H_K(\boldsymbol{p})$, but $H(X)$ is the entropy of a random variable X while $H_K(\boldsymbol{p})$ is a function of a probability vector $\boldsymbol{p} = (p_1, \ldots, p_K)$ with constraint $\sum_{i=1}^{K} p_i = 1$.

For $K = 2$, Figure 3.1 plots the entropy $H_2(\boldsymbol{p})$ as a function of p_1 and shows that the entropy function is concave and symmetric about $p_1 = 0.5$. $H_2(\boldsymbol{p})$ increases when $p_1 \leq 0.5$ and decreases when $p_1 \geq 0.5$. The maximal value $\log 2$ is achieved when $p_1 = 0.5$, and the minimal value 0 is attained at $p_1 = 0$ or 1.

Figure 3.1: Plot of $H_2(\boldsymbol{p})$ as a Function of p_1



3.2 Partition

Consider a Bayesian posterior density having the form of

$$\pi(\boldsymbol{\theta}|D) = \frac{q(\boldsymbol{\theta}|D)}{c(D)} = \frac{1}{c(D)} f(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

where D denotes data and the parameter $\boldsymbol{\theta}$ is an m-dimensional vector in the parameter space Ω, $f(D|\boldsymbol{\theta})$ is the likelihood of θ given D, $\pi(\boldsymbol{\theta})$ is the prior density of $\boldsymbol{\theta}$, $c(D) = \int f(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the normalizing constant, and $q(\boldsymbol{\theta}|D) = f(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ is called the kernel function in the Bayesian literature. Given $(p_1, \ldots, p_K)$ with constraint $\sum_{i=1}^{K} p_i = 1$, we intend to build a partition $(\Omega_1, \ldots, \Omega_K)$ for $\boldsymbol{\theta}$ with $\int_{\Omega_i} \pi(\boldsymbol{\theta}|\boldsymbol{D})d\boldsymbol{\theta} = p_i$ for $i = 1, \ldots, K$.

Remark 3.2.1. One way to construct a partition is to use the highest posterior density (HPD) regions. For a given probability p, the $100p\%$ highest posterior density region is the subset Ω_p^* of Ω, which takes the form of

$$\Omega_p^* = \{\boldsymbol{\theta} \in \Omega : \pi(\boldsymbol{\theta}|D) > k(p)\},$$

where $k(p)$ is the largest constant such that

$$P(\boldsymbol{\theta} \in \Omega_p^*) \geq p.$$

Given a vector of probabilities $(p_1, \ldots, p_K)$, we can construct a vector of HPD regions $(\Omega_1^*, \ldots, \Omega_K^*)$ such that Ω_i^* is the $100(\sum_{j=1}^{i} p_j)\%$ HPD region for $i = 1, ..., K$. With constraint $\sum_{i=1}^{K} p_i = 1$, we have:

$$\Omega_1 = \Omega_1^*,$$

$$\Omega_i = \Omega_i^* \cap (\Omega_{i-1}^*)^C, i = 2, \ldots, K.$$

Then $(\Omega_1, \ldots, \Omega_K)$ forms a partition on Ω.

3.3 Compatibility Measure

Our measure of compatibility is based on the posterior distributions. We assume that the data are more informative than the prior for the parameters. In other words, the posterior distribution should be more similar to the likelihood function than the prior distribution and the prior should be noninformative or essentially noninformative. Suppose we have two data sets D_1, D_2 with common parameters $\boldsymbol{\theta} \in R^m$ and a common prior $\pi(\boldsymbol{\theta})$. Then, under the Bayesian setting, we have

$$\pi(\boldsymbol{\theta}|D_1) = \frac{q(\boldsymbol{\theta}|D_1)}{c(D_1)} = \frac{1}{c(D_1)} f(D_1|\boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$
$$\pi(\boldsymbol{\theta}|D_2) = \frac{q(\boldsymbol{\theta}|D_2)}{c(D_2)} = \frac{1}{c(D_2)} f(D_2|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

We build a partition $(\Omega_{11}, \ldots, \Omega_{1K})$ on $\boldsymbol{\theta}$ such that $\int_{\Omega_{1i}} \pi(\boldsymbol{\theta}|D_1)d\boldsymbol{\theta} = p_{1i}$ for $i = 1, \ldots, K$ and $\boldsymbol{p_1} = (p_{11}, \ldots, p_{1K})$. Then we calculate $\boldsymbol{p_{2|1}} = (p_{2|1,1}, \ldots, p_{2|1,K})$, where $p_{2|1,i} = \int_{\Omega_{1i}} \pi(\boldsymbol{\theta}|D_2)d\boldsymbol{\theta}$ for $i = 1, \ldots, K$. We introduce a compatibility measure based on the entropy difference between $\boldsymbol{p_1}$ and $\boldsymbol{p_{2|1}}$.

Definition 3.3.1. The compatibility of data set D_2 based on data set D_1 with K partition subsets is defined as

$$M_K(D_2|D_1) = 100\left(1 - \frac{|H_K(\boldsymbol{p_1}) - H_K(\boldsymbol{p_{2|1}})|}{\log K}\right)\%.$$

Similarly, we can build a partition $(\Omega_{21}, \ldots, \Omega_{2K})$ with probabilities $\boldsymbol{p}_2 = (p_{21}, \ldots, p_{2K})$ on $\boldsymbol{\theta}$ such that $\int_{\Omega_{2i}} \pi(\boldsymbol{\theta}|D_2)d\boldsymbol{\theta} = p_{2i}$ for $i = 1, \ldots, K$. With $\boldsymbol{p_{1|2}} = (p_{1|2,1}, \ldots, p_{1|2,K})$, where $p_{1|2,i} = \int_{\Omega_{2i}} \pi(\boldsymbol{\theta}|D_1)d\boldsymbol{\theta}, i = 1, \ldots, K$, we have the following definition.

Definition 3.3.2. The compatibility of data set D_1 based on data set D_2 with K partition subsets is defined as

$$M_K(D_1|D_2) = 100\left(1 - \frac{|H_K(\boldsymbol{p_2}) - H_K(\boldsymbol{p_{1|2}})|}{\log K}\right)\%.$$

$M_K(D_2|D_1)$ is the compatibility of data set D_2 compared to data set D_1. In other words, we use D_1 as the base, and compare D_2 with it. Similarly $M_K(D_1|D_2)$ is the compatibility of D_1 compared to D_2. Then we take the average of these two to be the compatibility of D_1 and D_2.

Definition 3.3.3. The compatibility M of two data sets D_1, D_2 is

$$M_K(D_1, D_2) = \frac{M_K(D_2|D_1) + M_K(D_1|D_2)}{2}.$$

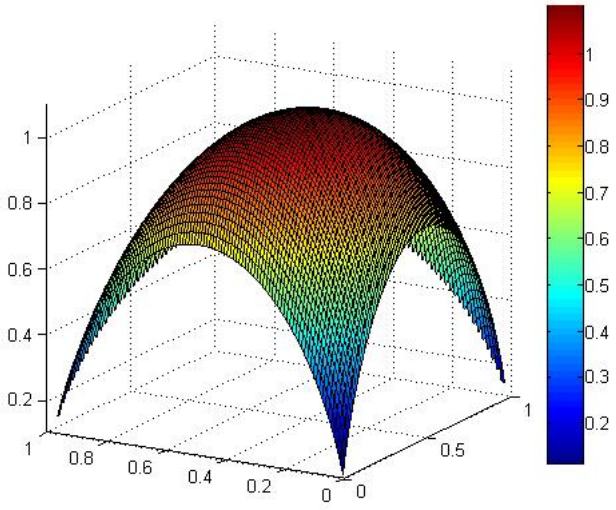
$M_K(D_1, D_2)$ is symmetric, while $M_K(D_2|D_1)$ and $M_K(D_1|D_2)$ are not.

$M_K(D_1, D_2)$ is a number from 0 to 100%. A higher compatibility value indicates the two data sets are more similar.

$M_K(D_1, D_2) = 0$ means that D_1 is incompatible with D_2 and is only achieved when $M_K(D_1|D_2) = M_K(D_2|D_1) = 0$. In other words, $\boldsymbol{p_1} = \boldsymbol{p_2} = (\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K})$ and each element of both $\boldsymbol{p_{2|1}}$ and $\boldsymbol{p_{1|2}}$ is either 0 or 1. This can happen only when the two distributions $\pi(\boldsymbol{\theta}|D_1)$ and $\pi(\boldsymbol{\theta}|D_2)$ each locates in the tail of the other.

$M_K(D_1, D_2) = 100\%$ is only achieved when $H_K(\boldsymbol{p_{2|1}}) = H_K(\boldsymbol{p_1})$ and $H_K(\boldsymbol{p_{1|2}}) = H_K(\boldsymbol{p_2})$. Only under $\boldsymbol{p_1} = \boldsymbol{p_2} = (\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K})$, we can get $\boldsymbol{p_{2|1}} = \boldsymbol{p_1}, \boldsymbol{p_{1|2}} = \boldsymbol{p_2}$ from $H_K(\boldsymbol{p_{2|1}}) = H_K(\boldsymbol{p_1})$ and $H_K(\boldsymbol{p_{1|2}}) = H_K(\boldsymbol{p_2})$.

Figure 3.2: Plot of $H_3(\boldsymbol{p})$



Remark 3.3.4. For K partition subsets, we recommend $\boldsymbol{p_1} = \boldsymbol{p_2} = (\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K})$ because the entropy is maximized at this choice of $\boldsymbol{p_1}, \boldsymbol{p_2}$ and other choices may have a duality problem. Figure 3.2 shows the entropy of a three dimensional vector. We can see when $p_1 = p_2 = p_3 = \frac{1}{3}$, $H_3(\boldsymbol{p})$ achieves its maximal value. Other choices of $\boldsymbol{p}$ will cause identification problems. For example, the posterior distribution given data D_1 is $N(0, 1)$ and the posterior distribution given data D_2 is $N(1, 1)$. If we use two partition subsets and choose $p_1 = 0.5976, p_2 = 0.5976$, then $M_2(D_1, D_2) = 0$, which means the two posterior distributions are identical. This, of course, is not true. The reason is that $p_{2|1} = p_{1|2} = 0.4023$, and $H_2(0.5976) - H_2(0.4023) = 0$. When $p_{2|1} = 1 - p_1$, inference is confounded due to the duality of the entropy function. Choosing $p_1 = 0.5$ solves the problem. We thus recommend using $p_1 = p_2 = 0.5$ under two partitions and

$\boldsymbol{p_1} = \boldsymbol{p_2} = (\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K})$ under K partitions. Table 3.1 shows detailed results for two partition subsets with different choices of p_1 under normal distributions.

As long as $p_1 \neq 0.5$, there is always a case when $M_2(D_2|D_1)$ is close to 0 while $p_{2|1} \neq p_1$ (Table 3.1). When comparing $N(0,1)$ with $N(1,1)$, it is $p_1 = 0.598$ that $M_2(D_2|D_1) = 0$ but $p_{2|1} \neq p_1$. When comparing $N(0,1)$ with $N(3,1)$ and $N(0,2)$ with $N(1,1)$, it is $p_1 = 0.907$ and $p_1 = 0.444$ respectively. This means that the compatibility measure based on the HPD intervals with an asymmetrical bipartition fails to detect the difference between D_1 and D_2. To avoid such problems, we suggest to set $p_1 = 0.5$ for bipartitions and $\boldsymbol{p_1} = (\frac{1}{K}, \frac{1}{K}, ..., \frac{1}{K})$ with K partition subsets.

