



# Adaptive thresholding estimator for differential association structures in two independent contingency tables

Seyed Kamran Ghoreishi \*<sup>1</sup>, Jingjing Wu <sup>2</sup>

<sup>1</sup>*Department of Statistics, Faculty of Science, University of Qom, Qom, Iran*

<sup>2</sup>*Department of Mathematics and Statistics, University of Calgary, Canada*

## Abstract

In this paper, we consider an adaptive thresholding procedure to estimate the difference of association structures in two independent two-way contingency tables of the same order. Here, we assume that the class of paired association structures have an approximately sparse difference. Under  $L_1$  and  $L_2$  loss functions, we establish the corresponding risk's upper bounds for our differential association adaptive thresholding estimators. Moreover, we show that these estimators perform well in a simulated setting. In this line, we carry out a simulation study and compare two well-known independent social mobility datasets.

**Mathematics Subject Classification (2010).** 62F35, 62F10, 62J99

**Keywords.** Association orthogonal components, intrinsic association parameters, adaptive thresholding estimator, sparsity

## 1. Introduction

Association studies in categorical data frequently occur in a wide range of applications. So far, there has been a great deal of work on association modeling in two-way and multi-way contingency tables with ordinal categorical variables. Among others, see [2, 3, 9–11].

In the literature, there are several works concentrated on the comparison of two contingency tables with fixed categorical levels, for more details see [4, 14, 15]. The last paper assigns some known scores, so-called orthogonal coefficients, to the classification categories via orthogonal polynomials. We follow this idea by decomposing of the cells' probabilities into orthogonal polynomial components.

Although the current methods seem rich to cover the subject, our methodology, established based on adaptive thresholding approach, enjoys optimal risk advantage. In Sections 3 and 4, in addition to theoretical support, we numerically illustrate its performance under different scenarios.

To the best of our knowledge, it is the first work which considers a model-based thresholding approach to compare association structures in two contingency tables, such that not only it is to compare the class of paired association structures having approximately

\*Corresponding Author.

Email addresses: atty\_ghoreishi@yahoo.com (S.K. Ghoreishi), jinwu@ucalgary.ca (J. Wu)

Received: 03.05.2019; Accepted: 25.04.2020

sparse difference (ASD) but also it can be adapted with the levels' size of the corresponding categorical variables. However, our main idea in choosing this approach is to tolerate to detect the deference for the components which have a negligible difference in the association structures of both tables. Equivalently, in the case when an association component has two really different measures in both tables, this component can be detected as an effective component to cause a difference in the association structures. In practice, some thresholding functions, like the soft, hard, and adaptive lasso thresholding functions, are used for this purpose.

The main goals of this paper are

- Motivated by several applications, we consider the model-based differential association adaptive thresholding estimators (DAATE) with ASD. Since these are new estimators, their risk properties are obtained under  $L_1$  and  $L_2$  loss functions.
- Ideally, an adaptive inference should have its upper risk bound automatically adjusted to the dimension's size of the contingency tables. We show that the established bounds have this property.
- Although we provide theoretical support for the optimum risk of adaptive thresholding estimators, we also numerically investigate that these estimators outperform the naive estimators.

The paper is organized as follows. In Section 2, we review some basic concepts and define necessary notations and definitions. Moreover, we present in detail the adaptive thresholding procedure for estimating the differential association structure. The main results of this work are presented in Section 3. We carry out a simulation study to evaluate the performance of the DAATE in Section 4 and illustrate the theoretical results to compare two well-known independent social mobility datasets in Section 2. Necessary technical proofs are given in the appendix.

## 2. Preliminaries

### 2.1. Notations and Definitions

Suppose we have two independent  $I \times J$  contingency tables, labeled with 1 and 2. The row categorical variable is denoted by  $X$  and the column categorical variable by  $Y$ . Let  $p_{ijt}$  be the probability in  $(i, j)$ th cell of table  $t (= 1, 2)$ . In each table, the row scores  $\mu_{ik}$  and the column scores  $\nu_{jk'}$  are known orthogonal polynomial contrast coefficients of order  $k (= 1, 2, \dots, I - 1)$  and  $k' (= 1, \dots, J - 1)$ , respectively. These coefficients satisfy the constraints

$$\begin{aligned} \sum_{i=1}^I \mu_{ik} &= 0, & \sum_{j=1}^J \nu_{jk'} &= 0, & \sum_{i=1}^I \mu_{ik}^2 &= 1, & \sum_{j=1}^J \nu_{jk'}^2 &= 1, \\ \sum_{i=1}^I \mu_{ik} \mu_{il} &= 0 \quad (k \neq l), & \sum_{j=1}^J \nu_{jk'} \nu_{jl'} &= 0 \quad (k' \neq l'). \end{aligned} \quad (2.1)$$

Definition of these scores is extensively discussed in [9, 15].

Indeed, we may use these scores to decompose the association structure into  $2 \times 3 = 6$  orthogonal polynomial components. Moreover, suppose  $n_{ijt}$  is the observed frequency in the  $(i, j)$ th cell, and  $n_t$  is the corresponding total sample size in table  $t (= 1, 2)$ . Consequently, the naive estimator of  $p_{ijt}$  is given by  $\hat{p}_{ijt} = \frac{n_{ijt}}{n_t}$ , see [1].

In table  $t$ , we define the intrinsic association parameters  $\phi_{kk't}$ , ( $k = 1, 2, \dots, I - 1$ ;  $k' = 1, 2, \dots, J - 1$ ), based on the logarithmic link as

$$\phi_{kk't} := \sum_{i=1}^I \sum_{j=1}^J \mu_{ik} \nu_{jk'} \log(p_{ijt}). \quad (2.2)$$

From (2.2), if we replace  $\log p_{ijt}$  with usual log-linear model  $\log p_{ijt} = \lambda_{0t} + \lambda_{it}^X + \lambda_{jt}^Y + \lambda_{ijt}^{XY}$ , we see that

$$\phi_{kk't} = \sum_{i=1}^I \sum_{j=1}^J \mu_{ik} \nu_{jk'} \lambda_{ijt}^{XY}.$$

This means that the  $\phi_{kk't}$ s, indeed, are to decompose the association structure, measured by association factor  $\lambda_{ijt}^{XY}$ .

