

## A discrete probabilistic model for analyzing pairwise comparison matrices

Kurata, Sumito  
Graduate School of Engineering Science, Osaka University

Hamada, Etsuo  
Graduate School of Engineering Science, Osaka University

<https://hdl.handle.net/2324/7179469>

---

出版情報 : Communications in Statistics - Theory and Methods. 48 (15), pp.3801-3815, 2018-10-29. Taylor and Francis  
バージョン :  
権利関係 :



# A DISCRETE PROBABILISTIC MODEL FOR ANALYZING PAIRWISE COMPARISON MATRICES

Sumito Kurata and Etsuo Hamada

Graduate School of Engineering Science

Osaka University

1-3 Machikaneyama, Toyonaka, Osaka 560-8531, JAPAN.

kurata@sigmath.es.osaka-u.ac.jp

Key Words: Pairwise comparison matrix; Multiple-criteria decision analysis; Analytic hierarchy process; BHHJ divergence; Robustness.

## Abstract

The pairwise comparison matrix is often used for the estimation of the priorities in the analytic hierarchy process. In this paper, we propose an estimation method based on the discrete probabilistic expression of each choice. Moreover, we show numerical examples to compare our method with commonly used ones. As a result, it is shown that, using a robust divergence measure for the estimation, the proposed method can extract the priorities more stably even if some outlying observations are included.

## 1 Introduction

The analytic hierarchy process (AHP) and the analytic network process (ANP) established by Saaty (1977) [20], (2005) [23] are herein developed based on the pairwise comparison matrix (PCM) whose elements indicate the degree of *priority* (importance or excellence) of an evaluation object compared with another object. Generally, it is complex and difficult to determine the relative degrees of the priority of each object for the answerers (decision makers), whereas one-to-one comparisons can be made much more easily. Thus, we often use the PCM to extract the priority ratios of evaluation objects from pairwise comparisons. Let  $O_1, \dots, O_M$  be the objects that we need to assign priorities. These objects correspond

to the evaluation criteria or alternatives in the AHP and ANP. In this paper, we denote by  $\tilde{\pi} = (\pi_1, \dots, \pi_M)^T$  the priority vector of the objects  $O_1, \dots, O_M$ . Since the observations are generated on the basis of uncertain human decisions, it is natural that there exist some outliers (distinctive, unusual, or mistaken data that is distant from other observations). Accordingly, the analysis procedure should extract the priority stably even if the data are contaminated. The aim of this study is to estimate the priority vector  $\tilde{\pi}$  more robustly.

Let  $\mathbf{A} = (a_{i,j})_{i,j}$  be the PCM. Each  $a_{i,j}$  indicates the relative importance of object  $O_i$  compared with  $O_j$ . Since  $a_{i,i} = 1$  and  $a_{j,i} = 1/a_{i,j}$  must always hold, we should consider the  $a_{i,j}$ 's for only  $i < j$ . Note that we are interested in not the individual values of the object priorities themselves but their fractions of the total priority. The value of  $\tilde{\pi}$  is invariant to multiplication by a constant. In many studies about the AHP and ANP, a constraint  $\pi_1 + \dots + \pi_M = 1$  is given, each priority is corrected to be an integer, or one element of the priority vector is fixed as 1.

A representative estimation method for the PCMs is the eigenvalue method by Saaty (1977) [20]. When *consistency* holds, namely, the PCM  $\mathbf{A}$  is equal to  $\bar{\mathbf{A}} = (\pi_i/\pi_j)_{i,j}$ , the relation  $\bar{\mathbf{A}}\tilde{\pi} = M\tilde{\pi}$  holds, where the priority vector  $\tilde{\pi}$  can be regarded as the eigenvectors. From this, we can estimate the priority vector by calculating the eigenvectors that correspond to the maximum eigenvalue of the observed PCM  $\mathbf{A}$ .

Another well-known method is the geometric mean method (Saaty and Vargas (1984) [24]). Each priority  $\pi_m$  is estimated by a calculation using the following equation:

$$\frac{\sqrt[M]{a_{m,1} a_{m,2} \cdots a_{m,M}}}{\sum_{m'=1}^M \sqrt[M]{a_{m',1} a_{m',2} \cdots a_{m',M}}} \quad (m = 1, \dots, M).$$

A method similar to the geometric mean method is the harmonic mean method. In this method, priorities are estimated by calculating the harmonic mean:

$$\frac{\frac{1}{a_{m,1}} + \frac{1}{a_{m,2}} + \cdots + \frac{1}{a_{m,M}}}{\sum_{m'=1}^M \left( \frac{1}{a_{m',1}} + \frac{1}{a_{m',2}} + \cdots + \frac{1}{a_{m',M}} \right)} \quad (m = 1, \dots, M).$$

For details, see Kato (2013) [11].

In addition, Lipovetsky and Conklin (2002) [14] proposed using the pairwise proportion matrix (PPM),  $(\pi_i/(\pi_i + \pi_j))_{i,j}$ , instead of the PCM. They established a procedure to estimate

the priority for “diminishing the influence of unusual values” relative to the eigenvalue or geometric mean method empirically. The estimator is obtained by calculating the eigenvector corresponding to the maximum eigenvalue of the matrix  $\mathbf{B} + \text{diag}(\mathbf{B}\mathbf{1}_M)$ , where  $\mathbf{B}$  is the observed PPM:  $(b_{i,j})_{i,j} = (a_{i,j}/(1 + a_{i,j}))_{i,j}$  and  $\mathbf{1}_M$  is the  $M$ -dimensional vector whose elements are all equal to 1.

Note that pairwise comparison is not always conducted by only one person. To apply these previous methods, the summarized PCM or PPMs obtained by taking the geometric mean of the individual matrices are often used. Now, we assume that there are  $N$  ( $\geq 2$ ) answerers who make the decision individually; thus, there are also  $N$  different PCMs. We denote by  $\mathbf{A}^{(n)} = (a_{i,j}^{(n)})_{i,j}$  the PCM of the  $n$ -th answerer ( $n = 1, \dots, N$ ).

Basak (1989) [2], (2002) [3] introduced probability distributions into the PCM problem and designed an algorithm for obtaining the maximum likelihood estimator (MLE) of  $\pi_i$ 's. Specifically, the pairwise comparison values  $a_{i,j}^{(n)}$  by the  $n$ -th answerer were considered to have been drawn from the following distribution structure:

$$a_{i,j}^{(n)} = \frac{\pi_i}{\pi_j} \epsilon_{i,j}^{(n)} \quad (i < j, n = 1, \dots, N) \quad (1)$$

where each  $\epsilon_{i,j}^{(n)}$  follows a certain (positive-valued) distribution independently. In particular, the estimator is equal to the result of the geometric mean method if we assume that the  $\epsilon_{i,j}^{(n)}$ 's have the log-normal distribution and we estimate the priority vector by the maximum likelihood method.

In the geometric mean method and Basak's method, it is assumed that each observed pairwise comparison value  $a_{i,j}^{(n)}$  is the product of the true priority ratio  $\pi_i/\pi_j$  and the error term, and that the error is continuously distributed. However, in fact, since each  $a_{i,j}^{(n)}$  is one element in a discrete choice set  $\mathcal{C}$  in most cases, there exists some gap between the assumption and the observed answers (we will describe the choice set in the next section). In contrast, using a continuous choice set (e.g., interval  $[1/9, 9]$ ) to correct the gap could cause trouble for the answerers because of the complexity. Thus, in this paper, we define a stochastic model named the discrete probabilistic model (DPM) that gives the answered probability of each choice based on the discrete distribution.

