An Estimation of Generalized Bradley-Terry Models Based on the \textit{em} Algorithm

Yu Fujimoto  
yu.fujimoto@it.aoyama.ac.jp  
Department of Integrated Information Technology, Aoyama Gakuin University,  
Chuo, Sagamihara, Kanagawa 252-5258, Japan

Hideitsu Hino  
hideitsu.hino@toki.waseda.jp

Noboru Murata  
noboru.murata@eb.waseda.jp  
School of Science and Engineering, Waseda University, Shinjuku,  
Tokyo 169-8555, Japan

The Bradley-Terry model is a statistical representation for one’s preference or ranking data by using pairwise comparison results of items. For estimation of the model, several methods based on the sum of weighted Kullback-Leibler divergences have been proposed from various contexts. The purpose of this letter is to interpret an estimation mechanism of the Bradley-Terry model from the viewpoint of flatness, a fundamental notion used in information geometry. Based on this point of view, a new estimation method is proposed on a framework of the \textit{em} algorithm. The proposed method is different in its objective function from that of conventional methods, especially in treating unobserved comparisons, and it is consistently interpreted in a probability simplex. An estimation method with weight adaptation is also proposed from a viewpoint of the sensitivity. Experimental results show that the proposed method works appropriately, and weight adaptation improves accuracy of the estimate.

1 Introduction

The Bradley-Terry (BT) model (Bradley & Terry, 1952) is a basic tool for representing user preference or ranking data from pairwise comparisons of many items. Let us consider, for example, rating a large number of wine brands. In general, it is difficult to establish a complete ranking of all the brands; however, for any given pair, we may choose a preferred one after comparatively tasting them. In such a situation, results are obtained as pairwise comparisons, as shown in Table 1. For instance, in a comparison of brands A and B, six tasters prefer brand A, and eight tasters prefer brand B. We can stochastically interpret the result that A is chosen 6 times
Table 1: Pairwise Comparison Results of Wine Brands.

<table>
<thead>
<tr>
<th>Preferred Brand</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>—</td>
<td>6</td>
<td>8</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>—</td>
<td>7</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>9</td>
<td>—</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>—</td>
<td>7</td>
</tr>
<tr>
<td>E</td>
<td>7</td>
<td>8</td>
<td>5</td>
<td>8</td>
<td>—</td>
</tr>
</tbody>
</table>

in 14 comparisons. The BT model aims at estimating the overall ranking of items (such as wine brands) based on the pairwise preference data. Statistical ranking models related to the BT model are useful for studying the mechanism of human preference in the brain and have attracted the attention of researchers over many years (Plackett, 1975; Huang, Weng, & Lin, 2006; Hino, Fujimoto, & Murata, 2010). BT models have also been used for formulating classification problems (Hastie & Tibshirani, 1998; Zadrozny, 2002). Estimation methods of the BT model have been discussed from several contexts, most of them based on the sum of weighted Kullback-Leibler (KL) divergences (Hastie & Tibshirani, 1998; Huang et al., 2006). However, their objective functions are somewhat contrived, because the functions are focused simultaneously on different domains of probability distributions.

We first interpret an estimation mechanism of the BT model from the viewpoint of information geometry (Amari & Nagaoka, 2000). In our formulation, observations for the BT model form a kind of flat data manifold; flatness and data manifold are fundamental notions used in information geometry. With these notions, we reorganize the KL-based estimation method within a framework of the \textit{em} algorithm (Amari, 1995), an information-geometrical alternative of the EM algorithm (McLachlan & Krishnan, 1996). Then an estimation process of the BT model is geometrically interpreted as a sequence of projections in a probability simplex. An estimation result based on our formulation is intuitive and tractable because it is derived consistently in a discussion of a single space of probability distributions given by a simplex. In the context of a classification problem, a similar attempt has been proposed by Ikeda (2006) to integrate binary classifiers. Our formulation is applicable not only to binary models but also to other multiclass models associated with similar flat data manifolds.

In addition, we discuss the selection of a set of weights for the divergence, which is very important to avoid overfitting. Based on sensitivity, is an important concept in robust statistics (Hampel, Ronchetti, Rousseuw, & Stahel, 1986), we introduce an estimation method with weight adaptation. In general, the property of divergence associated with the estimation is deeply related to the low-quality observations of the estimator, for example, an extremely small data set or a data set with outliers (Beran, 1977; Basu,
Harris, Hjort, & Jones, 1998; Murata, Takenouchi, Kanamori, & Eguchi, 2004; Fujimoto & Murata, 2007). Thus, another interpretation of weight adaptation is given from the perspective of the $\eta$-divergence (Takenouchi & Eguchi 2004; Takenouchi, Eguchi, Murata, & Kanamori, 2008)—a type of the Bregman divergence (Bregman, 1967) known to be robust against small sample sets and outliers. We experimentally show that our estimation method with weight adaptation works appropriately and improves the accuracy of estimating.

This letter is composed as follows. In section 2, we briefly explain the BT model and show its geometrical interpretation. In section 3, we introduce important ideas used in information geometry and propose an estimation method based on flatness and projections. We address a problem of weight selection and propose a modified estimation algorithm with weight adaptation in section 4. In section 5, we set out experimental results about our estimation procedure. Our concluding remarks are given in section 6.

2 Bradley-Terry Model and Its Geometrical Interpretation

In this section, we introduce the Bradley-Terry (BT) model and show its geometrical structure in a probability simplex.

2.1 The Bradley-Terry Model. Let $\Lambda = \{i\}$ be a finite index set of items and let $Q$ be a parameter set:

$$ Q = \{\theta_i\}_{i \in \Lambda} \quad \text{where} \quad \sum_{i \in \Lambda} \theta_i = 1, \ \theta_i > 0. $$

Note that $Q$ is also identified as a categorical distribution. We will use the symbol $|\Lambda|$ to denote the cardinality of the set $\Lambda$, which indicates the number of components in it. Consider a trial in which a user evaluates two items $i, j \in \Lambda$, and chooses a preferred one from $i$ and $j$ according to $Q$. Let $i \succ j$ denote an event that the item $i$ is chosen in this pairwise comparison of $i$ and $j$. In the conventional BT model, the probability of the event $i \succ j$ is given by

$$ \Pr(i \succ j) = \frac{\theta_i}{\theta_i + \theta_j} \quad (i \neq j). $$

(2.1)

Now assume that multiple users individually compare $i$ and $j$. Let $n_{ij}$ and $n_{ji}$ be the number of observed events $i \succ j$ and $j \succ i$, respectively. If $n_{ij} + n_{ji}$ comparisons are independent, equation 2.1 suggests that the empirical ratio of $\theta_i$ to $\theta_j$ is given by

$$ \theta_i : \theta_j = n_{ij} : n_{ji}. $$
Assume that evaluations for different pairs are also independent. Then the log likelihood of the BT model for pairwise comparisons by multiple users is given as

$$L(Q) = \sum_{i \neq j} n_{ij} \log \frac{\theta_i}{\theta_i + \theta_j},$$

(2.2)

where $\sum_{i \neq j}$ is the sum over pairs $i, j \in \Lambda$ except $i = j$. Generally the estimate $\tilde{Q}$ is obtained as a solution of the following optimization problem,

$$\tilde{Q} = \arg\max_Q L(Q) \quad \text{subject to} \quad \sum_{i \in \Lambda} \theta_i = 1, \theta_i > 0,$$

(2.3)

and fast algorithms are given in, for example, Hastie and Tibshirani (1998) and Huang et al. (2006).

2.2 An Estimation of the BT Model on a Probability Simplex from the Geometrical Viewpoint. We reformulate the BT estimation from the viewpoint of information geometry. Let $\mathcal{M}$ be a space of categorical distributions that is given by

$$\mathcal{M} = \left\{ Q = \{\theta_i\}_{i \in \Lambda} \left| \sum_{i \in \Lambda} \theta_i = 1, \theta_i > 0 \right. \right\}. $$

Figure 1a shows a geometrical interpretation of $\mathcal{M}$ for $\Lambda = \{1, 2, 3\}$, that is, a 2-simplex in the three-dimensional space equivalent to a regular triangle. In this letter, the simplex $\mathcal{M}$ is also called the model manifold.
Next, we consider a set of probabilities \( P = \{ \pi_i \}_{i \in \Lambda} \) for expressing pairwise comparisons in \( M \). A pairwise comparison indicates the observed ratio; for example, a comparison result \((n_{12}, n_{21})\) for \(1 \succ 2\) and \(2 \succ 1\) indicates a restriction for probabilities in the BT model, that is,

\[
\pi_1 : \pi_2 = n_{12} : n_{21}.
\]

We let \( D \) be a subspace of \( M \), which satisfies the observed ratio, that is,

\[
D = \left\{ P = \{ \pi_i \}_{i \in \Lambda} \in M \mid \pi_i : \pi_j = n_{ij} : n_{ji} \right\}.
\]

A constrained submanifold \( D \) is uniquely defined with a pairwise comparison of any \( i \) and \( j \), and in the following, \( D \) is called the data manifold.

As shown in Figure 1, the model manifold \( M \) and the data manifold \( D \) are visualized as a \(|\Lambda| - 1\) simplex and a hyperplane in it. Figure 1b shows an example of equation 2.4 for \( \Lambda = \{1, 2, 3\} \): \( D \) is shown as a line in the triangle that is drawn from \((0, 0, 1)\) onto the edge between \((1, 0, 0)\) and \((0, 1, 0)\). In the case of \(|\Lambda| > 3\), we need a careful treatment for visualization. For example, Figure 1c shows the model manifold \( M \) in the case of \(|\Lambda| = 4\), which is given as a 3-simplex, and Figure 1d shows an observed ratio \( \pi_1 : \pi_4 = n_{14} : n_{41} \), which is given as a cutting surface of a regular tetrahedron.

Although Figure 1c no longer shows coordinates of the parameter set \( Q \) directly, Figure 1d shows an observed ratio by distances between vertices. Henceforth, we use the 3-simplex and its cutting plane for a representation of \(|\Lambda| = 4\) case when it is necessary.