Using (2.1) and (2.2), it is obvious to see

$$p_{ijt} = \alpha_i \beta_j e^{\sum_{k=1}^{I-1} \sum_{k'=1}^{J-1} \phi_{kk't} \mu_{ik} \nu_{jk'}},$$

where  $\alpha_i$  and  $\beta_j$  are some positive quantities. Despite different interpretations, the above model is very similar to the Goodman's row-column association model of order  $M$ ,  $RC(M)$ . In a saturated model  $M = \min\{I - 1, J - 1\}$ , see [12, 13].

Borrowing the idea of detecting various polynomial trends of the factorial effects in an analysis of variance model, where the treatments have ordinal levels, Equation (2.2) establishes a similar analysis in contingency tables. Thus breaking down the association measure into Linear by Linear component (linear trend along the column variable at each level of the row variable and vice versa), Linear By Quadratic component (quadratic trend along the column variable at each level of the row variable and Linear trend along the row variable at each level of the column variable), and so on, will indicate various trends.

### 2.2. Differential association structure

Let us generalize our discussion by replacing the logarithmic link in (2.2) with a differentiable function  $\psi : (0, 1) \rightarrow (-\infty, \infty)$  as

$$\phi_{kk't} = \sum_{i=1}^I \sum_{j=1}^J \mu_{ik} \nu_{jk'} \psi(p_{ijt}).$$

Comparison of association structures in two tables can be achieved by examining the independent discrepancy parameters

$$d_{kk'} = \phi_{kk'1} - \phi_{kk'2}, \quad k = 1, \dots, I - 1; \quad k' = 1, \dots, J - 1. \tag{2.3}$$

Thus,  $d_{kk'} = 0$  for all  $k$  and  $k'$  implies the equality of association structures in two independent contingency tables. Any discrepancy between the components of association structures would result in  $d_{kk'} \neq 0$  for some  $k$  and  $k'$ s. This fact can help interpret the observed differences between tables in terms of the strength of their association components.

A naive approach to estimating the differential association structures  $d_{kk'} = \phi_{kk'1} - \phi_{kk'2}$  is to first estimate the various orthogonal components of association separately, and define

$$\hat{d}_{kk'} = \hat{\phi}_{kk'1} - \hat{\phi}_{kk'2} = \sum_{i,j} \mu_{ik} \nu_{jk'} \{\psi(\hat{p}_{ij1}) - \psi(\hat{p}_{ij2})\}. \tag{2.4}$$

Another approach that we propose in this paper is an differential association adaptive thresholding estimator,  $\hat{d}_{kk'}^{DAATE}$ . For the differential association component  $d_{kk'} = \phi_{kk'1} - \phi_{kk'2}$ , it is given by

$$\hat{d}_{kk'}^{DAATE} = S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2}),$$

where  $S_\lambda(\cdot)$  is a class of thresholding function illustrated in Subsection (2.5). Here, we assume that  $d_{kk'} = 0$ , for many  $k$  and  $k'$ s, i.e., the ASD property mentioned above. Theoretical and numerical performance of  $\hat{d}_{kk'}^{DAATE}$  are investigated in Sections 3 and 4.

### 2.3. Intrinsic association estimator

Given the row and column polynomial scores  $\mu_{ik}$  and  $\nu_{jk'}$  and the continuous and differentiable function  $\psi : (0, 1) \rightarrow (-\infty, \infty)$ , the  $(k, k')$ th association component's maximum likelihood estimator in the table  $t(= 1, 2)$  is defined as

$$\hat{\phi}_{kk't} = \sum_{i,j} \mu_{ik} \nu_{jk'} \psi(\hat{p}_{ijt}). \tag{2.5}$$

Using delta-method, we have

$$\psi(\hat{p}_{ijt}) = \psi(p_{ijt}) + \psi'(p_{ijt}^*)(\hat{p}_{ijt} - p_{ijt})$$

where  $p_{ijt}^*$  is in between  $\hat{p}_{ijt}$  and  $p_{ijt}$ . Therefore, (2.5) can be rewritten as

$$\hat{\phi}_{kk't} = \sum_{i,j} \mu_{ik} \nu_{jk'} \{ \psi(p_{ijt}) + \psi'(p_{ijt}^*)(\hat{p}_{ijt} - p_{ijt}) \},$$

or equivalently,

$$\hat{\phi}_{kk't} - \phi_{kk't} = \sum_{i,j} \mu_{ik} \nu_{jk'} \psi'(p_{ijt}^*)(\hat{p}_{ijt} - p_{ijt}). \tag{2.6}$$

Assuming that  $\psi'$  is a bounded function, we define  $M = \sup_{x \in (0,1)} \psi'(x)$ .

### 2.4. Sub-Gaussian random variables assumption

It is well-known that  $Z_{ijt} = \hat{p}_{ijt} - p_{ijt}$  have asymptotic normal distribution. However our focus in this work is to prove some theorems which cover any arbitrary sample size. Therefore, since  $Z_{ijt}$  are bounded, we assume that

$$Z_{ijt} = \hat{p}_{ijt} - p_{ijt} = \frac{n_{ijt}}{n_t} - p_{ijt}$$

are sub-Gaussian random variables. For more details on sub-Gaussian random variables see [6].