Section 2 describes the DPM, including its asymptotic properties. In Section 3, we show examples of the method in comparison with conventional methods. By using estimation based on a family of divergence measures, it can extract the priorities of the evaluation objects robustly. Section 4 summarizes this study. Some proofs of theorems are in the Appendix.

## 2 The discrete probabilistic model (DPM)

In this section, we introduce a probabilistic model for the PCM and describe the method to obtain an estimator of the priority vector.

### 2.1 A discrete representation of the selected probability

To evaluate how important (or excellent) one object is relative to another, we need a set of numbers that indicate importance (or excellence). As the set of choices for the answerers,

$$\begin{aligned}\mathcal{C}_{S7} &= \{1/7, 1/5, 1/3, 1, 3, 5, 7\} , \\ \mathcal{C}_{S9} &= \{1/9, 1/7, 1/5, 1/3, 1, 3, 5, 7, 9\} ,\end{aligned}$$

or a detailed version that has 17 grades

$$\mathcal{C}_S = \{1/9, 1/8, 1/7, \dots, 1/2, 1, 2, \dots, 7, 8, 9\}$$

are often used (e.g., Saaty (2004) [22]). The integers are regarded as how many times more important, and their reciprocals represent the unimportance.

Alternatively, the power-type set

$$\mathcal{C}_{(\psi, \omega)} = \{\psi^p \mid p = -\omega, \dots, -1, 0, 1, \dots, \omega\}$$

can also be used (e.g., Harker and Vargas (1987) [10]). The logarithmic conversion of elements in this set implies a regular interval  $\log \psi$ .

Generally, the set of choices comprises phrases like “Equal importance” and “Much more important” rather than numbers. Saaty (2004) [22] classified the degrees of “importance”

into “moderate”, “strong”, “very strong”, and “extreme”. When we use  $\mathcal{C}_{(\psi,3)}$ , the degrees of “importance” can be divided into “slightly more important”, “more important”, and “much more important”, and when  $\mathcal{C}_{(\psi,1)}$  is used, the choices become the following three: “unimportant”, “equal”, and “important”.

Now, we consider that the answered probability (frequency) of each choice is based on the “difference” from the true priority ratio  $\pi_i/\pi_j$ . As the “difference” between each  $c \in \mathcal{C}$  and the true  $\pi_i/\pi_j$ , for instance,  $|c - \pi_i/\pi_j|$ ,  $(\log c - \log \pi_i/\pi_j)^4$ ,  $\exp\{(c - \pi_i/\pi_j)^2\}$ , and the like are usable. Next, we give the definition of the probability that each choice is answered by the answerers and the models.

**Definition 1.** *We define the answered probability of each choice  $c \in \mathcal{C}$  of the comparison between the evaluation object  $O_i$  and  $O_j$  as*

$$p_{i,j}(c) = p_{i,j}(c | \tilde{\pi}) := \mathbf{P}\{c \text{ is answered} \mid \text{given } \pi_i/\pi_j\} \quad (i < j) .$$

*To construct the answered probability, we introduce a family of functions*

$$\tilde{p}_{i,j}(c) = \tilde{p}_{i,j}(c | \pi_i/\pi_j) := \eta \left( (\log c - \log \pi_i/\pi_j)^2 \right) ,$$

*where  $\eta$  is from the class of real functions  $\eta(z)$  ( $z > 0$ ) that are continuous, twice differentiable, and monotonically decreasing. Because the function  $\tilde{p}_{i,j}(\cdot)$  is not always a probability function, we must normalize it as  $p_{i,j}(c) = \tilde{p}_{i,j}(c) / \sum_{c' \in \mathcal{C}} \tilde{p}_{i,j}(c')$ .*

As the function  $\tilde{p}_{i,j}$ , for example, we can define the following:

$$\begin{aligned} \tilde{p}_{i,j}^2(c) &= \frac{1}{1 + (\log c - \log \pi_i/\pi_j)^2} , \\ \tilde{p}_{i,j}^4(c) &= \frac{1}{1 + (\log c - \log \pi_i/\pi_j)^4} , \\ \tilde{p}_{i,j}^e(c) &= \exp \left\{ - (\log c - \log \pi_i/\pi_j)^2 \right\} . \end{aligned} \tag{2}$$

These forms correspond to  $\eta(z) = 1/(1+z)$ ,  $1/(1+z^2)$ , and  $\exp(-z)$ , respectively. In applying such a model, it is preferable to use  $\mathcal{C}_{(\psi,\omega)}$  as the choice set, because its elements are distributed at regular intervals after taking the logarithm. In the rest of this subsection,

Table 1: Elements (choices) in  $\mathcal{C}_{(2,3)}$  and their definitions

Element	Definition
1/8	Much more unimportant than
1/4	More unimportant than
1/2	Slightly more unimportant than
1	Two objects have equal importance
2	Slightly more important than
4	More important than
8	Much more important than

we use  $\mathcal{C}_{(2,3)}$ . The elements and their definitions (given to the answerers) are shown in Table 1.

The answered probability is defined to express how often each choice is answered based on a given true priority ratio. For instance, if the true ratio  $\pi_i/\pi_j$  is equal to 1, it is reasonable to suppose that the choice “1 (two objects  $O_i$  and  $O_j$  have equal importance)” will be frequently answered, and the choices “1/8 ( $O_i$  is much more unimportant than  $O_j$ )” and “8 ( $O_i$  is much more important than  $O_j$ )” will be seldom answered. Similarly, many people may answer “1/4 ( $O_i$  is more unimportant than  $O_j$ )” or “1/2 ( $O_i$  is slightly more unimportant than  $O_j$ )” if the true ratio is 1/3, and few people will answer “1/8”, “1/4”, or “1/2” if the true ratio is 10. The shapes of these models are shown in Figures 1, 2, and 3.

## 2.2 Estimation of the priority vector

We assume that each pairwise comparison is made independently by the answerers. Since the true ratio differs depending on the compared objects, the probability distribution of each

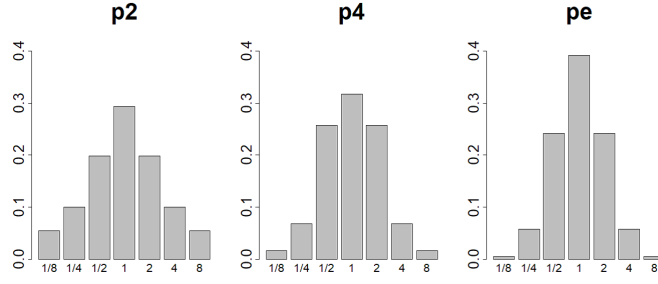


Figure 1: True ratio = 1

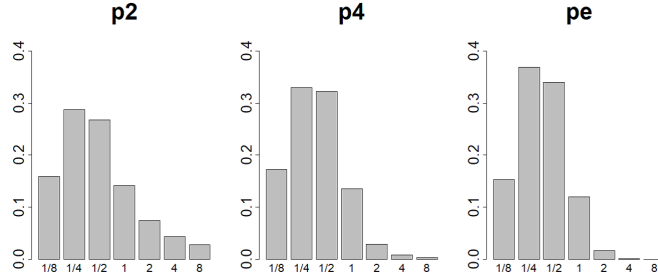


Figure 2: True ratio = 1/3

element of a PCM is nonhomogeneous:

$$\begin{aligned}
 a_{1,2}^{(1)}, \dots, a_{1,2}^{(N)} &\stackrel{i.i.d.}{\sim} \text{Multinomial} \left( 1, (p_{1,2}(c | \tilde{\pi}))_{c \in \mathcal{C}} \right), \\
 a_{1,3}^{(1)}, \dots, a_{1,3}^{(N)} &\stackrel{i.i.d.}{\sim} \text{Multinomial} \left( 1, (p_{1,3}(c | \tilde{\pi}))_{c \in \mathcal{C}} \right), \\
 &\vdots \\
 a_{M-1,M}^{(1)}, \dots, a_{M-1,M}^{(N)} &\stackrel{i.i.d.}{\sim} \text{Multinomial} \left( 1, (p_{M-1,M}(c | \tilde{\pi}))_{c \in \mathcal{C}} \right).
 \end{aligned}$$

