New Partition based Measures for Data Compatibility and Information Gain

Ming-Hui Chen

Department of Statistics, University of Connecticut
E-Mail: ming-hui.chen@uconn.edu

Joint Work with Daoyuan Shi, Lynn Kuo, and Paul Lewis

Present in 2022 ISBA World Meeting

Montreal, CA

June 28, 2022

Outline

1 Introduction

2 Data Compatibility

3 Information Gain

4 Applications

A Case Study on the Benchmark Approach 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 comprises 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,

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

The Two Benchmark Studies in Toxicology

- 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 (National Toxicology Program 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.
- In this analysis, the data were from liver tumor (neoplastic nodule) incidences of female rats from both studies.

Benchmark 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)

- Are Kociba data and NTP data comparable?
- Which of these two data sets has more information?
- If we treat the Kociba study as the historical data and the NTP study the current data, how much information can we leverage from the Kociba data while analyzing the NTP data?

Six Cities Data

- This is a well known environmental dataset from a study of the health effects of respiratory function in children (Ware et al., 1984).
- The binary response is the wheezing status of a child at age 11 with $y = 0$ representing no wheeze and $y = 1$ if wheeze.
- The wheezing status is modeled as a function of the city of residence (x_1) and the smoking status of the mother (x_2).
- The city of residence 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 by 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%.

Summary of Six Cities Data

y	x_1	x_2	Count	y	x_1	x_2	Count
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

The Models for Six Cities Data

- We use a logistic regression model for $[y|x_1, x_2]$, i.e.,

$$P(y_i = 1|x_{1i}, x_{2i}, \boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}.$$

- We further model $[x_1|x_2]$ by a logistic regression to handle the missing data problem. Specifically, let $[x_{1i}|x_{2i}]$ have independent Bernoulli distributions each with probability

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

  where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$.

- Is the complete case analysis adequate? How much information do the missing data contribute?

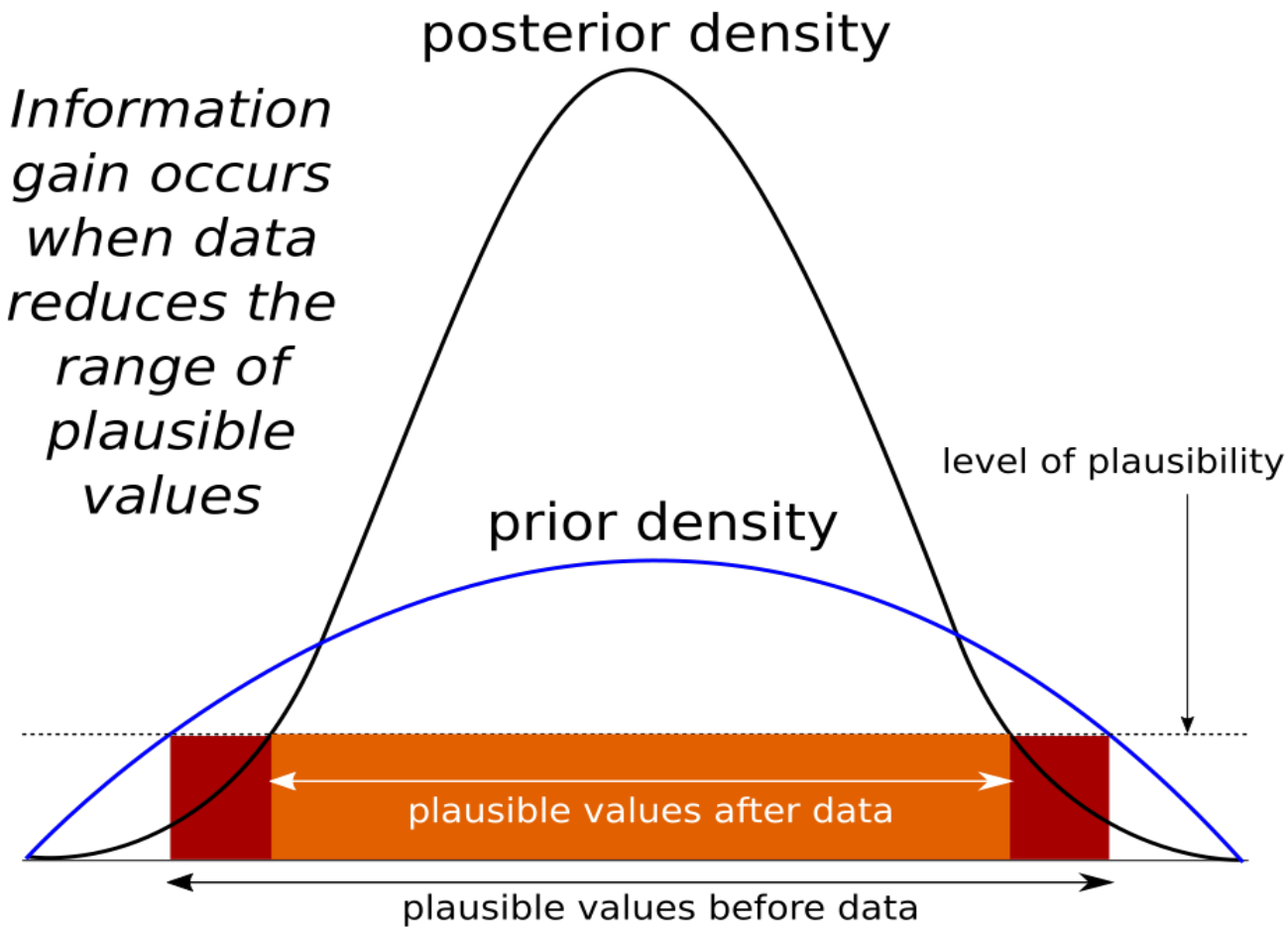
Why Comptibility?