### 2.5. Thresholding functions

In the present setting, we apply the idea of adaptive thresholding to the difference of association structure using a proper thresholding function  $S_\lambda$ . We assume the following three conditions for  $S_\lambda$

- i)  $|S_\lambda(z)| \leq c|y|$  for all  $z$  and  $y$  such that  $|z - y| \leq \lambda$ .
- ii)  $|S_\lambda(z)| = 0$  for  $|z| \leq \lambda$ .
- iii)  $|S_\lambda(z) - z| \leq \lambda$  for all  $z \in \mathbb{R}$ ,

where  $c > 0$  is a constant, see [7, 8]. From a practical point of view, in this work, thresholding function  $S_\lambda(d_{kk'})$  sets the differential association structure  $d_{kk'}$  to zero which is less than the thresholding value  $\lambda$  and reserves it when is greater than  $\lambda$  at the same time.

Conditions (i)-(iii) are met by several well-known thresholding functions, which are used in practice, [16]. They include

- The soft thresholding rule  $S_\lambda(z) = \text{sgn}(z) \max\{z - \lambda, 0\}$ , where  $\text{sgn}(z)$  is the sign function.
- The hard thresholding rule  $S_\lambda(z) = z.I_{|z| \geq \lambda}(z)$ , where  $I_A(z)$  is 1 if  $z \in A$  and 0 otherwise.
- The adaptive lasso rule  $S_\lambda(z) = z \max\{1 - \frac{\lambda}{|z|}^\eta, 0\}$  with  $\eta \geq 1$ .

Here, we propose to estimate the sparse association structure by the soft thresholding estimator

$$\hat{d}_{kk'}^{DAATE} = S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2}), \tag{2.7}$$

with

$$\lambda = M \frac{I + J}{2} \left\{ \max_{i,j} |\hat{p}_{ij1} - p_{ij1}| + \max_{i,j} |\hat{p}_{ij2} - p_{ij2}| \right\}. \tag{2.8}$$

### 3. Theoretical results

In this section we establish the main results containing two theorems which are important to evaluate the performance of the differential association estimators including the regular estimators ( $\hat{d}_{kk'}$ ) and the thresholding estimators ( $\hat{d}_{kk'}^{DAATE}$ ). For establishing these theorems, the assumption

- A) The quantities  $\hat{p}_{ijt} - p_{ijt}$  are sub-Gaussian random variables.
- and the condition
- B) The paired differential association structures have approximately sparse differences,
- are required.

#### 3.1. Upper bound for the risk of $\hat{d}_{kk'}$

The following theorem provides an upper bound for the risk of the estimators  $\hat{d}_{kk'} = \hat{\phi}_{kk'1} - \hat{\phi}_{kk'2}$ , the differential intrinsic association component, based on the  $I \times J$  contingency tables 1 and 2.

**Theorem 3.1.** *Under Assumption A and Condition B, the upper bound for the risk of estimator (2.4) satisfies*

- i) Under  $L_1$  norm loss

$$\begin{aligned} \sup_{\mu, \nu} E |\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2} - (\phi_{kk'1} - \phi_{kk'2})| \\ \leq C(I + J) \sqrt{1 + \log(2IJ)} \left( \frac{1}{n_1^{0.5}} + \frac{1}{n_2^{0.5}} \right), \end{aligned}$$

- ii) Under  $L_2$  norm loss

$$\begin{aligned} \sup_{\mu, \nu} E |\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2} - (\phi_{kk'1} - \phi_{kk'2})|^2 \\ \leq C(IJ) \left( \frac{1}{n_1} + \frac{1}{n_2} \right), \end{aligned}$$

for some constant  $C > 0$  that does not depend on  $I, J, n_1,$  and  $n_2$ , where  $\mu = \{(\mu_{1k}, \dots, \mu_{Ik})^T, k = 1, \dots, I-1\}$  and  $\nu = \{(\nu_{1k'}, \dots, \nu_{Jk'})^T, k' = 1, \dots, J-1\}$  are orthonormal bases satisfy in (2.1).

#### 3.2. Upper bound for the risk of $\hat{d}_{kk'}^{DAATE}$

In order to investigate an upper bound for the risk of thresholding estimators (2.7), we consider the following class of paired association structures in  $\mathbb{R}^{(I-1) \times (J-1)}$  with approximately sparse difference (ASD)

$$\begin{aligned} \mathcal{U}_q(T_q(I, J)) = \{(\phi_{kk'1}, \phi_{kk'2}); k = 1, \dots, I-1; k' = 1, \dots, J-1; \\ \sum_{k,k'} |\phi_{kk'1} - \phi_{kk'2}|^q \leq T_q(I, J)\}, \end{aligned} \tag{3.1}$$

for some  $0 \leq q < 1$ . For all  $(\phi_{kk'1}, \phi_{kk'2}) \in \mathcal{U}_q(T_q(I, J))$ , their differences are approximately zero in the sense that all intrinsic association differences, in two contingency tables lies in the  $L_q$  ball with radius  $T_q(I, J)$ . Therefore, the following theorem provides an upper bound for the total risk of the thresholding estimators (2.7).

**Theorem 3.2.** Under the framework of Theorem (3.1),  $\lambda$  given by (2.8), and the class of paired association differences (3.1), the upper bound for the total risk of the thresholding estimators (2.7) satisfies

i) under  $L_1$  norm loss:

$$\begin{aligned} & \sup_{\mu, \nu, \mathcal{U}_q} E \sum_{k, k'} |\{S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2}) - (\phi_{kk'1} - \phi_{kk'2})\}| \\ & \leq CT_q(I, J)(I + J)^{1-q}(1 + \log(2IJ))^{\frac{1-q}{2}} \left\{ \frac{1}{n_1} + \frac{1}{n_2} \right\}^{\frac{1-q}{2}} \end{aligned}$$

ii) under  $L_2$  norm loss:

$$\begin{aligned} & \sup_{\mu, \nu, \mathcal{U}_q} E \sum_{k, k'} |\{S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2}) - (\phi_{kk'1} - \phi_{kk'2})\}|^2 \\ & \leq CT_q(I, J)(I + J)^{2-q}(1 + \log(2IJ))^{\frac{2-q}{4}} \left\{ \frac{1}{n_1} + \frac{1}{n_2} \right\}^{\frac{2-q}{4}} \end{aligned}$$

for some constant  $C > 0$  that does not depend on  $I$ ,  $J$ ,  $n_1$ , and  $n_2$ . Moreover,  $\mu$  and  $\nu$  are the class of orthonormal bases defined earlier.

