Geometric Descriptions for Hierarchical Log-Linear Models

C. S. Hong, H. J. Choi, and M. G. Oh *

Abstract

We suggest a geometrical method to describe the test results of the hierarchical log-linear models. In order to represent the relationship of a set of hierarchical log-linear models, some shapes of right-angled triangles, tetrahedrons, and polyhedrons are allowed to draw. These geometric descriptions can be applied to the model selection method for given hierarchical models.

Key Words: Likelihoods ratio statistics; Model selection; Polyhedron plot; Right-angled plot; triangle plot.

1 INTRODUCTION

Fienberg and Gilbert (1970) developed the geometry of measure for 2×2 contingency tables, which allows for the visualization of the properties of the various models in terms of the loci of the tetrahedron. For the interpretation of the relationships among the factors in a given log-linear model, a simple and undirected graph has been explored by Darroch, Lauritzen and Speed (1980), Goodman (1971, 1973) and others. By this graph, they also defined the class of graphical models. Edwards and Kreiner (1983) gave an overview of the use of graphical models. Especially, they proposed the strategies for model selection based on this class of models. These previous works have focused on the interactions of the models by the notion of the conditional independence. However, they have not considered the measured degrees of associations and the magnitudes of the goodness of fit statistics.

In this article, we propose an alternative geometric description to represent the relationships of the generalized likelihood ratio statistics, G^2, corresponding to given hierarchical log-linear models. If the values of the likelihood ratio statistics were regarded as the squared norms of the vectors, we could evaluate visually the relationship between two hierarchical models through a right-angled triangle. A tetrahedron can describe the relationship of three models. Also, one can

*C. S. Hong is Professor, Department of Statistic, SungKyunKwan University, Seoul, 110-745, Korea(E-mail: cshong@skku.ac.kr). H. J. Choi is Full-time Lecturer, Department of Applied Information Statistics, Kyonggi University, Suwon, 442-760, Korea(E-mail: hjchoi@stat.kyonggi.ac.kr). M. G. Oh is Lecturer, Department of Statistic, SungKyunKwan University, Seoul, 110-745, Korea.

consider several right-angled triangles sequentially to explain the relationships among more than four hierarchical models, which will be called the polyhedron plot. We can identify visually the goodness-of-fit test results among hierarchical log-linear models for a given contingency table by using these geometries. These geometric descriptions are discussed in Section 2, 3, and 4, respectively, and an illustrative example for finding the best model in a given hierarchical structure is presented by using this suggested plot.

2 RIGHT-ANGLED TRIANGLE PLOT

In this paper our attention is restricted to the hierarchical structure for the complete contingency tables. Consider two log-linear models (a) and (b), where model (a) is the special case of model (b). Define $G^2(a)$ and $G^2(b)$ as the generalized likelihood ratio (GLR) test statistics for model (a) and (b), respectively. Under this hierarchical structure, it satisfies that $G^2(a) \geq G^2(b)$, and we have an equation

$$\begin{aligned} G^2(a) &= [G^2(a) - G^2(b)] + G^2(b) \\ &\equiv G^2(a|b) + G^2(b). \end{aligned} \tag{1}$$

$G^2(a)$ is partitioned into two parts, where the first component of the right-hand-side in equation (1) can be regarded as the variation explained by model (a) given model (b). Hence as Christensen (1990) noted, equation (1) can be rewritten such that

$$SST = SSR + SSE.$$

In linear regression analysis, this equation can be expressed as a right-angled triangle in terms of an orthogonal projection. In particular, Schey (1993) proposed geometric methods to investigate the relationship between the norms of $SSR(x_2)$ and $SSR(x_2|x_1)$. The three components in equation (1) are regarded as the squared norms of the vectors V_a, $V_{a|b}$, and V_b, where we define

$$G^2(a) = |V_a|^2, G^2(a|b) = |V_{a|b}|^2, \quad \text{and} \quad G^2(b) = |V_b|^2.$$

With these vectors, one could draw a right-angled triangle as the one in Figure 1. The angle θ_1 between two vectors V_a and V_b is defined as

$$cos\theta_1 = \frac{|V_b|}{|V_a|}. \tag{2}$$

The triangle in Figure 1 contains all the information concerning the goodness of fits for model (a) and (b), and we can evaluate the information simply by comparing the lengths of the vectors V_a, V_b, and $V_{a|b}$. A dotted line means that the G^2 value of the corresponding model is so high that its p-value is less than a

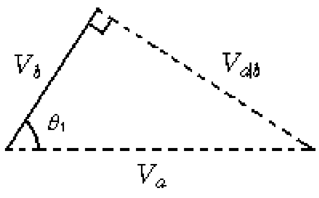

Figure 1: Right-angled triangle plot

given significant level, say 5%. A solid line means that the corresponding model is not significant but well-fitted. This right-angled triangle tells us that model (a) does not explain the given data but model (b) does. Moreover, there exists a significant difference between model (a) and (b) by observing $V_{a|b}$.