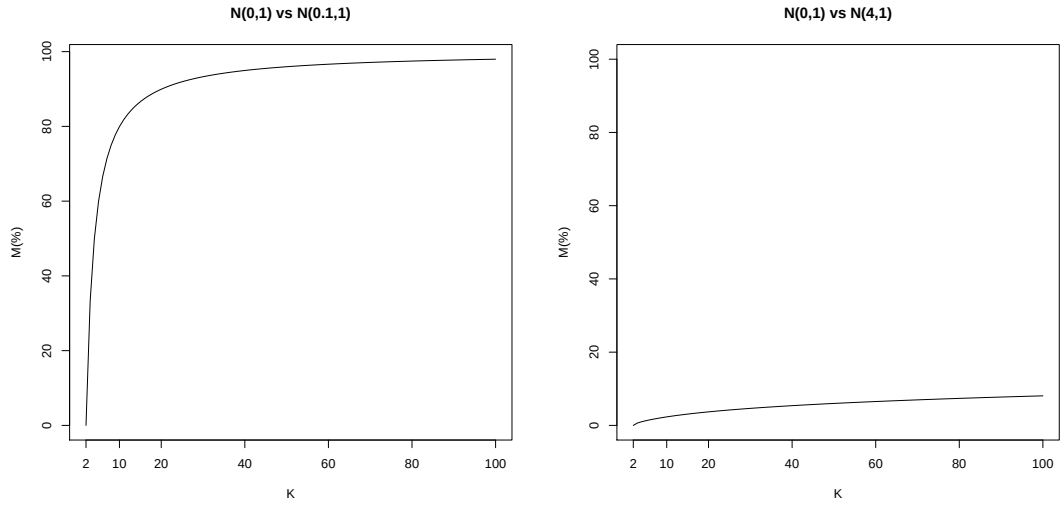
Table 3.1: Two Partition Subsets under Normal Distributions

μ_1	μ_2	σ_1	σ_2	p_1	$p_{2\|1}$	$M_2(D_2\|D_1)(\%)$
0	1	1	1	0.1	0.061	13.8
0	1	1	1	0.3	0.186	18.7
0	1	1	1	0.5	0.325	9.0
0	1	1	1	0.7	0.493	11.9
0	1	1	1	0.9	0.736	36.3
0	1	1	1	0.598	0.403	0.0
0	3	1	1	0.1	0.001	45.6
0	3	1	1	0.3	0.004	84.3
0	3	1	1	0.5	0.010	92.0
0	3	1	1	0.7	0.025	71.4
0	3	1	1	0.9	0.088	4.0
0	3	1	1	0.907	0.093	0.1
0	1	2	1	0.1	0.122	6.5
0	1	2	1	0.3	0.371	7.0
0	1	2	1	0.5	0.627	4.7
0	1	2	1	0.7	0.857	29.0
0	1	2	1	0.9	0.989	38.1
0	1	2	1	0.444	0.556	0.0

Remark 3.3.5. To choose the number of the partition subsets K, our recommendation is around 10. Although the limiting value and the convergence rate of the compatibility measure M depend on the distributions, M is already quite stable when K is around 10 to 20. Figure 3.3 shows the compatibility measure M for comparing normal distribution

$N(0,1)$ with $N(0.1,1)$ and $N(4,1)$ with different number of partition subsets K (up to 100).

Figure 3.3: Compatibility Measure for Comparing Normal Distributions with Different Numbers of Partition Subsets



Although in both cases the compatibility value is increasing all the time, the increasing rate slows down after $K = 20$. We can easily figure out whether the two distributions are compatible or not even at small K values.

3.4 Algorithm

With the previous settings, the following algorithm is developed to calculate the compatibility measure.

Step 1. Draw an MCMC sample $\{\boldsymbol{\theta_1}^{(t)}\}_{t=1,\ldots,N_1}$ from $\pi(\boldsymbol{\theta}|D_1)$, and independently draw an MCMC sample $\{\boldsymbol{\theta_2}^{(t)}\}_{t=1,\ldots,N_2}$ from $\pi(\boldsymbol{\theta}|D_2)$.

Step 2. Sort the N_1 values of $\{\pi(\boldsymbol{\theta_1}^{(t)}|D_1)\}$ from small to large. Let $a_t = \pi(\boldsymbol{\theta_1}^{(t)}|D_1)$. Then the sorted values are $\{a_{(1)}, \ldots, a_{(N_1)}\}$. Based on $\boldsymbol{p_1} = (p_{11}, \ldots, p_{1K})$, calculate

the HPD regions $\hat{\boldsymbol{\Omega}}_{\mathbf{1}}^{*} = (\hat{\Omega}_{11}^{*}, \ldots, \hat{\Omega}_{1K}^{*})$ such that

$$\hat{\Omega}_{1i}^{*} = \{\boldsymbol{\theta} : \pi(\boldsymbol{\theta}|D_1) \in [a_{(N_1 - [N_1 \sum_{j=1}^{i} p_{1j}])}, a_{(N_1)}]\}, i = 1, \ldots, K.$$

Step 3. Construct partition $\hat{\boldsymbol{\Omega}}_{\mathbf{1}} = (\hat{\Omega}_{11}, \ldots, \hat{\Omega}_{1K})$, where

$$\hat{\Omega}_{11} = \hat{\Omega}_{11}^{*},$$

$$\hat{\Omega}_{1i} = \hat{\Omega}_{1i}^{*} \cap (\hat{\Omega}_{1,i-1}^{*})^{C}, i = 2, \ldots, K.$$

Step 4. Calculate the proportion $\hat{\boldsymbol{p}}_{\mathbf{2|1}} = (\hat{p}_{2|1,1}, \ldots, \hat{p}_{2|1,K})$ of $\{\boldsymbol{\theta}_{\mathbf{2}}^{(t)}\}$ in each corresponding region of $\hat{\boldsymbol{\Omega}}_{\mathbf{1}}$ by

$$\hat{p}_{2|1,i} = \frac{\text{Number of } (\boldsymbol{\theta}_{\mathbf{2}}^{(t)} \in \hat{\Omega}_{1i})}{N_2}, i = 1, \ldots, K.$$

Step 5. Calculate the compatibility measure $M_K(D_2|D_1)$, which is

$$M_K(D_2|D_1) = 100\left(1 - \frac{|H_K(\boldsymbol{p}_{\mathbf{1}}) - H_K(\hat{\boldsymbol{p}}_{\mathbf{2|1}})|}{\log K}\right)\%.$$

Step 6. Sort the N_2 values of $\{\pi(\boldsymbol{\theta}_{\mathbf{2}}^{(t)}|D_2)\}$ from small to large. Let $b_t = \pi(\boldsymbol{\theta}_{\mathbf{2}}^{(t)}|D_2)$. Then the sorted values are $\{b_{(1)}, \ldots, b_{(N_2)}\}$. Based on $\boldsymbol{p}_{\mathbf{2}} = (p_{21}, \ldots, p_{2K})$ calculate the HPD Regions $\hat{\boldsymbol{\Omega}}_{\mathbf{2}}^{*} = (\hat{\Omega}_{21}^{*}, \ldots, \hat{\Omega}_{2K}^{*})$ such that

$$\hat{\Omega}_{2i}^{*} = \{\boldsymbol{\theta} : \pi(\boldsymbol{\theta}|D_2) \in [b_{(N_2 - [N_2 \sum_{j=1}^{i} p_{2j}])}, b_{(N_2)}]\}, i = 1, \ldots, K.$$

Step 7. Construct partition $\hat{\boldsymbol{\Omega}}_{\mathbf{2}} = (\hat{\Omega}_{21}, \ldots, \hat{\Omega}_{2K})$, where

$$\hat{\Omega}_{21} = \hat{\Omega}_{21}^{*},$$

$$\hat{\Omega}_{2i} = \hat{\Omega}_{2i}^{*} \cap (\hat{\Omega}_{2,i-1}^{*})^{C}, i = 2, \ldots, K.$$

Step 8. Calculate the proportion $\hat{\boldsymbol{p}}_{\mathbf{1|2}} = (\hat{p}_{1|2,1}, \ldots, \hat{p}_{1|2,K})$ of $\{\boldsymbol{\theta}_{\mathbf{1}}^{(t)}\}$ in each corresponding region of $\hat{\boldsymbol{\Omega}}_{\mathbf{2}}$, where

$$\hat{p}_{1|2,i} = \frac{\text{Number of } (\boldsymbol{\theta}_{\mathbf{1}}^{(t)} \in \hat{\Omega}_{2i})}{N_1}, i = 1, \ldots, K.$$

Step 9. Calculate the compatibility measure $M_K(D_1|D_2)$, which is

$$M_K(D_1|D_2) = 100\left(1 - \frac{|H_K(\boldsymbol{p_2}) - H_K(\hat{\boldsymbol{p}}_{\boldsymbol{1|2}})|}{\log K}\right)\%.$$

Step 10. Calculate $M_K(D_1, D_2)$, which is

$$M_K(D_1, D_2) = \frac{M_K(D_2|D_1) + M_K(D_1|D_2)}{2}.$$

Remark 3.4.1. In Step 2 and Step 6, we can directly work on the kernels $q(\boldsymbol{\theta}|D_1), q(\boldsymbol{\theta}|D_2)$ instead of $\pi(\boldsymbol{\theta}|D_1), \pi(\boldsymbol{\theta}|D_2)$. This allows us to avoid the calculation of the normalizing constants $c(D_1), c(D_2)$.

3.5 Partial Compatibility

Sometimes two distributions based on different data sets may be compatible only with respect to some, but not all, parameters. In other words, they share some common parameters, but not all. In other cases, we are only interested in some parameters instead of all parameters. To compare a subset of parameters, we introduce a new concept called the partial compatibility. We define $\boldsymbol{\theta}$ to include all parameters in data set D_1, $\boldsymbol{\theta}^*$ to include all parameters in data set D_2, $\boldsymbol{\theta}_1$ to include the parameters of interest (common to both D_1 and D_2), $\boldsymbol{\theta}_2$ to include all remaining parameters in D_1, and $\boldsymbol{\theta}_2^*$ to include all remaining parameters in D_2. Here $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_2^*$ can be the same or different, but we are only interested in $\boldsymbol{\theta}_1$. We calculate the marginal distributions of $\boldsymbol{\theta}_1$ from D_1, D_2

$$\pi(\boldsymbol{\theta}_1|D_1) = \int \pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|D_1)d\boldsymbol{\theta}_2,$$
$$\pi(\boldsymbol{\theta}_1|D_2) = \int \pi(\boldsymbol{\theta_1}, \boldsymbol{\theta}_2^*|D_2)d\boldsymbol{\theta}_2^*.$$

Applying our method to $\pi(\boldsymbol{\theta}_1|D_1)$ and $\pi(\boldsymbol{\theta}_1|D_2)$ provides the partial compatibility of the common parameters $\boldsymbol{\theta}_1$ between D_1, D_2.

Calculating the marginal probability based on a single MCMC chain normally is not straightforward. A nonparametric kernel density estimator can be used. Although it is easily implemented and requires no further assumptions, it may be less efficient and computational expensive in high dimensional cases because it dose not use the information from the sample of non-focal parameters and the known structure of the posterior distributions. Gelfand et al. (1992) proposed the conditional marginal density estimator but is analytically intractable as shown in Chen (1994). Chen (1994) proposed the importance weighted marginal density estimation method as a generalization of the conditional marginal density estimator. However the optimal weight function in his method is unavailable in most cases. Wang et al. (2017) provided an adaptive partition weighted marginal density estimator. We use the same notation as described in Wang et al. (2017) to illustrate their method. Let $(\boldsymbol{\theta}, \boldsymbol{\xi})$ be a v-dimensional vector of parameters, where $\boldsymbol{\theta}$ is a vector of parameters of interest. Knowing the joint posterior kernel $q(\boldsymbol{\theta}, \boldsymbol{\xi}|D)$, we want to calculate the marginal posterior distribution $\pi(\boldsymbol{\theta}|D)$. The adaptive partition weighted marginal density estimator given an MCMC sample $\{(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t), t = 1, \ldots, T\}$ has the form of

$$\hat{\pi}(\boldsymbol{\theta}_0|D) = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{q(\boldsymbol{\theta}_k^*, \boldsymbol{\xi}_t)1\{\boldsymbol{\theta}_t \in A_k(\boldsymbol{\xi}_t)\}q(\boldsymbol{\theta}_0, \boldsymbol{\xi}_t)}{\sum_{k=1}^{K} q(\boldsymbol{\theta}_k^*, \boldsymbol{\xi}_t)V(A_k(\boldsymbol{\xi}_t))q(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t)},$$

where $\{A_k(\boldsymbol{\xi}_t), k = 1, \ldots, K\}$ is the partition of the conditional working parameter space $\tilde{\Theta}_{\boldsymbol{\xi}_t}$, $\boldsymbol{\theta}_k^*$ is a fixed point in $A_k(\boldsymbol{\xi}_t)$ and $V(A_k(\boldsymbol{\xi}_t))$ is the volume of $A_k(\boldsymbol{\xi}_t)$. The conditional working parameter space is chosen such that

$$\int_{\tilde{\Theta}_{\boldsymbol{\xi}_t}} q(\boldsymbol{\theta}, \boldsymbol{\xi}_t)d\boldsymbol{\theta} > 0.$$

It is called the adaptive partition weighted marginal density estimator because the partition changes at each t.

With this adaptive partition weighted marginal density method, we can estimate the marginal posterior density values $\pi(\boldsymbol{\theta}_1|D_1)$ and $\pi(\boldsymbol{\theta}_1|D_2)$ just from an MCMC sample.