Indeed, theorem (3.2) yields a maximum rate of convergence for estimating  $(\phi_{kk'1} - \phi_{kk'2})$  with  $(\phi_{kk'1}, \phi_{kk'2}) \in \mathcal{U}_q(T_q(I, J))$ , and shows that the thresholding estimators  $\hat{d}_{kk'}^{DAATE} = S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2})$ , having its upper risk's bound automatically adjusted to the dimension's size of the contingency tables, have an optimal rate of convergence.

#### 4. Numerical studies

The main purpose of this section is to study the performance of the adaptive thresholding estimator of differential association structure through simulation study. Here, our serious challenge is how to estimate the thresholding level  $\lambda$  defined in (2.8). Our suggestion, here, is to replace its maximum terms with an upper bound, similar to the upper bound which is made in confidence intervals for  $p_{ijt}$ , as follows

$$\begin{aligned} \lambda &= M \frac{I + J}{2} \left\{ \max_{i,j} |\hat{p}_{ij1} - p_{ij1}| + \max_{i,j} |\hat{p}_{ij2} - p_{ij2}| \right\} \\ &\leq M \frac{I + J}{2} \left\{ \max_{i,j} \tau_0 \sqrt{\frac{p_{ij1}(1 - p_{ij1})}{n_1}} + \max_{i,j} \tau_0 \sqrt{\frac{p_{ij2}(1 - p_{ij2})}{n_2}} \right\} \\ &\leq \tau_0 M \frac{I + J}{4} \left\{ \sqrt{\frac{1}{n_1}} + \sqrt{\frac{1}{n_2}} \right\} = \tau \frac{I + J}{4} \left\{ \sqrt{\frac{1}{n_1}} + \sqrt{\frac{1}{n_2}} \right\}, \end{aligned}$$

where  $\tau = \tau_0 M$  is a positive unknown constant. It seems that the theoretical choice of  $\tau$  may not be optimal in finite sample performance. Empirically, we propose to estimate this quantity by using a data-driven choice based on cross-validation discussed below.

To proceed, for two given contingency tables of the same order  $I = J (= 4, 6, 8)$ , let  $\phi_{kk'1}$  and  $\phi_{kk'2}$  ( $k = 1, \dots, I$ , and  $k' = 1, \dots, J$ ) be intrinsic association parameters according to the following two scenarios

- Scenario I: we have  $\phi_{kk'1} = \phi_{kk'2} = 0$ , for  $k \neq k'$  and  $\phi_{kk1} = k (\neq 0) \pmod 3$ ,  $\phi_{kk2} = (-1)^k \phi_{kk1}$ ,
- Scenario II: we assume  $\phi_{kk'2} = k + k'$  and  $\phi_{kk'1} = (-1)^{k+k'} \phi_{kk'2}$ , for  $k, k' \leq 2$ . and  $\phi_{kk'2} = 0$ , otherwise.

Although the matrices  $\{\phi_{kk'1}\}$  and  $\{\phi_{kk'2}\}$  may not be sparse, their difference is more sparse. Given  $\phi_{kk'1}$  and  $\phi_{kk'2}$ , the probability in  $(i, j)$ th cell for  $t$ th population is obtained

**Table 1.** Comparison of  $S_\lambda(\hat{\phi}_1 - \hat{\phi}_2)$  and  $\hat{\phi}_1 - \hat{\phi}_2$ , under under  $L_1$  and  $L_2$  norms, for Scenarios I and II

	$n_1 = n_2$	$I = J$	$L_2$ norm		$L_1$ norm	
			$S_\lambda(\hat{\phi}_1 - \hat{\phi}_2)$	$\hat{\phi}_1 - \hat{\phi}_2$	$S_\lambda(\hat{\phi}_1 - \hat{\phi}_2)$	$\hat{\phi}_1 - \hat{\phi}_2$
Scenario I	500	4	0.02516	0.03097	0.02887	0.03505
	500	6	0.15752	0.19767	0.16009	0.20184
	500	8	1.39714	1.66039	1.40061	1.66587
	5000	4	0.00069	0.00131	0.00196	0.00227
	5000	6	0.00136	0.00527	0.00148	0.00638
	5000	8	0.00257	0.01597	0.00262	0.01720
Scenario II	500	4	0.80675	0.85076	0.81988	0.86502
	500	6	1.46072	1.60797	1.46390	1.61280
	500	8	2.68677	3.10026	2.69129	3.10690
	5000	4	0.00165	0.00238	0.00192	0.00320
	5000	6	0.00338	0.00732	0.00347	0.00830
	5000	8	0.00473	0.01861	0.00480	0.01984

as

$$p_{ijt} \propto \exp\left(\sum_{k=1}^{I-1} \sum_{k'=1}^{J-1} \phi_{kk't} \mu_{ik} \nu_{jk'}\right).$$