In the BT model, \( N = \binom{|\Lambda|}{2} \) comparisons are obtained from all pairs, and each pairwise comparison gives a corresponding data manifold. In the following discussion, let \( n = 1, \ldots, N \) be the renumbering of pairs \( \{ (i, j) \} \), and let \( \{ D^n \} = \{ D^n \}_{n=1}^N \) be \( N = \frac{|\Lambda|(|\Lambda| - 1)}{2} \) hyperplanes. If the observed data set has consistency,

\[
\bigcap_{n=1}^N D^n = \{ Q \},
\]

then the estimate \( \hat{Q} \) is obtained as a unique intersection point of all the hyperplanes in \( M \) where all the observed ratios exactly hold. Figure 2a shows an example of the case of \(|\Lambda| = 3\). In practical situations, however, the observed ratios lack strict consistency in general, that is,

\[
\bigcap_{n=1}^N D^n = \emptyset.
\]

Equation 2.6 suggests that all the hyperplanes do not intersect at a single point as in Figure 2b. In these inconsistent cases, the optimal model should
Figure 2: Intuitive interpretation of an estimator. (a) Consistent estimate in the ideal case: all the hyperplanes intersect at a unique point. (b) Typical case of inconsistent observed ratios: all the hyperplanes do not intersect at a unique point. (c) A segment of iterative estimation. A dotted arrow shows the nearest point on each data manifold from the current estimate, and solid arrows show the updated estimate.

satisfy constraints of pairwise comparisons as much as possible; a good estimate for the BT model is obtained as the nearest point from $N$ hyperplanes in the simplex. To determine an estimate $\hat{Q}$, we need a measure of closeness between a hyperplane $D$ and a point $Q$ on the simplex $M$. Here, we introduce the Kullback-Leibler (KL) divergence as a measure. Since elements of $M$ are identified as probability distributions, the KL divergence between points $P = \{\pi_i\}$ and $Q = \{\theta_i\}$ is given as

$$D_{KL}(P, Q) = \sum_{i \in \Lambda} \pi_i \log \frac{\pi_i}{\theta_i}.$$ 

Then we define the KL divergence between a hyperplane $D$ and a point $Q$ as

$$D_{KL}(D, Q) = \min_{P \in D} D_{KL}(P, Q). \quad (2.7)$$

Now we illustrate an estimate $\hat{Q}$ for multiple pairwise comparisons in a geometrical approach. Let $w^n$ be a weight for $D^n$, defined as

$$\{w^n\} = \{w^n\}_{n=1}^N, \quad \text{where} \quad w^n > 0, \quad \sum_{n=1}^N w^n = 1 \quad (n = 1, \ldots, N).$$

Intuitively speaking, $w^n$ indicates the relative importance of the data manifold $D^n$ in the estimation.\(^1\) There are some plausible candidates for a specific

---

\(^1\)A weight value itself does not have an essential meaning, so we can use $w^n = (0, \infty)$ in the same framework without loss of generality. We use the normalized weight only for convenience.
Table 2: Example of Pairwise Comparisons.

<table>
<thead>
<tr>
<th>i</th>
<th>j</th>
<th>Pr(i &gt; j)</th>
<th>Pr(i &lt; j)</th>
<th>Majority Rule</th>
<th>( \tilde{Q} )</th>
<th>( \hat{Q} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.56</td>
<td>0.44</td>
<td>1 ( &gt; ) 2</td>
<td>( \times )</td>
<td>( \checkmark )</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.51</td>
<td>0.49</td>
<td>1 ( &gt; ) 3</td>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.60</td>
<td>0.40</td>
<td>1 ( &gt; ) 4</td>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.96</td>
<td>0.04</td>
<td>2 ( &gt; ) 3</td>
<td>( \checkmark )</td>
<td>( \checkmark )</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.44</td>
<td>0.56</td>
<td>2 ( &lt; ) 4</td>
<td>( \times )</td>
<td>( \times )</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.59</td>
<td>0.41</td>
<td>3 ( &gt; ) 4</td>
<td>( \times )</td>
<td>( \times )</td>
</tr>
</tbody>
</table>

Notes: \( \checkmark \) and \( \times \) indicate whether \( \theta_i > \theta_j \) holds when \( i \succ j \) is given by majority rule. \( \tilde{Q} = \{0.29, 0.34, 0.16, 0.21\} \) and \( \hat{Q} = \{0.32, 0.29, 0.15, 0.23\} \).

value of \( w^n \)—for example, the ratio of observations for construction of the \( n \)th data manifold or \( w^n = \frac{1}{N} \) for all \( n \). In section 4, we discuss meanings of weights in detail. An objective function for the parameter estimation on the probability simplex \( \mathcal{M} \) is expressed as a weighted average of the KL divergences between \( D^n \) and \( Q \) as follows:

\[
F(Q) = \sum_{n=1}^{N} w^n D_{KL}(D^n, Q) = \sum_{n=1}^{N} w^n \min_{P \in D^n} D_{KL}(P, Q).
\]  

(2.8)

The minimizer of equation 2.8 with respect to \( Q \) is obtained by solving the following nested optimization problem:

\[
\hat{Q} = \arg\min_{Q \in \mathcal{M}} \left\{ \sum_{n=1}^{N} w^n \min_{P \in D^n} D_{KL}(P, Q) \right\}.
\]  

(2.9)

The relationship between equations 2.3 and 2.9 is discussed in section 2.3.

Since the minimizer \( \hat{Q} \) of the nested optimization problem is hard to obtain directly, we try to estimate it by an iterative process, as shown in Figure 2c. Intuitively, we find the closest point on each data manifold from a current estimate on a simplex (shown as dotted arrows in the figure), update the estimate by concatenating those closest points (shown as solid arrows), and repeat these steps until convergence. In the next section, we generalize the problem and provide a formal definition of this method to solve the nested optimization given by equation 2.9.

### 2.3 Comparison with a Conventional Estimation Setup.

Before describing optimization steps, we discuss the difference between the conventional setup and ours, which are given by equations 2.3 and 2.9, respectively.

Table 2 shows an example of pairwise comparisons found in Hastie and Tibshirani (1998). The table also shows the results derived with the majority
rule; \(i \succ j\) means that the relation \(\Pr(i \succ j) > \Pr(i \prec j)\) holds. Apparently there is no unique ranking that satisfies all majority rules. Here, we try to obtain a unique ranking through a relative relation in \(\{\hat{\theta}_i\}_{i \in \Lambda}\) with the BT setting. Hastie and Tibshirani (1998), give the estimate of equation 2.2 with \(\{w^n = \frac{1}{N}\}^N_{n=1}\) as \(\hat{Q} = \{0.29, 0.34, 0.16, 0.21\}\). The estimate of equation 2.9 with the same weights is \(\hat{Q} = \{0.32, 0.29, 0.15, 0.23\}\). In Table 2, we show whether majority rules are satisfied or violated by \(\hat{Q}\) and \(\hat{Q}\). In this case, the ranking in \(\hat{Q}\) is more consistent than that in \(\hat{Q}\), because the former supports \(1 \succ 2\) but the latter does not. The example shows that our method works better than the conventional one.

Differences of estimation results are geometrically explained by differences of objective functions. In the conventional setup estimating the BT model, an objective function is given as equation 2.2. Let \(P^n(b)\) and \(Q^n(b)\) be Bernoulli distributions for an index pair \(n = (i, j)\) associated with the observed data and the model \(Q\):

\[
P^n(b) = \left\{ \frac{n_{ij}}{n_{ij} + n_{ji}}, \frac{n_{ji}}{n_{ij} + n_{ji}} \right\}, \quad Q^n(b) = \left\{ \frac{\theta_i}{\theta_i + \theta_j}, \frac{\theta_j}{\theta_i + \theta_j} \right\}.
\]

These distributions give ratios between \(i\) and \(j\). Hence, they are indicated as points on the corresponding edge of the simplex as shown in Figure 3. Figure 3a shows the distribution \(P^n(b)\), which is uniquely determined as a point on the edge between \(i\) and \(j\) based on \(D^n\). In Figure 3b, \(Q^n(b)\) is shown as a similar point on the edge based on a corresponding gray line that passes through the given point \(Q\). Assume that the weight \(w^n\) for a comparison of a pair \(n = (i, j)\) is given by the number of observations as \(w^n = \sum_{i' \neq j'}n_{i'j'}\); the maximizer of equation 2.2 is equivalent to the minimizer of the sum of weighted KL divergences for evaluating \(P^n(b)\) and \(Q^n(b)\), given as

\[
F^\dagger(Q) = \sum_{n=1}^N w^n D_{KL}(P^n(b), Q^n(b)). \tag{2.10}
\]

The conventional method roughly evaluates the consistency of the parameter only at the edges of the probability simplex and does not exploit a property that the parameter set \(Q\) is identified as a \(|\Lambda|\)-dimensional categorical distribution. Figure 3c depicts this geometrical intuition. An arrow on an edge of the simplex indicates the KL divergence from \(P^n(b)\) to \(Q^n(b)\), and equation 2.10 minimizes the weighted sum of these arrow lengths.

Figure 3d shows an interpretation of our objective function given by equation 2.8. Our method minimizes the weighted sum of KL divergences from all the data manifolds to the point \(Q\). These interpretations show that our proposed setup naturally evaluates the point \(Q\) as a \(|\Lambda|\)-dimensional categorical distribution, though the conventional setup evaluates only the
distributions \( \{ Q_n^{(b)} \}_{n=1}^N \) projected on the corresponding edges. Therefore, our method is expected to be more appropriate to obtain the estimate \( \hat{Q} \) as a \(|\Lambda|\)-dimensional categorical distribution (e.g., ranking models) and multiclass categorization by using binary classifiers.

After introducing some additional notations, we discuss their difference again in section 6.

3 Estimation Based on Flat Data Manifolds

In this section, we generalize the setup of the BT model and propose an estimation method based on geometrical interpretation. To achieve these
goals, we introduce information geometrical notions such as flatness and projection.

3.1 Generalization of Data Manifolds. We generalize a data manifold $D^n$ for a one-versus-one pairwise comparison defined by equation 2.5 in order to describe more complex comparisons in which ratios between groups of probabilities are constrained. Let us assume that $N$ different comparisons are observed, and in the $n$th comparison, items are divided into $M^n$ groups and compared by these groups. Let $\Lambda^n_m \subseteq \Lambda$ be a set of observed indices in $n$th comparison group, which satisfies $\Lambda^n_0 = \Lambda$ and $\Lambda^n_m \cap \Lambda^n_{m'} = \emptyset$ ($m \neq m'$).

and let $\Lambda^n_+ = \bigcup_{m=1}^{M^n} \Lambda^n_m$ be a set of all the observed indices. We also assume uniqueness of the order of index subsets; that is, given two subsets $\Lambda^n_m$ and $\Lambda^n_{m'}$, $m < m'$ holds for $m, m' = 1, \ldots, M^n$ when $\min(i \in \Lambda^n_m) < \min(i' \in \Lambda^n_{m'})$ holds. Let $\Lambda^n_0$ be a set of unobserved indices, namely, $\Lambda^n_0 = \Lambda - \Lambda^n_+$; hence, a set $\{\Lambda^n_m\} = \{\Lambda^n_0, \Lambda^n_1, \ldots, \Lambda^n_{M^n}\}$ satisfies $\bigcup_{m=0}^{M^n} \Lambda^n_m = \Lambda^n_+ \cup \Lambda^n_0 = \Lambda$. For example, a comparison between any $i$ and $j > i$ in the original BT model is described as $\Lambda^n_1 = \{i\}$, $\Lambda^n_2 = \{j\}$, and $\Lambda^n_0 = \Lambda - \{i, j\}$ in this expression.