However, when $\boldsymbol{\theta}_1$ is composed of just one parameter, Chen and Shao (1999, Theorem 2) showed that the HPD regions may be obtained directly from the posterior samples when the posterior distribution is continuous and unimodal. This even obviates the need to calculate the value of the posterior kernel. Step 2 now can be simplified to

Step 2*. Sort the N_1 samples $\{\theta_1^{(t)}\}$ as $\{\theta_{1(1)}, \ldots, \theta_{1(N_1)}\}$. Choose j^* such that

$$\theta_{1(j^*+[N_1\sum_{k=1}^{i} p_{1k}])} - \theta_{1(j^*)} = \min_{1\le j\le N_1 - N_1\sum_{k=1}^{i} p_{1k}} (\theta_{1(j+[N_1\sum_{k=1}^{i} p_{1k}])} - \theta_{1(j)}),$$

$$\Omega_{1i}^* = (\theta_{1(j^*)}, \theta_{1(j^*+[N_1\sum_{k=1}^{i} p_{1k}])}).$$

By using this simplified method, we can directly obtain HPD regions and partition subsets from MCMC samples without calculating kernel values. Step 6 can also be simplified in a similar way.

3.6 Simulation Studies

Example 3.6.1. Compare Three Normal Distributions

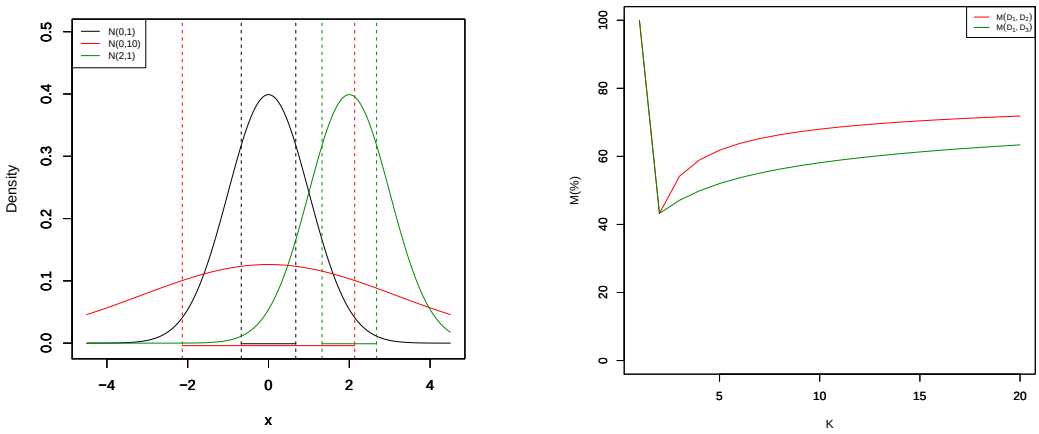
Suppose the posterior distribution given data D_1 is a normal distribution $N(0, 1)$, the posterior distribution given data D_2 is $N(0, 10)$, and the posterior distribution given data D_3 is $N(2, 1)$.

Figure 3.4(a) plots these three normal distributions and their 50% HPD intervals are indicated by dashed lines. $M_2(D_1, D_2) = M_2(D_1, D_3) = 0.432$ means that two partition subsets cannot detect the difference between D_2 and D_3. The black and red regions share the same location while the black and green regions do not. This example shows that a

partition comprising a small number of subsets sometimes can not detect real differences between data sets.

Figure 3.4(b) plots the compatibility measures $M_K(D_1, D_2)$ and $M_K(D_1, D_3)$ as a function of K. As the number of partition subsets increases, $M_K(D_1, D_2)$ and $M_K(D_1, D_3)$ differ from each other and become stable.

Figure 3.4: Compatibility for Comparing Three Normal Distributions



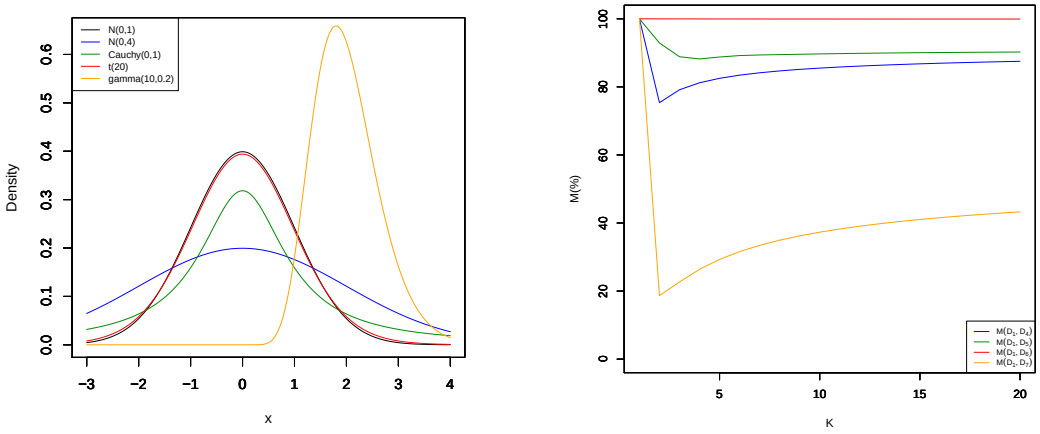
(a) Three Normal Distribution Densities (b) Compatibility with K Partition Subsets

Example 3.6.2. Location Compatibility

We can also compare the parameter which describes the similar behavior across different distributions. For example, we want to consider the location compatibility of several distributions. Suppose the posterior distribution given data set D_1 is a normal distribution $N(0, 1)$, the posterior distribution given D_4 is $N(0, 4)$, the posterior distribution given D_5 is a standard Cauchy distribution, the posterior distribution given D_6 is a t-distribution with 20 degrees of freedom, and the posterior distribution given D_7 is a gamma distribution Gamma(10,0.2) with mean 2. Although in each distribution, there is a different parameter describing the location and some distributions even do not have a specific parameter. For

the normal distribution, parameter μ controls the location. For the Cauchy distribution, parameter x_0 controls the location. For the t-distribution, there is no parameter to describe the location. For the gamma(α, β) distribution, the product $\alpha\beta$ determines the location.

Figure 3.5: Compatibility for Comparing Locations across Different Distributions



(a) Posterior Distributions

(b) Compatibility with K Partition Subsets

Figure 3.5(a) plots these posterior densities. Figure 3.5(b) plots $M_K(D_1, D_i), i = 4, \ldots, 7$ as a function of K. They show that D_1, D_4 are almost identical, so their compatibility is near 100%. D_5, D_6 share the same mean with D_1 but have different variances. So their compatibility is relatively large. D_7 has a different mean compared to D_1 so its compatibility is relatively small.

3.7 Analysis of Benchmark Dose Data in Toxicology

The benchmark approach is a useful tool in toxicology. The benchmark dose (BMD) is defined as the dose of an environmental toxicant that corresponds to a prescribed change in response compared with the background response level. The toxicological data comprise n binomial responses $\boldsymbol{y} = (y_1, \ldots, y_n)$ with $y_i \sim B(n_i, p_i)$, where n_i is the number of animals

tested at dose level x_i and p_i is the probability that an animal gives an adverse response at dose level x_i

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}, i = 1, \ldots, n.$$

The Kociba study (Kociba et al., 1978) is a lifetime feeding study of both female and male Sprague Dawley rats, with 50 rats tested in each group at doses of 0, 1, 10, and 100 ng/kg/day. Inferences derived from the Kociba study have been widely used as the basis for risk assessments for 2,3,7,8-tetrachlorodibenzodioxin (TCDD). The National Toxicology Program (NTP) study (NTP, 1982) is a study in which groups of 50 male rats, 50 female rats, and 50 male mice received TCDD as a suspension in 9:1 corn oil:acetone by gavage twice each week to achieve doses of 0, 10, 50, or 500 ng/kg/week for two years. These exposures correspond to daily averaged doses of 0, 1.4, 7.1, or 71 ng/kg/day for rats. Liver tumor (neoplastic nodule) incidences of female rats from both studies, shown in Table 3.2, are chosen as the data for this study. Shao and Small (2011) showed that while the historical data (Kociba study) $D_0 = (n_0, \boldsymbol{y_0}, \boldsymbol{x_0})$ and the current data (NTP study) $D_1 = (n_1, \boldsymbol{y_1}, \boldsymbol{x_1})$ are not compatible in terms of all parameters, they are compatible in terms of one common parameter. Our method confirms their conclusion. Here are some basic settings.

$$\pi(\beta_0) \sim N(0, 10000),$$

$$\pi(\beta_1) \sim N(0, 10000),$$

$$\pi(\beta_0, \beta_1 | D) \propto \pi(\beta_0)\pi(\beta_1) \prod_i \left[\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right]^{y_i} \left[\frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)}\right]^{n_i - y_i}.$$

Table 3.3 shows the compatibility measure for parameters β_0 and β_1. When considering (β_0, β_1) together, the compatibility measure returns exactly 0 with 2,4,6 partition subsets, which means that they are not compatible at all. When only considering the intercept

Table 3.2: Benchmark Dose Data Summary and Parameter Estimates

Study	TCDD(ng/kg/day) and Response				Estimates	
Kociba	Control (or 0)	1	10	100	β_0(SD)	β_1(SD)
	9/86	3/50	18/50	34/48	-1.785(0.210)	0.028(0.004)
NTP	Control (or 0)	1.4	7.1	71	β_0(SD)	β_1(SD)
	5/75	1/49	3/50	12/49	-3.030(0.366)	0.026(0.007)

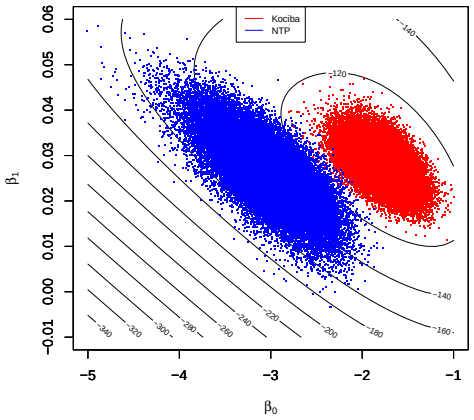
term, β_0, the compatibility measure also returns values close to 0. However, the data are very compatible with respect to the slope term β_1. Even when data sets are not compatible with respect to the full parameter vector, it is possible that they are compatible under a reduced set of parameters. For all the calculations, we build the partition with HPD regions and $\boldsymbol{p_1} = \boldsymbol{p_2} = (\frac{1}{K}, \ldots, \frac{1}{K})$.

Table 3.3: Compatibility for Parameters with Different Partition Subsets for Benchmark Dose Data

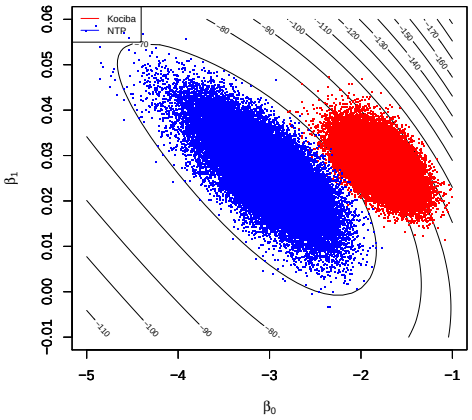
K	(β_0, β_1)	β_0	β_1
2	0	0.18%	85%
4	0	0.37%	88%
6	0	0.48%	89%

Figure 3.6(a)(b) plot the MCMC samples of (β_0, β_1) from the log posterior kernel given each data set. Red points are from the Kociba study and blue points are from the NTP study. Figure 3.6(a) has the contour lines based on the Kociba data, whereas Figure 3.6(b) has contour lines based on the NTP data. When we consider (β_0, β_1) together, we can easily distinguish two data sets using the estimated log posterior kernel values. Figure 3.6(c) plots the estimated marginal density of β_0 for these two data sets. They have different locations and thus are not compatible with each other. Figure 3.6(d) plots the estimated marginal density of β_1 for these two data sets. They share a similar location and are compatible with each other.