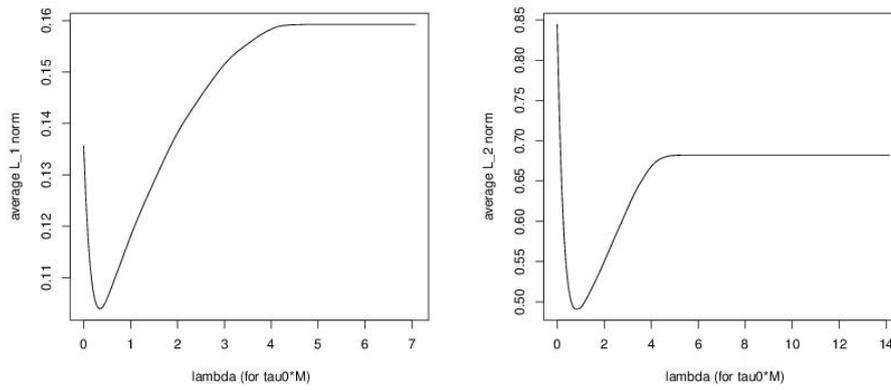
In this numerical study, we use the logarithmic link  $\psi(x) = \log x$ .

Suppose, we generate  $m_1 = 2K = 100$  independent samples from  $\mathbf{X}^1 \sim \text{Multi}(n_1, \{p_{ij1}\})$  and  $m_2 = 2K = 100$  independent samples from  $\mathbf{X}^2 \sim \text{Multi}(n_2, \{p_{ij2}\})$ , where  $\mathbf{X}^1 = \{\mathbf{X}_1^1, \mathbf{X}_2^1, \dots, \mathbf{X}_{m_1}^1\}$  and  $\mathbf{X}^2 = \{\mathbf{X}_1^2, \mathbf{X}_2^2, \dots, \mathbf{X}_{m_2}^2\}$ . Here, we assume  $n_1 = n_2 (= 500, 5000)$ . For various values of  $\tau \in [0, 3]$ , we implement the proposed method with the soft thresholding and repeat the experiment for 50 times. Moreover, we partition the interval  $[0, 3]$  into  $T + 1$  equi-spaced grid  $\{t \frac{3}{T}; t = 0, 1, \dots, T\}$ . In order to empirically choose  $\tau$ , we first divide both samples  $\mathbf{X}^1 = \{\mathbf{X}_1^1, \mathbf{X}_2^1, \dots, \mathbf{X}_{m_1}^1\}$  and  $\mathbf{X}^2 = \{\mathbf{X}_1^2, \mathbf{X}_2^2, \dots, \mathbf{X}_{m_2}^2\}$  randomly into  $K$  groups with two samples. Suppose  $d_{kk',l}^{DAATE}$  is the thresholding estimator for given thresholding constant  $\tau \in \{0, \frac{3}{T}, \dots, 3\}$ , based on the first sample, and  $\phi_{kk',l}^{(1)}$  and  $\phi_{kk',l}^{(2)}$  are the corresponding association components for the second sample in group  $l$ , for populations 1 and 2. Consequently, our empirical choice of  $\tau$  is given as

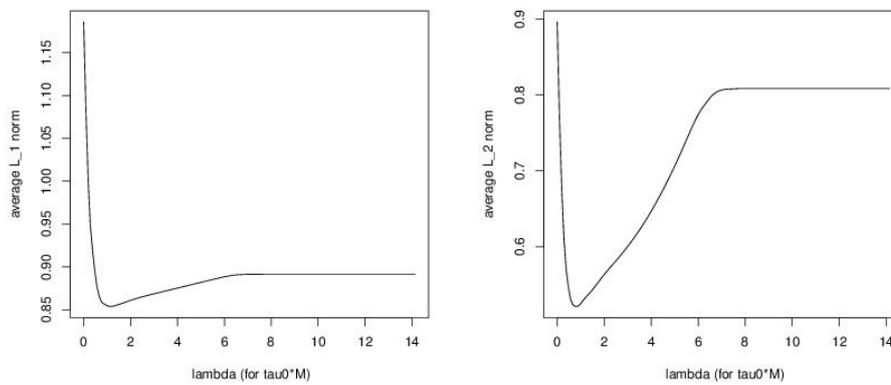
$$\hat{\tau} = \min_{\tau \in \{0, \frac{3}{T}, \dots, 3\}} \mathcal{A}(p) = \min_{\tau \in \{0, \frac{3}{T}, \dots, 3\}} \frac{1}{K} \sum_{l=1}^K \sum_{k=1}^{I-1} \sum_{k'=1}^{J-1} |d_{kk',l}^{DAATE} - (\phi_{kk',l}^{(1)} - \phi_{kk',l}^{(2)})|^p,$$

where  $p = 1$  or  $2$ , corresponding to  $L_1$  and  $L_2$  norms. The corresponding average loss,  $\mathcal{A}(1)$ , and  $\mathcal{A}(2)$ , are shown in Figure 1 when  $I = J = 8$  and  $n_1 = n_2 = 5000$ . These plots help us to determine the empirical value of  $\hat{\tau}$ .

Given above assumptions, the adaptive thresholding estimator is easy to implement for generated data under Scenarios (I) and (II), illustrated previously. For these scenarios, the numerical results are summarized in Table (1). In this table, we can easily compare the performance of two estimators of  $\phi_1 - \phi_2 = [\phi_{kk'1} - \phi_{kk'2}]$  including the naive estimator  $\hat{\phi}_1 - \hat{\phi}_2 = [\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2}]$  and the thresholding estimator  $S_\lambda(\hat{\phi}_1 - \hat{\phi}_2) = [S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2})]$ , while their performances are evaluated by  $L_1$  and  $L_2$  norms for two sample sizes. For both scenarios, we see that the thresholding estimator  $S_\lambda(\hat{\phi}_1 - \hat{\phi}_2)$  outperforms the naive estimator  $\hat{\phi}_1 - \hat{\phi}_2$ , for all  $k$  and  $k'$ .