Let $\{\alpha^n_m\}$ be an $M^n$-dimensional vector representing observed ratios, given by

$$\{\alpha^n_m\} = \{\alpha_1^n, \ldots, \alpha_{M^n}^n\}, \quad \text{where} \quad \alpha^n_m > 0, \quad \sum_{m=1}^{M^n} \alpha^n_m = 1.$$

Under the condition that total probabilities in index sets $\Lambda^n_m$ are proportional to $\{\alpha^n_m\}$,

$$\sum_{i \in \Lambda^n_1} \pi_i : \sum_{i \in \Lambda^n_2} \pi_i : \cdots : \sum_{i \in \Lambda^n_{M^n}} \pi_i = \alpha_1^n : \alpha_2^n : \cdots : \alpha_{M^n}^n,$$

categorical distributions compose a generalized data manifold, which is written as

$$D^n = \left\{ P = \{\pi_i\} \in \mathcal{M} \mid \sum_{i \in \Lambda^n_{m'}} \pi_i = 1 - \psi^n, \quad \sum_{i \in \Lambda^n_m} \pi_i = \psi^n \alpha^n_m \quad (m = 1, \ldots, M^n), \quad \psi^n \in (0, 1] \right\}. \quad (3.1)$$
Note that here we introduce a parameter $\psi^n$ to represent the proportion of observed items in $D^n$.

The generalized data manifold $D^n$ introduced by equation 3.1 has the following property:

**Theorem 1.** $m$-Flatness of generalized data manifold: Let $P = \{\pi_i\}_{i \in \Lambda}$ and $P' = \{\pi'_i\}_{i \in \Lambda}$ be two points in a $|\Lambda| - 1$ simplex, and $R(t)$ is an internally dividing point of $P$ and $P'$, given as

$$R(t) = \{(1 - t)\pi_i + t\pi'_i\}_{i \in \Lambda}, \quad 0 \leq t \leq 1.$$

If $P$, $P'$ belong to $D^n$, then $R(t)$ also belongs to $D^n$ for $0 \leq t \leq 1$.

**Proof.** $P$, $P'$, and $R(t)$ have the following property with the definition of $D^n$ given by equation 3.1:

$$\sum_{i \in \Lambda_m^n} \pi_i = \sum_{i \in \Lambda_m^n} \pi'_i = \sum_{i \in \Lambda_m^n} ((1 - t)\pi_i + t\pi'_i) = \alpha^n_m, \quad m = 1, \ldots, M^n.$$

This property is called $m$-flatness in the literature on information geometry (Amari & Nagaoka, 2000; Murata et al., 2004); thus, generalized data manifolds are called $m$-flat in this letter.

With these notations, we can extend the setup for the BT model to other types of $m$-flat data manifolds as follows:

**Example 1:** One-Versus-One Type. For the one-versus-one type pairwise comparison used in the BT model, an observed ratio is defined by a set of index subsets $\{\Lambda^n_m\}$ and a ratio set $\{\alpha^n_m\}$. In the case of equation 2.4 with $|\Lambda| = 3$, $\{\Lambda^n_m\}$ and $\{\alpha^n_m\}$ are given as

$$\{\Lambda^n_0\} = \{3\}, \quad \{\Lambda^n_1\} = \{1\}, \quad \{\Lambda^n_2\} = \{2\},$$

$$\{\alpha^n_m\} = \left\{\frac{n_{12}}{n_{12} + n_{21}}, \frac{n_{21}}{n_{12} + n_{21}}\right\}.$$

Its geometrical interpretation is shown in Figure 1b.

**Example 2:** Group-Versus-Group Type. The group-versus-group type of comparison proposed by Huang et al. (2006) is expressed by grouping some indices into one category of $\Lambda^n_{1,}, \ldots, \Lambda^n_{M^n}$. Let $\Lambda^n_i > \Lambda^n_j$ be an event that the item set $\Lambda^n_i$ is chosen in the comparison of two item sets $\Lambda^n_i$ and $\Lambda^n_j$. 
(a) One-vs-group. (b) Group-vs-group. (c) Triplet. (d) Home-field.

Figure 4: Examples of $m$-flat manifolds. (a) An $m$-flat manifold defined by $\Lambda_0^n = \emptyset, \Lambda_1^n = \{1\}$, and $\Lambda_2^n = \{2, 3\}$. (b) An $m$-flat manifold defined by $\Lambda_0^n = \emptyset, \Lambda_1^n = \{1, 2\}$, and $\Lambda_2^n = \{3, 4\}$. (c) An $m$-flat manifold defined by $\Lambda_0^n = \{1\}, \Lambda_1^n = \{2\}$, $\Lambda_2^n = \{3\}$, and $\Lambda_3^n = \{4\}$. (d) Two $m$-flat manifolds defined by $\Lambda_0^n = \{3\}, \Lambda_1^n = \{1\}$, and $\Lambda_2^n = \{2\}$.

Figure 4a is a typical data manifold for this type of comparison, defined as

$$\{\Lambda_m^n\} = \{\Lambda_0^n = \emptyset, \Lambda_1^n = \{1\}, \Lambda_2^n = \{2, 3\}\},$$

$$\{\alpha_m^n\} = \left\{ \alpha_1^n = \frac{n_{12}'}{n_{12}' + n_{21}'}, \alpha_2^n = \frac{n_{21}'}{n_{12}' + n_{21}'} \right\},$$

where $n_{ij}'$ is the number of observed event $\Lambda_i^n > \Lambda_j^n$. Figure 4b is another example of group-versus-group comparison in the case of $|\Lambda| = 4$, defined as

$$\{\Lambda_m^n\} = \{\Lambda_0^n = \emptyset, \Lambda_1^n = \{1, 2\}, \Lambda_2^n = \{3, 4\}\},$$

$$\{\alpha_m^n\} = \left\{ \alpha_1^n = \frac{n_{12}'}{n_{12}' + n_{21}'}, \alpha_2^n = \frac{n_{21}'}{n_{12}' + n_{21}'} \right\}.$$

For example, if one plays tennis in many doubles with several partners, the results of the matches are observed as group-versus-group types of comparisons. We can evaluate relative individual skills in playing tennis by estimating $Q = \{\theta_i\}$.

In Ikeda (2006), the estimation algorithm is proposed to handle these data manifolds where $\Lambda_0^n = \emptyset$ for $n = 1, \ldots, N$. Our setup permits having $\Lambda_0^n \neq \emptyset$ and gives a generalized viewpoint of Ikeda (2006).

Example 3: Multiple Groups. Typical models for comparison between multiple groups, including the so-called Pendergrass model (Pendergrass & Bradley, 1960) or the Plackett-Luce model (Plackett, 1975; Luce, 1959), can
also be explained with \( m \)-flat data manifolds. For example, the Pendergrass model for a triplet is formulated as a multistage BT model as

\[
\Pr(i > j > k) = \frac{\theta_i}{\theta_i + \theta_j + \theta_k} \cdot \frac{\theta_j}{\theta_j + \theta_k} \quad (i \neq j \neq k).
\]

The first term indicates the pairwise comparison result of \( i > \{j, k\} \), the second term indicates that of \( j > k \), and the equation shows that the data manifold of \( i > j > k \) is given as a submanifold of the data manifold \( i > \{j, k\} \). For example, in the case of \( \Lambda = \{1, 2, 3, 4\} \) and \( (i, j, k) = (2, 3, 4) \), \( \{\Lambda^n_m\} \) and \( \{\alpha^n_m\} \) for \( 2 > 3 > 4 \) is given as follows,

\[
\{\Lambda^n_m\} = \{\Lambda^n_0 = \{1\}, \Lambda^n_1 = \{2\}, \Lambda^n_2 = \{3\}, \Lambda^n_3 = \{4\}\},
\]

\[
\{\alpha^n_m\} = \left\{ \begin{array}{l}
\alpha^n_1 = \frac{n_{234} + n_{243}}{n_{234} + n_{243} + n_{342} + n_{432}}, \\
\alpha^n_2 = \frac{(1 - \alpha^n_1)n_{34}}{n_{34} + n_{43}}, \\
\alpha^n_3 = \frac{(1 - \alpha^n_1)n_{43}}{n_{34} + n_{43}}
\end{array} \right. ,
\]

where \( n_{ijk} \) is the number of observed event \( i > j > k \). Figure 4c shows an intuitive interpretation for this observation.

**Example 4:** Home Field Advantage. The home field advantage model, given in Agresti (2002), is a generalization of the BT model for representing a home field advantage. Here, we introduce another idea for representing the difference between “home” and “away.” In our formulation, the results of home teams and those of away teams can be expressed as different \( m \)-flat manifolds for a common \( \{\Lambda^n_m\} \). Assume that \( \Lambda = \{1, 2, 3\} \), and \( \{\Lambda^n_m\} \) is given as

\[
\{\Lambda^n_m\} = \{\Lambda^n_0 = \{3\}, \Lambda^n_1 = \{1\}, \Lambda^n_2 = \{2\}\}.
\]

Then two manifolds that represent the results of home and away for \( \{\Lambda^n_m\} \) are defined with \( \{\alpha^n_m\} \) and \( \{\alpha^{nh}_m\} \), given as

\[
\{\alpha^{nh}_m\} = \left\{ \begin{array}{l}
\alpha^{nh}_1 = \frac{n^{h}_{12}}{n^{h}_{12} + n^{h}_{21}}, \\
\alpha^{nh}_2 = \frac{n^{h}_{21}}{n^{h}_{12} + n^{h}_{21}}
\end{array} \right. ,
\]

\[
\{\alpha^{na}_m\} = \left\{ \begin{array}{l}
\alpha^{na}_1 = \frac{n^{a}_{12}}{n^{a}_{12} + n^{a}_{21}}, \\
\alpha^{na}_2 = \frac{n^{a}_{21}}{n^{a}_{12} + n^{a}_{21}}
\end{array} \right. ,
\]

where \( n^{h}_{ij} \) is the number of observed event \( i > j \) in the case where \( i \) indicates a home team and \( n^{a}_{ij} \) is that in the case where \( i \) indicates an away team. Their geometrical interpretations are shown in Figure 4d.
We stress that these models have been proposed based on equation 2.2 and are also naturally handled in our framework.

### 3.2 Pythagorean Relation on a Probability Simplex

To discuss the optimization problem $\min_{P \in \mathcal{D}} D_{KL}(P, Q)$ in equation 2.9, we first show a property of the KL divergence called the Pythagorean relation.

**Theorem 2.** Pythagorean theorem (Amari & Nagaoka, 2000): Let $P = \{\theta_i\}_{i \in \Lambda}$, $Q = \{\theta'_i\}_{i \in \Lambda}$, and $R = \{\theta''_i\}_{i \in \Lambda}$ be in $\mathcal{M}$. If the inner product of two $|\Lambda|$-dimensional vectors $R - P$ and $\log(Q) - \log(P)$ has the property given as

$$\sum_{i \in \Lambda} (\theta''_i - \theta_i)(\log \theta_i - \log \theta'_i) = 0,$$

then the following relation holds:

$$D_{KL}(R, Q) = D_{KL}(R, P) + D_{KL}(P, Q).$$

**Proof.** For any $P, Q, R \in \mathcal{M}$,

$$D_{KL}(R, Q) - D_{KL}(R, P) - D_{KL}(P, Q) = \sum_{i \in \Lambda} (\theta''_i - \theta_i)(\log \theta_i - \log \theta'_i)$$

holds. Then we have equation 3.3 if equation 3.2 holds.