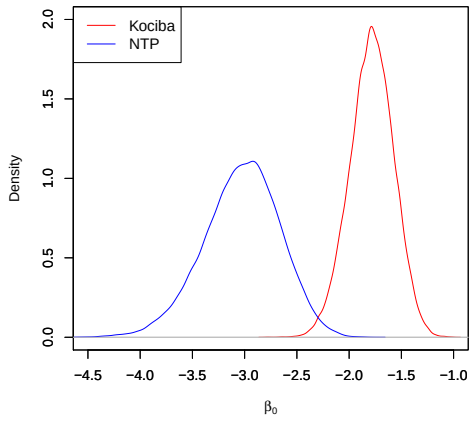
Figure 3.6: Benchmark Dose Data Plots



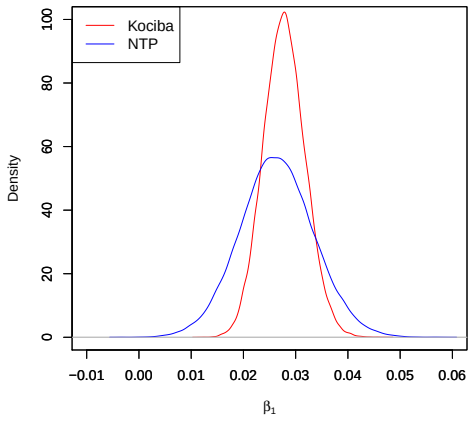
(a) Contour Plot Based on the Kociba Study



(b) Contour Plot Based on the NTP Study



(c) Marginal Density Plot of β_0



(d) Marginal Density Plot of β_1

Chapter 4

Information Gain

4.1 Information Gain Measure

When we already have some data D_1 to analyze, some extra knowledge D_0 comes in. This extra knowledge may be a historical data, a follow-up data, a partial missing data or some expert information. With this extra knowledge D_0, we will first evaluate whether D_1 and D_0 are compatible. If they are compatible, then we would ask how much information we can gain by combining them. Let $\boldsymbol{\theta}$ be the focal parameters, $\pi(\boldsymbol{\theta}|D_1)$ be the posterior distribution based on the data set D_1, $\pi(\boldsymbol{\theta}|D_1, D_0)$ be the posterior distribution combined with the extra knowledge D_0, $\boldsymbol{p_1} = (p_{11}, \ldots, p_{1K})$ be a pre-determined probability vector, $(\Omega_{11}, \ldots, \Omega_{1K})$ be the partition subsets built based on $\pi(\boldsymbol{\theta}|D_1)$. We calculate $\boldsymbol{p_{10|1}} = (p_{10|1,1}, \ldots, p_{10|1,K})$, where $p_{10|1,i} = \int_{\Omega_{1i}} \pi(\boldsymbol{\theta}|D_1, D_0)d\boldsymbol{\theta}$ for $i = 1, \ldots, K$.

Definition 4.1.1. Let I be the information gain when adding extra knowledge

$$I_{K,\alpha} = \zeta(\alpha)100\left(\frac{|H_K(\boldsymbol{p_1}) - H_K(\boldsymbol{p_{10|1}})|}{\log K}\right)\%,$$

where $\zeta(\alpha)$ is a sign function that determines the direction of information gain.

One choice of $\zeta(\alpha)$ is $\text{sign}\big(\int_{\Omega_{1\alpha}} \pi(\boldsymbol{\theta}|D_1, D_0)d\boldsymbol{\theta} - \alpha\big)$ where $\Omega_{1\alpha}$ is the $100\alpha\%$ HPD interval of $\pi(\boldsymbol{\theta}|D_1)$. With this setting, as long as the posterior distribution given combined data is more concentrated under level α, the information is gained.

$I_{K,\alpha}$ ranges from -100% to 100%. A positive $I_{K,\alpha}$ indicates that adding extra knowledge makes the combined distribution more concentrated. If the extra knowledge D_0 is really compatible with D_1, then combining them can reduce the variance of parameters $\boldsymbol{\theta}$. The posterior distribution given the combined data set should be more concentrated than the original posterior distribution. A negative $I_{K,\alpha}$ means adding extra knowledge leads to a mean-shift or less concentrated combined distribution. Therefore it indicates an information loss.

Remark 4.1.2. Here we still prefer to set $\boldsymbol{p_1} = (\frac{1}{K}, \ldots, \frac{1}{K})$. Under this setting, $I_{K,\alpha} = 0$ only when $\boldsymbol{p_{0|1}} = \boldsymbol{p_1}$. Also we can simplify the information gain under $\boldsymbol{p_1} = (\frac{1}{K}, \ldots, \frac{1}{K})$ as

$$I_{K,\alpha} = \zeta(\alpha)100\left(1 - \frac{H(\boldsymbol{p_{10|1}})}{\log K}\right)\%.$$

Proposition 4.1.3. When the posterior distributions of the original data and the combined data are both normal distributions with equal mean, under a two-subset partition with $\boldsymbol{p}_1 = (0.5, 0.5)$, among all the partition subsets of the form of $\Omega_1 = (a, b), \Omega_2 = (a, b)^C$, a HPD interval based partition maximizes the information gain.

The proof of this proposition is given in Appendix A. This proposition suggests that HPD interval construction is the most desirable choice for constructing partition subsets for measuring information gain.

Next we are going to apply this information gain measure on some real data examples to see how it works under two different scenarios, combining the historical data and incorporating the partial missing data.

4.2 Analysis of Benchmark Dose Data

Now we are going to visit the benchmark dose example again. Because we know β_1 is compatible between the Kociba study and the NTP study, we may combine the data sets together and quantify how much information we can gain. We also calculate the information gain of β_0 and (β_0, β_1) for comparison with β_1. When these two data sets are combined, we use the power prior method with the weight a_0. If we assume two models for those two data sets have the same parameters, we have the posterior kernel function

$$\pi(\beta_0, \beta_1, a_0|D_0, D_1) \propto \pi(\beta_0)\pi(\beta_1)[L(\beta_0, \beta_1|D_0)]^{a_0} L(\beta_0, \beta_1|D_1).$$

If we assume they share the same β_0, but have different β_1 values, then we use β_{10} for data D_0, β_{11} for data D_1, and we have

$$\pi(\beta_0, a_0|D_0, D_1) \propto \int \pi(\beta_0)\pi(\beta_{10})\pi(\beta_{11})[L(\beta_0, \beta_{10}|D_0)]^{a_0} L(\beta_0, \beta_{11}|D_1) d\beta_{10} d\beta_{11}.$$

If we assume they share the same β_1, but have different β_0 values, then we use β_{00} for data D_0, β_{01} for data D_1, and we have

$$\pi(\beta_1, a_0|D_0, D_1) \propto \int \pi(\beta_1)\pi(\beta_{00})\pi(\beta_{01})[L(\beta_{00}, \beta_1|D_0)]^{a_0} L(\beta_{01}, \beta_1|D_1) d\beta_{00} d\beta_{01}.$$

Information gain measure results are shown in Table 4.1 along with the parameter estimates given different choices of a_0.

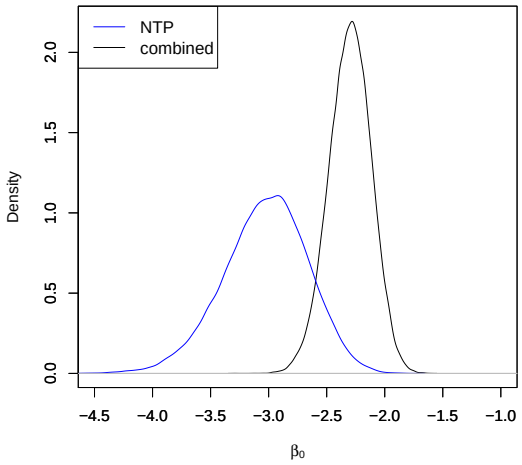
Table 4.1 shows that, for β_1 alone, combining the two data sets gains information and the standard deviation (SD) is reduced. In contrast, considering (β_0, β_1) jointly or β_0

Table 4.1: Combining Two Benchmark Dose Data Sets with Different Weights

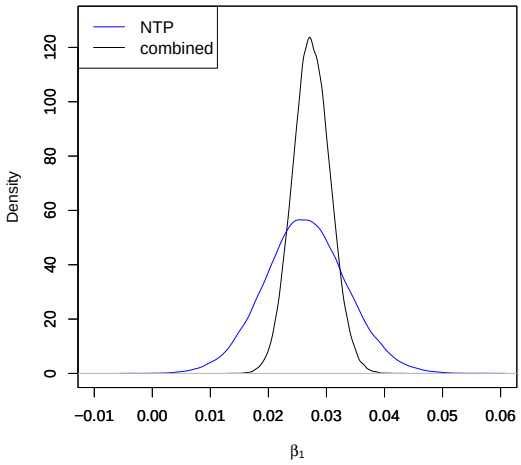
a_0		(β_0, β_1)	β_0	β_1	a_0		(β_0, β_1)	β_0	β_1
1	Estimate	(-2.298,0.027)	-2.214	0.027	0.5	Estimate	(-2.517,0.028)	-2.421	0.027
	SD	(0.181,0.003)	0.179	0.003		SD	(0.231,0.004)	0.226	0.004
	$I_{2,0.5}$	-99.9%	-98.0%	29.3%		$I_{2,0.5}$	-86.0%	-62.6%	12.7%
	$I_{4,0.25}$	-99.5%	-91.3%	18.5%		$I_{4,0.25}$	-73.4%	-46.7%	8.80%
	$I_{6,0.17}$	-99.0%	-88.4%	15.3%		$I_{6,0.17}$	-69.2%	-43.3%	7.46%
0.9	Estimate	(-2.332,0.027)	-2.245	0.027	0.4	Estimate	(-2.584,0.028)	-2.489	0.027
	SD	(0.190,0.003)	0.186	0.003		SD	(0.247,0.004)	0.241	0.005
	$I_{2,0.5}$	-99.8%	-95.8%	26.3%		$I_{2,0.5}$	-68.1%	-43.8%	9.05%
	$I_{4,0.25}$	-98.8%	-86.4%	16.9%		$I_{4,0.25}$	-54.2%	-30.7%	6.45%
	$I_{6,0.17}$	-98.0%	-83.1%	13.9%		$I_{6,0.17}$	-50.5%	-28.4%	5.54%
0.8	Estimate	(-2.369,0.027)	-2.278	0.027	0.3	Estimate	(-2.66,0.028)	-2.574	0.027
	SD	(0.197,0.003)	0.194	0.004		SD	(0.266,0.005)	0.261	0.005
	$I_{2,0.5}$	-99.5%	-93.0%	22.9%		$I_{2,0.5}$	-41.3%	-22.5%	5.78%
	$I_{4,0.25}$	-97.2%	-80.6%	15.0%		$I_{4,0.25}$	-30.1%	-14.9%	4.30%
	$I_{6,0.17}$	-95.7%	-76.5%	12.5%		$I_{6,0.17}$	-27.4%	-14.3%	3.81%
0.7	Estimate	(-2.412,0.027)	-2.320	0.027	0.2	Estimate	(-2.76,0.027)	-2.680	0.027
	SD	(0.206,0.004)	0.204	0.004		SD	(0.290,0.005)	0.283	0.006
	$I_{2,0.5}$	-98.3%	-86.2%	19.2%		$I_{2,0.5}$	-13.2%	-6.47%	2.97%
	$I_{4,0.25}$	-93.6%	-71.6%	12.8%		$I_{4,0.25}$	-8.61%	-3.97%	2.26%
	$I_{6,0.17}$	-91.2%	-67.5%	10.7%		$I_{6,0.17}$	-8.13%	-4.20%	2.04%
0.6	Estimate	(-2.460,0.028)	-2.365	0.027	0.1	Estimate	(-2.88,0.027)	-2.824	0.027
	SD	(0.218,0.004)	0.214	0.004		SD	(0.321,0.006)	0.316	0.006
	$I_{2,0.5}$	-94.7%	-77.4%	16.0%		$I_{2,0.5}$	-0.350%	-0.243%	0.754%
	$I_{4,0.25}$	-86.3%	-60.2%	11.0%		$I_{4,0.25}$	-0.190%	-0.144%	0.641%
	$I_{6,0.17}$	-82.8%	-56.4%	9.25%		$I_{6,0.17}$	-0.208%	-0.289%	0.605%

alone, combining two data sets loses information. Figure 4.1(a)(b) plots the estimated marginal density of β_0 and β_1 based on the combined data sets compared to the current NTP data respectively. Figure 4.2 plots the information gain as the weight a_0 changes from 0 to 1. We can draw the same conclusion from Figure 4.2 as from Table 4.1.

