Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 1, 2021 (New Series), pp 417–429

Unsaturated Log-Linear Model Selection for Categorical Data Analysis

Subir Ghosh[1] **and Arnab Chowdhury**[2]

[1]*Department of Statistics*
University of California, Riverside, California, USA
[2]*Department of Computational and Quantitative Medicine*
City of Hope National Medical Center, Duarte, California, USA

Received: 25 February 2021; Revised: 23 March 2021; Accepted: 25 March 2021

Abstract

The paper proposes a new metric SAVE for finding the best fitted unsaturated log-linear model to describe the categorical data in a contingency table with m categorical variables. Two kinds of extensions, standard and orthogonal, of an unsaturated log-linear model to the saturated model are the foundation of SAVE. The performance of SAVE in terms of the correct model parameter(s) detection is comparable with or even better than the commonly used metrics: Deviance, AIC, and BIC, as demonstrated in simulation studies.

Key words: Categorical; Log-linear; Model selection; Orthogonal extension; Unsaturated.

AMS Subject Classifications: 62H17, 62R07, 62B10.

1. Introduction

Let $X_1, \ldots, X_m$ denote m categorical variables, X_i with I_i categories, $i = 1, \ldots, m$. The n subjects selected in a study using a multinomial sample are cross-classified into $N = I_1 \times \cdots \times I_m$ possible combinations on m categorical variables $X_1, \ldots, X_m$ in a contingency table. The wth combination is represented by $(w_1, \ldots, w_m)$, where w_u is the level of X_u; $w_u = 0, \ldots, I_u - 1; u = 1, \ldots, m$. The number of subjects for the wth combination is a random variable Y_w having the observed value y_w and the expected value $E(Y_w) = \mu_w = np_w$, where p_w and μ_w are unknown parameters. The μ_w is the cell mean and p_w is the cell probability for the cell represented by the wth combination. We have $Y_w \geq 0$ and $Y_1 + \cdots + Y_N = n$, $p_1 + \cdots + p_N = 1$ and $\mu_1 + \cdots + \mu_N = n$. Also, $y_w \geq 0$, $w = 1, \ldots, N$, and $y_1 + \cdots + y_N = n$. The saturated log-linear model is

$$log(p_w) = \lambda + \delta_1 \lambda_{w_1}^{X_1} + \cdots + \delta_m \lambda_{w_m}^{X_m} + \delta_1 \delta_2 \lambda_{w_1 w_2}^{X_1 X_2} + \cdots + \delta_1 \delta_2 \delta_3 \lambda_{w_1 w_2 w_3}^{X_1 X_2 X_3} + \cdots + \delta_1 \ldots \delta_m \lambda_{w_1 \ldots w_m}^{X_1 \ldots X_m}, \quad (1)$$

where $\{\lambda_{w_{i_1} w_{i_2}}^{X_{i_1} X_{i_2}}\}, \{\lambda_{w_{i_1} w_{i_2} w_{i_3}}^{X_{i_1} X_{i_2} X_{i_3}}\}, \ldots,$ and $\lambda_{w_1 \ldots w_m}^{X_1 \ldots X_m}$, are the unknown association parameters. The $\{\lambda_{w_i}^{X_i}\}$ are the unknown effect parameters. The λ is the unknown overall effect parameter.

Correponding Author: Subir Ghosh
Email: subir.ghosh@ucr.edu

The $\delta_u, u = 1, \ldots, m$, are

$$\delta_u = \begin{cases} 0 & \text{if} \quad w_u = 0, \\ 1 & \text{if} \quad w_u \neq 0. \end{cases}$$

When at least one association parameter is zero or absent in the saturated model, the model becomes an unsaturated model in presence of the overall effect and the effect parameters. The unsaturated models considered in this paper consist of the overall effect, the effect parameters, and one or more association parameters. When all association parameters are absent in the unsaturated model, the categorical variables become mutually independent. "In practice, unsaturated models are preferable since their fit smoothes the sample data and has simpler interpretations" (page 341, Agresti (2013)). On the one hand, the over-fitted saturated model is unnecessary, but on the other hand, an under-fitted unsaturated model is deficient for describing the data. We propose a new method of finding the best fitted unsaturated log-linear model using the association parameters absent in the model considered but present in the saturated model. We compare the proposed method with the standard measures such as AIC, BIC, and Deviance using the 100,000 realizations of simulated data.

In Section 2, we present two saturated representations of standard and orthogonal extensions of unsaturated log-linear models. In Section 3, we explain the saturated representations with two illustrative examples in Sections 3.1 and 3.2. The data on the use of automobile seat-belt for lowering fatal injury is in Section 4. We propose the new metric, SAVE, in Section 5. We compare the new metric with the other available metrics AIC, BIC, and MDI in Section 5.1. Section 6 presents their performance comparison for the 100,000 simulated data from each of the six data-generating models. We conclude in Section 7 with some remarks.

