Journal of Data Science **5**(2007), 297-313

Linear Information Models: An Introduction

Philip E. Cheng, Jiun W. Liou, Michelle Liou and John A. D. Aston
Academia Sinica

Abstract: Relative entropy identities yield basic decompositions of categorical data log-likelihoods. These naturally lead to the development of information models in contrast to the hierarchical log-linear models. A recent study by the authors clarified the principal difference in the data likelihood analysis between the two model types. The proposed scheme of log-likelihood decomposition introduces a prototype of linear information models, with which a basic scheme of model selection can be formulated accordingly. Empirical studies with high-way contingency tables are exemplified to illustrate the natural selections of information models in contrast to hierarchical log-linear models.

Key words: Contingency tables, log-linear models, information models, model selection, mutual information.

1. Introduction

Analysis of contingency tables with multi-way classifications has been a fundamental area of research in the history of statistics. From testing hypothesis of independence in a 2×2 table (Pearson, 1904; Yule, 1911; Fisher, 1934; Yates, 1934) to testing interaction across a strata of 2×2 tables (Bartlett, 1935), many discussions had emerged in the literature to build up the foundation of statistical inference of categorical data analysis. In this vital field of applied statistics, three closely related topics have gradually developed and are still theoretically incomplete even after half a century of investigation.

The first and primary topic is born out of the initial hypothesis testing for independence in a 2×2 table. From 1960's utill the 1980's, Fisher's exact test repeatedly received criticism for being conservative due to discrete nature (Berkson, 1978; Yates, 1984). Although the arguments in favor of the unconditional tests essentially focused on using the unconditional exact tests in the recent decades, the reasons for preferred p values and sensitivity of the unconditional tests have not been assured in theory. In this respect, a recent approach via information theory proved that the power analysis of unconditional tests is not suitable for testing independence, but simply for equal and unequal binomial rates. Consequently, the long term ambiguous criticism of the exact test was finally proved to be logically vacuous (Cheng *et al.*, 2005). In retrospect, it is seen that noncentral hypergeometric distributions cannot determine power evaluations at arbitrary alternative 2 × 2 tables (Fisher, 1962).

Extending to several 2×2 tables, Bartlett (1935) addressed the topic of testing interaction and derived an estimate of the common odds ratio. Norton (1945), Simpson (1951) and Roy and Kastenbaum (1956) discussed interpretations of interactions and showed that Bartlett's test is a conditional maximum likelihood estimation (MLE) given the table margins. For the same data, the celebrated CMH test for association (Cochran, 1954; Mantel and Haenszel, 1959) has been applied extensively in the fields of biology, education, engineering, medicine, sociology and psychology. However, it was implicit that an inferential flaw lies in the estimating-equation design of the CMH score test (Woolf, 1955; Goodman, 1969; Mellenberg, 1982). In addition, the probability at alternate interactions given the observed data, that is, the power analysis at alternatives to the null, has only recently been discussed in the literature. A remedy to such statistical inference was recently provided by an analysis of invariant information identities (Cheng, *et al.*, 2007). The solution, as an extension of the power analysis for a single 2×2 table, will also be useful for testing hypothesis with high-way contingency tables, which is the topic of this study to be discussed below.

Analysis of variance (ANOVA, Fisher, 1925) inspired discussions of partitioning chi-squares within the contingency tables (Mood, 1950; Lancaster, 1951; Snedecor, 1958; Claringbold, 1961). It inspired in turn the development of log-linear models (Kullback, 1959; Darroch, 1962; Lewis, 1962; Birch, 1964; and Goodman, 1964). Hierarchical log-linear models were thereby formulated to analyze general aspects of contingency tables (cf. Goodman, 1970; Bishop *et al.*, 1975), and since then, have been widely used in the literature (Hagenaars, 1993;, Christensen, 1997; Agresti, 2002). A drawback of inference with the the test statistics by Lancaster, Kullback and Claringbold was remarked by Plackett (1962).

It was recently found that a flaw of inference exists with a likelihood ratio test for association (Roy and Kastenbaum, 1956) and another for testing interaction (Darroch, 1962) and again, the data likelihood identities provide appropriate explanations (Cheng, *et al.*, 2006). These data information identities also indicate that analysis of variance may not be designed and used to measure deviations from uniform association, or varied interactions between the categorical variables, which are simply defined by likelihood factorizations. A prototype of the linear information models will be formulated below and compared to the hierarchical log-linear model.