In the maximum likelihood method, we obtain the MLE by maximizing the likelihood function, in other words, minimizing the KL divergence (Kullback and Leibler (1951) [12]). Additionally, there are many statistical divergence measures besides the KL divergence. One of these is the BHHJ divergence family (Basu *et al.* (1998) [4], Ghosh and Basu (2013) [9]), which has more robustness than the KL divergence when estimating unknown parameters. We will next introduce the definition of the BHHJ divergence family. Hereafter, we denote by  $r_{i,j}$  the probability function of the true distribution that generates the observed answers



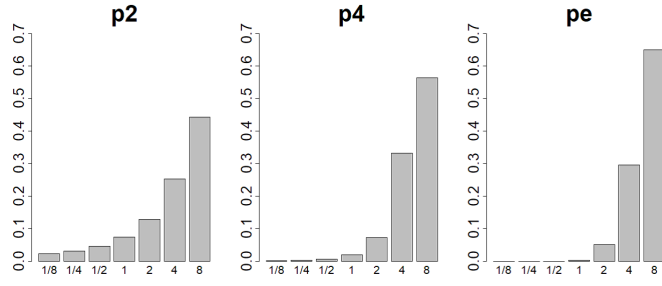


Figure 3: True ratio = 10

and that we want to estimate by using the model.

**Definition 2.** *The BHHJ divergence family (including the KL divergence) is defined as follows:*

$$\begin{aligned} & \frac{2}{M(M-1)} \sum_{i < j} \sum_{c \in \mathcal{C}} r_{i,j}(c) \log \frac{r_{i,j}(c)}{p_{i,j}(c | \tilde{\pi})} \quad (\alpha = 0), \\ & \frac{2}{M(M-1)} \sum_{i < j} \sum_{c \in \mathcal{C}} \left\{ p_{i,j}(c | \tilde{\pi})^{\alpha+1} - \frac{\alpha+1}{\alpha} p_{i,j}(c | \tilde{\pi})^{\alpha} r_{i,j}(c) + \frac{1}{\alpha} r_{i,j}(c)^{\alpha+1} \right\} \quad (\alpha > 0). \end{aligned}$$

Basu *et al.* (1998) [4] developed this divergence family for its robust parameter estimation. This family uses a nonnegative tuning parameter  $\alpha$  to control the trade-off between efficiency (small  $\alpha$ ) and robustness (large  $\alpha$ ) of the estimation. The family can be regarded as a generalization of the KL divergence, because the divergence of this family tends to the KL divergence as  $\alpha$  goes to 0. Thus, the estimator is close to the MLE (which estimates the unknown parameters efficiently but sensitively) when  $\alpha$  is close to 0. On the other hand, this becomes the minimum  $L^2$ -distance estimator, which can estimate the parameters robustly (but inefficiently) when  $\alpha$  is equal to 1.

Note that there are other famous divergence families that include the KL divergence, such as the power divergence (Cressie and Read (1984) [6]) and the  $\varphi$ -divergence (e.g., Csiszár and Shields (2004) [7], Pardo (2005) [19]), as well as more general families that include these families (e.g., Vonta *et al.* (2012) [25]).

The minimization function for obtaining the estimator for the BHHJ family is as follows:

$$\begin{aligned}
H_0(\tilde{\boldsymbol{\pi}}) &= -\frac{2}{N M (M-1)} \sum_{i < j} \sum_{n=1}^N \log p_{i,j}(a_{i,j}^{(n)} | \tilde{\boldsymbol{\pi}}), \\
H_\alpha(\tilde{\boldsymbol{\pi}}) &= \frac{2}{N M (M-1)} \sum_{i < j} \sum_{n=1}^N \left\{ \sum_{c \in \mathcal{C}} p_{i,j}(c | \tilde{\boldsymbol{\pi}})^{\alpha+1} - \frac{\alpha+1}{\alpha} p_{i,j}(a_{i,j}^{(n)} | \tilde{\boldsymbol{\pi}})^\alpha \right\} \quad (\alpha > 0).
\end{aligned}$$

The above minimization functions are obtained by removing from the definition of BHHJ divergence the terms that are independent of the priority vector and replacing the unknown true distribution with the empirical distribution based on the observations  $(a_{i,j}^{(n)})$ .

## 2.3 Asymptotic properties

In this paper, to ensure the uniqueness of the values of priority parameters, we fix the first priority  $\pi_1$  as 1 and estimate the other  $(M-1)$  free parameters. Hereafter, we denote by  $\boldsymbol{\pi} = (\pi_2, \dots, \pi_M)^T$  the vector of  $(M-1)$  free parameters, and let  $\Pi$  be the parameter space that is a subset of the  $(M-1)$ -dimensional positive real number subspace.

We denote the minimum BHHJ divergence estimator of  $\boldsymbol{\pi}$  with  $\alpha = 0$  (MLE) and  $\alpha > 0$  by  $\hat{\boldsymbol{\pi}}_0$  and  $\hat{\boldsymbol{\pi}}_\alpha$ , respectively. Additional notation is as follows. We denote by  $c_1, \dots, c_K$  the elements of the choice set  $\mathcal{C}$ . Let  $\mathbf{r}_{i,j} = (r_{i,j}(c_k))_{k=1,\dots,K}$  be the vector whose elements are each the true answered probability of a choice and let  $\mathbf{p}_{i,j}(\boldsymbol{\pi}) = (p_{i,j}(c_k | \boldsymbol{\pi}))_{k=1,\dots,K}$  be the answered probabilities of the model. Moreover, we define the  $KM(M-1)/2 \times (M-1)$  matrix  $\mathcal{J}(\boldsymbol{\pi})$  as follows:

$$\mathcal{J}(\boldsymbol{\pi}) = \begin{pmatrix} \mathcal{J}_{1,2}(\boldsymbol{\pi}) \\ \mathcal{J}_{1,3}(\boldsymbol{\pi}) \\ \vdots \\ \mathcal{J}_{M-1,M}(\boldsymbol{\pi}) \end{pmatrix}, \quad \text{where} \quad \mathcal{J}_{i,j}(\boldsymbol{\pi}) = \nabla \mathbf{p}_{i,j}(\boldsymbol{\pi}) \quad (i < j),$$

where  $\nabla$  denotes differentiation with respect to  $\boldsymbol{\pi}$ .

Herein, we assume the following regularity conditions.

- (A1) A “true” priority vector  $\boldsymbol{\pi}_*$  exists in the interior of  $\Pi$  such that  $p_{i,j}(c_k | \boldsymbol{\pi}_*) = r_{i,j}(c_k)$  for any  $k = 1, \dots, K$  and the pair  $(i, j)$ ,  $i < j$ .

- (A2) Each element of the true priority vector  $\pi_*$  is non-zero, positive, and finite.
- (A3) There exists  $\mathcal{V}_*$ , a convex neighborhood of  $\pi_*$  such that the probabilistic model  $p_{i,j}(\pi)$  is continuous with respect to  $\pi$  and is twice continuously differentiable with respect to  $\pi$  in  $\mathcal{V}_*$ , for any  $i < j$ .
- (A4) Each vector  $p_{i,j}(\pi)$  is injective with respect to  $\pi$ , and the inverse is continuous at  $p_{i,j}(\pi)$ ,  $\pi \in \mathcal{V}_*$ .
- (A5) The matrix  $\mathcal{J}(\pi)$  is full rank (i.e., the rank is equal to  $(M-1)$ ) in  $\mathcal{V}_*$ .