2. Two Saturated Representations : $S1$ and $S2$

Let $\boldsymbol{p} = (p_1, \ldots, p_N)^\top$ be the column vector of expected counts for the N cells of the contingency table, $\boldsymbol{\lambda}^{(1)}$ $(k_1 \times 1)$ be the vector of the overall effect, the effect parameters, and the one or more association parameters in an unsaturated model considered for fitting to the collected data, and $\boldsymbol{X}_1$ $(N \times k_1)$ be the model matrix generated from the indicator variables for the parameters in $\boldsymbol{\lambda}_1^{(1)}$. Let $\boldsymbol{\lambda}_2$ $(k_2 \times 1)$ be the vector of association parameters that are absent in $\boldsymbol{\lambda}^{(1)}$ and $\boldsymbol{X}_2$ $(N \times k_2)$ be the model matrix generated from the indicator variables for the parameters in $\boldsymbol{\lambda}_2$. In the saturated model (1), the parameters in both $\boldsymbol{\lambda}^{(1)}$ and $\boldsymbol{\lambda}_2$ are present. The unsaturated model consists of the parameters in $\boldsymbol{\lambda}^{(1)}$ but not the parameters in $\boldsymbol{\lambda}_2$. The matrix representation of the unsaturated model considered is

$$log\boldsymbol{p} = \boldsymbol{X}_1\boldsymbol{\lambda}_1^{(1)}, \tag{2}$$

where rank$(\boldsymbol{X}_1) = k_1$. We consider two representations of the saturated model. The first representation is the standard saturated model and we denote it by $S1$. The second representation is the orthogonal extension of the assumed unsaturated model in (2) and it is denoted by $S2$ (Klimova, Rudas and Dobra (2012), Klimova and Rudas (2016), Rudas (2018)). The standard representation $S1$ of the saturated log-linear model is

$$log\boldsymbol{p} = \boldsymbol{X}_1\boldsymbol{\lambda}_1^{(1)} + \boldsymbol{X}_2\boldsymbol{\lambda}_2, \tag{3}$$

where rank$(\boldsymbol{X}_1,\, \boldsymbol{X}_2) = k_1 + k_2 = N$.

Let $\boldsymbol{D}$ $(N \times k_2)$ be a matrix which satisfies

$$rank(\boldsymbol{D}) = k_2, \boldsymbol{X}_1^{\top}\boldsymbol{D} = \boldsymbol{0}. \tag{4}$$

The matrix $\boldsymbol{D}$ is not unique. A simple form of the matrix $\boldsymbol{D}$ satisfying (4) is

$$\boldsymbol{D} = [\boldsymbol{I}_N - \boldsymbol{X}_1(\boldsymbol{X}_1^{\top}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\top}]\boldsymbol{X}_2, \tag{5}$$

where $\boldsymbol{I}_N$ is the $(N \times N)$ identity matrix. Note that rank$([\boldsymbol{I}_N - \boldsymbol{X}_1(\boldsymbol{X}_1^{\top}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\top}]) = N - k_1 = k_2 =$ rank$(\boldsymbol{X}_2) =$ rank$(\boldsymbol{D})$. From (5), it can be seen

$$\begin{aligned} \boldsymbol{D}\boldsymbol{\lambda}_2 &= [\boldsymbol{I}_N - \boldsymbol{X}_1(\boldsymbol{X}_1^{\top}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\top}]\boldsymbol{X}_2\boldsymbol{\lambda}_2, \\ \boldsymbol{X}_2\boldsymbol{\lambda}_2 &= \boldsymbol{D}\boldsymbol{\lambda}_2 + \boldsymbol{X}_1(\boldsymbol{X}_1^{\top}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\top}\boldsymbol{X}_2\boldsymbol{\lambda}_2. \end{aligned} \tag{6}$$

Let

$$\boldsymbol{\lambda}_1^{(2)} = \boldsymbol{\lambda}_1^{(1)} + (\boldsymbol{X}_1^{\top}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\top}\boldsymbol{X}_2\boldsymbol{\lambda}_2. \tag{7}$$

The orthogonal saturated extension of the unsaturated model in (2), $S2$, is obtained from (3) and (7) as

$$\begin{aligned} log\boldsymbol{p} &= \boldsymbol{X}_1\boldsymbol{\lambda}_1^{(1)} + \boldsymbol{X}_2\boldsymbol{\lambda}_2 \\ &= \boldsymbol{X}_1\boldsymbol{\lambda}_1^{(1)} + \boldsymbol{D}\boldsymbol{\lambda}_2 + \boldsymbol{X}_1(\boldsymbol{X}_1^{\top}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\top}\boldsymbol{X}_2\boldsymbol{\lambda}_2 \\ &= \boldsymbol{X}_1\left(\boldsymbol{\lambda}_1^{(1)} + (\boldsymbol{X}_1^{\top}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\top}\boldsymbol{X}_2\boldsymbol{\lambda}_2\right) + \boldsymbol{D}\boldsymbol{\lambda}_2 \\ &= \boldsymbol{X}_1\boldsymbol{\lambda}_1^{(2)} + \boldsymbol{D}\boldsymbol{\lambda}_2. \end{aligned} \tag{8}$$

From (4) and (8), it follows that

$$\begin{aligned} \boldsymbol{\lambda}_1^{(2)} &= (\boldsymbol{X}_1^{\top}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\top}log\boldsymbol{p} \\ \boldsymbol{\lambda}_2 &= (\boldsymbol{D}^{\top}\boldsymbol{D})^{-1}\boldsymbol{D}^{\top}log\boldsymbol{p}. \end{aligned} \tag{9}$$

Klimova, Rudas and Dobra (2012), Klimova and Rudas (2016), Rudas (2018) defined two kinds of relational models, dual and non-dual. For a dual representation of a relational model, we have $\boldsymbol{D}^{\top}log\boldsymbol{p} = 0$. In other words, from (9), $\boldsymbol{\lambda}_2 = 0$. Hence, the unsaturated model in (2) has a dual representation. On the other hand, for a non-dual representation of a relational model, we have $\boldsymbol{D}^{\top}log\boldsymbol{p} \neq 0$. Therefore, the saturated model in (3) has a non-dual representation.