Figure 4.1: Benchmark Dose Combined Data Compared to NTP Data

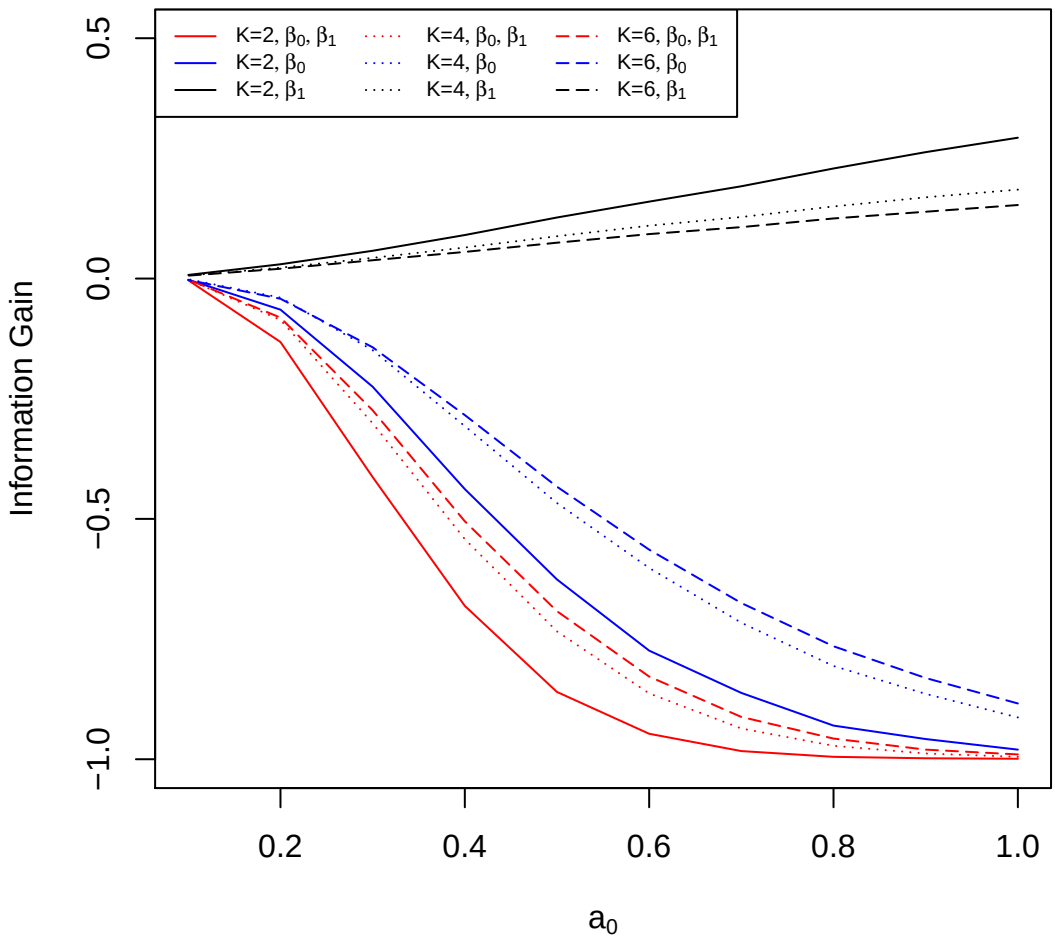


(a) Marginal Density Plot of β_0



(b) Marginal Density Plot of β_1

Figure 4.2: Information Gain with Different Weights for Benchmark Dose Data



4.3 Analysis of Six Cities Data

We consider the six cities longitudinal study of the health effects of respiratory function in children (Ware et al., 1984). This is a well known environmental dataset that has been analyzed extensively in the literature. The binary response is the wheezing status ($y = 0$ if no wheeze, $y = 1$ if wheeze) of a child at age 11. The wheezing status is modeled as a function of the city of residence (x_1) and smoking status of the mother (x_2). The covariate x_1 is a binary covariate which equals 1 if the child lived in Kingston–Harriman, Tennessee, the more polluted city, and 0 if the child lived in Portage, Wisconsin. The covariate x_2 is maternal cigarette smoking measured in number of cigarettes per day. There are 2394 subjects in the dataset. The covariate x_1 is missing for 32.8% of the cases, and x_2 is missing for 3.3% of the cases, and the total missing data fraction is 35.0%. Details of the dataset can be found in the original paper, we just present a brief summary of the data in Table 4.2.

Table 4.2: Summary of Six Cities Data

y	x_1	x_2	Frequency	y	x_1	x_2	Frequency
0	0	0	418	1	0	0	127
0	1	0	323	1	1	0	106
0	0	≥ 1	226	1	0	≥ 1	72
0	1	≥ 1	201	1	1	≥ 1	83
0	NA	NA	18	1	NA	NA	8
0	0	NA	19	1	0	NA	0
0	NA	0	369	1	NA	0	86
0	1	NA	24	1	1	NA	10
0	NA	≥ 1	229	1	NA	≥ 1	75

We only focus on a subset (2315 subjects) of the data, which includes all complete cases and the ones with only x_1 missing.

We use a logistic regression model for $[y|x_1, x_2]$, and thus take

$$\text{logit}\left(P(y_i = 1|x_{1i}, x_{2i}, \boldsymbol{\beta})\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}.$$

We further model $[x_1|x_2]$ by a logistic regression for handling the missing data problem. Specifically, we take $(x_{1i}|x_{2i})$ to have independent Bernoulli distributions each with success probability

$$P(x_{1i} = 1|x_{i2}, \boldsymbol{\alpha}) = \frac{\exp(\alpha_0 + \alpha_1 x_{2i})}{1 + \exp(\alpha_0 + \alpha_1 x_{2i})},$$

for $i = 1, \ldots, n$ where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$. The proposed joint prior distribution for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \pi(\boldsymbol{\alpha}|d_0)\pi(\boldsymbol{\beta}|c_0),$$

where $\boldsymbol{\alpha}|d_0 \sim N_2(\mathbf{0}, d_0 I_2)$ and $\boldsymbol{\beta}|c_0 \sim N_3(\mathbf{0}, c_0 I_3)$. Let r_i be the missing indicator and $r_i = 1$ if x_{1i} is missing, otherwise, $r_i = 0$.

Then the joint likelihood function based on complete case data D_{cc} is

$$\begin{aligned}
L(\boldsymbol{\alpha}, \boldsymbol{\beta}|D_{cc}) &= \prod_{i:r_i=0} f(y_i|x_{1i}, x_{2i}, \boldsymbol{\beta}) f(x_{1i}|x_{2i}, \boldsymbol{\alpha}), \text{ where} \\
f(y_i|x_{1i}, x_{2i}, \boldsymbol{\beta}) &= \frac{\exp[y_i(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})]}{1 + \exp(\beta_0 + \beta_2 x_{1i} + \beta_2 x_{2i})}, \\
f(x_{1i}|x_{2i}, \boldsymbol{\alpha}) &= \frac{\exp[x_{1i}(\alpha_0 + \alpha_1 x_{2i})]}{1 + \exp(\alpha_0 + \alpha_1 x_{2i})}.
\end{aligned}$$

The joint likelihood function based on all data D_{ac} is

$$\begin{aligned}
L(\boldsymbol{\alpha}, \boldsymbol{\beta}|D_{ac}) &= L(\boldsymbol{\alpha}, \boldsymbol{\beta}|D_{cc}) L_1(\boldsymbol{\alpha}, \boldsymbol{\beta}|D_{ac} \setminus D_{cc}), \text{ where} \\
L_1(\boldsymbol{\alpha}, \boldsymbol{\beta}|D_{ac} \setminus D_{cc}) &= \prod_{i:r_i=1} [f(y_i|x_{1i} = 1, x_{2i}, \boldsymbol{\beta}) f(x_{1i} = 1|x_{2i}, \boldsymbol{\alpha}) \\
&\quad + f(y_i|x_{1i} = 0, x_{2i}, \boldsymbol{\beta}) f(x_{1i} = 0|x_{2i}, \boldsymbol{\alpha})].
\end{aligned}$$

For the missing part, we just sum over all the possible values of x_{1i}. The posterior distributions of complete cases $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}|D_{cc})$ and all cases $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}|D_{ac})$ are defined as:

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}|D_{cc}) \propto L(\boldsymbol{\alpha}, \boldsymbol{\beta}|D_{cc})\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}),$$

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}|D_{ac}) \propto L(\boldsymbol{\alpha}, \boldsymbol{\beta}|D_{ac})\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}).$$

Two MCMC samples of size 100000 are generated from two posterior kernels, respectively, under $c_0 = d_0 = 100$. A summary of the estimates is given in Table 4.3. We can calculate the compatibility and information gain for each parameter based on the MCMC samples (Table 4.4).

Table 4.3: Posterior Mean and Standard Deviation of Parameters for Six Cities Data

D_{cc}			D_{ac}		
	Mean	SD		Mean	SD
β_0	-1.252	0.090	β_0	-1.336	0.081
β_1	0.144	0.118	β_1	0.144	0.118
β_2	0.011	0.005	β_2	0.013	0.004

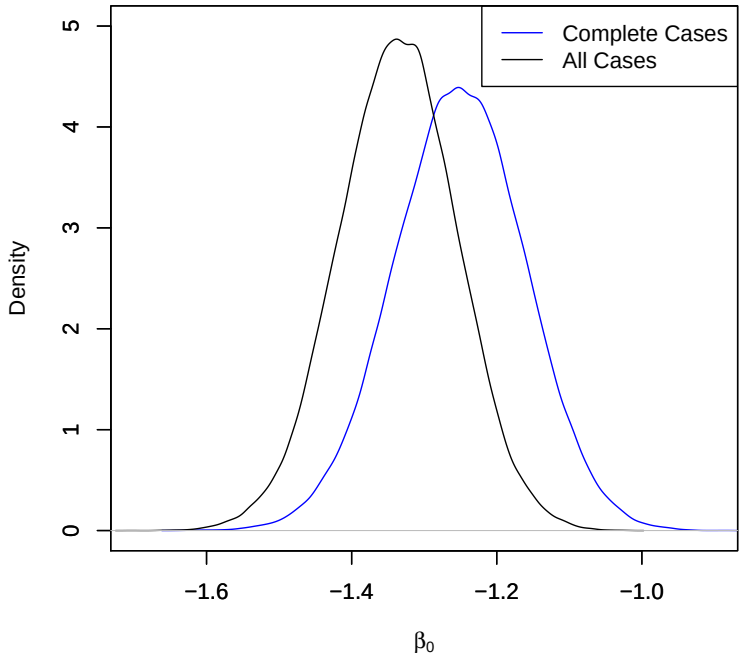
Table 4.4: Compatibility and Information Gain of Parameters for Six Cities Data

	β_0	β_1	β_2
M_2	91.30%	100.00%	98.08%
M_4	92.14%	100.00%	98.38%
M_6	92.77%	100.00%	98.47%
$I_{2,0.5}$	-7.49%	0.00%	0.39%
$I_{4,0.25}$	-5.84%	0.00%	0.37%
$I_{6,0.17}$	-5.24%	0.00%	0.33%

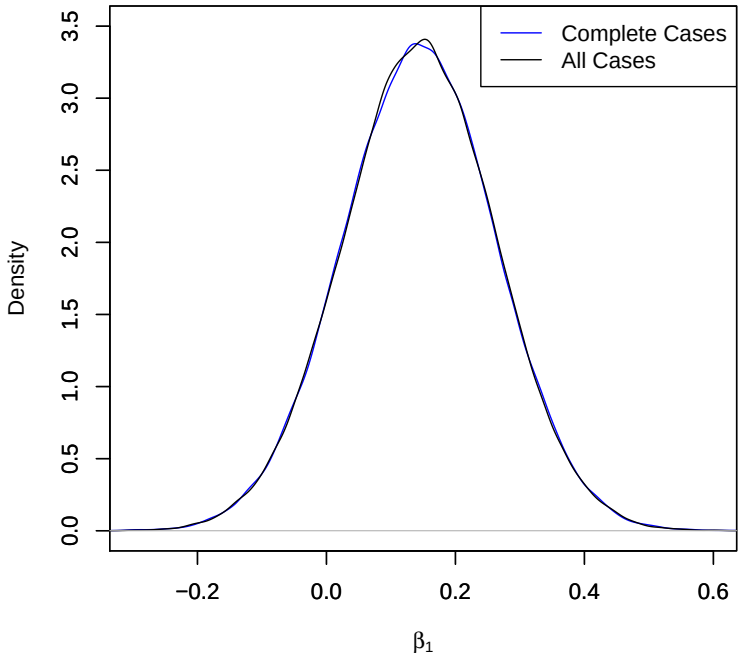
Figure 4.3(a)(b)(c) plots the estimated marginal density of β_0, β_1 and β_2 based on the complete cases compared to all cases respectively. We can see that β_1 and β_2 from complete cases and all cases are very similar, therefore their compatibility measures are over 98% and their information gain are less than 0.4%. For β_0, there is some difference

between the estimates from complete cases and all cases. The compatibility for β_0 is around 92% and we have about 5% information loss. Overall, we don't gain much for adding 32.7% missing data with only x_1 missing. The only difference is in the intercept term.