Under Assumptions (A1)–(A5), the following asymptotic properties are established for the minimum BHHJ divergence estimator (the proof of each theorem is in the Appendix).

**Theorem 1.** *Let  $\hat{p}_{i,j} = (\hat{p}_{i,j}(c_k))_{k=1,\dots,K}$  be the relative frequency vector for  $i < j$ . Then, for any  $\alpha \geq 0$ , the minimum BHHJ divergence estimator  $\hat{\pi}_\alpha$  can be expanded asymptotically as follow:*

$$\hat{\pi}_\alpha - \pi_* = -\mathcal{B}(\pi_*)^{-1} \sum_{i < j} \Upsilon_{i,j}(\pi_*)^T \text{diag} (p_{i,j}(\pi_*)^{(\alpha-1)/2}) (\hat{p}_{i,j} - p_{i,j}(\pi_*)) + o_P \left( \frac{1}{\sqrt{N}} \right) \quad (3)$$

as the number of answerers  $N$  goes to  $+\infty$ , where

$$\begin{aligned} \mathcal{B}(\pi) &= \sum_{i < j} \Upsilon_{i,j}(\pi)^T \Upsilon_{i,j}(\pi), \\ \Upsilon_{i,j}(\pi) &= \text{diag} (p_{i,j}(\pi)^{(\alpha-1)/2}) \mathcal{J}_{i,j}(\pi) \quad (i < j). \end{aligned}$$

**Theorem 2.** *The minimum BHHJ divergence estimator  $\hat{\pi}_\alpha$  is a consistent estimator of the true priorities and has asymptotic normality:*

$$\sqrt{N} (\hat{\pi}_\alpha - \pi_*) \xrightarrow{L} N(\mathbf{0}_{M-1}, \mathcal{B}(\pi_*)^{-1} \mathcal{A}(\pi_*) \mathcal{B}(\pi_*)^{-1})$$

as  $N \rightarrow +\infty$ , where

$$\begin{aligned} \mathcal{A}(\pi) &= \sum_{i < j} \Upsilon_{i,j}(\pi)^T \text{diag} (p_{i,j}(\pi)^{(\alpha-1)/2}) \Sigma_{i,j}(\pi) \text{diag} (p_{i,j}(\pi)^{(\alpha-1)/2}) \Upsilon_{i,j}(\pi), \\ \Sigma_{i,j}(\pi) &= \text{diag} (p_{i,j}(\pi)) - p_{i,j}(\pi) p_{i,j}(\pi)^T \quad (i < j). \end{aligned}$$

The minimum BHHJ divergence estimator corresponds to the MLE with  $\alpha = 0$ . On the other hand, the estimator with a somewhat large  $\alpha$  (e.g., 0.20 or 0.50) has robustness. Note that the tuning parameter has no theoretical upper limit, but the estimation sometimes becomes unreliable because of the large asymptotic variance when too large an  $\alpha$  is used. The BHHJ divergence has been applied to various situations. Basu *et al.* (2013) [5] applied this divergence to robust test statistics, and Mattheou *et al.* (2009) [17] and Mantalos *et al.* (2010a) [15] proposed the divergence information criterion (DIC) and its modification (MDIC) based on the BHHJ divergence when the observations are independent and have an identical normal distribution by using the approximation with small  $\alpha$ . Moreover, Kurata and Hamada (2018) [13] derived a robust model selection criterion family, BHHJ-C, that can be used even when the observations do not all have the same distribution.

In practice, since the occurrence of outliers cannot be completely prevented, robust estimation methods are desirable. Particularly, in the field of decision making, it is often observed that some data violate the presupposition. In the next section, we will show some examples to confirm the performance of the proposed DPM method and compare it with conventional methods.

### 3 Examples

We instantiate the methods that we have explained in previous sections by a real-data analysis and several numerical simulations.

### 3.1 Analysis of real data

We tried to estimate the priority vector of the PCM given in Saaty (1990) [21]:

$$\begin{array}{c} \begin{array}{c} \text{a} \\ \text{b} \\ \text{c} \\ \text{d} \\ \text{e} \\ \text{f} \\ \text{g} \\ \text{h} \end{array} \begin{pmatrix} & \text{a} & \text{b} & \text{c} & \text{d} & \text{e} & \text{f} & \text{g} & \text{h} \\ \text{a} & 1 & 5 & 3 & 7 & 6 & 6 & 1/3 & 1/4 \\ \text{b} & 1/5 & 1 & 1/3 & 5 & 3 & 3 & 1/5 & 1/7 \\ \text{c} & 1/3 & 3 & 1 & 6 & 3 & 4 & 6 & 1/5 \\ \text{d} & 1/7 & 1/5 & 1/6 & 1 & 1/3 & 1/4 & 1/7 & 1/8 \\ \text{e} & 1/6 & 1/3 & 1/3 & 3 & 1 & 1/2 & 1/5 & 1/6 \\ \text{f} & 1/6 & 1/3 & 1/4 & 4 & 2 & 1 & 1/5 & 1/6 \\ \text{g} & 3 & 5 & 1/6 & 7 & 5 & 5 & 1 & 1/2 \\ \text{h} & 4 & 7 & 5 & 8 & 6 & 6 & 2 & 1 \end{pmatrix} \end{pmatrix}.$$

These  $M = 8$  objects are the evaluation criteria for choosing the best house (for details, see Saaty (1990) [21]). The pairwise comparisons are based on Saaty's 17-grade set  $\mathcal{C}_S$ . As the answered probability of the DPM method, we use three types of models,  $\tilde{p}^2$ ,  $\tilde{p}^4$ , and  $\tilde{p}^e$ , given in (2).

The estimated values of the priority vector by the eigenvalue, the geometric mean (which is the same as Basak's method using the log-normal distribution), the harmonic mean, the PPM, and the DPM methods are shown in Table 2. The first priority  $\pi_a$  was fixed as 1 in each method. In the rightmost column for the DPM method, we list the value of the AIC (Akaike (1974) [1]) or the BHHJ-C (Kurata and Hamada (2018) [13]; the definition is in the Appendix) to compare the probabilistic models. Note that the values of the bias terms of the AIC ( $\alpha = 0$ ) are all the same because the number of free parameters for each model is equal to  $(M - 1)$ , whereas the bias of the BHHJ-C ( $\alpha > 0$ ) is not always uniform even if the models have the same number of free parameters.

In this case, the  $\alpha$ 's larger than 0.50 had a tendency to make the estimation unreliable due to inefficiency and the small sample size. Generally, the BHHJ divergence method with large  $\alpha$  has tended to perform well when the sample size was large (see Kurata and Hamada (2018) [13]). However, in the AHP and ANP, pairwise comparisons are not often made by a large number of answerers; thus, it is considered that using  $\alpha \leq 0.50$  is valid in this field.