(a) The best choice of  $\tau$  for Scenario I



(b) The best choice of  $\tau$  for Scenario II

Figure 1. Average loss in  $L_1$  and  $L_2$  norms for  $\tau \in [0, 3]$ ,  $n_1 = n_2 = 5000$ ,  $I = J = 8$ .

### 5. Application on real data

In this section, we analyze two social mobility datasets in the form of contingency tables made up of the row categorical variable "Father's Job" and the column categorical variable "Son's Job" each at 5 levels. The observed frequencies are given in Table (2). In each cell, the upper value belongs to England and the lower value to Denmark, [5]. Without loss of generality, suppose the row and column categorical variables are ordinal.

The estimates of intrinsic association parameters ( $\phi_{kk't}$ ) and their standard errors are given in Table 2. These results reveal that there are statistically significant "Linear by Linear" association component of 4.02, "Linear by Quadratic" association component of -0.85, and "4th by 4th degree" association component of 0.22 between "father's Job" and "son's Job" in England data. Similarly, there are statistically significant "Linear by Linear" association component of 4.22, "Linear by Cubic" association component of -0.86, "Quadratic by Quadratic" association component of 1.06, and finally "4th by 4th degree" association component of 0.37 between "father's Job" and "son's Job" in Denmark data. The significant components are marked with one asterisk for England data and two asterisks for Denmark Data. In this analysis, we see a sparse significant association structure between father's Job and Son's Job in each population. To analyze the difference in association structure, we apply our adaptive thresholding method with cross-validation to

**Table 2.** Observed frequencies of father’s and son’s Job in England(E) and Denmark(D)

		Son’s Job					
		1	2	3	4	5	
Father’s Job	1	E=50	45	8	18	8	
		D=18	17	16	4	2	
	2		28	174	84	154	55
			24	105	109	59	21
	3		11	78	110	223	96
			23	84	289	217	95
	4		14	150	185	714	447
			8	49	175	348	198
	5		3	42	72	320	411
			6	8	69	201	246

**Table 3.** The estimates of intrinsic association parameters and their standard errors inside the parentheses

	Linear	Quadratic	Cubic	4th – degree
Linear	4.02*(0.292)	-0.85*(0.279)	0.11(0.195)	0.25(0.206)
	4.22**(0.378)	0.25(0.334)	-0.86*(0.322)	-0.02(0.250)
Quadratic	-0.50(0.272)	1.00(0.555)	0.03(0.179)	-0.20(0.182)
	-0.46(0.330)	1.06**(0.291)	-0.29(0.278)	0.30(0.215)
Cubic	0.19(0.204)	-0.39(0.189)	0.20(0.132)	0.07(0.128)
	-0.20(0.266)	0.06(0.233)	0.00(0.204)	0.00(0.155)
4th – degree	0.03(0.192)	0.05(0.171)	-0.14(0.122)	0.22*(0.101)
	-0.10(0.195)	0.18(0.170)	-0.23(0.139)	0.37**(0.102)

**Table 4.** The estimates of  $\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2}$  (the upper) and  $S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2})$ (the under)

	Linear	Quadratic	Cubic	4th – degree
Linear	-0.20	-1.10*	0.97*	0.27
	0.00	-1.04**	0.93**	0.00
Quadratic	-0.04	-0.06	0.31	-0.50*
	0.00	0.00	0.00	0.00
Cubic	0.39	-0.45*	0.20	0.07
	0.00	0.00	0.00	0.00
4th – degree	0.13	-0.13	0.09	-0.15
	0.00	0.00	0.00	0.00

estimate differential association components. The results of comparison of differential association structure are given in Table (4). These differences are  $\hat{d}_{kk'} = \hat{\phi}_{kk'1} - \hat{\phi}_{kk'2}$  and  $\hat{d}_{kk'}^{DAATE} = S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2})$ . More interestingly, we only find two significant differential components " Linear by Quadratic" and "Linear by Cubic" between father’s Job and Son’s Job association structure in two countries.

To illustrate the difference between our method with the log-linear model fit, we compared the differences of the estimates of interaction effects in two populations. The results are presented in Table (5). As we can see, the comparison of the estimates of interaction parameters, resulting from the fit of two log-linear models to the two datasets, is a usual comparison and does not provide any clear interpretation for the differences in the pattern of variations for two categorical variables.

**Table 5.** Comparison table for estimated interaction differences for social mobility data based on log-linear model fit

		Son's Job				
		1	2	3	4	5
Father's Job	1	0.91(1.02)	-1.36(0.86)	-1.42(0.85)	0.24(0.90)	-
	2	0.33(0.77)	-1.54**(0.44)	-0.74*(0.35)	0.05(0.32)	-
	3	0.41(0.77)	-1.18**(0.44)	-0.50(0.26)	0.07(0.40)	-
	4	0.85(0.80)	-0.80(0.43)	-0.29(0.23)	-0.05(0.16)	-
	5	-	-	-	-	-