The basic log-linear models in three variables will be reviewed in Section 2, where the notations and parameters defined by ANOVA decompositions may require careful interpretations. Next, information identities of three-way tables are fully discussed to characterize the corresponding information models, which may differ from the log-linear models only in some representations. In Section 3, the prototypes of information models with four-way and high-way tables begin to indicate the essential difference from that of the log-linear models, in particular, an elementary scheme of model selection can be formulated with easily justified tests of significance. Section 4 will provide empirical study of four- and five-way tables, which have been analyzed with log-linear models in the literature. Comparisons between the log-linear models and the proposed information models will be discussed, and obvious advantages over the log-linear modeling are easily shown through the information models selection. In conclusion, remarks on criteria of information models selection are noted for further useful research.

2. Elementary Log-likelihood Equations

The representations of data log-likelihood can be formulated in various ways, depending on the methods of likelihood decomposition. There are numerous ways of decomposing the data likelihood with high-way tables. It is elementary and instructive to discuss the case of three-way tables, which allow only a few different expressions of log-likelihood equations. The three-way log-linear models is first reviewed.

2.1 Basic log-linear models

Suppose that individuals of a sample are classified according to three categorical variables $\{X\}$, $\{Y\}$, $\{Z\}$ with classified levels: $i = 1, ..., I$, $j = 1, ..., J$, $k = 1, ..., K$, respectively. Denote the joint probability density by

$$f_{ijk}(= f(i,j,k)) = P(X = i, Y = j, Z = k), \tag{2.1}$$

where $\sum_{ijk} f_{ijk} = 1$. The full (saturated) log-linear model (Goodman, 1970; Bishop, *et al.*, 1975) is defined as

$$\log f_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ}. \tag{2.2}$$

The full model can be reduced to be a submodel of no three-way interaction, or zero conditional association, which is denoted by (XY, YZ, XZ). The model permits all three pairs to be conditionally dependent, that is, no pair is conditionally independent (Agresti, 2002); and, the corresponding log-linear model is formulated as

$$\log f_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ}. \tag{2.3}$$

In case, exactly one pair of factors is conditionally independent given the third factor, say, $\{X, Z\}$ given Y, then the model is expressed as

$$\log f_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}. \tag{2.4}$$

With two pairs of conditionally independent factors, model (2.4) reduces to independence between $\{X, Y\}$ and Z, denoted (XY, Z), and written as

$$\log f_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}. \tag{2.5}$$

The final reduction to three pairs of conditional independence is denoted by (X, Y, Z), and expressed as

$$\log f_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z. \tag{2.6}$$

Indeed, equation (2.6) defines the mutual independence between the three factors, that is, zero mutual information (Cheng *et al.*, 2006).

Equation (2.3) obtains from (2.2) by checking the magnitude of the three-way interaction using the iterated proportional fitting (Deming and Stephens, 1940). Equation (2.4) is however not derived from (2.3), but directly computed from (2.2) by deleting the conditional association between $\{X, Z\}$ given Y, which is the sum of the unique three-way interaction and the conditional uniform association between $\{X, Z\}$ given Y. This will be further explained using equation (2.12) below. The parameters in the log-linear models, besides the normalizing constant λ, are scaled log-likelihood ratios, or the logarithmic odds ratios. It is obvious that the one-way and two-way parameters in each of the models (2.4) to (2.6) are statistically independent using disjoint data information. Such a desirable property is lost between the two-way and three-way terms in the models (2.2) and (2.3), which indicates an intrinsic difference in the interpretation of these log-linear models.

2.2 Basic information models

It may be useful to take a different look at the formulation of hierarchical log-linear models, in order to understand the intrinsic difference between dependent and independent likelihood decompositions. Basic identities of data likelihood are examined, and a prototype of linear information models in three variables will be considered. Define the marginal probabilities as

$$f_i = f_{i\cdot\cdot} = \sum_{j=1}^{J} \sum_{k=1}^{K} f_{ijk}, \;\; f_{ij\cdot} = \sum_{k=1}^{K} f_{ijk}; \tag{2.1}$$

and analogously, f_j, f_k; $f_{\cdot jk}$, $f_{i \cdot k}$. Also, let n_{ijk} denote the number of individuals classified to the cell $\{X = i, Y = j, Z = k\}$, and similar notations $n = n_{\dots}$, $n_i = n_{i \cdot \cdot}$, and $n_{ij \cdot}$, denote the total cell frequency and marginal totals, respectively. Some convenient terminology will be borrowed from our previous study (Cheng, *et al.*, 2006, Section 2). The entropy identity in three variables is expressed as