Table 2: Estimated priorities by each method (EV: eigenvalue; GM: geometric mean; LN: log-normal distribution; HM: harmonic mean; PPM: pairwise proportion matrix; DPM: discrete probabilistic model)

	$\pi_a$	$\pi_b$	$\pi_c$	$\pi_d$	$\pi_e$	$\pi_f$	$\pi_g$	$\pi_h$	
EV	1.000	0.312	1.087	0.101	0.179	0.210	0.964	1.926	
GM/LN	1.000	0.358	0.850	0.111	0.203	0.242	0.955	2.000	
HM	1.000	0.412	0.879	0.220	0.342	0.350	0.894	3.531	
PPM	1.000	0.360	0.943	0.147	0.247	0.272	1.092	2.622	
DPM ( $\alpha = 0$ )									AIC
$\tilde{p}^2$	1.000	0.294	0.591	0.062	0.150	0.190	1.201	2.593	133.7
$\tilde{p}^4$	1.000	0.326	0.753	0.080	0.159	0.189	1.227	2.107	125.3
$\tilde{p}^e$	1.000	0.328	0.844	0.071	0.172	0.211	0.953	2.417	<u>124.0</u>
DPM ( $\alpha = 0.20$ )									BHHJ-C
$\tilde{p}^2$	1.000	0.297	0.569	0.066	0.152	0.193	1.256	2.588	-177.1
$\tilde{p}^4$	1.000	0.335	0.646	0.087	0.174	0.208	1.453	2.124	-182.7
$\tilde{p}^e$	1.000	0.328	0.658	0.073	0.173	0.214	1.212	2.396	<u>-183.2</u>
DPM ( $\alpha = 0.50$ )									BHHJ-C
$\tilde{p}^2$	1.000	0.298	0.545	0.064	0.153	0.196	1.307	2.571	-37.6
$\tilde{p}^4$	1.000	0.338	0.582	0.095	0.188	0.220	1.601	2.188	-40.8
$\tilde{p}^e$	1.000	0.327	0.550	0.075	0.173	0.215	1.496	2.458	<u>-41.3</u>

We will discuss this problem in more detail in the next subsection.

### 3.2 Numerical simulations

We designed several numerical simulations. In this simulation, we used  $\mathcal{C}_{(2,3)}$  as the choice set (see Table 1). We defined the true priority vector of the  $M = 5$  evaluation objects as

$$(\pi_1^*, \pi_2^*, \pi_3^*, \pi_4^*, \pi_5^*)^T = (1, 2, 1/2, 1/4, 1/2)^T,$$

and we set the number of answerers as  $N$ . We tried the following settings to verify the performance of each method (the first priority  $\pi_1$  was fixed as 1 when we estimated by each method).

- I.**  $N = 1$ , non-contaminated.
- II.**  $N = 10$ , non-contaminated.
- III.**  $N = 50$ , non-contaminated.
- IV.**  $N = 10$ , based on the assumption of the GM and the LN method (see equation (1)), using the log-normal distribution  $L_N(0, 1)$ .
- V.**  $N = 10$ , uniform outlier on  $\mathcal{C}_{(2,3)}$  is contaminated with probability 20%.
- VI.**  $N = 10$ , two answerers use only “1/8 (much more unimportant)” and “8 (much more important)”.
- VII.**  $N = 10$ , five answerers made intransitive PCMs.
- VIII.**  $N = 10$ , all answerers made intransitive PCMs.

We use  $\tilde{p}^e$  as the model of the DPM method, since this recorded the best (minimum) AIC and BHHJ-C values in the real-data example. Except for case IV, the data were based on the assumption of the DPM method using the model  $\tilde{p}^e$ . Note that in case IV, the continuous values were corrected to the nearest elements of  $\mathcal{C}_{(2,3)}$ .

These experiments were carried out to examine changes in the estimation accuracy when the sample size, corresponding to the number of answerers, is changed (cases I–III), as well as to observe the effects of relaxing the assumption of the method or allowing some peculiar answers in the PCM. Case V was designed under the assumptions that some mistakes might occur. Case VI assumed that some answerers would give peculiar answers that affected the PCM. Furthermore, intransitive preference is often observed in practice; therefore, we

designed cases VII and VIII. Specifically, we rewrote the original observations of a certain comparison as their reciprocal values, as follows:

$$\begin{matrix} & O_1 & O_2 & O_3 & O_4 & O_5 \\ O_1 & \left( \begin{array}{ccccc} 1 & 1/2 & 4 & 4 & 2 \end{array} \right) \\ O_2 & \left( \begin{array}{ccccc} 2 & 1 & 4 & 8 & 4 \end{array} \right) \\ O_3 & \left( \begin{array}{ccccc} 1/4 & 1/4 & 1 & 2 & 1 \end{array} \right) \\ O_4 & \left( \begin{array}{ccccc} 1/4 & 1/8 & 1/2 & 1 & 1/2 \end{array} \right) \\ O_5 & \left( \begin{array}{ccccc} 1/2 & 1/4 & 1 & 2 & 1 \end{array} \right) \end{matrix} \rightarrow \begin{matrix} & O_1 & O_2 & O_3 & O_4 & O_5 \\ O_1 & \left( \begin{array}{ccccc} 1 & 1/2 & 4 & \underline{1/4} & 2 \end{array} \right) \\ O_2 & \left( \begin{array}{ccccc} 2 & 1 & 4 & 8 & 4 \end{array} \right) \\ O_3 & \left( \begin{array}{ccccc} 1/4 & 1/4 & 1 & 2 & 1 \end{array} \right) \\ O_4 & \left( \begin{array}{ccccc} \underline{4} & 1/8 & 1/2 & 1 & 1/2 \end{array} \right) \\ O_5 & \left( \begin{array}{ccccc} 1/2 & 1/4 & 1 & 2 & 1 \end{array} \right) \end{matrix}.$$

Hereby, the “three-way deadlock” situation was established artificially.

We evaluated the estimation methods by the sum of absolute errors:

$$\sum_{m=2}^M |\hat{\pi}_m - \pi_m^*|, \quad (4)$$

where  $\hat{\pi}_m$  is the estimator of the  $m$ -th priority. Table 3 shows the mean of (4) over 200 iterations. In addition to, Figure 4 shows the sums of absolute errors (4) for cases II and V–VIII. These cases had the same sample size and the true probabilistic structure, but there exist some outlying answers in cases V–VIII.

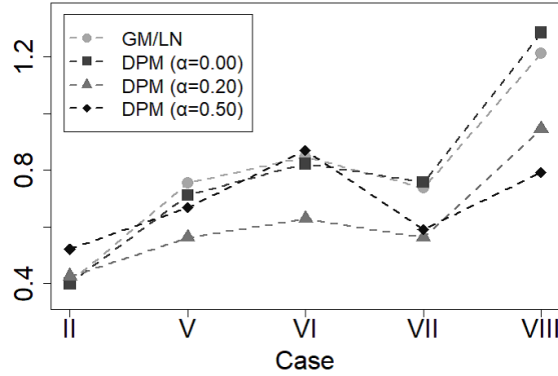


Figure 4: Means of the sums of absolute errors in a non-contaminated case (II) and contaminated cases (V–VIII)

The earlier methods, namely, the eigenvalue, geometric mean, harmonic mean, and PPM methods, worked well in the non-contaminated cases (cases I–IV); however, their accura-



Table 3: Mean of the sum of absolute errors by each method (EV: eigenvalue; GM: geometric mean; LN: log-normal distribution; HM: harmonic mean; PPM: pairwise proportion matrix; DPM: discrete probabilistic model)

	I	II	III	IV	V	VI	VII	VIII
EV	1.254	0.404	0.279	0.502	0.757	0.849	0.729	1.297
GM/LN	1.253	0.404	0.282	0.505	0.755	0.847	0.738	1.213
HM	1.654	0.474	0.263	0.548	0.780	0.882	0.836	1.507
PPM	1.338	0.420	0.253	0.508	0.765	0.857	0.791	1.236
DPM ( $\alpha$ )								
(0.00)	1.770	0.400	0.184	0.499	0.711	0.823	0.757	1.285
(0.01)	1.767	0.400	0.185	0.501	0.695	0.788	0.743	1.275
(0.10)	1.750	0.410	0.194	0.533	0.591	0.612	0.628	1.139
(0.20)	1.770	0.425	0.211	0.608	0.563	0.629	0.563	0.945
(0.30)	1.789	0.448	0.238	0.726	0.579	0.692	0.553	0.835
(0.40)	1.841	0.479	0.274	0.886	0.617	0.769	0.566	0.791
(0.50)	1.959	0.521	0.320	1.095	0.668	0.869	0.591	0.793

cies decreased considerably when applied to cases V–VIII. This result suggests that these conventional methods might be easily influenced by outlying answers.