The Pythagorean relation defines an orthogonality at the point $P$ in a simplex. Here, we introduce another type of subspace in a probability simplex based on the orthogonality:

**Definition 1.** e-flat subspace in a probability simplex: Let $\mathcal{D}$ be an $m$-flat manifold in $\mathcal{M}$. Assume that a subspace $S \subset \mathcal{M}$ includes a point $P \in \mathcal{D}$ and is orthogonal to $\mathcal{D}$, that is,

$$S(P) = \{Q \in \mathcal{M} | D_{KL}(R, Q) = D_{KL}(R, P) + D_{KL}(P, Q), \forall R \in \mathcal{D}\}.$$

Then the subspace $S(P)$ is called e-flat.

Note that $\mathcal{D}$ and $S(P)$ are orthogonal at $P$, and for any $R \in \mathcal{D}$ and $Q \in S(P)$, the orthogonal relation equation 3.3 holds. An intuitive interpretation of $S(P)$ is shown in Figure 5. For a given $m$-flat data manifold $\mathcal{D}$, a set of e-flat subspaces constructs the following structure,

$$\{S(P) | \forall P \in \mathcal{D}\} = \mathcal{M}, \quad S(P) \cap S(P') = \emptyset \quad (P \neq P')$$
Figure 5: Pythagorean relation in 2- and 3-simplices. The relation defines an orthogonality at $P \in \mathcal{D}$. Note that we can consider an \( e \)-flat subspace \( S(P) \) for any $P \in \mathcal{D}$ in any dimensional simplex \( \mathcal{M} \).

which is called the orthogonal foliation (Murata et al., 2004). The structure indicates that there exists an \( e \)-flat subspace \( S(P) \) in \( \mathcal{M} \) for an arbitrary point $P \in \mathcal{D}$. Conversely, it indicates that an arbitrary point $Q \in \mathcal{M}$ has a unique point $P \in \mathcal{D}$ such that $Q \in S(P)$ holds.

Now we discuss the inner minimization in equation 2.7. We introduce a minimization problem called the \( e \)-projection.

**Definition 2.** \( e \)-projection: Let $\mathcal{D}$ be a subspace of $\mathcal{M}$. The minimization problem,

\[
\hat{P} = \arg\min_{P \in \mathcal{D}} D_{KL}(P, Q),
\]

is called the \( e \)-projection (Amari, 1995; Amari & Nagaoka, 2000).

In general, the existence of a unique solution for minimization problem 3.4 is not guaranteed; however, if $\mathcal{D}$ is $m$-flat, the following property holds:

**Theorem 3.** \( e \)-projection onto an $m$-flat manifold (Amari, 1995): If the manifold $\mathcal{D}$ is $m$-flat, the minimization problem 3.4 has a unique solution.

**Proof.** Let $P$ be a point in $\mathcal{D}$ that satisfies $Q \in S(P)$. Such a point $P$ is uniquely determined from the orthogonal foliation structure. For any point $R \in \mathcal{D}$, $D_{KL}(R, Q) \geq D_{KL}(P, Q)$ holds because of the Pythagorean theorem and the positivity of the KL divergence $D_{KL}(R, P) \geq 0$. This proves the theorem.

By using $m$-flat data manifolds \( \{ \mathcal{D}^n \} \) given by equation 3.1, we can obtain a unique point $P \in \mathcal{D}^n$ for each manifold by the \( e \)-projection from a certain point $Q \in \mathcal{M}$. 

3.3 Decomposition of Distributions. To derive a concrete form of the $\epsilon$-projection on a simplex $\mathcal{M}$, we first introduce the decomposition of distributions $Q = \{\theta_i\}_{i \in \Lambda} \in \mathcal{M}$ and $P = \{\pi_i\}_{i \in \Lambda} \in \mathcal{D}$. Since we discuss a specific $\mathcal{D}$ henceforth, we omit the subscript $n$ in $\{\Lambda_n\}_{m=1}^M$, $\{\alpha_n\}_{m=1}^M$, and $\psi_n$ as for notational simplicity. Let

$$D = \left\{ P = \{\pi_i\} \mid \sum_{i \in \Lambda_0} \pi_i = 1 - \psi, \right.$$ \hspace{1cm} \left. \sum_{i \in \Lambda_m} \pi_i = \psi \alpha_m \ (m = 1, \ldots, M), \ \psi \in (0, 1) \right\},$$

for notational simplicity. Let

$$p_m = \sum_{i \in \Lambda_m} \pi_i, \quad q_m = \sum_{i \in \Lambda_m} \theta_i \quad (m = 0, \ldots, M),$$

be the sums of probabilities over the given index subset $\Lambda_m$ and

$$p_+ = \sum_{m=1}^M p_m = \psi, \quad q_+ = \sum_{m=1}^M q_m$$

be sums over the observed index subsets $\Lambda_+ = \{\Lambda_m\}_{m=1}^M$. Note that $p_0 + p_+ = 1$ and $q_0 + q_+ = 1$ hold. Here we define the following decomposed distributions:

**Definition 3.** Decomposed distributions on $\mathcal{M}$: Given $\{\Lambda_m\}$, $Q = \{\theta_i\}_{i \in \Lambda} \in \mathcal{M}$ is decomposed as

$$(Q; \{\Lambda_m\}) \leftrightarrow (\{Q^{(w)}_m\}_{m=0}^M, Q^{(b)}, Q^{(r)}; \{\Lambda_m\}),$$

where

$$Q^{(w)}_m = \left\{ \frac{\theta_i}{q_m} \right\}_{i \in \Lambda_m} = \{\theta^{(w)}_{i|m}\}_{i \in \Lambda_m} \quad \text{(within-group distribution)},$$

$$Q^{(b)} = \left\{ \frac{q_m}{q_+} \right\}_{m=1}^M = \{\theta^{(b)}_m\}_{m=1}^M \quad \text{(between-group distribution)},$$

$$Q^{(r)} = \{q_0, q_+\} = \{\theta^{(r)}_0, \theta^{(r)}_+\} \quad \text{(rated or unrated distribution)}.$$
Estimation of Generalized Bradley-Terry Models 1639

\[ m = 0 \text{ where } \sum_{m=1}^{M} \theta_m^{(b)} = 1 \text{ holds, and the rated or unrated distribution } Q^{(r)} \text{ is a Bernoulli distribution for } \Lambda_0 \text{ and the rest, where } \theta_0^{(r)} + \theta_+^{(r)} = 1 \text{ holds.} \]

Since degrees of freedom of \( Q, Q_m^{(w)}, Q_m^{(b)}, \text{ and } Q^{(r)} \) are \(|\Lambda| - 1, |\Lambda_m| - 1, M - 1, \text{ and } 1, \) respectively, and \( \sum_{m=0}^{M} |\Lambda_m| = |\Lambda| \text{ holds, decomposed distributions } \{ Q_m^{(w)}, Q_m^{(b)}, \text{ and } Q^{(r)} \text{ are nonredundant expression of } Q. \text{ Thus, these decomposed distributions are uniquely defined by any } Q \in M \text{ and a set of index subsets } \{ \Lambda_m \} \text{ and have the following relation:} \]

\[
\theta_i = \begin{cases} 
\theta_i^{(w)} \cdot \theta_0^{(r)} & (i \in \Lambda_0), \\
\theta_i^{(w)} \cdot \theta_m^{(b)} \cdot \theta_+^{(r)} & (i \in \Lambda_m \subset \Lambda_+). 
\end{cases} \tag{3.9}
\]

Furthermore, we define similar decomposition for \( P \in D \subset M \) as follows:

**Definition 4.** Decomposed distributions on \( D: \) Given \( \{ \Lambda_m \}, P = \{ \pi_i \}_{i \in \Lambda} \in D \subset M \text{ is decomposed as} \]

\[
(P; \{ \Lambda_m \}) \Leftrightarrow (\{ P_m^{(w)} \}_{m=0}^{M}, P^{(b)}, P^{(r)}; \{ \Lambda_m \}),
\]

where

\[
P_m^{(w)} = \left\{ \frac{\pi_i}{p_m} \right\}_{i \in \Lambda_m} = \{ \pi_i^{(w)} \}_{i \in \Lambda_m} \quad \text{(within-group distribution)}, \tag{3.10}
\]

\[
P^{(b)} = \left\{ \frac{p_m}{p_+} \right\}_{m=1}^{M} = \{ \pi_m^{(b)} \}_{m=1}^{M} = \{ \alpha_m \}_{m=1}^{M} \quad \text{(between-group distribution)}, \tag{3.11}
\]

\[
P^{(r)} = \{ p_0, p_+ \} = \{ \pi_0^{(r)}, \pi_+^{(r)} \} = \{ (1 - \psi), \psi \} \quad \text{(rated/unrated distribution).} \tag{3.12}
\]

In this case, the between-group distribution \( P^{(b)} \) and the rated or unrated distribution \( P^{(r)} \) are denoted with \( \{ \alpha_m \} \) and \( \psi. \) Note that the degree of freedom of \( P^{(b)} \) is 0 because \( \{ \alpha_m \} \) is given and that of \( P \) is \(|\Lambda| - M, \) so that these decomposed distributions \( \{ P_m^{(w)}, P^{(b)}, \text{ and } P^{(r)} \text{ are also nonredundant. And } \pi_i \in P \text{ is expressed with elements of decomposed distributions equations 3.10 to 3.12 as} \]

\[
\pi_i = \begin{cases} 
\pi_i^{(w)} \cdot \pi_0^{(r)} & (i \in \Lambda_0), \\
\pi_i^{(w)} \cdot \pi_m^{(b)} \cdot \pi_+^{(r)} & (i \in \Lambda_m \subset \Lambda_+). 
\end{cases} \tag{3.13}
\]
With decomposed distributions given by equations 3.6 to 3.8 and 3.10 to 3.12, the KL divergence $D_{KL}(P, Q)$ is decomposed as follows:

**Theorem 4.** Decomposition of KL divergence: Let $P, Q \in \mathcal{M}$ be $|\Lambda|$-dimensional probability vectors where $P = \{\pi_i\}_{i \in \Lambda}$ and $Q = \{\theta_i\}_{i \in \Lambda}$. Given $\{\Lambda_m\}$, the KL divergence between $P$ and $Q$ is decomposed as

$$D_{KL}(P, Q) = \sum_{m=0}^{M} p_m D_{KL}(P_m^{(w)}, Q_m^{(w)}) + p_+ D_{KL}(P^{(b)}, Q^{(b)}) + D_{KL}(P^{(r)}, Q^{(r)}).$$

**Proof.** With equations 3.9 and 3.13, $D_{KL}(P, Q)$ is deformed as

$$D_{KL}(P, Q) = \sum_{m=0}^{M} \sum_{i \in \Lambda_m} \pi_i \log \frac{\pi_i}{\theta_i} \pi_0 \log \frac{\pi_0}{\theta_0} + \sum_{m=1}^{M} \sum_{i \in \Lambda_m} \pi_i \log \frac{\pi_{i|m}}{\theta_{i|m}} \pi_0 \log \frac{\pi_0}{\theta_0} + \sum_{m=1}^{M} \pi_{i|m} \log \frac{\pi_{i|m}}{\theta_{i|m}}$$

$$+ \sum_{m=1}^{M} p_m \pi_{i|m} \log \frac{\pi_{i|m}}{\theta_{i|m}} + p_+ \log \frac{\pi_0}{\theta_0} + \log \frac{\pi_0}{\theta_0} + \log \frac{\pi_0}{\theta_0} + \log \frac{\pi_0}{\theta_0}$$

$$= \sum_{m=0}^{M} p_m D_{KL}(P_m^{(w)}, Q_m^{(w)}) + p_+ D_{KL}(P^{(b)}, Q^{(b)}) + D_{KL}(P^{(r)}, Q^{(r)}).$$

With theorem 4, the following corollary is obtained in the case of $P \in \mathcal{D} \subset \mathcal{M}$:

**Corollary 1.** Decomposed KL divergence for an m-flat data manifold: Let $P = \{\pi_i\}_{i \in \Lambda} \in \mathcal{D} \subset \mathcal{M}$ and $Q = \{\theta_i\}_{i \in \Lambda} \in \mathcal{M}$ be $|\Lambda|$-dimensional probability vectors. Given $\{\Lambda_m\}$, the KL divergence between $P$ and $Q$ is decomposed as

$$D_{KL}(P, Q) = \sum_{m=0}^{M} p_m D_{KL}(P_m^{(w)}, Q_m^{(w)}) + \psi D_{KL}(\{\alpha_m\}, \{\theta_m^{(b)}\})$$

$$+ D_{KL}(P^{(r)}, Q^{(r)}).$$

(3.15)
Proof. With equations 3.5, 3.7, and 3.11, equation 3.14 is deformed as equation 3.15.