$$H(X) + H(Y) + H(Z) = H(X, Y, Z) + I(X; Y; Z) \ , \tag{2.8}$$

where the marginal entropy, joint entropy, and the mutual information are defined respectively to be

$$\begin{aligned} H(X) &= -\sum_i f_i \log f_i \ , \ H(Y) = -\sum_j g_j \log g_j \ , \\ H(Z) &= -\sum_k h_k \log h_k \ , H(X, Y) = -\sum_{i,j} f_{ij\cdot} \log f_{ij\cdot} \ , \\ H(X, Y, Z) &= -\sum_{i,j,k} f_{ijk} \log f_{ijk} \\ I(X; Y; Z) &= \sum_{i,j,k} f_{ijk} \log \left(\frac{f_{ijk}}{f_i g_j h_k} \right) = D(f(x, y, z) || f(x)g(y)h(z)). \end{aligned} \tag{2.9}$$

The last entry of (2.9), the mutual information, is the Kullback-Leibler divergence between the joint density and the product marginal density of (X, Y, Z). The sample analogs of the entries in (2.9), the sample entropy and the sample mutual information, are defined in terms of the observed cell frequencies, which are the natural maximum likelihood estimates under the general model of multinomial distribution. For example, the sample analog of the mutual information $I(X; Y; Z)$ yields the likelihood ratio test statistic

$$\sum_{i,j,k} n_{ijk} \log \left(\frac{n_{ijk} n^2}{n_i n_j n_k} \right) , \tag{2.10}$$

such that twice of (2.10) is asymptotically chi-squared distributed with d.f. $IJK - (I + J + K) + 2$ under the null hypothesis of mutual independence. It is easily seen that data log-likelihood admits three equivalent information identities, which are the three possible saturated information models for a three-way table,

$$\begin{aligned} I(X; Y; Z) &= I(X; Y) + I(Y; Z) + I(X; Z|Y) \\ &= I(X; Y) + I(X; Z) + I(Y; Z|X) \\ &= I(X; Z) + I(Y; Z) + I(X; Y|Z). \end{aligned} \tag{2.11}$$

Being saturated models, the equations of (2.11) are different expressions of the log-linear full model (2.2). All three equations are reduced to model (2.3), if the common interaction term $Int(X;Y;Z)$ in the equations, say,

$$I(X;Z|Y) = Int(X;Y;Z) + I(X;Z||Y), \tag{2.12}$$

is removed (equal to zero). The remaining terms are not the same, for example, $I(X;Z||Y)$ characterizes the uniform association between X and Z, within the levels of Y. The unique three-way interaction, $Int(X;Y;Z)$ of (2.12), characterizes exactly the three-way term λ_{ijk}^{XYZ} of (2.2), and the latter expression is unique up to the normalizing constant λ.

The special case of $2 \times 2 \times K$ table, the simplest three-way table, has been a much disucussed topic in the literature. It was recently proposed that the two summands of (2.12) form a natural two-step likelihood ratio (LR) test, where the first step LR statistic, $Int(X;Y;Z)$, tests for no interaction between X and Z, across K strata of Y; and the second, $I(X;Z||Y)$, tests for no association between X and Z, within strata of Y (Cheng *et al.*, 2007). Logically, the second step is in use only when the first step is insignificant. The two-step LR test for no association across strata Y is shown to be asymptotically unbiased and more powerful than the one-step omnibus LR test, the conditional mutual information (CMI) $I(X;Z|Y)$ of (2.12), or Pearson's chi-square test. Likewise, the two-step LR test improves over the combination of the score tests, the Breslow-Day test (1980) and the CMH test (1954, 1959).

If the omnibus CMI $I(X;Z|Y)$ is insignificantly small, then the conditional odds ratios between X and Z, across the levels of Y, are close to 1, and the first equation of (2.11) may be reduced to the sum $I(X;Y) + I(Y;Z)$ which is model (2.4). Otherewise, this CMI is significantly large, in which the two independent summands, $Int(X;Y;Z)$ and $I(X;Z||Y)$, can be individually significantly large or small, with four possible combined cases. In this case, if the test for no interaction, $Int(X;Y;Z) = 0$, is rejected, it is logically consistent with the significant omnibus test; and, there is no need to perform the second step test; and clearly, it is insufficient to use only the second-step test for testing no association between X and Z, across Y. If the interaction is insignificantly small, then the second-step test statistic $I(X;Z||Y)$, also called the generalized CMH test (Agresti, 2002, p.325), is usually expected to be significantly large, that is also consistent with the significant omnibus test. Here, with a rare chance it may happen that both $Int(X;Y;Z)$ and $I(X;Z||Y)$ are insignificantly small, whereas the omnibus CMI is marginally significant; and, it is theoretically rare that the two-step test is not sufficiently sensitive. It is understood that such significant or insignificant tests are defined with a common fixed level of significance for each approximating chi-square distribution; and, it follows from (2.12) that four combinations of significance and insignificance patterns between the three LR tests are meaningful in practice. These information with varied statistical inference are not directly revealed between models (2.2) to (2.4) as mentioned above. The key point of fact, that equations (2.11) disallow the inclusion of all three two-way terms, is not clarified in the discussion of hierarchical log-linear models.

