The Korean Communications in Statistics
Vol. 13 No. 3, 2006, pp.755-764

Graphical Methods for Hierarchical Log-Linear Models

Chong Sun Hong[1] and Ui Ki Lee[2]

Abstract

Most graphical methods for categorical data can describe the structure of data and represent a measure of association among categorical variables. Among them the polyhedron plot represents sequential relationships among hierarchical log-linear models for a multidimensional contingency table. This kind of plot could be explored to describe the differences among sequential models. In this paper we suggest graphical methods, containing all the information, that reflect the relationship among all log-linear models in a certain hierarchical structure. We use the ideas of a correlation diagram.

Keywords : Goodness of fit; hierarchical model; likelihood ratio statistics; log-linear model; measure of association; odds ratio.

1. Introduction

There exist many graphical methods which describe the structure of categorical data. The probabilities or frequencies of several categories for one categorical variable might be expressed using a bar chart, a pie chart, and a star chart. Fienberg (1975) proposed the four-fold circular display which represents a 2×2 contingency table. There are more graphical methods for two dimensional cate -gorical data: the block chart, the mosaic plot (Hartigan and Kleiner 1981, 1984; Friendly 1992, 1994), the association plot (Cohen 1980; Friendly 1991), the grouped bar graph (Tufte 1983), the grouped dot plot and framed rectangle chart (Cleveland and McGill 1984), the trellis display (Becker, Cleveland and Shyu 1996), and the diamond graph (Li, Buechner, Tarwater and Munoz 2003), etc.

For $I\times J\times K$ three dimensional contingency tables, the separated $I\times J$ contingency tables can be analyzed by using a mosaic plot with respect to each category of the third variable. The contingency table of four dimensions or more can be explored by extending the mosaic plot (see Hartigan and Kleiner (1984), Friendly (1994b), and Wilkinson (1999) for more detail).

1) Professor, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, Korea. Correspondence : cshong@skku.ac.kr
2) Research Fellow, Research Institute of Applied Statistics, Sungkyunkwan University, Seoul, 110-745, Korea.

The mosaic plot and the four-fold circular display can also explain measures of association among categorical variables. Fienberg (1968), and Fienberg and Gilbert (1970) proposed a method to represent the association measure geometrically in terms of loci within a tetrahedron for a 2×2 contingency table. Tukey (1977) proposed the two-way plot which represents the goodness of fit for a two dimensional contingency table. Darroch, Lauritzen, and Speed (1980) proposed graphical models which could describe an independent model and a conditional independent model for multidimensional contingency tables. There exist two other methods which are based on odds ratios and their confidence intervals for 2×2 contingency tables: the contour plot (Doi, Nakamura and Yamamoto 2001; Yamamoto and Doi 2001) and the raindrop plot (Barrowman and Myers 2002, 2003).

In addition to these methods, Hong, Choi, and Oh (1999) proposed graphical methods to express the relationship among the goodness of fits of hierarchical log-linear models for a multidimensional contingency table. Their methods can describe information for any pair and sequential pairs of log-linear models in the hierarchy by constructing the right-angled triangle plot and the polyhedron plot. In this paper, we suggest an alternative method which contains information about each log-linear model and all possible pairs of models in a certain hierarchical structure. This method named by G^2 plot could reflect the relationship among all log-linear models in a hierarchy by using ideas of the correlation diagram introduced by Trosset (2005). The correlation diagram can be used to visualize a correlation coefficient matrix on a unit circle. The vectors of a correlation diagram denote the corresponding variables. The angles between the vectors have information about the correlation between the variables. The G^2 plot would use values of the generalized likelihood ratio statistics for hierarchical log-linear models rather than that of the correlation coefficients.

Section 2 provides relevant summaries of the polyhedron plot and the correlation diagram introduced by Hong et al. (1999) and Trosset (2005), respec -tively. In section 3, we propose the G^2 plot which is an alternative to represent relationships among all log-linear models in a hierarchical structure. Another G^2 plot and an extended G^2 plot for multidimensional categorical data are discussed in section 4. Section 5 concludes this paper.