One can also find that as the magnitude of the vector V_b gets smaller for a given fixed model (a), both the length of the vector $V_{a|b}$ and the value of the angle θ_1 are increasing. If θ_1 has a large value, it means that model (b) fits the data better than model (a) does, and that the variation due to the difference between model (a) and (b), $G^2(a|b)$, is significant. Hence, we can identify not only the goodness-of-fits for each log-linear models but also the difference between goodness-of-fits of model (a) and (b) via the right-angled triangle.

Suppose we regard model (a) as the smallest complete independent model. The coefficient of determination for model (b) defined by Christensen (1990) can be expressed with the angle θ_1 of this triangle such that

$$\begin{aligned} R_b^2 &= \frac{G^2(a) - G^2(b)}{G^2(a)} \\ &= \frac{|V_{a|b}|^2}{|V_a|^2} \quad = \quad 1 - \frac{|V_b|^2}{|V_a|^2} \quad = \quad 1 - cos^2\theta_1 \\ &= sin^2\theta_1. \end{aligned} \tag{3}$$

Therefore we find that the coefficient of determination from model (b) is the squared sine function of θ_1.

3 TETRAHEDRON PLOT

We can add model (c) satisfying $G^2(b) \geq G^2(c)$ in the above hierarchical structure. Now let us examine the relationships among model (a), (b), and (c). With equation (1), we also obtain the followings two equations:

$$G^2(b) = G^2(b|c) + G^2(c), \tag{4}$$
$$G^2(a) = G^2(a|c) + G^2(c). \tag{5}$$

The equations in (4) and (5) bring up the image of two right-angled triangles in Figure 2, respectively, where we denote

$$G^2(b|c) = |V_{b|c}|^2, \quad G^2(a|c) = |V_{a|c}|^2, \quad \text{and} \quad G^2(c) = |V_c|^2.$$

The first and second triangles show the relationships between models (b) and (c), and models (a) and (c) which can be described via the vectors $V_{b|c}$ and $V_{a|c}$, respectively. Readers can interpret the relationships with the similar arguments as we did in Figure 1.

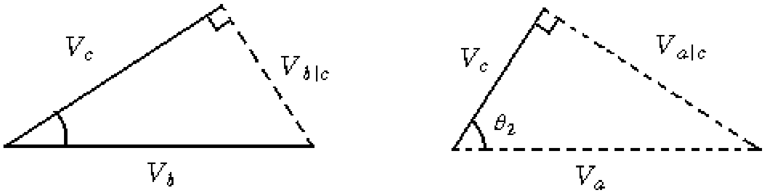

Figure 2: Right-angled triangle plots

From Figure 1 and Figure 2, model (a) does not fit the data but both model (b) and (c) explain the data well. And the difference between model (a) and (b), and model (b) and (c) are statistically significant. One might say that model (c) is better than model (b) which is also better than model (a).

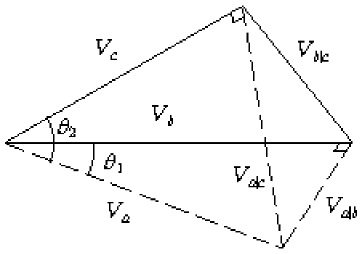

Figure 3: Tetrahedron plot

Three equations in (1), (4), and (5) can be interpreted as a tetrahedron like the one in Figure 3 in the three dimensional space. In the same context, we can induce the geometrical interpretation on the shape in Figure 3.

We will explore these plots with certain hierarchical log-linear models which may explain the well-known data in Table 1. This data was studied earlier by Ries and Smith (1963), and many others. This is a cross classification of the degree of softness of the water they used (soft, medium, hard) (var. 1), brand preferences (X or M) (var. 4), whether they had used the brand previously (yes, no) (var. 2), and the temperature of the laundry water used (high, low) (var.

3). For this $3 \times 2 \times 2 \times 2$ contingency table, we consider the following hierarchical structure listed in Table 2.

Figure 1, Figure 2, and Figure 3 show an exact description of the relationships of model (a), (b), and (c) of the above data. We obtain the following by using equation (2) and (3),

$$\theta_1 = cos^{-1}\sqrt{|V_b|\Big/|V_a|} = 43.83, \qquad \theta_2 = cos^{-1}\sqrt{|V_c|\Big/|V_a|} = 49.66.$$

and

$$R_b^2 = sin^2(\theta_1) = 0.48, \qquad R_c^2 = sin^2(\theta_2) = 0.58.$$

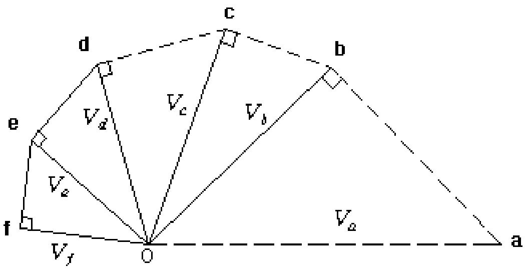

Figure 4: Polyhedron plot

4 POLYHEDRON PLOT

From this three dimensional tetrahedron plot, it is not easy to interpret the relationships of three log-linear models simultaneously. Moreover, it is impossible to draw and evaluate the relationships among four hierarchical log-linear models on three dimensional space. Nonetheless, we find that the relationships among more than four log-linear models could be explored sequentially.