- 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 conflicts among data sets are important to systematists (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 et al. 2003).

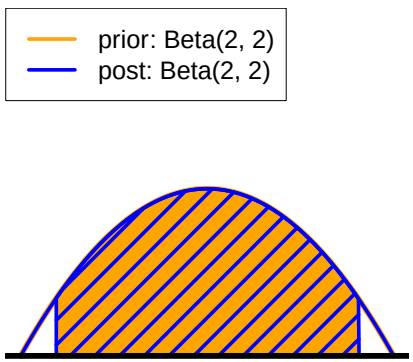
Concept of Information Gain

- When we add a new data set B to the historical data set A, information is gained if the posterior distribution given A+B has lower variability than the posterior distribution given A alone.
- In this case, the new data B allows some possible values of the parameters previously considered to be plausible (on the basis of data A alone) to be effectively excluded from consideration.
- Information is lost if the posterior distribution given A+B has higher variability than the posterior distribution given A alone, indicating that the effect of new data B has been to increase, not decrease, the number of plausible values of the parameters.
- Figure 1 shows a graphical illustration of information gain.

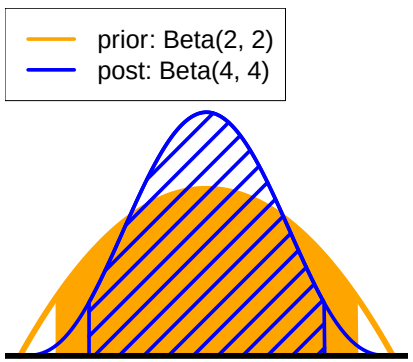
Graphical depiction of information gain



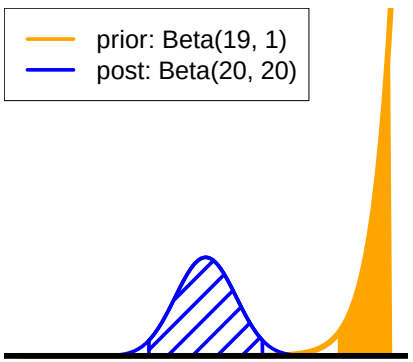
Examples of information gain



(a) No Change



(b) Gain



(c) Lost

- Reference: Shi et al.(2021, Bayesian Analysis).