3.4 Projections in Probability Simplex. In this section, we discuss concrete forms of optimization steps for our objectives.

First, we focus on the inner minimization of equation 2.9, that is, the minimization of $D_{\text{KL}}(P, Q)$ with respect to $P \in \mathcal{D}$ for fixed $Q$. The second term of equation 3.15, $D_{\text{KL}}(\tilde{P}^{(w)}_{m}, \tilde{Q}^{(w)}_{m})$, can be zero because arbitrary $\{\pi_{i|m}^{(w)}\}$ can be chosen for any given $\Lambda_{m}$, and $D_{\text{KL}}(\{\alpha_{m}\}, \{\theta_{m}^{(b)}\}) = D_{\text{KL}}(P^{(b)}, Q^{(b)})$ in equation 3.15 is a constant because $P^{(b)} = \{\alpha_{m}\}$ is fixed. Then, knowing that $D_{\text{KL}}(P^{(r)}, Q^{(r)}) = D_{\text{KL}}(1 - \psi, \psi, q_{0}, q_{+})$, the estimate $\psi$ is given by the solution of

$$\frac{\partial}{\partial \psi} D_{\text{KL}}(P, Q) = -\log \frac{1 - \psi}{q_{0}} + \log \frac{\psi}{q_{+}} + D_{\text{KL}}(P^{(b)}, Q^{(b)}) = 0,$$

namely,

$$\psi = \frac{\delta \cdot q_{+}}{q_{0} + \delta \cdot q_{+}} \quad (n = 1, \ldots, N), \quad (3.16)$$

where $\delta = \exp \left(-D_{\text{KL}}(P^{(b)}, Q^{(b)})\right)$. With equations 3.1 and 3.16, the e-projection $\tilde{P} = \{\tilde{\pi}_{i}\}$ is given by

$$\tilde{\pi}_{i} = \frac{\theta_{i}}{q_{m}} p_{m}(\psi) \quad (i \in \Lambda_{m})$$

$$= \begin{cases} \frac{\theta_{i}}{q_{0} + \delta \cdot q_{+}}, & (i \in \Lambda_{0}), \\ \frac{\alpha_{m} \cdot \theta_{i} \cdot \delta \cdot q_{+}}{q_{m} \cdot q_{0} + \delta \cdot q_{+}}, & (i \in \Lambda_{m} \subset \Lambda_{+}). \end{cases} \quad (3.17)$$

Figure 6a shows an example of $m$-flat data manifolds $\{\mathcal{D}^{n}\}$ on $|\Lambda| = 3$ and a point $Q$. Theorem 3.2 shows that e-projections from $Q$ onto $\{\mathcal{D}^{n}\}_{n=1}^{N}$ determine the $N$ points shown as Figure 6b.

Next we focus on the outer minimization of equation 2.9 with a set of $N$ fixed points $\{P^{n} = \{\pi_{i}^{n}\} \in \mathcal{D}^{n}\}_{n=1}^{N}$. We introduce another minimization problem as follows:

Definition 5. m-projection: The minimization problem,

$$\tilde{Q} = \arg \min_{Q \in \mathcal{M}} \sum_{n=1}^{N} w^{n} D_{\text{KL}}(P^{n}, Q), \quad (3.18)$$

$$= \begin{cases} \frac{\theta_{i}}{q_{0} + \delta \cdot q_{+}}, & (i \in \Lambda_{0}), \\ \frac{\alpha_{m} \cdot \theta_{i} \cdot \delta \cdot q_{+}}{q_{m} \cdot q_{0} + \delta \cdot q_{+}}, & (i \in \Lambda_{m} \subset \Lambda_{+}). \end{cases} \quad (3.17)$$
Figure 6: Intuitive interpretation of projections. (a) An example of $m$-flat manifolds $\{D^n\}$ and $Q$. (b) e-projection from $Q$ onto $\{D^n\}$. (c) $\hat{Q}$ is estimated based on $N$ points.

is called the m-projection (Genest & Zidek, 1986; Amari, 2007; Murata & Fujimoto, 2009).

The solution $\hat{Q} = \{\hat{\theta}_i\}$ of equation 3.18 is calculated by minimizing equation 2.8 with respect to $Q$ under the constraint of $\sum_{i \in \Lambda} \theta_i = 1$, which is given as a saddle point of the following Lagrangian,

$$L(Q, \lambda) = \sum_{n=1}^{N} w^n D_{KL}(P^n, Q) + \lambda \left( \sum_{i \in \Lambda} \theta_i - 1 \right)$$

$$= \sum_{n=1}^{N} w^n \sum_{i \in \Lambda} \pi^n_i \log \frac{\pi^n_i}{\theta_i} + \lambda \left( \sum_{i \in \Lambda} \theta_i - 1 \right),$$

where $\lambda$ is a Lagrange multiplier. Because $\sum_{n=1}^{N} w^n = 1$, a saddle point is uniquely calculated as

$$\hat{\theta}_i = \sum_{n=1}^{N} w^n \pi^n_i.$$

which means $\hat{Q}$ in equation 2.9 is obtained as a weighted average of $N$ points in $\{D^n\}$. Figure 6c shows an intuitive image.

3.5 The em Estimation. The estimate $\hat{Q}$ for the nested optimization problem 2.9 is not directly obtained; therefore, we use two projections introduced in equations 3.4 and 3.18 iteratively. The procedure is known as the em algorithm (Amari, 1995).
Algorithm 1: em Estimation for the m-Flat Data Manifolds

**input** restrictions \( \{\Lambda_{m}^{n}\}_{m=1}^{M} \) and \( \{\alpha^{n}\}_{n=1}^{N} \)

**initialize** choose an initial parameter set \( Q(0) \), and construct \( \{D^{n}\} \) from \( \{\Lambda^{n}\} \) and \( \{\alpha^{n}\} \)

**repeat** from \( t = 0 \), until convergence

**e-step** apply the e-projection from the previous estimate \( Q(t) \) to \( \{D^{n}\} \), and obtain \( \{P^{n}(t+1)\} \):

\[
P^{n}(t+1) = \arg\min_{P \in D^{n}} D_{KL}(P, Q(t)) \quad (n = 1, \ldots, N).
\]

**m-step** apply the m-projection from \( \{P^{n}(t+1)\} \) to \( M \), and obtain \( Q(t+1) \):

\[
Q(t+1) = \arg\min_{Q \in M} \sum_{n=1}^{N} w^{n} D_{KL}(P^{n}(t+1), Q).
\]

**output** converged estimate \( \hat{Q} \)

The detailed e-step and m-step are described as follows:

In the e-step, \( P^{n}(t+1) \) is derived by the e-projection, which is given as

\[
\pi_{i}^{n}(t+1) = \begin{cases} 
\frac{\theta_{i}(t)}{q_{0}^{n}(t) + \delta(t) \cdot q_{+}^{n}(t)} & (i \in \Lambda_{0}), \\
\frac{\alpha_{m}^{n} \cdot \theta_{i}(t) \cdot \delta(t) \cdot q_{0}^{n}(t)}{q_{m}^{n}(t) \cdot q_{0}^{n}(t) + \delta(t) \cdot q_{+}^{n}(t)} & (i \in \Lambda_{m} \subset \Lambda_{+}).
\end{cases}
\]

where \( \delta(t) = \exp(-D_{KL}(P^{n(b)}, Q^{n(b)}(t))) \). In the m-step, \( Q(t+1) \) is calculated by the m-projection, and it is given as

\[
\theta_{i}(t+1) = \sum_{n=1}^{N} w^{n} \pi_{i}^{n}(t+1).
\]

Next, we give the convergence property of the proposed algorithm.

3.6 Convergence Property. In general, the em estimation converges only to a local optimum depending on an initial parameter set, which is called the local minima problem. Hence, we have to choose an appropriate initial parameter set to obtain the global minimum. However, the solution obtained with algorithm 1 converges to a unique and global minimum under a mild condition because the objective function \( F(Q) \) is convex, as shown below.

First, we define parallel m-flat subspaces as follows:

**Definition 6.** Parallel m-flat subspace: Let \( D \) and \( D' \) be m-flat subspaces in \( M \) with \( \{\Lambda_{m}\}_{m=0}^{M} \). Then subspaces \( D \) and \( D' \) are called mutually parallel.
We also introduce an extended parameter set $\dot{Q}$,

$$\dot{Q} = \{\dot{\theta}_i\}_{i \in \Lambda} \quad \text{where} \quad \sum_{i \in \Lambda} \dot{\theta}_i = 1, \quad \dot{\theta}_i \geq 0,$$

for the following assumption.

**Assumption 1.** Data manifolds $\{D^n\}$ hold the following three conditions:

1. At least two elements in $\{D^n\}$ are not mutually parallel.
2. $|\bigcap_{n=1}^N D^n| \leq 1$, where $|\cdot|$ indicates the number of components in a set.
3. $F(\dot{Q}) = \infty$ for all $\dot{Q} \in \text{Cl}(\mathcal{M}) - \mathcal{M}$ where $\text{Cl}(\mathcal{M})$ is the closure of $\mathcal{M}$.

Next, we show the convexity of the KL divergence:

**Theorem 5.** Convexity of the KL divergence: Given two points $\overline{Q} = \{\overline{\theta}_i\}$ and $Q = \{\theta_i\}$ in $\mathcal{M}$, let $\overline{P} = (\overline{\pi}_i)$, $P = (\pi_i)$, $P^* = (\pi^*_i)$ be three points in an $m$-flat subspace $D$, given as

$$\overline{P} = \arg\min_{P \in D} D_{KL}(P, \overline{Q}),$$

$$P = \arg\min_{P \in D} D_{KL}(P, Q),$$

$$P^* = \arg\min_{P \in D} D_{KL}(P, t\overline{Q} + (1-t)Q),$$

where $0 < t < 1$. For any two points $\overline{Q} \neq Q$, the KL divergence has the following convexity:

$$D_{KL}(P^*, t\overline{Q} + (1-t)Q) \leq tD_{KL}(\overline{P}, \overline{Q}) + (1-t)D_{KL}(P, Q). \quad (3.19)$$

Equality holds iff both $\overline{Q}$ and $Q$ are in an $m$-flat subspace $D'$, which is parallel to $D$.