In contrast, although the proposal method performed worse in case I because of the sample size being too small, we can see that the proposed estimation system with many  $\alpha$ 's performed well not only when the observations were drawn from our probabilistic model and the sample size is somewhat large (cases II, III) but also in the consistent situation assumed in earlier studies (case IV). Moreover, when there existed some outlying answers (cases V–VIII), the estimation based on the BHHJ divergence performed well for most  $\alpha > 0$ . Particularly, large  $\alpha$  performed remarkably well in the “three-way deadlock” situation (cases VII, VIII).

From the above results, we may have no strong reason to use the DPM method with a large  $\alpha$  when the data are not contaminated; however, it is quite difficult to identify what outliers are present in practice, and  $\alpha$  needs to be large for complexly contaminated case

(especially, see case VIII). From our experiments, it is valid to use  $\alpha \in [0.20, 0.40]$  as the tuning parameter of BHHJ divergence in this field.

In many studies, the optimal choice of the tuning parameter has been discussed. For example, Mantalos *et al.* (2010b) [16] pointed out that their model selection criterion (MDIC) with  $\alpha = 0.25$  performed well in the determination of the order of the autoregressive model. Durio and Isaia (2011) [8] reported that a small (close to 0)  $\alpha$  is good in non-contaminated cases and a large  $\alpha$  (in an example, 0.75 performed best) is preferable in contaminated cases, in the estimation of the linear regression model. Kurata and Hamada (2018) [13] showed that  $\alpha = 1.25$  and 1.50 performed particularly well whether there exist outliers or not when selecting the order in the case of polynomial regression models.

## 4 Conclusion

In this paper, we proposed a discrete probabilistic model (DPM) and the estimation of the priorities of the PCM based on it. The proposed model is defined by dividing the answered probability of each choice on the basis of the difference between the choice and the true priority ratio. The model can be robustly estimated by using the BHHJ divergence. We also showed the asymptotic properties of the minimum divergence estimators. Moreover, we confirmed the performance of the method, especially its robustness against unexpected answers compared with conventional methods, through some numerical examples. Since human decisions are uncertain and it is too difficult to identify and remove the outliers, we consider that it is valid to apply the DPM method with the BHHJ divergence to the PCM problem.

## Acknowledgement

The authors would like to express gratitude to the reviewer and the editor-in-chief for their valuable comments, which remarkably improved our earlier draft.

## References

- [1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19** (6), 716–723.
- [2] Basak, I. (1989). Estimation of the multi-criteria worths of the alternatives in a hierarchical structure of comparisons. *Communications in Statistics - Theory and Methods*, **18** (10), 3719–3738.
- [3] Basak, I. (2002). On the use of information criteria in analytic hierarchy process. *European Journal of Operational Research*, **141** (1), 200–216.
- [4] Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85** (3), 549–559.
- [5] Basu, A., Mandal, A., Martin, N., and Pardo, L. (2013). Testing statistical hypotheses based on the density power divergence. *Annals of the Institute of Statistical Mathematics*, **65** (2), 319–348.
- [6] Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, **46** (3), 440–464.
- [7] Csiszár, I. and Shields, P. C. (2004). Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, **1** (4), 417–528.
- [8] Durio, A. and Isaia, E. D. (2011). The minimum density power divergence approach in building robust regression models. *Informatica*, **22** (1), 43–56.
- [9] Ghosh, A. and Basu, A. (2013). Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. *Electronic Journal of Statistics*, **7**, 2420–2456.

- [10] Harker, P. T. and Vargas, L. G. (1987). The theory of ratio scale estimation: Saaty's analytic hierarchy process. *Management science*, **33** (11), 1383–1403.
- [11] Kato, Y. (2013). *Examples and solutions of the AHP -foundation and application- (in Japanese)*. Minerva Shobo.
- [12] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22** (1), 79–86.
- [13] Kurata, S. and Hamada, E. (2018). A robust generalization and asymptotic properties of the model selection criterion family. *Communications in Statistics - Theory and Methods*, **47** (3), 532–547.
- [14] Lipovetsky, S. and Conklin, W. M. (2002). Robust estimation of priorities in the AHP. *European Journal of Operational Research*, **137** (1), 110–122.
- [15] Mantalos, P., Mattheou, K., and Karagrigoriou, A. (2010a). Forecasting ARMA models: a comparative study of information criteria focusing on MDIC. *Journal of Statistical Computation and Simulation*, **80** (1), 61–73.
- [16] Mantalos, P., Mattheou, K., and Karagrigoriou, A. (2010b). An improved divergence information criterion for the determination of the order of an AR process. *Communications in Statistics - Simulation and Computation*, **39** (5), 865–879.
- [17] Mattheou, K., Lee, S., and Karagrigoriou, A. (2009). A model selection criterion based on the BHHJ measure of divergence. *Journal of Statistical Planning and Inference*, **139** (2), 228–235.
- [18] Menéndez, M. L., Morales, D., Pardo, L., and Vajda, I. (2001). Minimum disparity estimators for discrete and continuous models. *Applications of Mathematics*, **46** (6), 439–466.
- [19] Pardo, L. (2005). *Statistical inference based on divergence measures*. CRC Press.

- [20] Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of mathematical psychology*, **15** (3), 234–281.
- [21] Saaty, T. L. (1990). How to make a decision: the analytic hierarchy process. *European journal of operational research*, **48** (1), 9–26.
- [22] Saaty, T. L. (2004). Decision making - the analytic hierarchy and network processes (AHP/ANP). *Journal of systems science and systems engineering*, **13** (1), 1–35.
- [23] Saaty, T. L. (2005). *Theory and applications of the analytic network process: decision making with benefits, opportunities, costs, and risks*. RWS publications.
- [24] Saaty, T. L. and Vargas, L. G. (1984). Comparison of eigenvalue, logarithmic least squares and least squares methods in estimating ratios. *Mathematical modelling*, **5** (5), 309–324.
- [25] Vonta, F., Mattheou, K., and Karagrigoriou, A. (2012). On properties of the  $(\Phi, a)$ -power divergence family with applications in goodness of fit tests. *Methodology and Computing in Applied Probability*, **14** (2), 335–356.

## A Appendix

First, we introduce a wide family of divergence measures. We extend the divergence by Vonta *et al.* (2012) [25] to nonhomogeneous settings such as the DPM model.