Outline

1 Introduction

2 Data Compatibility

3 Information Gain

4 Applications

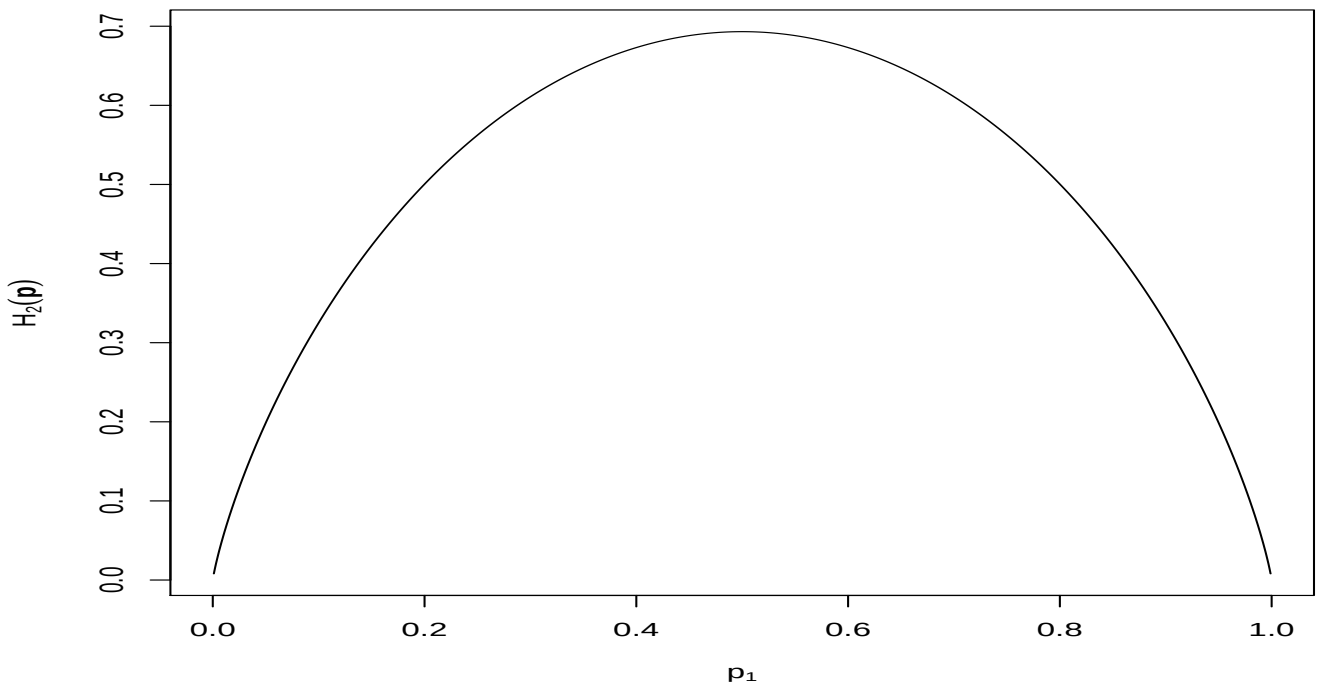
Entropy

- 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 entropy is defined as

$$H(X) = E[-\log f(X)] = -\sum_{i=1}^{K} f(x_i) \log f(x_i),$$

  where $0 \log 0 = 0$.
- 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 $\sum_{i=1}^{K} p_i = 1$.

Plot of $H_2(\boldsymbol{p})$ as a function of p_1 when $K = 2$



- $H_2(\boldsymbol{p})$ 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.

Partition and HPD

- Consider a Bayesian posterior density having the form

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

  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 the kernel function.

- 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 $i = 1, \ldots, K$.
- One way to construct a partition is to use the highest posterior density (HPD) regions.