3. Examples

3.1. Example 1

For a $2 \times 2 \times 2$ contingency table and the unsaturated model in (2) with three independent categorical variables X_1, X_2 and X_3, we have $m = 3$, $N = 8$, $k_1 = k_2 = 4$. Table 1 presents the cell representations.

Table 1: The cell representations for Example 1

Number w	Combination (w_1, w_2, w_3)	Probability p_w
1	(0, 0, 0)	p_1
2	(0, 0, 1)	p_2
3	(0, 1, 0)	p_3
4	(0, 1, 1)	p_4
5	(1, 0, 0)	p_5
6	(1, 0, 1)	p_6
7	(1, 1, 0)	p_7
8	(1, 1, 1)	p_8

The matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \boldsymbol{X}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \tag{10}$$

and the vectors $\boldsymbol{\lambda}_1^{(1)}$ and $\boldsymbol{\lambda}_2$ in (3) are

$$\boldsymbol{\lambda}_1^{(1)} = (\lambda, \lambda_1^{X_1}, \lambda_1^{X_2}, \lambda_1^{X_3})^\top, \boldsymbol{\lambda}_2 = (\lambda_{11}^{X_1X_2}, \lambda_{11}^{X_1X_3}, \lambda_{11}^{X_2X_3}, \lambda_{111}^{X_1X_2X_3})^\top. \tag{11}$$

Two $\boldsymbol{D}$ matrices, $\boldsymbol{D}_{(1)}$ and $\boldsymbol{D}_{(2)}$ in (12), are obtained by using (5) and (10). The last column of $\boldsymbol{D}_{(1)}$ is not orthogonal to its first three columns. The first three columns of $\boldsymbol{D}_{(1)}$ are mutually orthogonal. The first three columns of $\boldsymbol{D}_{(2)}$ are the same as the corresponding columns in $\boldsymbol{D}_{(1)}$. The four columns of $\boldsymbol{D}_{(2)}$ are mutually orthonormal. Thus, $\boldsymbol{D}_{(2)}^\top \boldsymbol{D}_{(2)} = \boldsymbol{I}_4$.

$$\boldsymbol{D}_{(1)} = (1/4) \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 0 \\ -1 & 1 & -1 & 0 \\ -1 & -1 & 1 & -1 \\ -1 & -1 & 1 & 0 \\ -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 2 \end{bmatrix}, \boldsymbol{D}_{(2)} = (1/8) \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{bmatrix}. \tag{12}$$

For a dual relational model by using the expression of $\boldsymbol{D}_{(2)}$ in (12) for $\boldsymbol{D}$, we find

$$\begin{aligned}(i).\ log\Big(\frac{p_1p_2p_7p_8}{p_3p_4p_5p_6}\Big) = 0,\ \ (ii).\ log\Big(\frac{p_1p_3p_6p_8}{p_2p_4p_5p_7}\Big) = 0,\\ (iii).\ log\Big(\frac{p_1p_4p_5p_8}{p_2p_3p_6p_7}\Big) = 0,\ \ (iv).\ log\Big(\frac{p_1p_4p_6p_7}{p_2p_3p_5p_8}\Big) = 0.\end{aligned} \tag{13}$$

From the equations (iii) and (iv) in (13), it can be seen

$$(i).\ log\Big(\frac{p_1p_4}{p_2p_3}\Big) = 0,\ \ (ii).\ log\Big(\frac{p_5p_8}{p_6p_7}\Big) = 0. \tag{14}$$

In Table 1, we observe that $X_1 = 0$ for $w = 1, 2, 3, 4$ and $X_1 = 1$ for $w = 5, 6, 7, 8$. The equation (i) in (14) implies the conditional independence between the categorical variables X_2 and X_3 given $X_1 = 0$. The equation (ii) in (14) implies the conditional independence between X_2 and X_3 given $X_1 = 1$.

From the equations (i) and (iv) in (13), we observe

$$(i).\ log\Big(\frac{p_1p_6}{p_2p_5}\Big) = 0,\ \ (ii).\ log\Big(\frac{p_4p_7}{p_3p_8}\Big) = 0. \tag{15}$$

In Table 1, we observe that $X_2 = 0$ for $w = 1, 2, 5, 6$ and $X_2 = 1$ for $w = 3, 4, 7, 8$. The equation (i) in (15) implies the conditional independence between the categorical variables X_1 and X_3 given $X_2 = 0$. The equation (ii) in (15) implies the conditional independence between X_1 and X_3 given $X_2 = 1$.

From the equations (i) and (iv) in (13), we find

$$(i).\ log\Big(\frac{p_1p_7}{p_3p_5}\Big) = 0,\ \ (ii).\ log\Big(\frac{p_2p_8}{p_4p_6}\Big) = 0. \tag{16}$$

In Table 1, we observe that $X_3 = 0$ for $w = 1, 3, 5, 7$ and $X_3 = 1$ for $w = 2, 4, 6, 8$. The equation (i) in (16) implies the conditional independence between the categorical variables X_1 and X_2 given $X_3 = 0$. The equation (ii) in (16) implies the conditional independence between X_1 and X_2 given $X_3 = 1$.

3.2. Example 2

For a 3×2 contingency table and the unsaturated model in (2) with two independent categorical variables X_1 and X_2, we have $m = 2$, $N = 6$, $k_1 = 4, k_2 = 2$. Table 2 presents the cell representations.

The matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 & 0 & 0 & 0\\ 1 & 0 & 0 & 1\\ 1 & 1 & 0 & 0\\ 1 & 1 & 0 & 1\\ 1 & 0 & 1 & 0\\ 1 & 0 & 1 & 1 \end{bmatrix}, \boldsymbol{X}_2 = \begin{bmatrix} 0 & 0\\ 0 & 0\\ 0 & 0\\ 1 & 0\\ 0 & 0\\ 0 & 1 \end{bmatrix}, \tag{17}$$

Table 2: The cell representations for Example 2

Number w	Combination (w_1, w_2)	Probability p_w
1	(0, 0)	p_1
2	(0, 1)	p_2
3	(1, 0)	p_3
4	(1, 1)	p_4
5	(2, 0)	p_5
6	(2, 1)	p_6

and the vectors $\boldsymbol{\lambda}_1^{(1)}$ and $\boldsymbol{\lambda}_2$ in (3) are

$$\boldsymbol{\lambda}_1^{(1)} = (\lambda, \lambda_1^{X_1}, \lambda_2^{X_1}, \lambda_1^{X_2})^\top, \boldsymbol{\lambda}_2 = (\lambda_{11}^{X_1X_2}, \lambda_{21}^{X_1X_2})^\top. \tag{18}$$

The matrices $\boldsymbol{D}_{(1)}$ and $\boldsymbol{D}_{(2)}$ in (19) are obtained by using (5) and (17). The two columns of $\boldsymbol{D}_{(1)}$ are not mutually orthogonal. The two columns of $\boldsymbol{D}_{(2)}$ are mutually orthonormal. Thus, $\boldsymbol{D}_{(2)}^\top \boldsymbol{D}_{(2)} = \boldsymbol{I}_2$.

$$\boldsymbol{D}_{(1)} = (1/6)\begin{bmatrix} 1 & 1 \\ -1 & -1 \\ -2 & 1 \\ 2 & -1 \\ 1 & -2 \\ -1 & 2 \end{bmatrix}, \boldsymbol{D}_{(2)} = \begin{bmatrix} 1 & 1 \\ -1 & -1 \\ -2 & 0 \\ 2 & 0 \\ 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} (1/2\sqrt{3}) & 0 \\ 0 & (1/2) \end{bmatrix}. \tag{19}$$

For a dual relational model by using the expression of $\boldsymbol{D}_{(2)}$ in (19) for $\boldsymbol{D}$, we find

$$(i).\ log\left(\frac{p_1p_5}{p_2p_6}\right) = 2 \times log\left(\frac{p_3}{p_4}\right),\ \ (ii).\ log\left(\frac{p_1p_6}{p_2p_5}\right) = 0. \tag{20}$$

4. A Real Data

A research investigation started with a question (Agresti (2013)): Does seat-belt use in automobiles reduce injury? The collected data in Table 4 were on the injury outcomes of 68,694 passengers in autos and light trucks involved in accidents one year in the state of Maine, USA. Three factors each at two levels displayed in Table 3 were three categorical variables ($m = 3$) for the Table 4 data.

For the vectors $\boldsymbol{\lambda}_1^{(1)}$ and $\boldsymbol{\lambda}_2$ in (3) as

$$\boldsymbol{\lambda}_1^{(1)} = (\lambda, \lambda_1^{X_1}, \lambda_1^{X_2}, \lambda_1^{X_3}, \lambda_{11}^{X_1X_3}, \lambda_{11}^{X_2X_3})^\top, \boldsymbol{\lambda}_2 = (\lambda_{11}^{X_1X_2}, \lambda_{111}^{X_1X_2X_3})^\top, \tag{21}$$

Table 3: Three factors and their levels

Factors/ Categories	X_i	Levels 0	 1
Location	X_1	Urban	Rural
Seat-belt use	X_2	No	Yes
Injury	X_3	No	Yes

Table 4: The number of subjects y_w

w	X_1, X_2, X_3	y_w
1	000	17,668
2	001	1,808
3	010	22,556
4	011	1,139
5	100	9,369
6	101	2,057
7	110	12,827
8	111	1,270

the matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ in (3), and $\boldsymbol{D}$ in (5) are

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \boldsymbol{X}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}, \boldsymbol{D} = \begin{bmatrix} 0.25 & 0.00 \\ 0.25 & 0.25 \\ -0.25 & 0.00 \\ -0.25 & -0.25 \\ -0.25 & 0.00 \\ -0.25 & -0.25 \\ 0.25 & 0.00 \\ 0.25 & 0.25 \end{bmatrix}. \tag{22}$$

5. SAVE - A New Model Selection Criterion

For the saturated log-linear model $S1$ in (3), assume

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{11} & \boldsymbol{X}_{12} \\ \boldsymbol{X}_{21} & \boldsymbol{X}_{22} \end{bmatrix}, \tag{23}$$

where the matrix $\boldsymbol{X}_{11}(k_1 \times k_1)$ has rank k_1 and $\boldsymbol{X}_1^\top \boldsymbol{X}_2 \neq \boldsymbol{0}$. Recall from (2) and (3) that rank$(\boldsymbol{X}_1) = k_1$ and rank$(\boldsymbol{X})= k_1 + k_2 = N$.