2. Preliminaries

2.1 Polyhedron Plot

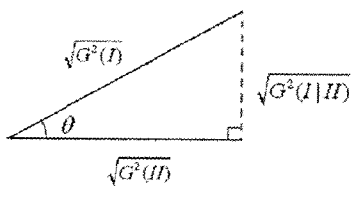
Christensen (1990, pp. 65, 90) considered four log-linear models for the three dimensional categorical data of 3182 people without cardiovascular disease. The data was cross-classified by three factors: A (personality), B (cholesterol), and C (diastolic blood pressure). The results of this data analysis are listed in <Table 2.1>. For these hierarchical log-linear models, Model Ⅰ is nested by Model Ⅱ which is nested by Model Ⅲ which is also nested by Model Ⅳ. Define $G^2(\mathrm{I})$, $G^2(\mathrm{II})$, $G^2(\mathrm{III})$, and $G^2(\mathrm{IV})$ as the generalized likelihood ratio test statistics for Model Ⅰ, Ⅱ, Ⅲ, and Ⅳ, respectively. Then this hierarchical structure satisfies the relationship such that $G^2(\mathrm{I}) \geq G^2(\mathrm{II}) \geq G^2(\mathrm{III}) \geq G^2(\mathrm{IV})$. For Model Ⅰ and Ⅱ, for example, we have the following equation:

$$\begin{aligned} G^2(\mathrm{I}) &= [G^2(\mathrm{I}) - G^2(\mathrm{II})] + G^2(\mathrm{II}) \\ &\equiv G^2(\mathrm{I}\,|\,\mathrm{II}) + G^2(\mathrm{II}). \end{aligned}$$

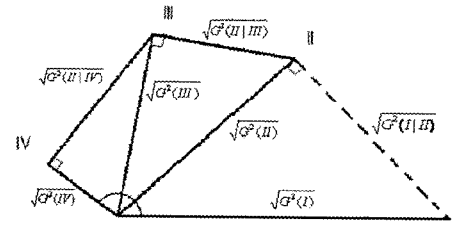
For any pair models among Model Ⅰ to Ⅳ, the above relationship is also established. Hong et al. (1999) proposed the right-angled triangle plot in <Figure 2.1>. This plot represents the relationship between Models Ⅰ and Ⅱ which is measured by the angle θ satisfying

$$\cos(\theta) = \sqrt{G^2(\mathrm{II})/G^2(\mathrm{I})}. \tag{2.1}$$

This triangle in <Figure 2.1> contains all informations concerning the goodness of fits for model Ⅰ and Ⅱ, and we can evaluate the information simply by observing and comparing the lengths of $\sqrt{G^2(\mathrm{I})}$, $\sqrt{G^2(\mathrm{II})}$, and $\sqrt{G^2(\mathrm{I}\,|\,\mathrm{II})}$. The lengths are an indication of whether the corresponding test statistic is significant. That is, a dotted line represents that the G^2 value of the corresponding model is so high that its p-value is less than a given significant level, say 5%, and a solid line means that the corresponding model is not significant but well-fitted. This right-angled triangle tells us that both model Ⅰ and Ⅱ explain the given data, and there exists a significant difference between model Ⅰ and Ⅱ by observing $\sqrt{G^2(\mathrm{I}\,|\,\mathrm{II})}$.



<Figure 2.1> Right-angled triangle Plot



<Figure 2.2> Polyhedron Plot

Hong et al. (1999) also proposed the polyhedron plot which could describe the sequential relationships among more than two log-linear models under a certain

hierarchical structure. The polyhedron plot in <Figure 2.2> consists of three right-angled triangles. This plot represents the relationships between hierarchical log-linear model Ⅰ and Ⅱ, model Ⅱ and Ⅲ, and model Ⅲ and Ⅳ, sequentially. With this plot, one can find that all of these models are well-fitted at 5% significant level. The difference between Ⅰ and Ⅱ is significant, but the differences between Ⅱ and Ⅲ, and between Ⅲ and Ⅳ are not. Christensen (1990, pp. 90) mentioned that the model Ⅱ:[AB][C] is the smallest model that adequately fits the data among this hierarchical structure. Hence this polyhedron plot might play an important role in selecting the best fitted log-linear model among hierarchical models by using the forward and the backward selection methods.