Partition and HPD (continued)

- For a given probability p, the $100p\%$ HPD region is the subset Ω_p^* of Ω, which takes the form of

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

  where $k(p)$ is the largest constant such that $\int_{\Omega_p^*} \pi(\boldsymbol{\theta}|D) d\boldsymbol{\theta} \geq p$.

- Given a vector of probabilities $(p_1, \ldots, p_K)$, we can construct a vector of HPD regions $(\Omega_{p_1}^*, \ldots, \Omega_{p_K}^*)$ 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

$$\Omega_1 = \Omega_{p_1}^*,$$
$$\Omega_i = \Omega_{p_i}^* \cap (\Omega_{p_{i-1}}^*)^C, \ i = 2, \ldots, K.$$

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

Posterior Distributions

- Our measure of compatibility is based on 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

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

Compatibility of D_2 based on D_1

- 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 $\boldsymbol{p_{2|1}}$.
- The compatibility of data set D_2 based on data set D_1 with K partition subsets is defined as

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

Compatibility of D_1 based on D_2

- 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 $\boldsymbol{p}_{\mathbf{1|2}} = (p_{1|2,1}, \ldots, p_{1|2,K})$, 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.
- The compatibility of data set D_1 based on data set D_2 with K subsets is defined as

$$M_K(D_1|D_2) = 100\left(1 - \frac{|H_K(\boldsymbol{p}_2) - H_K(\boldsymbol{p}_{\mathbf{1|2}})|}{\log K}\right).$$

- Here $M_K(D_2|D_1)$ 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.

Compatibility of D_1 and D_2

- The compatibility M of two data sets D_1, D_2 is

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

- Notice that $M_K(D_1, D_2)$ 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 that 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 $M_K(D_1|D_2) = M_K(D_2|D_1) = 0$. This can happen only when each of the two distributions $\pi(\boldsymbol{\theta}|D_1)$ and $\pi(\boldsymbol{\theta}|D_2)$ 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 $H_K(\boldsymbol{p_{1|2}}) = H_K(\boldsymbol{p_2})$.

Remarks

- For K partitions, 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.
- For example, let the posterior distribution given data D_1 be $N(0, 1)$ and the posterior distribution given data D_2 be $N(1, 1)$.
- If we have two partition subsets and choose $p_{11} = 0.5976, p_{21} = 0.5976$, then $M_2(D_1, D_2) = 1$, which means the two posterior distributions are compatible.
- This, of course, is not true. The reason is that $p_{2|1,1} = p_{1|2,1} = 0.4024$, and $H_2(0.5976, 0.4024) - H_2(0.4024, 0.5976) = 0$.
- When $p_{2|1,1} = 1 - p_{11}$, inference is confounded due to the duality of the entropy function.
- Choosing $p_{11} = 0.5$ solves the problem. We thus recommend using $p_{11} = p_{21} = 0.5$ under two partitions and $\boldsymbol{p_1} = \boldsymbol{p_2} = (\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K})$ under K partitions.

Computational Algorithm

- Step 1. Draw an MCMC sample $\{\boldsymbol{\theta}_{\mathbf{1}}^{(t)}\}_{t=1,\ldots,N_1}$ from $\pi(\boldsymbol{\theta}|D_1)$, and independently draw an MCMC sample $\{\boldsymbol{\theta}_{\mathbf{2}}^{(t)}\}_{t=1,\ldots,N_2}$ from $\pi(\boldsymbol{\theta}|D_2)$.
- Step 2. Sort the N_1 values of $\{\pi(\boldsymbol{\theta}_{\mathbf{1}}^{(t)}|D_1)\}$ from small to large. Let $a_t = \pi(\boldsymbol{\theta}_{\mathbf{1}}^{(t)}|D_1)$. 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

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

  where $[N_1 \sum_{j=1}^{i} p_{1j}]$ is the largest integer that is less or equal to $N_1 \sum_{j=1}^{i} p_{1j}$.
- 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 $\hat{\boldsymbol{\Omega}}_{\mathbf{1}}$:

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

- Step 5. Calculate $M_K(D_2|D_1) = 100\big(1 - \frac{|H_K(\boldsymbol{p_1}) - H_K(\hat{\boldsymbol{p}}_{\mathbf{2|1}})|}{\log K}\big)$.