For the saturated log-linear model $S2$ in (8), assume

$$\boldsymbol{X}^* = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{D} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{11} & \boldsymbol{D}_1 \\ \boldsymbol{X}_{21} & \boldsymbol{D}_2 \end{bmatrix}, \tag{24}$$

where $\operatorname{rank}(\boldsymbol{X}^*) = k_1 + k_2 = N$. Recall from (4) that $\operatorname{rank}(\boldsymbol{D}) = k_2$ and $\boldsymbol{X}_1^\top \boldsymbol{D} = \boldsymbol{0}$. Let $\boldsymbol{P}$ be an $(N \times N)$ lower-diagonal matrix

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{I}_{k_1} & \boldsymbol{0} \\ -\boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1} & \boldsymbol{I}_{k_2} \end{bmatrix}. \tag{25}$$

Pre-multiplying the matrices $\boldsymbol{X}$ in (23) and $\boldsymbol{X}^*$ in (24) by $\boldsymbol{P}$ in (25)

$$\boldsymbol{P}\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_{11} & \boldsymbol{X}_{12} \\ \boldsymbol{0} & \boldsymbol{X}_{22} - \boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1}\boldsymbol{X}_{12} \end{bmatrix}, \boldsymbol{P}\boldsymbol{X}^* = \begin{bmatrix} \boldsymbol{X}_{11} & \boldsymbol{D}_1 \\ \boldsymbol{0} & \boldsymbol{D}_2 - \boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1}\boldsymbol{D}_1 \end{bmatrix}. \tag{26}$$

Let $\widehat{\boldsymbol{\lambda}}_1^{(1)}$ be the estimator of $\boldsymbol{\lambda}_1^{(1)}$ and $\widehat{\boldsymbol{\lambda}}_2^{(1)}$ of $\boldsymbol{\lambda}_2$, for $S1$ in (3). Let $\widehat{\boldsymbol{\lambda}}_1^{(2)}$ be the estimator of $\boldsymbol{\lambda}_1^{(2)}$ and $\widehat{\boldsymbol{\lambda}}_2^{(2)}$ of $\boldsymbol{\lambda}_2$, for $S2$ in (8). From (8) and (26), it can be seen that

$$\begin{aligned} \boldsymbol{X}_{11}\widehat{\boldsymbol{\lambda}}_1^{(1)} + \boldsymbol{X}_{12}\widehat{\boldsymbol{\lambda}}_2^{(1)} &= \boldsymbol{X}_{11}\widehat{\boldsymbol{\lambda}}_1^{(2)} + \boldsymbol{D}_1\widehat{\boldsymbol{\lambda}}_2^{(2)}, \\ (\boldsymbol{X}_{22} - \boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1}\boldsymbol{X}_{12})\widehat{\boldsymbol{\lambda}}_2^{(1)} &= (\boldsymbol{D}_2 - \boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1}\boldsymbol{D}_1)\widehat{\boldsymbol{\lambda}}_2^{(2)}. \end{aligned} \tag{27}$$

Clearly from (27),

$$\begin{aligned} \widehat{\boldsymbol{\lambda}}_2^{(1)} &= (\boldsymbol{X}_{22} - \boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1}\boldsymbol{X}_{12})^{-1}(\boldsymbol{D}_2 - \boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1}\boldsymbol{D}_1)\widehat{\boldsymbol{\lambda}}_2^{(2)}, \\ \widehat{\boldsymbol{\lambda}}_1^{(2)} - \widehat{\boldsymbol{\lambda}}_1^{(1)} &= \boldsymbol{X}_{11}^{-1}(\boldsymbol{X}_{12}\widehat{\boldsymbol{\lambda}}_2^{(1)} - \boldsymbol{D}_1\widehat{\boldsymbol{\lambda}}_2^{(2)}). \end{aligned} \tag{28}$$

Theorem 1: For two matrices, $\boldsymbol{X}$ in (23) in the standard representation $S1$ of the saturated log-linear model in (3) and $\boldsymbol{X}^*$ in (24) in the orthogonal extension representation $S2$ of the saturated log-linear model in (5), the estimators $\widehat{\boldsymbol{\lambda}}_1^{(1)}$ of $\boldsymbol{\lambda}_1^{(1)}$ and $\widehat{\boldsymbol{\lambda}}_2^{(1)}$ of $\boldsymbol{\lambda}_2$ for $S1$ in (3), $\widehat{\boldsymbol{\lambda}}_1^{(2)}$ of $\boldsymbol{\lambda}_1^{(2)}$ and $\widehat{\boldsymbol{\lambda}}_2^{(2)}$ of $\boldsymbol{\lambda}_2$ for $S2$ in (8), satisfy
(i) $\widehat{\boldsymbol{\lambda}}_2^{(2)} = \widehat{\boldsymbol{\lambda}}_2^{(1)}$ if $(\boldsymbol{X}_{22} - \boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1}\boldsymbol{X}_{12}) = (\boldsymbol{D}_2 - \boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1}\boldsymbol{D}_1)$,
(ii) $\widehat{\boldsymbol{\lambda}}_1^{(2)} = \widehat{\boldsymbol{\lambda}}_1^{(1)}$ if and only if $\boldsymbol{X}_{12}\widehat{\boldsymbol{\lambda}}_2^{(1)} = \boldsymbol{D}_1\widehat{\boldsymbol{\lambda}}_2^{(2)}$.

Proof: The proof follows from (28).