**Proof.** The left-hand side of equation 3.19 has the following inequality,

$$D_{KL}(P^*, t\overline{Q} + (1-t)Q) \leq D_{KL}(t\overline{P} + (1-t)P, t\overline{Q} + (1-t)Q) \quad (3.20)$$

$$\leq tD_{KL}(\overline{P}, \overline{Q}) + (1-t)D_{KL}(P, Q). \quad (3.21)$$

In equation 3.20, equality holds iff $P^* = t\overline{P} + (1-t)P$ holds. And the relation given in equation 3.21 has been proved in Cover and Thomas (2006) with the log sum inequality, and equality holds iff $\overline{\pi}_i = \pi_i$ holds for all $i \in \Lambda$. 
Now, let $\bar{\delta}$ and $\delta$ be values given by between-group distributions, defined as

$$
\bar{\delta} = \exp \left( -D_{KL} \left( \left\{ \alpha_m \right\}_{m=1}^M, \left\{ \frac{q_m}{q^+} \right\}_{m=1}^M \right) \right),
$$

$$
\delta = \exp \left( -D_{KL} \left( \left\{ \alpha_m \right\}_{m=1}^M, \left\{ \frac{q_m}{q^+} \right\}_{m=1}^M \right) \right),
$$

where

$$
\bar{q}_m = \sum_{i \in \Lambda_m} \bar{\theta}_i, \quad q_m = \sum_{i \in \Lambda_m} \theta_i, \quad \bar{q}^+_m = \sum_{i \in \Lambda^+_m} \bar{\theta}_i, \quad q^+_m = \sum_{i \in \Lambda^+_m} \theta_i.
$$

With equation 3.17, the equality condition in equation 3.21 is given as follows:

$$
\frac{1}{\bar{q}_0 + \bar{\delta} q^+_0} = \frac{1}{q_0 + \delta q^+_0} \quad (i \in \Lambda_0),
$$

$$
\frac{1}{\bar{q}_m} : \frac{1}{\bar{q}_0 + \bar{\delta} q^+_0} = \frac{1}{q_m} : \frac{\delta q^+_m}{q_0 + \delta q^+_0} \quad (i \in \Lambda_m \subset \Lambda^+_m).
$$

Since $\bar{q}^+_m = \sum_{m=1}^M q^+_m$ and $q^+_m = \sum_{m=1}^M q^+_m$, the following equations are derived:

$$
\frac{\bar{q}_0 + \bar{\delta} q^+_0}{\bar{q}_m} = \frac{\bar{\delta} q^+_0}{q^+_m}, \quad \frac{\frac{\bar{\delta} q^+_0}{\bar{q}_m}}{\frac{\delta q^+_m}{q^+_m}} = \frac{\delta q^+_m}{q^+_m} \quad (m = 1, \ldots, M).
$$

These equations indicate that the equality in equation 3.21 holds iff $\bar{\delta} = \delta$ and $\bar{q}_m = q^+_m$ for $m = 0, \ldots, M$ simultaneously hold. These conditions mean that both $\bar{Q}$ and $Q$ are in an $m$-flat manifold with the same $\{\Lambda_m\}$ as $\mathcal{D}$. Under the same conditions, equality in equation 3.20 also holds because they satisfy $\pi^*_i = t \pi_i + (1-t)\pi^+_i$ for all $i \in \Lambda$. Therefore, the condition of equality in equation 3.19 is given with a parallel manifold: both $\bar{Q}$ and $\bar{Q}$ are in an $m$-flat subspace $\mathcal{D}'$, which is parallel to $\mathcal{D}$.

Now we give the following lemma with regard to $F(Q)$:

**Lemma 1.** Convexity of $F(Q)$: Under conditions 1 and 2 in assumption 1, $F(Q)$ given by equation 2.8 is strictly convex with regard to $Q$ for any $\{P^n\}$. 
Proof. With the convexity of the KL divergence and positivity of $w^n$, we have the following convexity for any different points $Q, Q' \in \mathcal{M}$;

$$
\sum_{n=1}^{N} w^n \min_{P \in \mathcal{D}^n} D_{KL}(P, tQ + (1 - t)Q') \\
\leq t \sum_{n=1}^{N} w^n \min_{P \in \mathcal{D}^n} D_{KL}(P, Q)+(1-t) \sum_{n=1}^{N} w^n \min_{P \in \mathcal{D}^n} D_{KL}(P, Q'),
$$

where $0 < t < 1$. From theorem 5, the equality holds iff $Q$ and $Q'$ are in an $m$-flat subspace that is parallel to all the elements in $\{\mathcal{D}^n\}$. Note that points $Q, Q' \in \bigcap_{n=1}^{N} \mathcal{D}^n$ also satisfy the equality, and so the strict convexity is ensured by eliminating cases given by the condition $|\bigcap_{n=1}^{N}| > 1$. This completes the proof.

Finally, we give the following theorem:

**Theorem 6.** Convergence property of algorithm 1: If assumption 1 holds, the converged result of algorithm 1 is the unique and global minimizer of $F(Q)$.

Proof. From lemma 1, conditions 1 and 2 in assumption 1 ensure the strict convexity of $F(Q)$. And condition 3 in assumption 1 apparently guarantees that the minimizer $\hat{Q}$ is in $\mathcal{M}$. With the convergence property of the em algorithm, the proof is complete.

Therefore, we assume the conditions given in assumption 1 in this letter.

4 Weight Adaptation

In this section, we discuss the role of weights $\{w^n\}$ and a way to adapt them in our proposed framework.

4.1 Weight Adaptation. Figure 2 shows intuitive examples of consistent and inconsistent observations. In the usual case of Figure 2b, selection of a set of weights $\{w^n\}$ is important because it dominates the quality of the estimate $\hat{Q}$. Typical examples of a weight $w^n$ for comparison between $i$ and $j$ are:

- $w^n = 1/N$: a uniform weight
- $w^n = c(n_{ij} + n_{ji})$: the number of observations
- $w^n = c\alpha_{ij}^{n_{ij}+n_{ji}}$, where $\alpha_{ij}^{n_{ij}} = \frac{n_{ij}}{n_{ij}+n_{ji}}$ and $\alpha_{ij}^{n_{ji}} = \frac{n_{ji}}{n_{ij}+n_{ji}}$: the weight used in Hastie and Tibshirani (1998),

where $c$ is a normalization constant to ensure $\sum_{n=1}^{N} w^n = 1$. Now we discuss the tuning of weights for reliable estimation. For example, we can formulate the weight tuning problem as follows:

$$
\min_{\{w^n\}} \sum_{n=1}^{N} w^n D_{KL}(P^n, Q)
$$

subject to $w^n > 0$ ($n = 1, \ldots, N$), $\sum_{n=1}^{N} w^n = 1$.

Since simultaneous optimization with regard to $\{P^n\}$, $Q$, and $\{w^n\}$ is difficult, we consider minimizing $\{w^n\}$ for given $\{P^n\}$ and $Q$ in this problem. This type of minimization is implemented by using linear programming with the em estimation result. An iterative sequence of equations 2.8 and 4.1 will minimize the objective function $\sum_{n=1}^{N} w^n D_{KL}(P^n, Q)$. However, the obtained weight set $\{w^n\}$ will be sparse in this case; only a few elements in $\{w^n\}$ will be large, and the rest will be close to zero. With such a weight set, the estimate $Q$ has low reliability because it is sensitive to the noise for constructing $D_n$. Therefore, we propose an adaptive weight tuning method for our em estimation from a viewpoint of sensitivity.

Suppose that a vector $\alpha^n = \{\alpha^n_1, \ldots, \alpha^n_M\}$, which is used for defining a data manifold $D_n$, is contaminated with an $M$-dimensional unit vector $\Delta_m$, whose components are all zero except the $m$th element, as follows,

$$
\alpha^n \rightarrow \alpha^n(m; \varepsilon) = (1 - \varepsilon)\alpha^n + \varepsilon \Delta_m
$$

$$
= \{(1 - \varepsilon)\alpha^n_1, \ldots, (1 - \varepsilon)\alpha^n_m + \varepsilon, \ldots, (1 - \varepsilon)\alpha^n_M\},
$$

where $0 < \varepsilon \ll 1$. Given any weights (e.g., $\{w^n = 1/N\}_{n=1}^{N}$), converged points $\hat{Q}$ and $\hat{P}_n = P^\ast_{n}(\alpha^n)$ are obtained after a sequence of em steps. We consider an influence of contamination $\Delta_m$ at point $\hat{P}_n(\alpha^n)$, which measures an impact of contamination on the estimator and is intuitively given as

$$
\lim_{\varepsilon \to 0} \frac{\hat{P}_n(\alpha^n(m; \varepsilon)) - \hat{P}_n(\alpha^n)}{\varepsilon},
$$

where

$$
\hat{P}_n(\alpha^n) = \arg\min_{P \in D^n(\alpha^n)} D_{KL}(P, \hat{Q}).
$$

Since $\hat{P}_n = \{\pi^n_{i}\}_{i \in A}$ has $|A| - 1$ nonredundant parameters and $\pi^n_{|A|} = \sum_{i \in A - \{i\}} \pi^n_i$ holds, we define a $|A| - 1$ dimensional column vector $IF(m)$
for approximating the influence by using equation 3.17 with fixed $\delta$:

$$IF(m)_i = \begin{cases} 0 & (i \in \Lambda_0^n, i \neq |\Lambda|) \\ 
\frac{1 - \alpha^n_m}{q'_m} \cdot \frac{\theta_i \cdot \delta \cdot q_+}{q_0 + \delta \cdot q_+} & (i \in \Lambda^n_m \subseteq \Lambda^n_+, i \neq |\Lambda|) \\ 
\frac{\alpha^n_m}{q'_m} \cdot \frac{\theta_i \cdot \delta \cdot q_+}{q_0 + \delta \cdot q_+} & (i \in \Lambda^n_{m'} \subseteq \Lambda^n_+ - \Lambda^n_m, i \neq |\Lambda|). 
\end{cases}$$  \quad (4.2)

With equation 4.2, we can calculate the information-standardized sensitivity (Hampel et al., 1986), given by

$$ISS^n = \max_m (IF(m)^T I(\hat{P}^n)^{-1} IF(m))^{\frac{1}{2}},$$  \quad (4.3)

where $I(\hat{P}^n)$ is a $(|\Lambda| - 1) \times (|\Lambda| - 1)$ Fisher information matrix calculated at $\hat{P}^n$ on $M$, whose components are given by

$$I(\hat{P}^n)_{ij} = \begin{cases} 
\frac{1}{\hat{\pi}_{i|\Lambda}} & (i \neq j, \{i, j\} \in \Lambda - \{\Lambda\}) \\
\frac{1}{\hat{\pi}_i} + \frac{1}{\hat{\pi}_{i|\Lambda}} & (i = j, \{i, j\} \in \Lambda - \{\Lambda\}) 
\end{cases}.$$