Computational Algorithm

- Step 6. Sort the N_2 values of $\{\pi(\boldsymbol{\theta}_2^{(t)}|D_2)\}$ from small to large. Let $b_t = \pi(\boldsymbol{\theta}_2^{(t)}|D_2)$. Then the sorted values are $\{b_{(1)}, \ldots, b_{(N_2)}\}$. Based on $\boldsymbol{p_2} = (p_{21}, \ldots, p_{2K})$ calculate the HPD Regions $\hat{\boldsymbol{\Omega}}_{\boldsymbol{2}}^{*} = (\hat{\Omega}_{21}^{*}, \ldots, \hat{\Omega}_{2K}^{*})$ such that

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

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

$$\hat{\Omega}_{21} = \hat{\Omega}_{21}^{*}, \hat{\Omega}_{2i} = \hat{\Omega}_{2i}^{*} \cap (\hat{\Omega}_{2,i-1}^{*})^{C}, \ i = 2, \ldots, K.$$

- Step 8. Calculate the proportion $\hat{\boldsymbol{p}}_{\boldsymbol{1|2}} = (\hat{p}_{1|2,1}, \ldots, \hat{p}_{1|2,K})$ of $\{\boldsymbol{\theta}_{\boldsymbol{1}}^{(t)}\}$ in each corresponding region of $\hat{\boldsymbol{\Omega}}_{\boldsymbol{2}}$, where $\hat{p}_{1|2,i} = \frac{\text{Number of } (\boldsymbol{\theta}_{\boldsymbol{1}}^{(t)} \in \hat{\Omega}_{2i})}{N_1}, i = 1, \ldots, K.$
- 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 $M_K(D_1, D_2) = \frac{M_K(D_2|D_1) + M_K(D_1|D_2)}{2}$.

Partial Compatibility

- Sometimes two data sets may be compatible only with respect to some, but not all, parameters in their models. Therefore, 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 let (ϕ, ξ) denote the parameters in the posterior distribution corresponding to the data set D_1, and also let (ϕ, ξ^*) denote the parameters in the posterior distribution corresponding to the data set D_2, where ϕ are the parameters of interest shared by these two posterior distributions.
- We calculate the marginal distribution of ϕ from D_1, D_2. Here ξ and ξ^* can be the same or different, but we are only interested in ϕ.

$$\pi(\phi|D_1) = \int \pi(\phi, \xi|D_1)d\xi, \quad \pi(\phi|D_2) = \int \pi(\phi, \xi^*|D_2)d\xi^*.$$

- Applying our method to $\pi(\phi|D_1)$ and $\pi(\phi|D_2)$ provides the partial compatibility of the common parameters ϕ between D_1 and D_2.

Outline

1 Introduction

2 Data Compatibility

3 Information Gain

4 Applications

Preliminary

- Let D_1 be the focal data set and let D_0 be the historical data. With the historical data 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 parameter(s), let $\pi(\boldsymbol{\theta}|D_1)$ be the posterior distribution based on the data set D_1, and let $\pi(\boldsymbol{\theta}|D_1, D_0)$ be the posterior distribution combined with the historical data D_0. Here $\boldsymbol{p_1} = (p_{11}, \ldots, p_{1K})$ is still a pre-determined probability vector, and $(\Omega_{11}, \ldots, \Omega_{1K})$ are 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 $p_{10|1,i} = \int_{\Omega_{1i}} \pi(\boldsymbol{\theta}|D_1, D_0)d\boldsymbol{\theta},\ i = 1, \ldots, K.$

Definition of Information Gain

- I is the information gained by adding background 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 the 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. With varying α, this $\zeta(\alpha)$ actually returns a consistent sign.