Theorem 2: For the orthogonal extension representation $S2$ of the saturated log-linear model in (8), the matrix $\boldsymbol{D}$ is not unique but $\boldsymbol{D}\widehat{\boldsymbol{\lambda}}_2^{(2)}$ is unique.

Proof: From (3), (8), (9), and the condition $\boldsymbol{X}_1^\top \boldsymbol{D} = \boldsymbol{0}$ in (4),

$$\begin{aligned}
log\widehat{\boldsymbol{p}} &= \boldsymbol{X}_1\widehat{\boldsymbol{\lambda}}_1^{(1)} + \boldsymbol{X}_2\widehat{\boldsymbol{\lambda}}_2^{(1)} \\
&= \boldsymbol{X}_1\widehat{\boldsymbol{\lambda}}_1^{(2)} + \boldsymbol{D}\widehat{\boldsymbol{\lambda}}_2^{(2)}, \\
\widehat{\boldsymbol{\lambda}}_2^{(2)} &= (\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{D}^\top\boldsymbol{X}_2\widehat{\boldsymbol{\lambda}}_2^{(2)} \\
&= (\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{D}^\top log\widehat{\boldsymbol{p}}, \\
\widehat{\boldsymbol{\lambda}}_1^{(2)} &= (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top log\widehat{\boldsymbol{p}}, \\
\boldsymbol{D}\widehat{\boldsymbol{\lambda}}_2^{(2)} &= log\widehat{\boldsymbol{\mu}} - \boldsymbol{X}_1\widehat{\boldsymbol{\lambda}}_1^{(2)} \\
&= [\boldsymbol{I}_N - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top]log\widehat{\boldsymbol{p}}.
\end{aligned} \tag{29}$$

The right hand side of $\boldsymbol{D}\widehat{\boldsymbol{\lambda}}_2^{(2)}$ in (29) depends only on $\boldsymbol{X}_1$ and $\widehat{\boldsymbol{p}}$ but not $\boldsymbol{D}$ since the elements of $\widehat{\boldsymbol{p}}$ are $y_w/n, w = 1, \ldots, N$. Hence, $\boldsymbol{D}\widehat{\boldsymbol{\lambda}}_2^{(2)}$ is unique.

Theorem 3: The sum of the elements in $\boldsymbol{D}\widehat{\boldsymbol{\lambda}}_2^{(2)}$ is zero.

Proof: Since the first column of $\boldsymbol{X}_1$ is an $(N \times 1)$ column vector $\boldsymbol{j}_N = (1, 1, \ldots, 1)^\top$ with the elements equal to one, it follows from (4) that $\boldsymbol{j}_N^\top \boldsymbol{D} = \boldsymbol{0}$ and therefore, $\boldsymbol{j}_N^\top \boldsymbol{D}\boldsymbol{\lambda}_2^{(2)} = 0$. In other words, the sum of elements of $\boldsymbol{D}\widehat{\boldsymbol{\lambda}}_2^{(2)}$ is zero.

It follows from Theorem 3 that the non-zero elements of $\boldsymbol{D}\widehat{\boldsymbol{\lambda}}_2^{(2)}$ are either positive or negative. Moreover, the sum of the positive elements is negative of the sum of the negative values. A new model comparison criterion is proposed as

$$\begin{aligned}
\text{SAVE} &= \text{The sum of the absolute values of the elements in } \boldsymbol{D}\widehat{\boldsymbol{\lambda}}_2^{(2)} \\
&= 2 \times \text{The sum of the positive elements in } \boldsymbol{D}\widehat{\boldsymbol{\lambda}}_2^{(2)}.
\end{aligned} \tag{30}$$

for comparing a class of unsaturated log-linear models. Smaller the value of SAVE for a model means the better fit to describe the data. The unsaturated model having the smallest value of SAVE means the elements of $\boldsymbol{D}\widehat{\boldsymbol{\lambda}}_2^{(2)}$ are overall individually small. In other words, the unsaturated model provides the closest fitted values of p_w to their corresponding observed values y_w/n, for $w = 1, \ldots, N$.

5.1. Comparison of unsaturated models fitted to the seat-belt use data

Table 5 compares the seven unsaturated models in fitting to the Section 4 data using the four criterion functions: AIC, BIC, MDI, and SAVE. From now on, $\lambda_1^{X_1}, \lambda_1^{X_2}, \lambda_1^{X_3}, \lambda_{11}^{X_1X_2}, \lambda_{11}^{X_1X_3}$, $\lambda_{11}^{X_2X_3}$, and $\lambda_{111}^{X_1X_2X_3}$ are denoted by $\lambda_1, \lambda_2, \lambda_3, \lambda_{12}, \lambda_{13}, \lambda_{23}$, and λ_{123}, respectively.

Table 5: The comparison of seven unsaturated log-linear models

	Model	AIC	BIC	MDI	SAVE
1	$\lambda_{123} = 0$	99.14	99.69	85.14	0.09
2	$\lambda_{23} = \lambda_{123} = 0$	878.96	879.43	866.96	1.50
3	$\lambda_{13} = \lambda_{123} = 0$	830.53	831.01	818.53	1.44
4	$\lambda_{12} = \lambda_{123} = 0$	111.70	112.18	99.70	0.09
5	$\lambda_{13} = \lambda_{23} = \lambda_{123} = 0$	1596.57	1596.96	1586.57	1.51
6	$\lambda_{12} = \lambda_{23} = \lambda_{123} = 0$	877.73	878.13	867.73	1.50
7	$\lambda_{12} = \lambda_{13} = \lambda_{123} = 0$	829.31	829.71	819.31	1.47

The criterion functions AIC and BIC (Akaike (1973), Schwarz (1978)), Konishi and Kitagawa (2008)) penalize the bigger model, while the Minimum Discrimination Information (MDI) (Kullback and Leibler (1951), Kullback (1959), Csiszár (1975), Gokhale and Kullback (1978), Haberman (1984), Kullback, Keegel, and Kullback (2013)) and SAVE do not. The best-fitted model having the smallest values of all four criterion functions is the model with $\lambda_{123} = 0$. The second-best model under all four criterion functions, is the model having $\lambda_{12} = \lambda_{123} = 0$, which means the conditional independence between X_1 and X_2 given X_3. The proposed criterion function SAVE does not discriminate visibly between the top two models by the other three criterion functions numerically for the data considered.