It is perceivable that extensions of identity (2.11) to high-way contingency tables will enhance interpretations of high-way effects and manifest further differences from the conventional log-linear models. The goal of this study is to make a natural and systematic extension of equation (2.11) to high-way contingency tables, which are coined information models. The difference between the proposed information models and the log-linear models will be illustrated using four-way and five-way data tables that have been analyzed in the literature. Subsequently, a prototype of the linear information model, defined by simple criteria of model selection, will be introduced to offer advantage of statistical inference over the conventional log-linear model.

3. Information Structures of High-Way Tables

The natural extensions of equation (2.11) consist of two major formulae for each saturated model of a K-way contingency table. The primary formula is that the log-likelihood decomposition is the sum of K main effects, $(K-1)$ two-way (MI) effects, $(K-2)$ three-way (CMI) effects, ..., two $(K-1)$-way CMI, and one K-way CMI. The secondary formula is that each t-way CMI $(t \leq K)$ is a mutual information between a pair of variables, conditioned upon other $t-2$ variables. A t-way CMI is always a sum of the t-way interaction and the uniform association between the pair of variables, conditioned on the remaining $t-2$ variables. It is plain that both formulae of data log-likelihood decompositions are built upon the notion of mutual information, not in terms of ANOVA as defined with the log-linear models.

3.1 A four-way information structure

It is observed from equation (2.11) that a common variable appears in each term of the decomposed three-way mutual informtion, and it is used as the conditioning variable (termed CV). This is not necessarily required of a four-way or a high-way table, however, the use of such a CV is always applicable and particularly useful. When any variable is of primary interest, like the response variable in a linear regression model, it is used as a CV for finding its relations to the remaining variables. In case of four variables, the analog of (2.8) is the entropy identity

$$H(W) + H(X) + H(Y) + H(Z) = H(W, X, Y, Z) + I(W; X; Y; Z). \qquad (3.1)$$

And, an analog of (2.11) among many equivalent identities may be expressed as

$$\begin{aligned} I(W;X;Y;Z) &= I(W;X;Y) + I((W,X,Y);Z) \\ &= I(W;X;Y) + I((X,Y);Z) + I(W;Z|X,Y) \\ &= I(W;Y) + I(X;Y) + I(W;X|Y) + I(Y;Z) \\ &\quad + I(X;Z|Y) + I(W;Z|X,Y). \end{aligned} \tag{3.2}$$

The last equation of (3.2) follows by using (2.11) with the variable Y as the CV. If there is no special CV of interest, then either a different identity of (2.11) may be used to express $I(W;X;Y)$, or another identity such as $I((X,Y);Z) = I(X;Z) + I(Y;Z|X)$ can be used, in the second equation of (3.2). For example, the last statement simply leads to different non-CV saturated models such as

$$\begin{aligned} I(W;X;Y;Z) &= I(W;Y) + I(X;Y) + I(W;X|Y) + I(X;Z) \\ &\quad + I(Y;Z|X) + I(W;Z|X,Y) \\ &= I(W;X) + I(W;Y) + I(X;Y|W) + I(X;Z) \\ &\quad + I(Y;Z|X) + I(W;Z|X,Y). \end{aligned} \tag{3.3}$$

It is worth noting with these saturated models that all the variables must appear at least once in the main effects, and also among the two-way MI terms, the three-way CMI terms, and a specified four-way CMI, respectively. For a three-way contingency table, exactly one of three CV models may be used according to equation (2.4); and, for a four-way table, there are exactly six distinct saturated models for each fixed CV, which yields 24 distinct saturated CV models. If both CV and non-CV models are included, the total number of saturated information models in four variables would be $72 = (4!) \times 3$.