<Table 2.1> Results for the three dimensional data

Model	d.f.	G^2	p-value
Model Ⅰ: [A][B][C]	4	8.72	.067
Model Ⅱ: [AB][C]	3	4.60	.207
Model Ⅲ: [AB][AC]	2	2.06	.358
Model Ⅳ: [AB][AC][BC]	1	0.61	.434

2.2 Correlation Diagram

Corsten and Gabriel (1976) introduced the h plot which is a standard method for visualizing product-moment correlation coefficients as angles, which is a fragment of the biplot of Gabriel (1971). Using ideas from the h plot, Trosset (2005) proposed a correlation diagram which could be used to visualize a $p \times p$ matrix of correlation coefficients, $\{R\} = (r_{jk})$. The corresponding variables are visualized as the vectors of the correlation diagram. The angles between the vectors convey information about the correlation between the variables. The correlation diagram follows the lead of Corsten and Gabriel (1976), but all vectors in the diagram of Trosset are set to unit length. We seek p points on the unit circle, identified with scalar angles $\theta_1, \theta_2, \cdots, \theta_p$ to construct a correlation diagram. The angle between vectors j and k is $\theta_j - \theta_k$. One finds $\Theta = (\theta_1, \theta_2, \cdots, \theta_p)$ for which

$$r_{jk} \approx \cos(\theta_j - \theta_k) \text{ for all } j \text{ and } k. \tag{2.2}$$

Thus, we solve an unconstrained optimization problem of the form

$$\min 2 \sum_{j<k} [\, r_{jk} - \cos(\theta_j - \theta_k) \,]^2. \tag{2.3}$$

Trosset (2005) used the S-Plus function, *nlminb*, a quasi-Newton algorithm developed by Gay (1983, 1984) to minimize the optimization problem in (2.3).

3. G^2 Plot

With similar arguments for defining hierarchical log-linear models in section 2.1, consider p log-linear models whose generalized likelihood ratio test statistics are defined as $G^2(i)$, $i=1, \cdots, p$. Then the model j is nested by the model k, and $G^2(j) \geq G^2(k)$ for $j<k$. The relationship between two hierarchical log-linear models j and k could be measured by the difference of scalar angles $\theta_j - \theta_k$, as we explained in (2.1). We follow the ideas of Trosset's correlation diagram and restrict our attention to the differences among the generalized likelihood ratio statistics for hierarchical log-linear models rather than the correlation coefficients discussed in section 2.2. For visualizing information for hierarchical log-linear models, we apply both equations (2.1) and (2.2) together. Therefore we can establish the following unconstrained optimization problem to seek p points on the circle, identified with scalar angles $\theta_1, \cdots, \theta_p$:

$$\min \sum_{j<k} \left[\sqrt{G^2(k)/G^2(j)} - \cos(\theta_k - \theta_j) \right]^2 \tag{3.1}$$

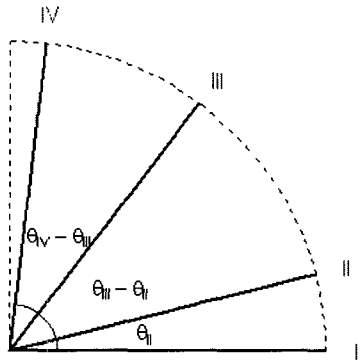
The optimization problem in (3.1) can also be minimized by *nlminb*. Note that since a possible maximum angle, $\sec \sqrt{\min G^2(i) \,/\, \max G^2(i)}$, is less than $\pi/2$, this plot is supposed to represent on a quarter-circle. To remove rotational indeterminacy, we set $\theta_1 = 0$. We propose this plot as the G^2 plot. Since the function, *nlminb*, is provided in *R*, our efforts for constructing the G^2 plot are made in *R*.

For example, consider the four log-linear models for three dimensional categorical data explained in section 2.1. Then, we find three scalar angles $\theta_{\text{II}}, \theta_{\text{III}}, \theta_{\text{IV}}$ $(\theta_{\text{I}} = 0)$ on the unit quarter-circle. With values of generalized likelihood ratio statistics for four hierarchical log-linear models in <Table 2.1>, the minimization process (3.1) results in the unique result. The objective value which is the local minimum value for the process (3.1) is equal to 0.1943. The G^2 plot in <Figure 3.1> displays the relationship among four log-linear models for this data.

One property of the polyhedron plot makes clear the differences between two sequential hierarchical log-linear models shown in <Figure 2.2>. However, the G^2 plot in <Figure 3.1> displays the total relationships of all possible pairs of log-linear models in a certain hierarchical structure. With a similar arguments for the polyhedron plot, one could express goodness of fits for corresponding log-linear models with dotted and solid lines. Since all of these models fit the data at 5% significant level, all vectors in <Figure 3.1> are expressed as solid lines. Since there is close relation between model I and II, and there exists a

significant difference between model II and III, we might say that model III:[AB][AC] is the best model among four models. Christensen (1990, pp. 90, 152) insisted that this model is the best based on Akaike's information criterion. This conclusion is the same as we obtain from this G^2 plot.