Let  $\varphi : (0, +\infty) \rightarrow \mathbf{R}^1$  be a continuous and twice differentiable function that is monotonically decreasing on  $(0, 1)$  and monotonically increasing on  $(1, +\infty)$ , and that satisfies  $\varphi(1) = \varphi'(1) = 0$ ,  $\varphi''(1) > 0$ . Then, for  $\alpha \geq 0$  and the two groups of discrete probability functions  $\mathcal{P} = (\mathbf{p}_{i,j})_{i < j} = (p_{i,j}(c_k))_{k=1, \dots, K, i < j}$  and  $\mathcal{Q} = (\mathbf{q}_{i,j})_{i < j} = (q_{i,j}(c_k))_{k=1, \dots, K, i < j}$ , the divergence is defined as

$$D_{\alpha}^{\varphi}(\mathcal{P}; \mathcal{Q}) = \frac{2}{M(M-1)} \sum_{i < j} \sum_{k=1}^K p_{i,j}(c_k)^{\alpha+1} \varphi \left( \frac{q_{i,j}(c_k)}{p_{i,j}(c_k)} \right) .$$

This includes the BHHJ divergence and the  $\varphi$ -disparity (e.g., Menéndez *et al.* (2001) [18], Pardo (2005) [19]). We denote by  $\hat{\pi}_\alpha^\varphi$  the minimum divergence estimator  $\arg \min_{\pi} D_\alpha^\varphi(\hat{\mathbf{P}}_N; \mathbf{P}(\pi))$ , where the  $(KM(M-1)/2)$ -dimensional vectors  $\hat{\mathbf{P}}_N = (\hat{\mathbf{p}}_{i,j})_{i < j} = (\hat{p}_{i,j}(c_k))_{k=1, \dots, K, i < j}$  and  $\mathbf{P}(\pi) = (\mathbf{p}_{i,j}(\pi))_{i < j} = (p_{i,j}(c_k | \pi))_{k=1, \dots, K, i < j}$  are the groups of probability functions of the relative frequencies and the model, respectively. If  $\varphi(x) = x^{\alpha+1} - \frac{\alpha+1}{\alpha} x^\alpha + \frac{1}{\alpha}$ , then this corresponds with the BHHJ divergence (Definition 2).

## A.1 Proof of Theorem 1

This theorem is proved by extending the asymptotic theorem about the  $\varphi$ -divergence family (Pardo (2005) [19], Chapter 5). In this proof, we denote by  $\mathbf{R} = (\mathbf{r}_{i,j})_{i < j}$  the true probability function. Let  $\mathcal{I}^K = (0, 1)^K$ , which is a superset of

$$\Delta_+^K = \left\{ (\rho_1, \dots, \rho_K)^T \mid \rho_1, \dots, \rho_K > 0; \rho_1 + \dots + \rho_K = 1 \right\}.$$

For any  $\bar{\mathbf{P}} = (\bar{p}_{i,j}(c_k))_{k=1, \dots, K, i < j} \in (\Delta_+^K)^{M(M-1)/2}$  and  $\boldsymbol{\pi} = (\pi_2, \dots, \pi_M)^T \in \mathcal{V}_*$  (see Assumptions (A3)–(A5)), we define the function  $\mathcal{F} = (\mathcal{F}_2, \dots, \mathcal{F}_M)^T : (\mathcal{I}^K)^{M(M-1)/2} \times \mathcal{V}_* \rightarrow \mathbf{R}^{M-1}$  as follows:

$$\mathcal{F}_s = \mathcal{F}_s(\bar{\mathbf{P}}, \boldsymbol{\pi}) = \frac{\partial D_\alpha^\varphi(\bar{\mathbf{P}}; \mathbf{P}(\boldsymbol{\pi}))}{\partial \pi_s} \quad (s = 2, \dots, M).$$

Since  $\mathbf{R} = \mathbf{P}(\boldsymbol{\pi}_*)$  by Assumption (A1) and  $\varphi(1) = \varphi'(1) = 0$ ,  $\mathcal{F}(\mathbf{P}(\boldsymbol{\pi}_*), \boldsymbol{\pi}_*) = \mathbf{0}_{M-1}$  holds.

By differentiating  $\mathcal{F}_s$  with respect to  $\pi_t$  ( $s, t = 2, \dots, M$ ), we obtain

$$\left. \frac{\partial \mathcal{F}}{\partial \boldsymbol{\pi}} \right|_{\bar{\mathbf{P}}=\mathbf{P}(\boldsymbol{\pi}_*), \boldsymbol{\pi}=\boldsymbol{\pi}_*} = \frac{2\varphi''(1)}{M(M-1)} \sum_{i < j} \boldsymbol{\Upsilon}_{i,j}(\boldsymbol{\pi}_*)^T \boldsymbol{\Upsilon}_{i,j}(\boldsymbol{\pi}_*). \quad (5)$$

Note that the right-hand side of (5) is nonsingular (positive definite) due to Assumption (A5).

Therefore, by the implicit function theorem, there is a neighborhood  $\mathcal{U}_*$  of  $(\mathbf{P}(\boldsymbol{\pi}_*), \boldsymbol{\pi}_*)$  such that  $\frac{\partial \mathcal{F}}{\partial \boldsymbol{\pi}}$  is positive definite. Moreover, there exists an open set  $\mathcal{W}_* \subset (\mathcal{I}^K)^{M(M-1)/2}$  including  $\mathbf{R} = \mathbf{P}(\boldsymbol{\pi}_*)$  and a continuously differentiable function  $\boldsymbol{\tau} : \mathcal{W}_* \rightarrow \mathbf{R}^{M-1}$  such that

$$\{(\bar{\mathbf{P}}, \boldsymbol{\pi}) \in \mathcal{U}_* \mid \mathcal{F}(\bar{\mathbf{P}}, \boldsymbol{\pi}) = \mathbf{0}_{M-1}\} = \{(\bar{\mathbf{P}}, \boldsymbol{\tau}(\bar{\mathbf{P}})) \in \mathcal{U}_* \mid \bar{\mathbf{P}} \in \mathcal{W}_*\}.$$

Thus, since  $\mathbf{P}(\boldsymbol{\pi}_*) \in \mathcal{W}_*$ ,

$$\begin{aligned} \mathcal{F}(\mathbf{P}(\boldsymbol{\pi}_*), \boldsymbol{\tau}(\mathbf{P}(\boldsymbol{\pi}_*))) &= \left. \frac{\partial D_\alpha^\varphi(\mathbf{P}(\boldsymbol{\pi}); \mathbf{P}(\boldsymbol{\tau}(\mathbf{P}(\boldsymbol{\pi}))))}{\partial \boldsymbol{\pi}} \right|_{\boldsymbol{\pi}=\boldsymbol{\pi}_*} = \mathbf{0}_{M-1}, \\ (\mathbf{P}(\boldsymbol{\pi}_*), \boldsymbol{\tau}(\mathbf{P}(\boldsymbol{\pi}_*))) &\in \mathcal{U}_* \end{aligned}$$

holds. Hence,  $\boldsymbol{\tau}(\mathbf{P}(\boldsymbol{\pi}_*)) = \boldsymbol{\pi}_*$  gives the local minimum value, because  $\frac{\partial \mathcal{F}}{\partial \boldsymbol{\pi}}$  is positive definite at  $\boldsymbol{\pi}_*$ . Moreover, as the relative frequency  $\hat{\mathbf{P}}_N$  is a consistent estimator of  $\mathbf{R} = \mathbf{P}(\boldsymbol{\pi}_*)$ , we obtain that for sufficiently large  $N$ ,  $\hat{\mathbf{P}}_N \in \mathcal{W}_*$ ,  $\mathcal{F}(\hat{\mathbf{P}}_N, \boldsymbol{\tau}(\hat{\mathbf{P}}_N)) = \frac{\partial D_\alpha^\varphi(\hat{\mathbf{P}}_N; \mathbf{P}(\boldsymbol{\tau}(\hat{\mathbf{P}}_N)))}{\partial \boldsymbol{\pi}} = \mathbf{0}_{M-1}$ , and  $(\hat{\mathbf{P}}_N, \boldsymbol{\tau}(\hat{\mathbf{P}}_N)) \in \mathcal{U}_*$ . Hence,  $\boldsymbol{\tau}(\hat{\mathbf{P}}_N)$  gives a local minimum value.