To calculate the concrete form of the inverse matrix, let $D$ be a $(|\Lambda| - 1) \times (|\Lambda| - 1)$ diagonal matrix with $\frac{1}{\hat{\pi}_i}$ in the $(i, i)$ element and $u$ be a $|\Lambda| - 1$ dimensional column vector such that all the elements are one. With these notations, we calculate the concrete form of $I(\hat{P}^n)^{-1}$ by using the Woodbury identity:

$$I(\hat{P})^{-1} = \left( D + u \frac{1}{\hat{\pi}_{i|\Lambda}} u^T \right)^{-1}$$

$$= D^{-1} - D^{-1} u (\hat{\pi}_{i|\Lambda} + u^T D^{-1} u)^{-1} u^T D^{-1}$$

$$= D^{-1} - D^{-1} u \left( \sum_{i \in \Lambda} \hat{\pi}_i \right)^{-1} u^T D^{-1}$$

$$= D^{-1} - D^{-1} uu^T D^{-1}.$$  \quad (4.4)

With equation 4.4, we obtain the following:

$$I(\hat{P}^n)^{-1} = \begin{cases} \hat{\pi}_i - \hat{\pi}_i^2 & (i = j) \\
-\hat{\pi}_i \hat{\pi}_j & (i \neq j).
\end{cases}$$


Intuitively the sensitivity shows unreliability derived from the worst-case scenario. For example, $I F (m)_i$ tends to be large when the data manifold $D^n$ is very close to a boundary of the probability simplex. In this case, some of the components in the vector $\alpha^n$ are extremely small values, and $1 - \alpha^n_\ell$ or $\alpha^n_\ell$ in equation 4.2 could be largely affected by the location of $\hat{Q}$. Our idea is that for stable estimation, weights $\{w^n\}$ should be determined so as to make overall sensitivity of $\{\hat{P}^n\}$ as small as possible. We define the sensitivity at $\hat{Q}$ as a function of $\{w^n\}$:

$$\text{Sensitivity}(w; \hat{Q}) = \sum_{n=1}^{N} (w^n \cdot \text{ISS}^n)^2. \quad (4.5)$$

Our final goal is to obtain minimizers $\{P^n\}$, $Q$, and $\{w^n\}$ for multicriteria equations 2.8 and 4.5 simultaneously. This type of optimization is known as the multiobjective optimization problem, and it has several Pareto solutions. Since we currently do not have any criteria to choose one of Pareto solutions, we attempt to obtain a preferable weight set $\{w^n\}$ from the viewpoint of equation 4.5 and minimizers $\{P^n\}$ and $Q$ in equation 2.8.

In our method, we first obtain em estimates $\{P^n\}$ and $Q$ with initial weights. Then we minimize equation 4.5 subject to the constraint $\sum_{n=1}^{N} w^n = 1$ with these estimates. The optimal point is given as a saddle point of the following Lagrangian,

$$L(w, \lambda) = \sum_{n=1}^{N} (w^n \cdot \text{ISS}^n)^2 + \lambda \left( 1 - \sum_{n=1}^{N} w^n \right), \quad (4.6)$$

which is a solution of the following equation,

$$\frac{\partial}{\partial w^n} \left\{ \sum_{n=1}^{N} (w^n \cdot \text{ISS}^n)^2 + \lambda \left( 1 - \sum_{n=1}^{N} w^n \right) \right\} = 0, \quad (4.7)$$

where $\lambda$ is a Lagrange multiplier. From equation 4.7, the minimizer $\{\hat{w}^n\}$ is given by

$$\hat{w}^n = \frac{\lambda'}{(\text{ISS}^n)^2}, \quad (4.8)$$

where $\lambda' = 2\lambda = \sum_{n=1}^{N} \frac{1}{(\text{ISS}^n)^2}$. Then we update $\{P^n\}$, $Q$ by the em estimation with a new weight set $\{w^n\}$ again. With equation 4.8, we formulate a modified estimation method, weight adaptation:
Algorithm 2: em Estimation with Weight Adaptation

**input** restrictions \( \{\Lambda^m\}_{n=1}^N = \{(\Lambda^m_n)_{m=1}^M\}_{n=1}^N \) and the number of iteration \( S \).

**initialize** choose an initial parameter set \( Q(0) \) and initial weights \( \{w_n(0)\} \), and construct \( D^m \) from \( \{\Lambda^m\} \) and \( \{\alpha^m\} \).

**em estimation** obtain converged distributions \( \{\hat{P}^n(1) \in D^m\} \) and \( \hat{Q}(1) \) with initial weights \( \{w_n(0)\} \) by em estimation.

**repeat** from \( s = 1 \), until \( s = S \).

**weight adaptation** calculate information-standardized sensitivities \( \{IS^n(s)\} \) with current distributions \( \{\hat{P}^n(s)\} \) as

\[
IS^n(s) = \max_m (IF(m)^T I(\hat{P}^n(s))^{-1} IF(m))^\frac{1}{2},
\]

and update \( \{w^n\} \) as

\[
w^n(s+1) = \frac{1/(IS^n(s))^2}{\sum_{n'=1}^N 1/(IS^n(s'))^2}.
\]

**em estimation** obtain converged distributions \( \{\hat{P}^n(s+1) \in D^m\} \) and \( \hat{Q}(s+1) \) with current weights \( \{w^n(s+1)\} \) by em estimation.

**output** final estimate \( \hat{Q} \)

At each weight adaptation step in algorithm 2, the updated weight \( \{w^n\} \) is uniquely obtained. Also at each em estimation step, the estimates \( \{P^n\} \) and \( Q \) are uniquely obtained. It is difficult to discuss the convergence property of this type of algorithm for a multiobjective optimization problem; however, in our experiments, weights stably converge to unique values from various initial weights within several iterations. In the numerical experiments shown in the next section, we use uniform weights as initial weights.

4.2 Weight Adaptation from a Viewpoint of Divergence. Now we consider an effect of the adapted weights \( \{\hat{w}^n\} \) from the viewpoint of employing other types of divergences instead of KL divergence. In section 4.1, we derived an influence of \( \Delta m \) in \( D^n \), denoted as \( IF(m) \), for calculating sensitivity. The sensitivity of the estimate is known to be deeply related to the divergence used for estimation. Here, we introduce a type of the Bregman divergence (Murata et al., 2004) called the \( \eta \)-divergence (Takenouchi & Eguchi, 2004; Takenouchi et al., 2008). The \( \eta \)-divergence between \( P = \{\pi_i\}_{i \in \Lambda} \in D \) and \( Q = \{\theta_i\}_{i \in \Lambda} \in M \) is given by

\[
D_\eta(P, Q) = \sum_{i \in \Lambda} \left( \frac{\pi_i - \eta}{|\Lambda|} \right) \log \frac{\pi_i - \frac{n}{|\Lambda|}}{\theta_i - \frac{n}{|\Lambda|}}.
\]

Let \( \rho_\eta(P) = \{(1-\eta)\pi_i + \frac{\eta}{|\Lambda|}\}_{i \in \Lambda} \) and \( \rho_\eta(Q) = \{(1-\eta)\theta_i + \frac{\eta}{|\Lambda|}\}_{i \in \Lambda} \) be two distributions derived from \( \hat{P} \) and \( \hat{Q} \); both \( \rho_\eta(P) \) and \( \rho_\eta(Q) \) approach the
uniform distribution as \( \eta \) goes to one. Then the \( \eta \)-divergence has a property:

\[
D_\eta(\rho_\eta(P), \rho_\eta(Q)) = \sum_{i \in \Lambda} (1 - \eta) \pi_i \log \frac{(1 - \eta) \pi_i}{(1 - \eta) \theta_i} - \sum_{i \in \Lambda} (1 - \eta) \pi_i \log \frac{\theta_i}{\pi_i} - (1 - \eta) D_{KL}(P, Q).
\] (4.9)

With this property, we can interpret adapted weights \( \{\hat{w}^n\} \) in the objective function equation 2.9 by substituting values \( \eta^n = 1 - w^n \) as follows:

\[
\hat{Q} = \arg\min_{Q \in M} \left\{ \sum_{n=1}^N (1 - \eta^n) \min_{P \in D^n} D_{KL}(P, Q) \right\}
\]

\[
= \arg\min_{Q \in M} \left\{ \sum_{n=1}^N \min_{P \in D^n} D_{\eta^n}(\rho_\eta(P), \rho_\eta(Q)) \right\}.
\]

In Takenouchi and Eguchi (2004), the empirical distribution \( \rho_\eta(P) \) is assumed to be a probability distribution \( P \) that is contaminated with 0/1 noises with an error probability \( \eta \), and a probability distribution \( Q \) is estimated with \( \rho_\eta(P) \). In our method, the distributions \( \rho_\eta(P) \) and \( \rho_\eta(Q) \) do not have explicit meanings like the effect of errors. However, we can give them another interpretation, as shown in Figure 7. Assume that \( Q' = \{(1 - \eta) \theta_i\}_{i \in \Lambda} \) is a point on a shrunk simplex \( M_\eta \), which is defined as

\[
M_\eta = \left\{ Q' = \{(1 - \eta) \theta_i\}_{i \in \Lambda} \left| \sum_{i \in \Lambda} \theta_i = 1, \theta_i > 0 \right. \right\},
\]

and \( P' = \{(1 - \eta) \pi_i\}_{i \in \Lambda} \) is a point on a manifold \( D_\eta \) in \( M_\eta \), defined as

\[
D_\eta = \left\{ P' = \{(1 - \eta) \pi_i\}_{i \in \Lambda} \in M_\eta \left| \sum_{i \in \Lambda_0} \pi_i = 1 - \psi, \sum_{i \in \Lambda_m} \pi_i = \psi \alpha_m, \right. \right. \psi \in (0, 1) \right\}.
\]

From equation 4.9, the weighted KL divergence \( (1 - \eta) D_{KL}(P, Q) \) in the original simplex \( M \) is equivalent to the KL divergence \( D_{KL}(P', Q') \) in a shrunk
1652 Y. Fujimoto, H. Hino, and N. Murata

Figure 7: Interpretation of weight from a viewpoint of the $\eta$-divergence. (a) An arrow from $P$ to $Q$ indicates $D_{KL}(P, Q)$ with weight $1 - \eta$. (b) An arrow from $P'$ to $Q'$ indicates $D_{KL}(P', Q')$ in a shrunk simplex. Note that $(1 - \eta)D_{KL}(P, Q) = D_{KL}(P', Q')$ holds for these figures.

simplex $\mathcal{M}_\eta$. Therefore, the proposed weight adaptation method has an aspect of selecting appropriate divergence for each data manifold from the viewpoint of its sensitivity. Since elements of $IF(m)$ of $D_\eta(\rho_\eta(P), \rho_\eta(Q))$ are obviously $(1 - \eta)$ times those of $D_{KL}(P, Q)$ from equation 4.9, an unreliable $P^n$ that has a large $ISS^n$ value is evaluated by the $\eta$-divergence with small $(1 - \eta)$ value, and vice versa.

5 Experimental Results

In this section, we compare the proposed methods with the conventional method and discuss some candidates for weights in the BT estimation experiments on typical pairwise comparison settings.