Properties of $I_{K,\alpha}$

- $I_{K,\alpha}$ ranges from -100% to 100%.
- A positive $I_{K,\alpha}$ indicates that adding background knowledge makes the combined distribution more concentrated.
- If the historical data 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 background knowledge leads to a mean-shift or less concentrated combined distribution. Therefore it indicates information loss.

Properties of $I_{K,\alpha}$

- 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 $\boldsymbol{p_{0|1}} = \boldsymbol{p_1}$, allowing us to simplify information gain as

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

- *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-based partition maximizes the information gain.*

Outline

1 Introduction

2 Data Compatibility

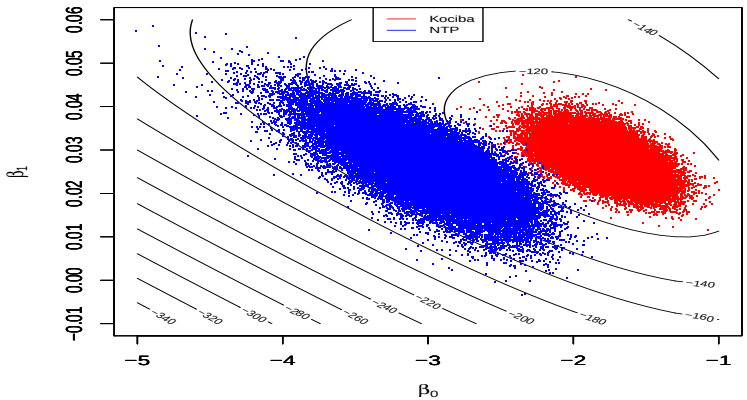
3 Information Gain

4 Applications

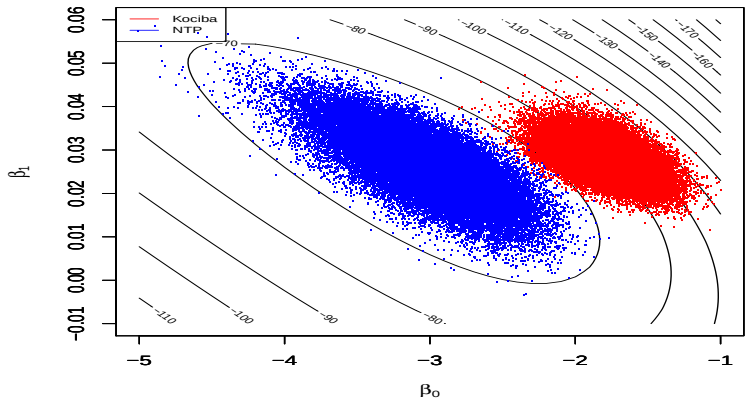
Analysis of Benchmark Dose Data

- When considering (β_0, β_1) together, the compatibility measure returns exactly 0 with $K = 2, 4, 6$ partition subsets, which means that they are not compatible at all.
- When only considering the intercept term, β_0, the compatibility measure also returns a small value: $M = 0.18\%$, 0.37%, and 0.48% when $K = 2$, 4, and 6, respectively.
- The two data sets are, however, very compatible with respect to the slope term β_1: $M = 85\%$, 88%, and 89% when $K = 2$, 4, and 6, respectively. Even when the 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 $\boldsymbol{p_1} = \boldsymbol{p_2} = (\frac{1}{K}, \ldots, \frac{1}{K})$.