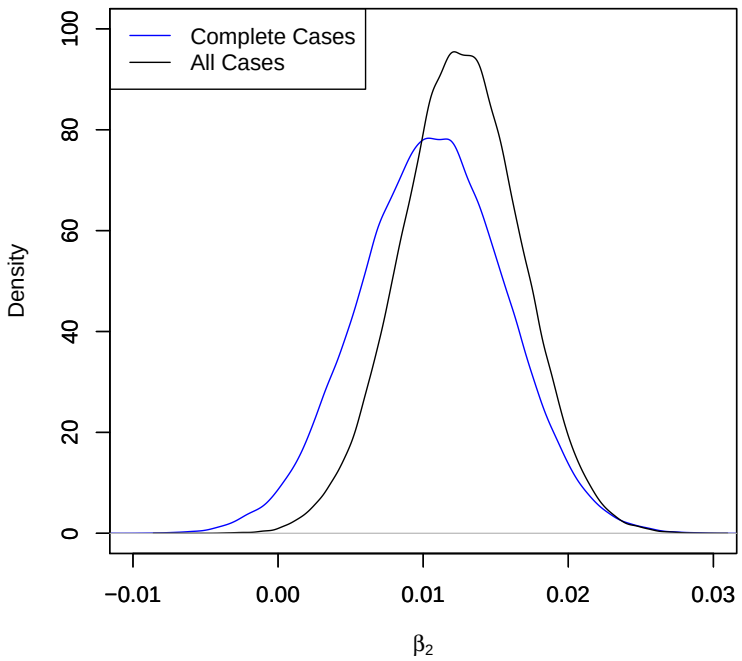
Figure 4.3: Posterior Densities from Complete Cases and All Cases for Six Cities Data



(a) Marginal Density Plot of β_0



(b) Marginal Density Plot of β_1



(c) Margianl Density Plot of β_2

4.4 Analysis of Melanoma Clinical Trials Data

Next we consider an Eastern Cooperative Oncology Group (ECOG) phase III melanoma clinical trials E1690 example. E1690 was intended as a confirmatory trial comparing high dose interferon (IFN) to observation (OBS). There are a total of 427 patients enrolled. Results of the E1690 trial have been published in Kirkwood et al. (2001). In Kirkwood et al. (2001), frequentist Cox regression analysis, deleting the cases with missing covariates, were used throughout. Models incorporating a surviving fraction have become quite common for cancers such as melanoma, since for these diseases, cure fractions can typically range from 30% to 60%. Chen et al. (2002) carried out a Bayesian analysis of E1690 based on the observed data posterior, i.e., the posterior distribution incorporating the missing covariates. The response variable is relapse-free survival (RFS) time y, which may be right censored. The covariates are treatment (x_1: IFN, OBS), age (x_2), sex (x_3), logarithm of Breslow depth of the tumor (x_4), logarithm of size of the primary tumor (x_5), and type of the primary tumor (x_6). A brief summary of the data is provided in Table 4.5. Covariates x_1, x_3 and x_6 are all binary covariates, whereas x_2, x_4 and x_5 are all continuous. The covariates x_1, x_2 and x_3 are completely observed, while x_4, x_5 and x_6 have missing values. The total missing data fraction is 28.6%, where 25.3% of the cases had exactly one missing covariate, 3.0% had exactly two missing covariates, 0.2% had exactly three missing covariates. The regression coefficients corresponding to the covariates $x_1, \ldots, x_6$ are denoted by $\beta_1, \ldots, \beta_6$, respectively. An intercept term is also included in the model, and let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_6)$.

Following Chen et al. (2002), with a finite partition of the time axis, $0 < s_1 < \ldots < s_J$, where $s_J > \max(y)$, we have J intervals $(0, s_1], (s_1, s_2], \ldots, (s_{J-1}, s_J]$. In the jth interval,

Table 4.5: Summary of Melanoma Clinical Trials Data

Complete Observed Variables		
y	Censored 186	Relapsed 241
x_1	OBS 212	IFN 215
x_2	Mean 47.9	SD 13.2
x_3	Male 268	Female 159
Missing Covariates		
x_4	Observed 417	Missing 10
x_5	Observed 347	Missing 80
x_6	Observed 380	Missing 47

we assume a constant hazard λ_j. The complete data likelihood function can be written as

$$
\begin{aligned}
L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_{cc}) &= \prod_{i=1}^{n} L(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{x}_i, y_i, N_i), \\
L(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{x}_i, y_i, N_i) &= \prod_{j=1}^{J} \exp\Big\{ -(N_i - v_i)\delta_{ij}\big[\lambda_j(y_j - s_{j-1}) + \sum_{g=1}^{j-1} \lambda_g(s_g - s_{g-1})\big]\Big\} \\
&\quad \times \prod_{j=1}^{J} (N_i\lambda_j)^{\delta_{ij}v_i} \exp\Big\{ -v_i\delta_{ij}\big[\lambda_j(y_i - s_{i-1}) + \sum_{g=1}^{j-1} \lambda_g(s_g - s_{g-1})\big]\Big\} \\
&\quad \times \exp[N_i\boldsymbol{x}_i'\boldsymbol{\beta} - \log(N_i!) - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})],
\end{aligned}
$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)$, N_i is the number of metastasis-competent cells for the i th subject, $\delta_{ij} = 1$ if the ith subject failed or was censored in the jth interval, and 0 otherwise, $\boldsymbol{x}_i' = (x_{i1}, \ldots, x_{i6})$ denotes the covariates for the ith subject.

The posterior kernel for the complete cases is

$$
p(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_{cc}) \propto L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_{cc})\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}),
$$

where we choose non-informative prior $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}) = 1$.

For model incorporating missing covariate values, we need to specify marginal densities. We take

$$\begin{aligned}
p(x_4, x_5, x_6|x_1, x_2, x_3, \boldsymbol{\alpha}) =& p(x_6|x_1, x_2, x_3, x_4, x_5, \boldsymbol{\alpha}_6) \\
& \times p(x_5|x_1, x_2, x_3, x_4, \boldsymbol{\alpha}_5)p(x_4|x_1, x_2, x_3, \boldsymbol{\alpha}_4), \\
P(x_6 = 1|x_1, x_2, x_3, x_4, x_5, \boldsymbol{\alpha}_6) =& \frac{\exp(\alpha_{60} + \alpha_{61}x_1 + \alpha_{62}x_2 + \alpha_{63}x_3 + \alpha_{64}x_4 + \alpha_{65}x_5)}{1 + \exp(\alpha_{60} + \alpha_{61}x_1 + \alpha_{62}x_2 + \alpha_{63}x_3 + \alpha_{64}x_4 + \alpha_{65}x_5)}, \\
x_5|x_1, x_2, x_3, x_4, \boldsymbol{\alpha}_5 \sim& N(\alpha_{50} + \alpha_{51}x_1 + \alpha_{52}x_2 + \alpha_{53}x_3 + \alpha_{54}x_4, \sigma_5^2), \\
x_4|x_1, x_2, x_3, \boldsymbol{\alpha}_4 \sim& N(\alpha_{40} + \alpha_{41}x_1 + \alpha_{42}x_2 + \alpha_{43}x_3, \sigma_4^2),
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\alpha} =& (\boldsymbol{\alpha}_4, \boldsymbol{\alpha}_5, \boldsymbol{\alpha}_6), \\
\boldsymbol{\alpha}_4 =& (\alpha_{40}, \ldots, \alpha_{43}, \sigma_4^2), \\
\boldsymbol{\alpha}_5 =& (\alpha_{50}, \ldots, \alpha_{54}, \sigma_5^2), \\
\boldsymbol{\alpha}_6 =& (\alpha_{60}, \ldots, \alpha_{65}).
\end{aligned}$$

Again we use the missing indicator and set $r_i = 1$ if at least one of x_{4i}, x_{5i}, x_{6i} is missing, otherwise, $r_i = 0$. Then the likelihood functions and posterior kernels of all cases are

$$\begin{aligned}
L(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\alpha}|D_{ac}) =& L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D_{cc})L_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\alpha}|D_{ac} \setminus D_{cc}), \text{where} \\
L_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\alpha}|D_{ac} \setminus D_{cc}) =& \prod_{r_i=1} \big[\int_{\boldsymbol{x_i}} \sum_{N_i} L(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{x}_i, y_i, N_i) \big] p(x_4, x_5, x_6|\boldsymbol{\alpha}) d\boldsymbol{x_i}, \\
\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\alpha}|D_{ac}) \propto& L(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\alpha}|D_{ac})\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\alpha}).
\end{aligned}$$

We choose prior $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}) \propto \sigma_4^{-7}\sigma_5^{-7}$.

Our main interest is in posterior inferences about $\boldsymbol{\beta}$, with $\boldsymbol{\lambda}, \boldsymbol{\alpha}$ being treated as nuisance parameters. With $J = 10$, we generate two MCMC samples of size 10000 from two posterior kernels. Table 4.6 shows results for $\boldsymbol{\beta}$ with complete cases and all cases.

Table 4.6: Posterior Mean and Standard Deviation of Parameters for Melanoma Clinical Trials Data

	D_{cc}			D_{ac}	
	Mean	SD		Mean	SD
β_0	0.150	0.153	β_0	0.131	0.134
β_1	-0.073	0.150	β_1	-0.205	0.133
β_2	0.025	0.078	β_2	0.097	0.066
β_3	-0.065	0.163	β_3	-0.145	0.139
β_4	0.025	0.086	β_4	0.026	0.074
β_5	0.081	0.075	β_5	0.101	0.075
β_6	-0.191	0.163	β_6	-0.034	0.149

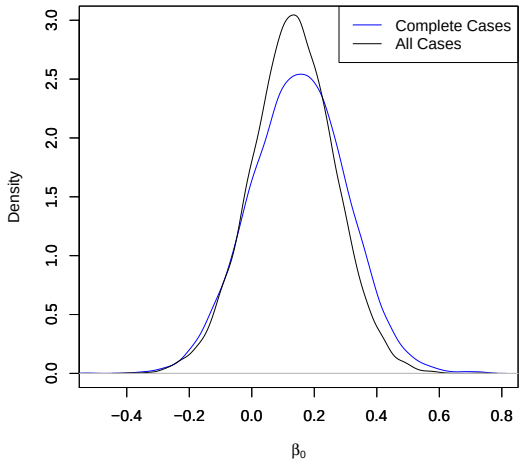
The marginal compatibility measure and information gain are given in Table 4.7.

Table 4.7: Compatibility and Information Gain of Parameters for Melanoma Clinical Trials Data

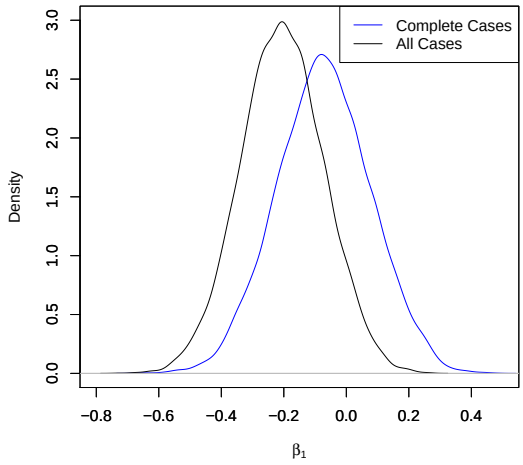
	β_0	β_1	β_2	β_3	β_4	β_5	β_6
M_2	98.73%	91.49%	89.95%	98.56%	99.12%	99.92%	88.58%
M_4	98.99%	93.42%	92.03%	98.46%	99.07%	99.88%	91.46%
M_6	99.15%	94.11%	92.75%	98.55%	99.13%	99.87%	91.43%
$I_{2,0.5}$	1.16%	-6.19%	-6.26%	0.00%	0.85%	-0.14%	-10.36%
$I_{4,0.25}$	0.94%	-4.12%	-4.77%	0.31%	0.94%	-0.11%	-6.25%
$I_{6,0.17}$	0.71%	-3.57%	-4.16%	0.15%	0.82%	-0.20%	-7.00%

Figure 4.4(a)(b)(c)(d)(e)(f)(g) plots the estimated marginal density of $\beta_0, \ldots, \beta_6$ based on the complete cases compared to all cases respectively. From Table 4.6 and Figure 4.4, we can see this time including the missing data really makes some differences in the coefficients, not only the intercept. For β_1, β_2 and β_6, the compatibility measure is around 90% and the information loss ranges from 3% to 10%, which indicating the missing mechanism may not be missing complete at random and we should model the missing pattern.