## References

- [1] A. Agresti, *Categorical Data Analysis*, J. Wiley, Hoboken, New Jersey, 2002.
- [2] F. Bartolucci and A. Forcina, *Extended RC association models allowing for order restrictions and marginal modeling*, J. Amer. Statist. Associ., **97**, 1192-1199, 2002.
- [3] M.P. Becker and A. Agresti, *Maximum likelihood estimation of the RC(M) association model*, Appl. Statist., **39**, 152-167, 1992.
- [4] M.P. Becker and C. Clogg, *Analysis of sets of two-way contingency tables using association models*, J. Amer. Statist. Asso., **84**, 142-151, 1989.
- [5] Y.M. Bishop, S.E. Fienberg, and P.W. Holland, *Discrete Multivariate Analysis: Theory and Applications*, Springer, 2007.
- [6] V.V. Buldygin and V.Y. Kozachnko, *Subgaussian random variables*, Ukrainian Math. J., **32**, 483-489, 1980.
- [7] T.T. Cai and A. Zhang, *Inferencial for high-dimensional differential correlation matrices*, J. of Multivariate Analysis, **143**, 107-126, 2016.
- [8] T.T. Cai and H.H. Zhou, *Optimal rates of convergence for sparse covariance matrix estimation*, Ann. Statist., **40**(5), 2389-2420 2012.
- [9] S.K. Ghoreishi and M.R. Meshkani, *Bayesian analysis of association (BANOAS) in contingency tables with ordinal and interval variables*, J. of Statistical Theory and Applications, **5**(2), 363-372, 2006.
- [10] S.K. Ghoreishi and M.R. Meshkani, *Asymptotic Maximum Likelihood and Bayesian Analysis of Shares of Various Weighted Trends in Association Models in Contingency Tables*, J. of Statistical Theory and Applications, **7**(2), 229-243, 2008.
- [11] L.O. Goodman, *Simple models for the analysis of association in cross-classifications having ordered categories*, J. Amer. Statist. Associ., **74**, 537-552, 1979.
- [12] L.O. Goodman, *The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models and asymmetry models for contingency tables with or without missing entries*, Ann. statist., **13**, 10-69, 1985.
- [13] L.O. Goodman, *Measures, models, and graphical displays in the analysis of cross-classified data*, J. Amer. Statist. Assoc., **86**, 1085-1111, 1991.
- [14] M. Kateri, T. Papaioannou, and R. Ahmad, *New association models for the analysis of sets of two-way contingency tables*, Statistica Applicata, **8**, 537-551, 1996.
- [15] M. Kateri, R. Ahmad, and T. Papaioannou, *New features in the class of association models*, Applied Stochastic models and Data Analysis, **14**, 125-136, 1998.
- [16] A. Rothman, E. Levina, and J. Zhu, *Generalized thresholding of large covariance matrices*, J. Amer. Statist. Associ., **104**, 177-186, 2009.

## Appendix

Here are given the proofs of the theorems. In proving theorems, we frequently use two inequalities  $\sqrt{ab} \leq \frac{a+b}{2}$  and  $(\sqrt{a} + \sqrt{b})^2 \leq 2(a + b)$ . Moreover, we carry the argument for bounding the maximum of (not necessarily independent) normal random variables over sub-Gaussian random variables through the following proposition.

**Proposition 5.1.** *Suppose  $X_1, \dots, X_p$  are (not necessarily independent) sub-Gaussian with scale factors bounded by a constant  $\sigma$ . Then*

$$E \max_{1 \leq i \leq p} |X_i| \leq \frac{3}{2} \sigma \sqrt{1 + \log(2p)}.$$

**Proof.** For each  $t > 0$

$$e^{tE(\max_{1 \leq i \leq p} |X_i|)} \leq E(\max_{1 \leq i \leq p} e^{t|X_i|}) \leq \sum_{i=1}^p \{E(e^{tx}) + E(e^{-tx})\} \leq 2pe^{\frac{1}{2}\sigma^2 t^2}.$$

The last inequality follows from the definition of sub-Gaussian random variable. By choosing  $t = \frac{\log(2p)}{\sigma}$ , the proof immediately follows.  $\square$

Now, we are ready to prove our theorems.

**Proof.** Proof of Theorem (3.1)

It is easy to see

$$\begin{aligned} |\hat{\phi}_{kk't} - \phi_{kk't}| &= \left| \sum_{i,j} \mu_{ik} \nu_{jk'} (\psi(\hat{p}_{ijt}) - \psi(p_{ijt})) \right| \\ &\leq \sum_{i,j} |\mu_{ik} \nu_{jk'}| |\psi'(p_{ijt}^*) (\hat{p}_{ijt} - p_{ijt})| \\ &\leq M \sum_{i,j} |\mu_{ik} \nu_{jk'}| |(\hat{p}_{ijt} - p_{ijt})|, \end{aligned}$$

Using Proposition (5.1), we have

$$\begin{aligned} E|\hat{\phi}_{kk't} - \phi_{kk't}| &\leq M \frac{I+J}{2} \cdot E(\max_{i,j} |\hat{p}_{ijt} - p_{ijt}|) \\ &\leq M \frac{I+J}{2} \frac{3}{2} \sqrt{\frac{1}{4n_t}} \sqrt{1 + \log(2IJ)}. \end{aligned} \tag{A.1}$$

(A.1) implies

$$E|\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2} - (\phi_{kk'1} - \phi_{kk'2})| \leq C(I+J) \sqrt{1 + \log(2IJ)} \left( \frac{1}{n_1^{0.5}} + \frac{1}{n_2^{0.5}} \right)$$

Similarly

$$\begin{aligned} E|\hat{\phi}_{kk't} - \phi_{kk't}|^2 &= \left( \sum_{i,j} \mu_{ik} \nu_{jk'} \psi'(p_{ijt}^*) (\hat{p}_{ijt} - p_{ijt}) \right)^2 \\ &\leq M^2 \sum_{i,j} \mu_{ik}^2 \nu_{jk'}^2 E \left\{ \sum_{i,j} (\hat{p}_{ijt} - p_{ijt})^2 \right\} \\ &\leq M^2 IJ \max_{i,j} E(\hat{p}_{ijt} - p_{ijt})^2 \\ &\leq M^2 IJ \max_{i,j} \frac{p_{ijt}(1 - p_{ijt})}{n_t} \\ &\leq M^2 IJ \frac{1}{4n_t}, \end{aligned}$$

which yields

$$E|\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2} - (\phi_{kk'1} - \phi_{kk'2})|^2 \leq \frac{M^2 IJ}{4} \left( \frac{1}{n_1} + \frac{1}{n_2} \right),$$

and the proof is complete  $\square$

**Proof.** Proof of Theorem (3.2). Similar to Theorem (3.1), We have

$$\begin{aligned} |\hat{\phi}_{kk't} - \phi_{kk't}| &\leq M \sum_{i,j} |\mu_{ik}\nu_{jk'}| \max_{i,j} |\hat{p}_{ijt} - p_{ijt}| \\ &\leq M \frac{I+J}{2} \max_{i,j} |\hat{p}_{ijt} - p_{ijt}| \end{aligned}$$