Consider hierarchical models (a), (b), (c), ... satisfying $G^2(a) \geq G^2(b) \geq G^2(c) \geq \cdots$. As we did in Figure 1 and the first plot in Figure 2, model (a) can be compared with model (b), which can be compared with (c), and so on. We might consider a plane geometry as in Figure 4 which consists of several right-angled triangles which represent the relationships between models (a) and (b), models (b) and (c), and models (c) and (d), etc. Since this figure looks like polyhedrons, we call it a polyhedron plot.

Bishop et al. (1975) and Fienberg (1983) have considered the six hierarchical log-linear models for the data in Table 1, so we consider the same hierarchy. The first three models are already in Table 2, and the other three are listed in Table 3. The polyhedron plot which puts together five right-angled triangles for the six log-linear models are presented in Figure 4, where this plot describes all goodness-of-fits test results of the models and the relationships among the models in Table 2 and Table 3.

It is very clear that this plot helps us to explore the given hierarchical model structure. Also, we could choose a well-fitted model based on this plot. The forward and the backward selection methods would be used with the one in Figure 4. With the polyhedron plot in Figure 4, the models (c), (d), and (e) all are well-fitted. Since the difference between (c) and (d) is significant, and the difference between (d) and (e) is not, we might conclude that the model (d) : [13][24][34] is the best one among this hierarchical structure. Therefore, this geometric description can be applied to the model selection method for given hierarchical models.

Table 1: Detergent brand preference data

A (Water softness)	B (Brand preference)	C (Previous use of Detergent Brand M)			
		Yes		No	
		D (Temperature)			
		High	Low	High	Low
Soft	X	19	57	29	63
	M	29	49	27	53
Medium	X	23	47	33	66
	M	47	55	23	50
Hard	X	24	37	42	68
	M	43	52	30	42

Table 2: Results of the GLR tests

ID	Model	d.f.	G^2	Difference	d.f.	G^2
(a)	[1][2][3][4]	18	42.93^*			
(b)	[1][3][24]	17	22.35	(a) and (b)	1	20.58^*
(c)	[1][24][34]	16	17.99	(b) and (c)	1	4.36^*

$*$ indicates that the p-value of the statistic is less than 5% significant level.

Table 3: Further results of the GLR tests

ID	Model	d.f.	G^2	Difference	d.f.	G^2
(d)	[24][34][13]	14	11.89	(c) and (d)	2	6.10^*
(e)	[13][234]	12	8.41	(d) and (e)	2	3.48
(f)	[123][234]	8	5.66	(e) and (f)	4	2.75

$*$ indicates that the p-value of the statistic is less than 5% significant level.

5 CONCLUSION

We develop some graphical descriptions to evaluate visually the relationships between given hierarchical log-linear models. In order to explain the relationship of two hierarchical log-linear models, a right-angled triangle plot is proposed. We make use of the shape of a tetrahedron for three log-linear models. And for more than three log-linear models, polyhedron plot is used. These geometric description plots help us to evaluate a certain hierarchical model structure, and can be applied to the model selection method to choose the best model among given hierarchical log-linear models.

Bibliography

[1] Bishop, Y. M. M, Fienberg, S. E., and Holland P. W. (1975), *Discrete Multivariate Analysis*, MIT Press.

[2] Christensen, R. (1990), *Log-Linear Model*, Springer-Velag.

[3] Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980), Markov-fields and log-linear models for contingency tables, *Annals of Statistics*, Vol. 8, 522-39.

[4] Edwards, D., and Kreiner, S. (1983), The analysis of contingency tables by graphical models, *Biometrika*, Vol. 70, 553-565.

[5] Fienberg, S. E. (1983), *The Analysis of Cross-Classified Categorical Data*, MIT Press.

[6] Fienberg, S. E., and Gilbert, J. P. (1970), The Geometry of a Two by Two Contingency Table, *Journal of the American Statistical Association*, Vol. 65, 694-701.

[7] Goodman, L. A. (1971), Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional tables, *Journal of the American Statistical Association*, Vol. 66, 339-344.

[8] Goodman, L. A. (1973), The analysis of contingency tables when some variables are posterior to others : A modified path analysis approach, *Biometrika*, Vol. 60, 179-192.

[9] Ries, P. N., and Smith, H. (1963), The use of chi-square for preference testing in multidimensional problems, *Chemical Engineering Progress*, Vol. 59, 39-43.

[10] Schey, H. M. (1993), The relationship Between the magnitudes of and : A Geometric Description, *The American Statistician*, Vol. 47, No. 1, 26-30.