Then, we differentiate  $\mathcal{F}(\bar{\mathbf{P}}, \boldsymbol{\tau}(\bar{\mathbf{P}}))$  with respect to the  $((K-1)M(M-1)/2)$ -dimensional vector  $\bar{\mathbf{P}}^\dagger = \left( \bar{\mathbf{p}}_{i,j}^\dagger \right)_{i < j} = (\bar{p}_{i,j}(c_k))_{k=1, \dots, K-1, i < j}$ . Note that the vectors  $\bar{\mathbf{p}}_{i,j} = (\bar{p}_{i,j}(c_1), \dots, \bar{p}_{i,j}(c_K))^T$  are each a probability function, so each last element  $\bar{p}_{i,j}(c_K)$  is determined automatically.

Then we can obtain the following  $(M-1) \times (K-1)M(M-1)/2$  matrix:

$$\left. \frac{\partial \mathcal{F}}{\partial \bar{\mathbf{P}}^\dagger} \right|_{\bar{\mathbf{P}}=\mathbf{P}(\boldsymbol{\pi}_*)} = \left( \frac{\partial \mathcal{F}}{\partial \bar{\mathbf{p}}_{1,2}^\dagger} : \frac{\partial \mathcal{F}}{\partial \bar{\mathbf{p}}_{1,3}^\dagger} : \dots : \frac{\partial \mathcal{F}}{\partial \bar{\mathbf{p}}_{M-1,M}^\dagger} \right) \Big|_{\bar{\mathbf{P}}=\mathbf{P}(\boldsymbol{\pi}_*)}, \quad (6)$$

where for each  $i < j$ ,

$$\left. \frac{\partial \mathcal{F}}{\partial \bar{\mathbf{p}}_{i,j}^\dagger} \right|_{\bar{\mathbf{P}}=\mathbf{P}(\boldsymbol{\pi}_*)} = -\mathcal{J}_{i,j}^\dagger(\boldsymbol{\pi}_*)^T \text{diag} \left( \mathbf{p}_{i,j}^\dagger(\boldsymbol{\pi}_*)^{\alpha-1} \right) + p_{i,j}^K(\boldsymbol{\pi}_*)^{\alpha-1} \frac{\partial p_{i,j}^K(\boldsymbol{\pi}_*)}{\partial \boldsymbol{\pi}} \mathbf{1}_{K-1}^T$$

$\mathcal{J}_{i,j}^\dagger(\boldsymbol{\pi}) = \nabla \mathbf{p}_{i,j}^\dagger(\boldsymbol{\pi})$ , and  $\mathbf{p}_{i,j}^\dagger(\boldsymbol{\pi}) = (p_{i,j}(c_1 | \boldsymbol{\pi}), \dots, p_{i,j}(c_{K-1} | \boldsymbol{\pi}))^T$ .

Furthermore, for  $\hat{\mathbf{p}}_{i,j}^\dagger = (\hat{p}_{i,j}(c_1), \dots, \hat{p}_{i,j}(c_{K-1}))^T$  ( $i < j$ ), it can be seen that

$$\begin{aligned} &\left\{ \mathcal{J}_{i,j}^\dagger(\boldsymbol{\pi}_*)^T \text{diag} \left( \mathbf{p}_{i,j}^\dagger(\boldsymbol{\pi}_*)^{\alpha-1} \right) - p_{i,j}^K(\boldsymbol{\pi}_*)^{\alpha-1} \frac{\partial p_{i,j}^K(\boldsymbol{\pi}_*)}{\partial \boldsymbol{\pi}} \mathbf{1}_{K-1}^T \right\} \left( \hat{\mathbf{p}}_{i,j}^\dagger - \mathbf{p}_{i,j}^\dagger(\boldsymbol{\pi}_*) \right) \\ &= \mathcal{J}_{i,j}(\boldsymbol{\pi}_*)^T \text{diag} \left( \mathbf{p}_{i,j}(\boldsymbol{\pi}_*)^{\alpha-1} \right) (\hat{\mathbf{p}}_{i,j} - \mathbf{p}_{i,j}(\boldsymbol{\pi}_*)) \quad (i < j). \end{aligned}$$

Thus, by equations (5) and (6), and the chain rule,

$$\left. \frac{\partial \mathcal{F}(\bar{\mathbf{P}}, \boldsymbol{\tau}(\bar{\mathbf{P}}))}{\partial \bar{\mathbf{P}}^\dagger} \right|_{\bar{\mathbf{P}}=\mathbf{P}(\boldsymbol{\pi}_*)} = \frac{\partial \mathcal{F}(\bar{\mathbf{P}}, \boldsymbol{\tau}(\bar{\mathbf{P}}))}{\partial \boldsymbol{\tau}(\bar{\mathbf{P}})} \frac{\partial \boldsymbol{\tau}(\bar{\mathbf{P}})}{\partial \bar{\mathbf{P}}^\dagger} \Big|_{\bar{\mathbf{P}}=\mathbf{P}(\boldsymbol{\pi}_*)},$$

we have the required result.

Note that, although the assumption  $\hat{\mathbf{P}}_N \in (\Delta_+^K)^{M(M-1)/2}$  is needed in general, this assumption is unnecessary for the BHHJ divergence because there is no fraction term in the definition of the BHHJ divergence.

## A.2 Proof of Theorem 2

The consistency is proved by Theorem 1. From the asymptotic expansion (3),

$$\begin{aligned} & \sqrt{N} (\hat{\pi}_\alpha^\varphi - \pi_*) \\ &= -\mathcal{B}(\pi_*)^{-1} \sum_{i < j} \Upsilon_{i,j}(\pi_*)^T \text{diag}(\mathbf{p}_{i,j}(\pi_*)^{(\alpha-1)/2}) \sqrt{N} (\hat{\mathbf{p}}_{i,j} - \mathbf{p}_{i,j}(\pi_*)) + o_P(1). \end{aligned}$$

Note that  $\mathcal{B}(\pi_*)$ ,  $\Upsilon_{i,j}(\pi_*)$ , and  $\text{diag}(\mathbf{p}_{i,j}(\pi_*)^{(\alpha-1)/2})$  ( $i < j$ ) are non-random. By the central limit theorem, the asymptotic distribution of each relative frequency  $\hat{\mathbf{p}}_{i,j}$  is the degenerate normal distribution:

$$\sqrt{N} (\hat{\mathbf{p}}_{i,j} - \mathbf{p}_{i,j}(\pi_*)) \xrightarrow{L} N_K(\mathbf{0}_K, \Sigma_{i,j}(\pi_*)) \quad (i < j).$$

Since every comparison is stochastic independent, the theorem is proved.

## A.3 BHHJ-C Family

Kurata and Hamada (2018) [13] proposed the family of model selection criteria BHHJ-C. This family is based on the BHHJ divergence family. For the discrete probabilistic model, it can be defined as

$$\text{BHHJ-C}_\alpha = \hat{H}_\alpha + \frac{2}{NM(M-1)} \text{tr} \{ \mathcal{B}_\alpha(\hat{\pi}_\alpha)^{-1} \mathcal{A}_\alpha(\hat{\pi}_\alpha) \} \quad (\alpha \geq 0),$$

where  $\hat{H}_\alpha$  is the minimum value of  $H_\alpha(\cdot)$  with respect to the priority vector, and  $\hat{\pi}_\alpha$  is the minimum BHHJ divergence estimator. Note that their values shown in Table 2 are corrected by multiplying by a constant to adjust to the order of the AIC (e.g., Akaike (1974) [1]).