Note that the set of scalar angles containing the relationship among hierarchical log-linear models is obtained as $\Theta = (0, 18.9883, 38.9450, 85.8784)$ for the G^2 plot in <Figure 3.1>.



<Figure 3.1> G^2 Plot

4. Multidimensional Categorical Data

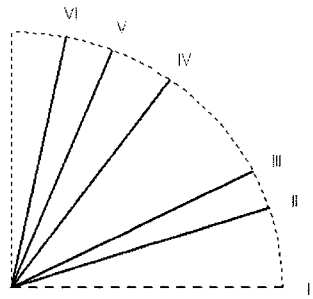
Bishop et al. (1975, pp. 142) and Fienberg (1983, pp. 71) have considered the six hierarchical log-linear models for four dimensional data on detergent preferences (Ries and Smith 1963). We consider the same hierarchy. These hierarchical models and the goodness of fits of these models are listed in <Table 4.1>.

<Table 4.1> Results of the Four Dimensional Data

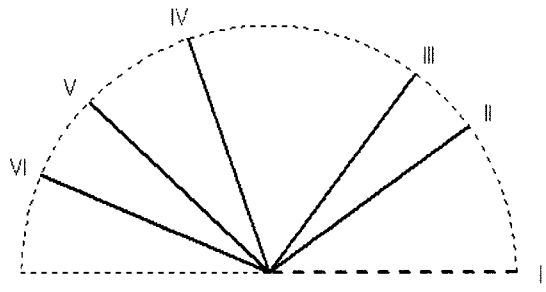
Model	d.f	G^2	p-value
Model I : [1][2][3][4]	18	42.93	0.0000
Model II : [1][3][24]	17	22.35	0.1717
Model III : [1][24][34]	16	17.99	0.347
Model IV : [13][24][34]	14	11.89	0.6154
Model V : [13][234]	12	8.41	0.7526
Model VI : [123][234]	8	5.66	0.6857

The relationship among the six log-linear models for this data is described with the G^2 plot in <Figure 4.1>. Minimization of the optimization problem in (3.1) results in the smallest objective value, 0.2696. The set of scalar angles is obtained as $\Theta = (0.00, 18.03, 26.80, 54.54, 68.23, 78.44)$ for the G^2 plot in <Figure 4.1>.

Note that the corresponding log-linear models are visualized as the vectors of the G^2 plot. The angles between the vectors convey the relationships among log-linear models. As number of hierarchical log-linear models increases, number of vectors in the G^2 plot is also increasing, but the angles between the vectors are decreasing. In this case, the G^2 plot may be extended to draw on the half-circle rather than on a quarter-circle and the angles between the vectors are doubled as in <Figure 4.2>.



<Figure 4.1> G^2 Plot



<Figure 4.2> Extended G^2 Plot

The log-linear models II to VI fit the data, so that the corresponding vectors in <Figure 4.1> and <Figure 4.2> are expressed as solid lines. Only the first vector corresponding to model I is drawn as a dotted line. There is close relation between model II and III, and there exist another close relationship among model IV, V, and VI. Since we could find a significant difference between model III and VI, we might say that model IV:[13][24][34] seems adequate. Bishop et al. (1975, pp. 166-167) and Fienberg (1983, pp. 76-77) mentioned that this model IV is the best among these hierarchy, which is also the same conclusion as we obtain from this G^2 plot.

5. Conclusion

The polyhedron plot introduced by Hong et al. (1999) could be expressed information of sequential pairs of hierarchical log-linear models. The G^2 plot proposed in this paper could be explored to represent the relationships among hierarchical log-linear models considering all possible pairs of log-linear models for multidimensional categorical data. This plot describes the goodness of fits of log-linear models with dotted and solid lines, since this plot is constructed based on all kinds of differences of the generalized likelihood ratio statistics. As we discussed in section 3 and section 4, the G^2 plot representing overall relationships among hierarchical log-linear models might be better to select the best model than

any other geometrical method including the polyhedron plot. Therefore, we might say that the G^2 plot could be applied to use for a best log-linear model selection method.

As the number of log-linear models in a hierarchy increases, the number of vectors which represent corresponding log-linear models is increasing. On the other hand, the differences of the angles, $\theta_{i+1} - \theta_i$, are decreasing. In this case, the G^2 plot composed on a quarter-circle might be shown to be complex. we may extend the region of the G^2 plot from $\pi/2$ to π or 2π, which means that the G^2 plot could be a half circle or a circle.

References

[1] Barrowman, N.J. and Myers, R.A. (2000). Still more spawner-recruitment curves: the hockey stick and its generalizations. *Canadian Journal of Fisheries and Aquatic Sciences*, Vol. 57, 665-676.