We prepare three types of probability distributions $P^*$ used in Hastie and Tibshirani (1998):

$$P^*_A = \left\{ \pi_i^* \mid \pi_1^* = \frac{1.5}{|\Lambda|}, \pi_j^* = \frac{1 - \pi_1^*}{|\Lambda| - 1} (j = 2, \ldots, |\Lambda|) \right\},$$

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/NECO_a_00129 by guest on 02 May 2021
\[
P_B^* = \left\{ \pi_i^* \mid \pi_1^* = 2.85, \pi_j^* = \frac{0.95 - \pi_1^*}{|\Lambda|/2 - 1} \quad (j = 2, \ldots, |\Lambda|/2) \right\},
\]
\[
\pi_j^* = \frac{0.05}{|\Lambda|/2} \quad (j = \left\lfloor \frac{|\Lambda|}{2} \right\rfloor + 1, \ldots, |\Lambda|)
\]
\[
P_C^* = \left\{ \pi_i^* \mid \pi_1^* = 0.7125, \pi_2^* = 0.2375, \pi_j^* = \frac{0.05}{|\Lambda| - 2} \quad (j = 3, \ldots, |\Lambda|) \right\}.
\]

For each distribution, we set \(m\)-flat manifolds \(\{D_n\}_{n=1}^{N}\) where \(N = \left\lfloor \frac{|\Lambda|}{2} \right\rfloor\), which are constructed based on all the one-versus-one observations, with \(\alpha^n\) for comparison of \(i\) and \(j\) as

\[
\alpha^n = \left\{ \alpha_1^n = \min\left(0.9999, \max\left(0.0001, \frac{\pi_i^*}{\pi_i^* + \pi_j^*} + 0.1z_{ij}\right)\right), \right.\]
\[
\alpha_2^n = 1 - \alpha_1^n \right\},
\]

where \(z_{ij}\) is a standard normal variate.

At first, we show the results of our estimation methods, given by algorithms 1 and 2. For algorithm 1, we prepare three types of weight for comparison between \(i\) and \(j\), defined as

\[
w_u^n = c_u, \quad w_o^n = c_o (\pi_i^* + \pi_j^*), \quad w_v^n = \frac{c_v}{\alpha_1^n \cdot \alpha_2^n},
\]

where \(c_u, c_o,\) and \(c_v\) are constants to hold \(\sum_{n=1}^{N} w^n = 1\). Note that \(w_u^n\) gives the uniform weight, \(w_o^n\) gives an ideal ratio related to \(i\) and \(j\) in all the items, and \(w_v^n\) gives a weight introduced in Hastie and Tibshirani (1998). Figure 8 shows estimation results with medians and quartiles of \(D_{KL}(P^*, \hat{Q})\) for 500 trials with \(|\Lambda| = 3, 5, 10, 15, 20\), respectively. As shown in the figure, the quality of the estimates based on our proposed method with the weights \(\{w_v^n\}\) tends to be low. The results indicate that the estimation with invalid weights causes an overfitting problem, which means excessive emphasis on unreliable data manifolds, and it easily leads to inaccurate results. The figure indicates, however, that the weight adaptation method works appropriately in most cases, especially in the case of \(P_C^*\), where few items with significantly large probabilities dominate observations and the other items have very low probabilities. Huang et al. (2006) mentioned that a typical situation in practical problems resembles the case of \(P_C^*\). Our experimental results suggest that the weight adaptation method gives better estimates in such a case. This is because the adapted weights tend to have a lower value for a
Figure 8: Plots of estimation results with typical weights. Medians and quartiles of $D_{KL}(P^*, \hat{Q})$ for 500 trials with various $|\Lambda|$ are plotted. A2: algorithm 2 with $S = 5$; A1,u, A1,o, and A1,v: algorithm 1 with $\{w^n_u\}$, $\{w^n_o\}$, and $\{w^n_v\}$, respectively.

Figure 9: Plots of estimation results obtained by conventional methods and our proposed method. $D_{KL}(P^*, \tilde{Q})$ and $D_{KL}(P^*, \hat{Q})$ for 500 trials are plotted. conv,u, conv,o and conv,v: conventional estimation with $\{w^n_u\}$, $\{w^n_o\}$, and $\{w^n_v\}$, respectively.

sensitive data manifold, which compares an item with very large probability with one with small probability.

Second, we show the comparison results of conventional estimation methods with $\{w^n_u\}$, $\{w^n_o\}$, $\{w^n_v\}$, and algorithm 2 in the same setting. As shown in Figure 9, the conventional method with $\{w^n_o\}$ ends up with low-quality estimates as well as our algorithm 1. The results show that
invalid weights in the conventional methods also cause the overfitting. In this respect, it seems that the weight adaptation method is possible to avoid overfitting, and the estimation accuracy of our proposed method is basically better than that of conventional methods.

Finally, we discuss our proposed method with weight adaptation from the viewpoint of robustness. Figure 10 shows a relation between estimation performance $D_{KL}(P^*, \tilde{Q})$ at $|\Lambda| = 20$ and a missing ratio of comparisons, that is, a certain ratio of manifolds in $\{D_n\}_{n=1}^N$ is randomly chosen and discarded. As shown in the figure, our weight adaptation seems a little sensitive to missing observations, but it is still better than the conventional methods, especially in the case of $P_C^*$.  

6 Discussion and Concluding Remarks

In this letter, we reinterpreted an estimation with the Bradley-Terry model and its related models in the space of categorical distributions and then proposed a new estimation method based on a framework of the information-geometric $em$ algorithm. Our proposed method is different in its objective function from conventional methods. With the KL decomposition given by equation 3.15, the objective function equation 2.8 is rewritten as

$$F(Q) = \sum_{n=1}^N w^n \psi^n D_{KL}(P^{n(b)}, Q^{n(b)}) + \sum_{n=1}^N w^n D_{KL}(P^{n(r)}, Q^{n(r)}),$$

(6.1)
which is the essential part of equation 3.15. There are two important differences between equations 2.10 and 6.1. The first one is $\psi^n$ in the first term of equation 6.1. Weight $w^n$ for the $n$th $m$-flat data manifold is adjusted by a hidden variable $\psi^n$, which is implicitly estimated in the $em$ steps. The other difference is the second term of equation 6.1. This term evaluates the weighted sum of KL divergences between \{P^n(r)\} and \{Q^n(r)\}. In this term, $D_{KL}(P^n(r), Q^n(r))$ indicates the difference between the ratio $(1 - \psi^n) : \psi^n$ defined by the $e$-projection from $Q$ onto $D^n$ and the ratio of directly summed parameters $\sum_{i \in \Lambda_0^n} \theta_i : \sum_{j \in \Lambda_+} \theta_j$. A geometrical interpretation of $D_{KL}(P^n(r), Q^n(r))$ in this term is shown in Figure 11. Intuitively, the KL divergence between categorical distributions is naturally decomposed into two different divergences: the divergence between distributions of compared items and the divergence between distributions associated with the ratio of uncompared items. Equation 6.1 evaluates these two kinds of consistency of the parameter set. The difference between equation 2.10 and 6.1 becomes large when point $Q$ is far from $D^n$. Therefore, point $Q$ tends to be close to $D^n$ not only in the conventional way, but also from the viewpoint of estimating a set of hidden variables $\{\psi^n\}$.

In addition, we discussed the selection of weights for divergences in the objective function and proposed an estimation method with weight adaptation from a viewpoint of sensitivity in the simplex. We also considered the relation between weights and divergences with a concrete type of the Bregman divergence. Experimental results showed that the proposed method with weight adaptation works appropriately and improves the accuracy of the estimates. As Râtsch, Smola, and Mika (2002) noted, the BT-like models generally have a problem with selecting reliable data manifolds, a code optimization problem in the context of multiclass classification. We believe that the proposed weight adaptation method shows another way to alleviate this problem.
Figure 12: A simple way to represent ties. (a) Representation of a data manifold defined by $\Lambda_0^n = \{2\}, \Lambda_1^n = \{T\}, \Lambda_2^n = \{1\},$ and $\Lambda_3^n = \{3\}$ in the extended simplex. (b) Projection of the estimate onto the original $|\Lambda| - 1$-dimensional simplex.

An important topic to extend our method is a formulation for observations of ties. When there are ties in pairwise comparisons, observed events in comparison between $i$ and $j$ are denoted as $i \succ j$ ($i$ is chosen), $i \prec j$ ($j$ is chosen), and $i = j$ (tie in the comparison of $i$ and $j$). In the conventional framework, there are some extensions for the BT model to treat ties by introducing new tie parameters (Rao & Kupper, 1967; Davidson & Beaver, 1977; Kuk, 1995). Here we sketch a possible way to treat ties in our framework.

Let $Q' = \{\theta'_{ij} > 0\}$ be a modified parameter set and $\tau > 0$ be a tie parameter where $\sum_{i \in \Lambda} \theta'_{ij} + \tau = 1$ holds. We can represent them in an $|\Lambda|$ probability simplex with a dummy item $T$ as shown in Figure 12a. In this simplex, $\{\Lambda_m^n\}$ and $\{\alpha_m^n\}$ for the comparison of two items $i, j \in \Lambda$ with tie observations is given as follows:

\[
\{\Lambda_m^n\} = \{\Lambda_0^n = \Lambda - \{i, j, T\}, \Lambda_1^n = \{T\}, \Lambda_2^n = \{i\}, \Lambda_3^n = \{j\}\},
\]

\[
\{\alpha_m^n\} = \left\{\alpha_1^n = \frac{t_{ij}}{n_{ij} + n_{ji} + t_{ij}}, \alpha_2^n = \frac{n_{ij}}{n_{ij} + n_{ji} + t_{ij}}, \alpha_3^n = \frac{n_{ji}}{n_{ij} + n_{ji} + t_{ij}}\right\}.
\]

With this type of data manifold, we can obtain estimates $\hat{\tau}$ and $\hat{Q'}$ for $\tau$ and $Q$. If we need $\hat{Q'}$ to be a probability distribution, a kind of projection onto the reduced dimensional simplex will help it, as shown in Figure 12b. The conventional extensions also should be compared with our extension, though it remains work for the future.

There are some other ways to improve our method. The first is the convergence speed of the $em$ estimation. The speed of convergence in the proposed method is slower than the conventional one, which is implemented with the fixed-point method used in Huang et al. (2006). It could
be a problem when online estimation is needed for large $|\Lambda|$. To improve convergence speed, we need to devise an acceleration for the $\text{em}$ algorithm, like the acceleration methods featured in McLachlan and Krishnan (1996) and Watanabe and Yamaguchi (2003) for the EM algorithm. The second is selection of weights and divergences. As we stated in section 4, if we have a criterion for choosing a Pareto solution, we should optimize the given multiobjective problem in the proper way. In addition, we also noted that weights and divergences in the objective function are mutually connected and deeply related to robustness. We expect that a more appropriate objective function can be designed in terms of robustness with, for example, the Bregman divergence, which is known to be a convenient class for $m$-flat data manifolds (Murata et al., 2004).

References

Estimation of Generalized Bradley-Terry Models


Received April 26, 2010; accepted November 2, 2010.