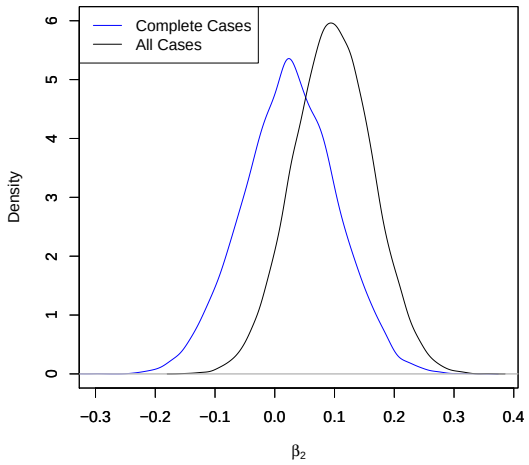
Figure 4.4: Posterior Densities from Complete Cases and All Cases for Melanoma Clinical Trials Data



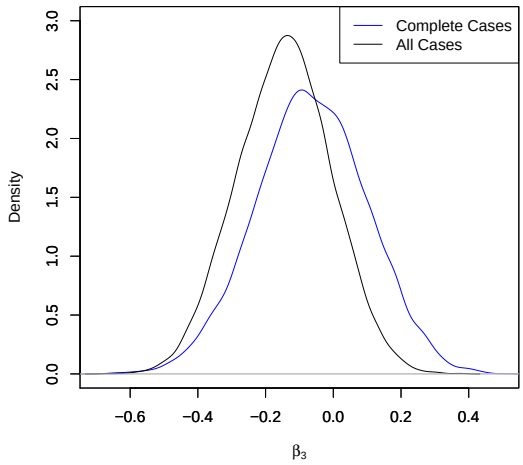
(a) Marginal Density Plot of β_0



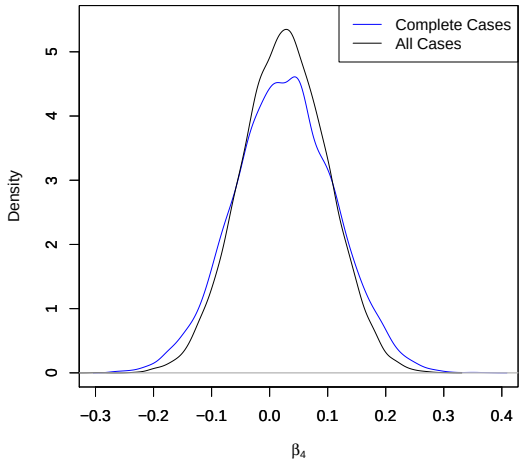
(b) Marginal Density Plot of β_1



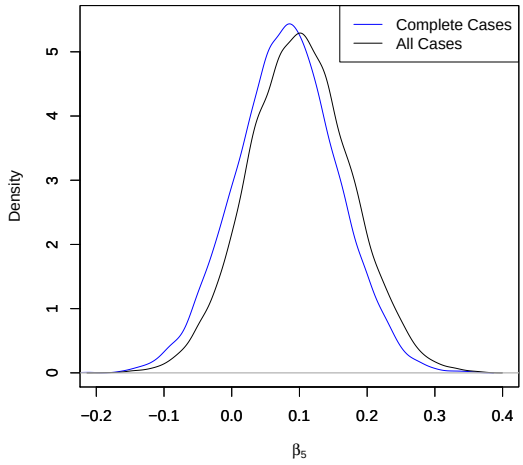
(c) Marginal Density Plot of β_2



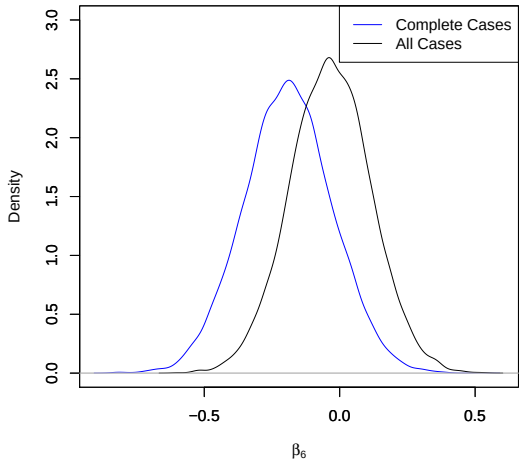
(d) Marginal Density Plot of β_3



(e) Marginal Density Plot of β_4



(f) Marginal Density Plot of β_5



(g) Marginal Density Plot of β_6

4.5 Further Applications

There are many fields that our measures can be applied to. In survival analysis, using the exact survival time usually gives us very precise estimates. However, when the exact survival time is not available, the interval censoring data has to be used. We can use the compatibility measure to compare the estimates from the exact data and the censored data and use the information gain measure to get the information loss due to the censoring. Let D_{exact} and D_{censor} be the exact survival time data and the censored data. By replacing D_1 to D_{censor} and (D_1, D_0) to D_{exact}, one can get the information gain from **Definition 4.1.1**. Similar techniques also can be applied to compare between censored data with different interval precisions. For example, if the information gain from the daily observed data comparing to the weekly observed data is not much, we can save the costs and get a similar result by just using the weekly observed data.

In the missing data problem, we can use information gain measure to test whether adding partial missing data will gain. If there is almost nothing gain, as is shown in **Section 4.3 Analysis of Six Cities Data**, the data is probably missing at random and we can just make inference based on the complete part and save the efforts to impute or marginalize the missing part. If there is some information gain, as is shown in **Section 4.4 Analysis of Melanoma Clinical Trials Data**, we should definitely include the missing part and make a better inference. If there is information loss, the missing mechanism probably is not missing at random and we should model that missing pattern.

In phylogenetics, methods have been proposed by Lewis et al. (2016) to measure information content with respect to the discrete tree topology component of the model;

however, evolutionary biologists need ways of inferring marginal information gain for particular continuous parameters of interest. Of special interest is divergence time estimation. Estimating the (calendar) time since a lineage split (e.g. determining the age of major lineages of organisms, such as the age of green plants, mammals, vertebrates, or even eukaryotes) can only be done in a Bayesian context because nucleotide sequence data only provides information about sequence divergence, which is the product of substitution rate and time. Complex interactions among separate prior distributions assumed for individual fossils result in a joint induced prior that is difficult to appreciate intuitively. Measuring the information gain from prior to posterior guards against making conclusions based solely on the prior in cases where the data provide little information about divergence times, and negative information gain would warn the researcher that the chosen prior conflicts with the data.

In clinical trials, when we want to study rare diseases, there may not be enough samples to perform trials. As we already showed in **Section 4.2 Analysis of Benchmark Dose Data**, it is important to borrow information from the historical data or similar trials. Our measures can tell whether we can borrow and how to borrow to get a more powerful analysis with limited samples.

Chapter 5

Concluding Remarks, Extensions and Future Research Works

5.1 Concluding Remarks

To overcome the undesirable negativity issue of the Shannon differential entropy, we introduce the fractional size adjusted entropy in **Section 2.1**. It has good properties, such as non-negativity, additivity and invariant under one-to-one linear transformation. It always exists with proper choices of k. To directly involve the scale parameter, we extend the previous concept to the generalized fractional size adjusted entropy. With proper choices of parameters k_1, k_2, it always exists and maintains to be non-negative. It also includes many measures as its special cases.

We also propose a partition-based data compatibility measure and an information gain measure. The partition method uses the posterior distribution of the parameters and constructs a partition on the parameter space using the HPD regions. Using the compatibility measure we can assess whether the two data sets are compatible in terms of all parameters or some common parameters. Our information gain measure is useful

in deciding whether to combine two data sets. It is especially useful when the current analysis needs to borrow some historical data. In the benchmark dose data example, we illustrate how the compatibility measure successfully picks the compatible parameter and the information gain measure provides suggestions on how to choose weights for the power prior method.

The compatibility measure and information gain can also be applied to compare two distributions directly. Although in most of our examples we are comparing data sets, the posterior distributions are the real components used in the measures. We can use the compatibility measure to compare whether two distributions are close, or use the partial compatibility measure to determine whether two marginal distributions are close.

5.2 Extension of the Compatibility Measure to the Multiple Data Comparison

When we have more than two data sets, we can first calculate the pairwise compatibility measure for all the combinations of the data sets. Then we can use different summary statistics to describe the overall compatibility. Choosing the maximal value indicating the largest compatibility among all the data sets while choosing the minimal value indicating the smallest compatibility among all the data sets. Choosing the mean or the median can describe the average compatible level. If we have weights for all the data sets, we can also use the weighted mean.

Another approach is to combine data sets. Suppose we have three data sets D_1, D_2, D_3, we can combine (D_1, D_2) to D_{12} and calculate the compatibility measure of D_{12} and D_3. We can do the similar calculations by combining (D_1, D_3) and (D_2, D_3). Then we can calculate the average of all the compatibility measures based on the combined data.

5.3 Extension of the Compatibility Measure and the Information Gain to the Bayesian Hierarchical Modeling

Bayesian hierarchical modeling is a statistical model written in multiple levels (hierarchical form) that estimates the parameters of the posterior distribution using the Bayesian method. Suppose we have two data sets D_1, D_2. By assigning different parameters, we can write their likelihood functions as $f(D_1|\boldsymbol{\theta}_1)$ and $f(D_2|\boldsymbol{\theta}_2)$. The Bayesian hierarchical model assumes that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ further follow a higher level distribution $\pi(\boldsymbol{\theta}_1|\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}_2|\boldsymbol{\theta})$. With a prior distribution $\pi(\boldsymbol{\theta})$, we can write the posterior distribution as

$$\pi(\boldsymbol{\theta}|D_1, D_2) \propto \int \pi(\boldsymbol{\theta}) \prod_{i=1}^{2} f(D_i|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i|\boldsymbol{\theta}) d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2. \tag{5.1}$$

To apply our compatibility measure, we consider another model

$$\pi(\boldsymbol{\theta}|D_1, D_2) \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^{2} f(D_i|\boldsymbol{\theta}). \tag{5.2}$$

In model (5.2) we assume two data sets sharing the exact same parameters while in model (5.1) we use the hierarchical structure to model the parameters. We can use our compatibility measure to directly work on the MCMC samples from these two posterior distributions of $\boldsymbol{\theta}$. The compatibility measure based on the hierarchical modeling can automatically apply to multiple data sets by simply letting $i = 1, \ldots, K$ in model (5.1) and (5.2).

The information gain measure can also be applied to the hierarchical modeling. Suppose we have the posterior distribution of $\boldsymbol{\theta}$ given in (5.1). Given another data set D_3, we can write the updated posterior distribution

$$\pi(\boldsymbol{\theta}|D_1, D_2, D_3) \propto \int \pi(\boldsymbol{\theta}) \prod_{i=1}^{3} f(D_i|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i|\boldsymbol{\theta}) d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2 d\boldsymbol{\theta}_3. \tag{5.3}$$

With (5.1) and (5.3), we can calculate how much information gain by adding D_3.

5.4 Future Research Works to the Choices of k value in the Fractional Size Adjusted Entropy and (k_1, k_2) values in the Generalized Fractional Size Adjusted Entropy

For the fractional size adjusted entropy, we suggest to start with $k = 1$. If under certain situations that the fractional size adjusted entropy doesn't exist, we can change k to other values. For the generalized fractional size adjusted entropy, if one already knows the closed form of the distribution, then k_1, k_2 can be chosen to satisfy equation (2.1). If the closed form is unknown, one has to try multiple combinations of k_1, k_2 to make the generalized fractional size adjusted entropy non-negative.

Our next step is to study the performance of the generalized size adjusted entropy with different combinations of k_1, k_2. With different k_1, k_2 values, the order of the generalized size adjusted entropies from different distributions is different. More theoretical properties have to be studied before we obtaining a guideline for the choices of k_1, k_2. For example, when the kernel value $f(x)$ is smaller than 1, a larger $k(> 1)$ will make the $f^k(x)$ smaller, while when $f(x)$ is larger than 1, $f^k(x)$ will be larger. For distributions with different shapes, we may assign different k_1, k_2 to reveal the different properties and return useful entropies.

5.5 Future Research Works to Using the Generalized Fractional Size Adjusted Entropy to Bayesian Prior Selection

Jaynes (1957) first introduced the principle of maximum entropy when he emphasized a natural correspondence between statistical mechanics and information theory. This principle is often used to obtain prior distributions for Bayesian inference. The maximum

entropy distribution represents the least informative information. For example, the exponential distribution is the maximum entropy distribution when the mean of the parameter is fixed, the normal distribution is the maximum entropy distribution when both the mean and the variance are fixed and the uniform distribution is the maximum entropy distribution when there is no constraint. Since the principle of maximum entropy is based on the Shannon entropy, we can apply it to the generalized fractional size adjusted entropy. We will study the least informative prior distributions with different constraints based on the generalized fractional size adjusted entropy.