Define

$$\lambda_t = M \frac{I+J}{2} \max_{i,j} |\hat{p}_{ijt} - p_{ijt}|.$$

For comparison of given two contingency tables, it is obvious

$$\begin{aligned} |\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2} - (\phi_{kk'1} - \phi_{kk'2})| &\leq M \frac{I+J}{2} \{ \max_{i,j} |\hat{p}_{ij1} - p_{ij1}| \\ &\quad + \max_{i,j} |\hat{p}_{ij2} - p_{ij2}| \}. \end{aligned} \tag{A.2}$$

We define the thresholding level  $\lambda$  as

$$\lambda = \lambda_1 + \lambda_2 = M \frac{I+J}{2} \{ \max_{i,j} |\hat{p}_{ij1} - p_{ij1}| + \max_{i,j} |\hat{p}_{ij2} - p_{ij2}| \}.$$

Next, by (A.2) and Conditions (i) and (iii) of  $S_\lambda(z)$ ,

$$\begin{aligned} |S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2}) - (\phi_{kk'1} - \phi_{kk'2})| &\leq |S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2})| + |\phi_{kk'1} - \phi_{kk'2}| \\ &\quad (1+C)|\phi_{kk'1} - \phi_{kk'2}|, \end{aligned}$$

and

$$\begin{aligned} |S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2}) - (\phi_{kk'1} - \phi_{kk'2})| &\leq |S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2}) - (\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2})| \\ &\quad + |\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2} - (\phi_{kk'1} - \phi_{kk'2})| \leq 2\lambda. \end{aligned}$$

These two inequalities imply

$$|S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2}) - (\phi_{kk'1} - \phi_{kk'2})| \leq (2\lambda)^{1-q} (1+C)^q |\phi_{kk'1} - \phi_{kk'2}|^q,$$

and

$$|S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2}) - (\phi_{kk'1} - \phi_{kk'2})|^2 \leq (2\lambda)^{2-2q} (1+C)^q |\phi_{kk'1} - \phi_{kk'2}|^q,$$

where  $0 \leq q < 1$ . Hence, for  $L_1$  loss, we have

$$\begin{aligned} E \sum_{k,k'} |S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2}) - (\phi_{kk'1} - \phi_{kk'2})| &\leq 2^{1-q} E(\lambda)^{1-q} (1+C)^q \\ &\quad \times \sum_{k,k'} |\phi_{kk'1} - \phi_{kk'2}|^q \\ &= 2^{1-q} E(\lambda)^{1-q} (1+C)^q T_q(I, J) \end{aligned}$$

where

$$\begin{aligned} E(\lambda)^{1-q} &= (M \frac{I+J}{2})^{1-q} E \{ \max_{i,j} |\hat{p}_{ij1} - p_{ij1}| + \max_{i,j} |\hat{p}_{ij2} - p_{ij2}| \}^{1-q} \\ &\leq (M \frac{I+J}{2})^{1-q} \{ E \max_{i,j} |\hat{p}_{ij1} - p_{ij1}| + E \max_{i,j} |\hat{p}_{ij2} - p_{ij2}| \}^{1-q} \\ &\leq (M \frac{I+J}{2})^{1-q} (\frac{3}{2})^{1-q} 2^{\frac{1-q}{2}} (1 + \log(2IJ))^{\frac{1-q}{2}} \{ \frac{1}{n_1} + \frac{1}{n_2} \}^{\frac{1-q}{2}}. \end{aligned}$$

Moreover for  $L_2$  loss, we have

$$\begin{aligned} E \sum_{k,k'} |S_\lambda(\hat{\phi}_{kk'1} - \hat{\phi}_{kk'2}) - (\phi_{kk'1} - \phi_{kk'2})|^2 &\leq 2^{2-q} E(\lambda)^{2-q} (1+C)^q \\ &\quad \times \sum_{k,k'} |\phi_{kk'1} - \phi_{kk'2}|^q \\ &= 2^{2-q} E(\lambda)^{2-q} (1+C)^q T_q(I, J) \end{aligned}$$

where

$$\begin{aligned} E(\lambda)^{2-q} &= \left(M \frac{I+J}{2}\right)^{2-q} E \left\{ \max_{i,j} |\hat{p}_{ij1} - p_{ij1}| + \max_{i,j} |\hat{p}_{ij2} - p_{ij2}| \right\}^{2-q} \\ &\leq \left(M \frac{I+J}{2}\right)^{2-q} 2^{2-q} \left\{ E \max_{i,j} |\hat{p}_{ij1} - p_{ij1}|^2 + E \max_{i,j} |\hat{p}_{ij2} - p_{ij2}|^2 \right\}^{1-\frac{q}{2}} \\ &\leq \left(M \frac{I+J}{2}\right)^{2-q} 2^{2-q} \left\{ E \max_{i,j} |\hat{p}_{ij1} - p_{ij1}| + E \max_{i,j} |\hat{p}_{ij2} - p_{ij2}| \right\}^{1-\frac{q}{2}} \\ &\leq \left(M \frac{I+J}{2}\right)^{2-q} 2^{2-q} \left(\frac{3}{2}\right)^{1-\frac{q}{2}} (1 + \log(2IJ))^{\frac{2-q}{4}} \left\{ \frac{1}{n_1} + \frac{1}{n_2} \right\}^{\frac{2-q}{4}}, \end{aligned}$$

which finishes the proof of Theorem (3.2). □