6. A Performance Evaluation Simulation Study for a $2\times2\times2$ Contingency Table

The 100,000 multinomial random samples are generated from the six log-linear models satisfying (1) for a $2 \times 2 \times 2$ contingency table. The eight λ values for the data generating six models are given in Table 6 so that the sum of $p_w, w = 1, \ldots, 8$, is 1. The p_w values are displayed in Table 7.

Table 6: The λ parameters of the six data generating models

Parameters	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$
λ	-2.4654	-4.3262	-1.3008	-2.0844	-0.7839	-3.9759
λ_1	-1.6094	0.5000	-1.6094	0.5000	-1.6094	-1.6094
λ_2	-0.9163	-0.9163	-0.9163	-0.9163	-0.9163	-0.9163
λ_3	-1.2040	-1.2040	-1.2040	-1.2040	-1.2040	-1.2040
λ_{12}	0.0100	0.0100	0.0100	0.0100	0.0100	0.0100
λ_{13}	3.2834	3.2834	0.0500	0.0500	0.0150	3.2834
λ_{23}	2.3434	2.3434	2.3434	2.3434	0.0200	2.3434
λ_{123}	0.0300	0.0300	0.0300	0.0300	0.0300	1.9738

Table 8 displays the unsaturated models fitted to the 100,000 datasets generated using each model in Table 6. The best-fitted models satisfy the criterion functions Deviance, AIC, BIC, and SAVE.

The number or proportion of times a parameter appearing or not-appearing in the best-fitted models is a measure of correct detection. For the data generating six models in Table 6, the values of λ_{12} are identical, smallest, and close to zero. Hence, smaller the number or proportion of times λ_{12} appearing in the best-fitted models is better and larger the number or proportion of times λ_{12} not-appearing in the best fitted models is better, are two equivalent measures of correct detection. Table 9 provides the comparison between Deviance Statistic/AIC/BIC and SAVE in terms of the number of times λ_{12} does not appear in the best fitted models of three groups ($g = 1, 2, 3$ in Table 8) to 100,000 datasets generated by the six models ($M_i, i = 1, \ldots, 6$, in Table 7). Table 9 demonstrates that the number of times λ_{12} does not appear in the best fitted models using the criterion function SAVE, is greater than or equal to the corresponding number which is the common value of the criterion functions Deviance, AIC, and BIC. In other words, the new criterion function SAVE makes the correct detection more frequently than the three popular criterion functions: Deviance, AIC, and BIC.

Table 7: The cell probabilities p_w of the six data generating models

w	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$
(0,0,0)	0.0850	0.0132	0.2803	0.1244	0.4566	0.0188
(0,0,1)	0.0255	0.0040	0.0841	0.0373	0.1370	0.0056
(0,1,0)	0.0340	0.0053	0.1121	0.0498	0.1826	0.0075
(0,1,1)	0.1062	0.0165	0.3504	0.1555	0.0559	0.0235
(1,0,0)	0.0170	0.0218	0.0561	0.2051	0.0913	0.0038
(1,0,1)	0.1360	0.1743	0.0177	0.0647	0.0278	0.0300
(1,1,0)	0.0069	0.0088	0.0227	0.0829	0.0369	0.0015
(1,1,1)	0.5895	0.7561	0.0767	0.2825	0.0118	0.9094

For the data generating six models $M1, \ldots, M6$, the values of λ_{13} are equal and largest among the association parameters for $M1$, $M2$, and $M6$. Therefore, larger the number of times λ_{13} appearing in the best fitted models is better. Table 10 presents the comparison between Deviance Statistic/AIC/BIC and SAVE with respect to the number of times λ_{13} appears in the best fitted models of three groups($g = 1, 2, 3$ in Table 8) to 100,000 datasets generated by $M1$, $M2$, and $M6$. The SAVE makes the correct detection more frequently than Deviance/AIC/BIC for the datasets generated by $M1$ in the group $g = 1$ and for the datasets generated by $M6$ in the group $g = 2$. The performances are equal for the other cases in Table 10. The Deviance/AIC/BIC makes the correct detection more frequently than SAVE for the datasets generated by $M2$ in the group $g = 1$. Overall, SAVE performs better than Deviance/AIC/BIC.