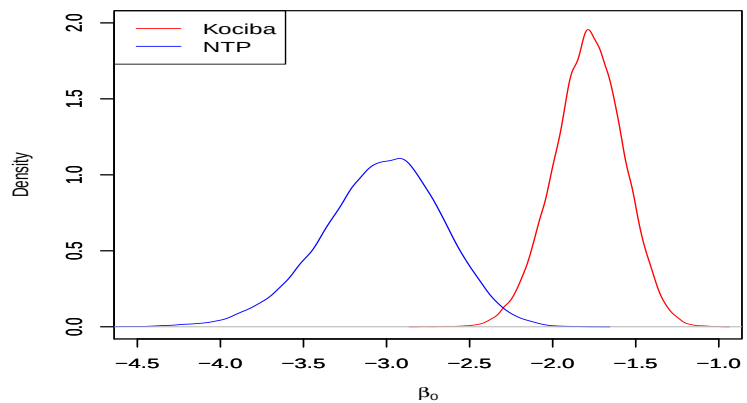
Benchmark data plots



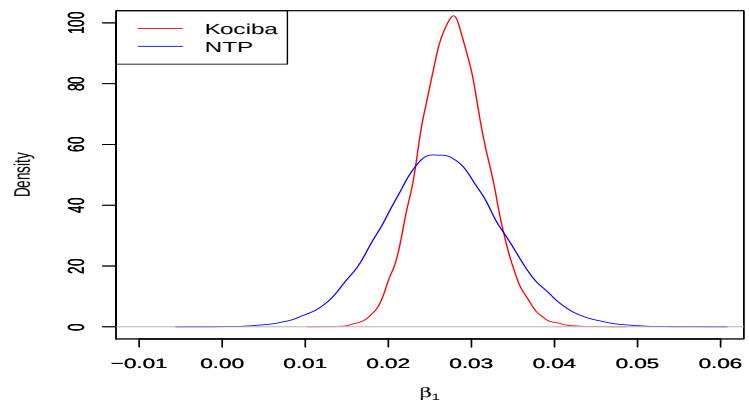
(a) Contour plot based on Kociba study



(b) Contour plot based on NTP study



(c) Density plot of β_0



(d) Density plot of β_1

Information Gain from Kociba

- Since we know β_1 is compatible between these two data sets, we may combine the data sets together and quantify information gain.
- We also calculate the information gain measure for 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 weight a_0.
- If we assume they have the same parameters, we have the kernel function:

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

  where D_1 denotes the NTP data and D_0 is the Kociba data.
- 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

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

- 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

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

Combining two 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	Est	(-2.298,0.027)	-2.214	0.027	0.5	Est	(-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%
	c	0.003	0.033	0.896		c	0.027	0.144	0.910
0.8	Est	(-2.369,0.027)	-2.278	0.027	0.3	Est	(-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%
	c	0.007	0.056	0.896		c	0.101	0.308	0.921
0.6	Est	(-2.460,0.028)	-2.365	0.027	0.1	Est	(-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%
	c	0.017	0.106	0.907		c	0.467	0.670	0.943

c is proposed by Presanis (2013) (small $\rightarrow$ two posteriors are different.)

Analysis of Six Cities Data

- 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 assume that

$$f(y_i|x_{1i}, x_{2i}, \boldsymbol{\beta}) = \frac{\exp[y_i(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})]}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})},$$
$$f(x_{1i}|x_{2i}, \boldsymbol{\alpha}) = \frac{\exp[x_{1i}(\alpha_0 + \alpha_1 x_{2i})]}{1 + \exp(\alpha_0 + \alpha_1 x_{2i})}.$$

- The posterior distributions of complete cases $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}|D_{cc})$ and all cases $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}|D_{ac})$ are given by

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

  Independent $N(0, 100)$ priors are specified for all of the model parameters.

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%
c	0.48	1.00	0.76

Concluding Remarks

- We 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.
- When constructing the partitions, HPD regions are preferred because of the simplicity for the algorithm.
- In the information gain measurement, for $\zeta(\alpha)$ we recommend $\alpha = \frac{1}{K}$, and we show in Appendix III that $\zeta(\alpha)$ is fairly robust to the choice of α.
- In practice, we recommend $K \geq 4$ partitions for both compatibility and information gain measures.

Concluding Remarks

- 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.
- This presentation is based on

  Shi, D., Chen, M.-H., Kuo, L., and Lewis, P.O. (2021). New partition based measures for data compatibility and information gain. *Statistics in Medicine, 40*, 3560-3581.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. DEB-1354146. Dr. M.-H. Chen's research was also partially supported by NIH grants #GM70335 and #P01CA142538.

Thank you !