Appendix A

Proofs of Theorems

Proof of Property 2.1.2.

Proof. For a random variable X with a PDF or PMF $f(x)$, its FASE is $-E(\log \frac{f^k(X)}{E[f^k(x)]})$. Since log function is concave function, $-\log$ is a convex function. By Jenson's inequality, we have

$$-E(\log \frac{f^k(X)}{E[f^k(X)]}) \geq -\log E[\frac{f^k(X)}{E[f^k(X)]}] = -\log 1 = 0.$$

The equality sign holds only if X is point mass or uniformly distributed. □

Proof of Property 2.1.3.

Proof. Suppose random variables X, Y are independent and they have correspondingly PDF or PMF $g(x), h(y)$. Their joint PDF or PMF is $f(x, y)$, where $f(x, y) = g(x)h(y)$.

$$\begin{aligned}
FSAE(X, Y, k) &= -E\Big[\log \frac{f^k(X, Y)}{E[f^k(X, Y)]}\Big] \\
&= -E\Big[\log \frac{g^k(X)h^k(Y)}{E[g^k(X)h^k(Y)]}\Big] \\
&= -E[\log g^k(X)] - E[\log h^k(Y)] + \log E[g^k(X)] + \log E[h^k(Y)] \\
&= -E\Big[\log \frac{g^k(X)}{E[g^k(X)]}\Big] - E\Big[\log \frac{h^k(Y)}{E[h^k(Y)]}\Big] \\
&= FSAE(X, k) + FSAE(Y, k).
\end{aligned}$$

□

Proof of Property 2.1.4.

Proof. Suppose X is a random variable with PDF or PMF $f(x)$. And let $Y = aX+b, a \neq 0$ with PDF or PMY $g(y)$. When X is discrete, we have

$$\begin{aligned} g(Y) &= f\Big(\frac{Y-b}{a}\Big) \\ FSAE(Y,k) &= -\sum \Big(\log \frac{g^k(y)}{\sum g^k(y)g(y)}\Big)g(y) \\ &= -\sum \Big(\log \frac{f^k(\frac{y-b}{a})}{\sum f^k(\frac{y-b}{a})f(\frac{y-b}{a})}\Big)f(\frac{y-b}{a}) \\ &= FSAE(X,k). \end{aligned}$$

When X is continuous, we have

$$\begin{aligned} g(Y) &= \frac{1}{a}f\Big(\frac{Y-b}{a}\Big) \\ FSAE(Y,k) &= -\int \log \frac{g^k(y)}{\int g^k(y)g(y)dy}g(y)dy \\ &= -\int \log \frac{\frac{f^k(\frac{y-b}{a})}{a^k}}{\int \frac{f^k(\frac{y-b}{a})}{a^k}f(\frac{y-b}{a})d(\frac{y-b}{a})}f(\frac{y-b}{a})d(\frac{y-b}{a}) \\ &= -\int \log \frac{f^k(x)}{\int f^k(x)f(x)dx}f(x)dx \\ &= FSAE(X,k). \end{aligned}$$

□

Proof of Proposition 4.1.3.

Proof. Suppose $\pi(\theta|D_1) \sim N(\mu, \sigma_1^2)$, $\pi(\theta|D_1, D_0) \sim N(\mu, \sigma_2^2)$ and $\sigma_1^2 > \sigma_2^2$.

Assuming a symmetric bipartition $\boldsymbol{p_1} = (0.5, 0.5)$, we can build the partition as

$$\Omega_1 = (\mu + z_{\alpha_1}\sigma_1, \mu + z_{\alpha_2}\sigma_1),$$

where z_{α_1} and z_{α_2} are the corresponding α_1, α_2 percentiles of normal distribution and $\alpha_2 - \alpha_1 = 0.5$. With $K = 2$ and $\alpha = 0.5$, $\boldsymbol{p_{10|1}} = (p_{10|1,1}, p_{10|1,2})$, where

$$p_{10|1,1} = \Phi(\frac{\sigma_1}{\sigma_2} z_{\alpha_2}) - \Phi(\frac{\sigma_1}{\sigma_2} z_{\alpha_1}).$$

Since $\frac{\sigma_1}{\sigma_2} > 1$, $p_{10|1,1} > 0.5$, we have

$$\begin{aligned}
I_{K,\alpha} &= \zeta(\alpha) 100 \left(\frac{|H_K(\boldsymbol{p_1}) - H_K(\boldsymbol{p_{10|1}}|)}{\log K} \right)\% \\
&= 100 \left(\frac{|H_2(0.5, 0.5) - H_2(p_{10|1,1}, p_{10|1,2})|}{\log 2} \right)\% \\
&= 1 - \frac{H_2(p_{10|1,1}, p_{10|1,2})}{\log 2} \\
&= 1 + \frac{p_{10|1,1} \log(p_{10|1,1}) + (1 - p_{10|1,1}) \log(1 - p_{10|1,1})}{\log 2}.
\end{aligned}$$

From Figure 3.1, maximizing I is equivalent to maximizing $p_{10|1,1}$. To simplify the notation, we let $c = \frac{\sigma_1}{\sigma_2}, x_1 = z_{\alpha_1}, x_2 = z_{\alpha_2}$. Since $p_{10|1,1} = \Phi(cx_2) - \Phi(cx_1)$, we obtain

$$\begin{aligned}
\frac{d}{dx_1} p_{10|1,1} &= \frac{d}{dx_1} (\Phi(cx_2) - \Phi(cx_1)) \\
&= c\phi(cx_2) \frac{dx_2}{dx_1} - c\phi(cx_1) \\
&= \frac{c}{\sqrt{2\pi}} \left[\exp(-\frac{x_1^2 + c^2 x_2^2 - x_2^2}{2}) - \exp(-\frac{c^2 x_1^2}{2}) \right].
\end{aligned}$$

Observing that $\Phi(x_2) = \Phi(x_1) + 0.5$, we can compute $\frac{dx_2}{dx_1}$.

If $x_1^2 + c^2 x_2^2 - x_2^2 > c^2 x_1^2$, $x_2^2 > x_1^2$ and then $\frac{d}{dx_1} p_{10|1,1} < 0$.

If $x_1^2 + c^2 x_2^2 - x_2^2 < c^2 x_1^2$, $x_2^2 < x_1^2$ and then $\frac{d}{dx_1} p_{10|1,1} > 0$.

Thus, I is maximized when $x_1^2 = x_2^2$. In other words, when the intervals are symmetric (HPD), we have the maximal information gain. $\square$

Bibliography

Ali, S. M., and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*: 131-142.

Awad, A. M., and Alawneh, A. J.(1987). Application of entropy to a life-time model. *IMA Journal of Mathematical Control and Information*, 4(2): 143-148.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3): 200-217.

Chen, M. H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *Journal of the American Statistical Association*, 89(427): 818-824.

Chen, M.-H., and Shao, Q.-M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, 8(1): 69-92.

Chen, M. H., Ibrahim, J. G., and Lipsitz, S. R. (2002). Bayesian methods for missing covariates in cure rate models. *Lifetime Data Analysis*, 8(2): 117-146.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432): 1313-1321.

Chib, S., and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96(453): 270-281.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37(3/4): 256-266.

Conway, K. P., et al. (2014). Data compatibility in the addiction sciences: an examination of measure commonality. *Drug and Alcohol Dependence*, 141: 153-158.

Cover, T. M. and Thomas, J. A. (2012). *Elements of Information Theory*. John Wiley & Sons.

Csiszár, I., and Shields, P. C. (2004). *Information Theory and Statistics: A Tutorial*. Foundations and Trends® in Communications and Information Theory, 1(4): 147-528.

Lefebvre, G., Steele, R., and Vandal, A. (2010). A path sampling identity for computing the Kullback-Leibler and J divergences. *Computational Statistics and Data Analysis*, 54(7): 17191731.

Gelfand, A. E., and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*: 501-514.

Gelfand, A. E., Smith, A. F., and Lee, T. M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, 87(418): 523-532.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society* :1317-1339.

Hartley, R. V. (1928). Transmission of information. *Bell Labs Technical Journal*, 7(3): 535-563.

Higgins, J., and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11): 1539-1558.

Higgins, J., et al. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, 327: 557-560.

Ibrahim, J. G, and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science* :46-60.

Ibrahim, J. G., Chen, M. H., Gwon, Y., and Chen, F. (2015). The power prior: theory and applications. *Statistics in Medicine*, 34(28): 3724-3749.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4):620-630.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3): 227-241.

Kluge, A. G., and Farris, J. S. (1969). Quantitative phyletics and the evolution of anurans. *Systematic Biology*, 18(1): 1-32.

Kirkwood, J. M., Ibrahim, J. G., Sosman, J. A., Sondak, V. K., Agarwala, S. S., Ernstoff, M. S., and Rao, U. (2001). High-dose interferon alfa-2b significantly prolongs relapse-free and overall survival compared with the GM2-KLH/QS-21 vaccine in patients with resected stage IIB-III melanoma: results of intergroup trial E1694/S9512/C509801. *Journal of Clinical Oncology*, 19(9): 2370-2380.

Kittaneh, O. A., Khan, M. A., Akbar, M., and Bayoud, H. A. (2016). Average entropy: a new uncertainty measure with application to image segmentation. *The American Statistician*, 70(1): 18-24.

Kociba, R.J., et al. (1978). Results of a two-year chronic toxicity and oncogenicity study of 2, 3, 7, 8-tetrachlorodibenzo-p-dioxin in rats. *Toxicology and Applied Pharmacology*, 46(2): 279-303.

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1): 79-86.

Lartillot, N., and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology*, 55(2): 195-207.

Lewis, L. A., and Trainor, F. R. (2012). Survival of Protosiphon botryoides (Chlorophyceae, Chlorophyta) from a Connecticut soil dried for 43 years. *Phycologia*, 51(6): 662-665.

Lewis, P. O., Chen, M. H., Kuo, L., Lewis, L. A., Fukov, K., Neupane, S., Wang, Y. B., and Shi, D.(2016). Estimating Bayesian phylogenetic information content. *Systematic Biology*, 65(6): 1009-1023.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*: 986-1005.

Meacham, C. A. (1981). A manual method for character compatibility analysis. *Taxon*: 591-600.

Morimoto, T. (1963). Markov processes and the H-theorem. *Journal of the Physical Society of Japan*, 18(3): 328-331.

Newton, M. A., and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*: 3-48.

National Toxicology Program. (1982). Carcinogenesis bioassay of 2, 3, 7, 8-tetrachlorodibenzo-p-dioxin (CAS No. 1746-01-6) in Osborne-Mendel rats and B6C3F1 mice (Gavage Study). *National Toxicology Program Technical Report Series*, 209: 1.

Petris, G., and Tardella, L. (2003). A geometric approach to transdimensional Markov chain Monte Carlo. *Canadian Journal of Statistics*, 31(4): 469-482.

Rao, M., Chen, Y., Vemuri, B. C., and Wang, F. (2004). Cumulative residual entropy: a new measure of information. *IEEE Transactions on Information Theory*, 50(6): 1220-1228.

Rényi, A. (1961). On measures of entropy and information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Robert, C. P., and Wraith, D. (2009). Computational methods for Bayesian model choice. In *AIP Conference Proceedings*, 1193(1): 251-262.

Shannon, C. E. (1948). A mathematical theory of communication, Part I, Part II. *Bell Syst. Tech. J.*, 27: 623-656.

Shao, K., and Small, M. J. (2011). Potential uncertainty reduction in model-averaged benchmark dose estimates informed by an additional dose study. *Risk Analysis*, 31(10), 1561-1575.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M. H. (2010). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, 60(2): 150-160.

Wang, Y. B., Chen, M. H., Kuo, L., and Lewis, P. O. (2016). Adaptive Partition Weighted Monte Carlo Estimation.

Wang, Y. B., Chen, M. H., Kuo, L., and Lewis, P. O.(2017). A New Monte Carlo Method for Estimating Marginal Likelihoods. *Bayesian Analysis*.

Ware, J. H., Dockery, D. W., Spiro III, A., Speizer, F. E., and Ferris Jr, B. G. (1984). Passive Smoking, Gas Cooking, and Respiratory Health of Children Living in Six Cities 1-3. *American Review of Respiratory Disease*, 129(3):366-374.