[2] Barrowman, N.J. and Myers, R.A. (2003). Raindrop plots: a new way to display collections of likelihoods and distributions. *American Statistician*, Vol. 57, 268-274.

[3] Becker, R.A., Cleveland, W.S., and Shyu, M.J. (1996). The visual design and control of trellis display. *Journal of Computational and Graphical Statistics*. Vol. 6(2), 123-155.

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

[5] Christensen, R. (1990). *Log-Linear Models and Logistic Regression*, Springer, New York.

[6] Cleveland, W.S. and McGill, R. (1984). Graphical perception: theory, experimen -tation, and application to the development of graphical methods. *Journal of American Statistical Association,* Vol. 79(387), 531-554.

[7] Cohen, A. (1980). On the graphical display of the significant components in a two-way contingency table. *Communications in Statistics-Theory and Methods*, Vol. 9, 1025-1041.

[8] Corsten, L.C.A.. and Gabriel, K.R. (1976). Graphical exploration in comparing variance matrice. *Biometrics*, Vol. 32, 851-863.

[9] 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.

[10] Doi, M., Nakamura, T., and Yamamoto, E. (2001). Conservative tendency of the crude odds ratio. *Journal of Japan Statistical Society,* Vol. 31(1),

53-65.

[11] Fienberg, S.E. (1968). *The Estimation of Exponential Probabilities in Two-way Contingency Tables*, Ph. D. Thesis, Department of Statistics, Harvard University.

[12] Fienberg, S.E. and Gilbert, J.P. (1970). The geometry of a 2×2 contingency tables. *Journal of American Statistical Association*, Vol. 65, 694-701.

[13] Fienberg, S.E. (1975). Perspective canada as a social report. *Social Indicators Research*, Vol. 2, 154-174.

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

[15] Friendly, M. (1991). *The SAS System for Statistical Graphics*, SAS Institute Inc.

[16] Friendly, M. (1992). Mosaic displays for log-linear models, proceedings of the statistical graphics section. *American Statistical Association,* 61-68.

[17] Friendly, M. (1994a). Mosaic displays for multi-way contingency tables. *Journal of American Statistical Association*, Vol. 89, 190-200.

[18] Friendly, M. (1994b). SAS/IML graphics for fourfold displays. *Observations*, Vol. 3(4), 47-56.

[19] Gabriel, K.R. (1971). The biplot graphical display of matrices with applications to principal component analysis. *Biometrika*, Vol. 58, 453-467.

[20] Gay, D.M. (1983). Algorithm 611. subroutines for unconstrained minimization using a model / trust-region approach. *ACM Transactions on Mathe -matical Software*, Vol. 9, 503-524.

[21] Gay, D.M. (1984). A trust region approach to linearly constrained optimization, *Numerical Analysis*, Proceedings, Dundee 1983, ed. F.A. Lootsma, Berlin: Springer, 171-189.

[22] Hartigan, J.A. and Kleiner, B. (1981). Mosaic for contingency tables. *Com -puter Science and Statistics*, Proceedings of the 13th Symposium on the Interface, ED. W. F. Eddy, New York : Springer-Verlag, 268-273.

[23] Hartigan, J.A. and Kleiner, B. (1984). A mosaic of the television ratings. *American Statisticians*, Vol. 38, 32-35.

[24] Hong, C.S., Choi. H.J., and Oh, M.G. (1999). Geometric descriptions for hierarchical log-linear models, *InterStat*, September, 1999. *InterStat on the Internet.*

[25] Li, X., Buechner, J.M., Tarwater, P.M., and Munoz, A. (2003). A diamond -shaped equiponderant graphical display of the effects of two categorical predictors on continuous outcomes. *American Statisticians*, Vol. 57, 193-199.

[26] 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.

[27] Trosset. M.W. (2005). Visualizing correlation. *Journal of Computational and Graphical Statistics,* Vol. 14, 1-19.

[28] Tufte, E.R. (1983). *The Visual Display of Quantitative Information*, Graphics Press. Cheshire, Connecticut.

[29] Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison-Wesley Publishing Company.

[30] Wilkinson, L. (1999). *The Grammar of Graphics*, Springer, New York.

[31] Yamamoto, E. and Doi, M. (2001). Noncollapsibility of common odds ratios without/with confounding. *Bulletin of The 53rd Session of the Inter -national Statistical Institute,* Book 3, 39-40.

[Received September 2006, Accepted October 2006]