7. Concluding Remarks

We constructed the new metric SAVE from the standard and orthogonal extensions of the unsaturated models. The construction process is simple and meaningful. We made the comparison of the metric SAVE with its competitors Deviance, AIC, and BIC. The SAVE

Table 8: The fitted models for $k = 1,\ 2,$ and 3

g	h	The fitted	The common λ	The other λ	The λ
		Model $g.h$	parameters present	parameters present	parameters absent
1	1	1.1	$\lambda,\ \lambda_1,\ \lambda_2,\ \lambda_3$	λ_{123}	$\lambda_{12},\ \lambda_{13},\ \lambda_{23}$
	2	1.2	$\lambda,\ \lambda_1,\ \lambda_2,\ \lambda_3$	λ_{12}	$\lambda_{123},\ \lambda_{13},\ \lambda_{23}$
	3	1.3	$\lambda,\ \lambda_1,\ \lambda_2,\ \lambda_3$	λ_{13}	$\lambda_{123},\ \lambda_{12},\ \lambda_{23}$
	4	1.4	$\lambda,\ \lambda_1,\ \lambda_2,\ \lambda_3$	λ_{23}	$\lambda_{123},\ \lambda_{13},\ \lambda_{12}$
2	1	2.1	$\lambda,\ \lambda_1,\ \lambda_2,\ \lambda_3$	$\lambda_{123}, \lambda_{23}$	$\lambda_{12},\ \lambda_{13}$
	2	2.2	$\lambda,\ \lambda_1,\ \lambda_2,\ \lambda_3$	$\lambda_{123}, \lambda_{13}$	$\lambda_{12},\ \lambda_{23}$
	3	2.3	$\lambda,\ \lambda_1,\ \lambda_2,\ \lambda_3$	$\lambda_{123}, \lambda_{12}$	$\lambda_{13},\ \lambda_{23}$
	4	2.4	$\lambda,\ \lambda_1,\ \lambda_2,\ \lambda_3$	$\lambda_{12}, \lambda_{13}$	$\lambda_{123},\ \lambda_{23}$
	5	2.5	$\lambda,\ \lambda_1,\ \lambda_2,\ \lambda_3$	$\lambda_{12}, \lambda_{23}$	$\lambda_{123},\ \lambda_{13}$
	6	2.6	$\lambda,\ \lambda_1,\ \lambda_2,\ \lambda_3$	$\lambda_{13}, \lambda_{23}$	$\lambda_{123},\ \lambda_{12}$
3	1	3.1	$\lambda,\ \lambda_1,\ \lambda_2,\ \lambda_3$	$\lambda_{123},\ \lambda_{13},\ \lambda_{23}$	λ_{12}
	2	3.2	$\lambda,\ \lambda_1,\ \lambda_2,\ \lambda_3$	$\lambda_{123},\ \lambda_{12},\ \lambda_{23}$	λ_{13}
	3	3.3	$\lambda,\ \lambda_1,\ \lambda_2,\ \lambda_3$	$\lambda_{123},\ \lambda_{13},\ \lambda_{12}$	λ_{23}
	4	3.4	$\lambda,\ \lambda_1,\ \lambda_2,\ \lambda_3$	$\lambda_{12},\ \lambda_{13},\ \lambda_{23}$	λ_{123}

Table 9: The number of best fitted unsaturated models without λ_{12}

g	Data	Deviance/	SAVE	g	Data	Deviance/	SAVE
	generated by	AIC/BIC			generated by	AIC/BIC	
1	$M1$	100,000	100,000	2	$M1$	100,000	100,000
	$M2$	100,000	100,000		$M2$	100,000	100,000
	$M3$	100,000	100,000		$M3$	97,211	98,658
	$M4$	100,000	100,000		$M4$	94,223	96,230
	$M5$	100,000	100,000		$M5$	56,993	62,540
	$M6$	100,000	100,000		$M6$	99,987	100,000
3	$M1$	51,143	62,773				
	$M2$	55,043	58,533				
	$M3$	46,476	55,344				
	$M4$	40,095	49,677				
	$M5$	27,104	32,107				
	$M6$	100,000	100,000				

Table 10: The number of best fitted unsaturated models including λ_{13}

g	Data	Deviance/	SAVE	g	Data	Deviance/	SAVE
	generated by	AIC/BIC			generated by	AIC/BIC	
1	$M1$	0	36,684	2	$M1$	100,000	100,000
	$M2$	100,000	29, 153		$M2$	100,000	100,000
	$M6$	100,000	100,000		$M6$	99,987	100,000
3	$M1$	100,000	100,000				
	$M2$	100,000	100,000				
	$M6$	100,000	100,000				

performed as well as or even better than Deviance, AIC, and BIC. We compared them in terms of the correct identification of parameters of unsaturated log-linear models.

Acknowledgements

We thank a reviewer for the careful reading of an earlier version of our paper, the thoughtful comments, and the editors for their valuable suggestions.

References

Agresti, A. (2013). *Categorical Data Analysis.* Wiley, New York, Third Edition.

Csiszár, I. (1975). i-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, **3(1),** 146-158.

Gokhale, D. V. and Kullback, S. (1978). *The Information in Contingency Tables.* Marcel Dekker, New York.

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Annals of Statistics,* **12(3),** 971-988.

Klimova, A., Rudas, T., and Dobra, A. (2012). Relational models for contingency tables. *Journal of Multivariate Analysis,* **104,** 159-173.

Klimova, A. and Rudas, T. (2016). On the closure of relational models. *Journal of Multivariate Analysis,* **143** 440-452.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling.* Springer, New York.

Kullback, S., Keegel, J., and Kullback, J. (2013). *Topics in Statistical Information Theory.* Springer, New York.

Kullback, S. (1959). *Information Theory and Statistics.* Wiley, New York. (Dover edition, 1997.)

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Rudas, T. (2018). *Lectures on Categorical Data Analysis.* Springer, New York.