3.2 Elementary model selection schemes

Without loss of generality, the information model selection scheme is illustrated with a four-way contingency table. The question is how to select a meaningful and parsimonious model among those seventy-two candidate linear information models. A selection scheme based on equation (3.2) using a CV is first outlined below. This is organized as a four-step procedure for ease of exposition.

Step 1: Select the CV, say Y, either because it is a variable of focus; or, among the four variables $\{W, X, Y, Z\}$, Y yields the maximal significant (in p value, for a fixed nominal level) sum of three two-way effects, say, twice the sample analog of $I(W;Y) + I(X;Y) + I(Y;Z)$.

Step 2: Find the *maximal insignificant* (in p value) among the insignificant four-way CMI (between certain two variables, conditioned upon the chosen CV, Y, and the remaining variable). Otherwise, choose the *minimal significant*, when the three available four-way CMI are all significantly large. Suppose the chosen four-way CMI is $I(W; Z|X, Y)$, then, the two candidate three-way CMI, $I(W; X|Y)$ and $I(Z; X|Y)$, are directly obtained by reading the appeared variables from the chosen four-way CMI. Since the three two-way (MI) terms have already been selected in Step 1, a saturated CV information model selection in the four-variable case is essentially complete. In other words, each variable must appear at least once among the two-way terms of a saturated model before parsimonious selection, whether a common CV is in use or not.

Step 3: To confirm the selected MI and CMI terms, each one of the ten selected terms (including four main-effect terms) must be individually and separately tested against the same nominal level, say, 0.05 in common practice. The sum of all the insignificantly (small in p value) terms, among the ten terms, is taken as the tentative remainder which is then tested against the total (chi-square) d.f., to the same nominal level. This step is asymptotically correct by the orthogonal likelihood decomposition. If this test is insignificant, then the tentative remainder is insignificantly small and deleted as an insignificant residual, so that the tentative model is accepted. Otherwise, the remainder is significantly large (near or over a 95th percentile of the associated chi-square distribution) and the tentative model may be lack of sufficient information. In this case, a simple remedy is recommended. A maximal insignificant (or, a minimal significant) high-way CMI term can be replaced by its next term, the second-maximal insignificant (or, second-minimal significant) CMI; and then, the same selection procedure is continued with the renewed model modified in Step 2. This remedy as a supplementary scheme is easily used in practice, because it can always choose the next insignificant (significant) term whenever needed, from high-way to low-way subtables, while modifying the information decomposition.

Step 4. To conclude a parsimonious model selection after performing the above three steps, it is often of extra interest, though not necessary, to test against the summands of each selected CMI term, an interaction term and a uniform association term (cf. (2.12)), under the same nominal level. Finally, the new remainder is likewise tested to be a negligible residual, as shown in Step 3, to yield a more parsimonious model.

It is understood that a general scheme may select either a CV model (3.2), or a non-CV model (3.3). If there is no need to fix a CV of particular interest, Step 1 is bypassed, and Step 2 is generalized without requiring a fixed CV throughout the selection scheme; and, Steps 3 and 4 are kept unchanged. Thus, it is expected that such a general scheme often results in selecting a non-CV model, more balanced in selecting the variables among the CMI terms.

On the other hand, there are alternate ways of selecting a model like (3.2) or (3.3). As an alternative choice in Step 2, selecting a minimal insignificant high-way CMI term may be preferred, if any; otherwise, select the minimal (or maximal) significant high-way CMI term when all such CMI terms are significantly large. These alternatives to Step 2 may sometimes yield less high-way CMI terms compared to the original Step 2, particularly with high-way tables. However, they usually lead to selecting more terms at end, and sacrifices model parsimony. It is remarkable that the principal idea in formulaing the models (3.2) and (3.3) is to delete more high-way CMI terms, compared to the deletion of high-way interaction terms as a common practice in the selection of a log-linear model. This will be illustrated below in Section 4, in particular, Example 4.1. The above three- or four-step model selection scheme can be easily extended to high-way contingency tables. An application to five-way data table will also be illustrated in Section 4.

4. Applications to High-Way Tables

Four-way and five-way contingency data tables in the literature will be examined. The proposed information modeling and the four-step model selection scheme of Section 3 will be applied to both four-way and five-way data below. It is entertaining that the proposed method yields easily interpretable and more parsimonious models compared to those obtained from the hierarchical log-linear modeling.

Example 4.1: A four-way contingency table

A $3 \times 3 \times 3 \times 3$ four-way data frequency table (Agresti, 2002, Table 8.19) is exemplified for a comparison study. The data consist of three-level individual's opinions on each of four variables of government spending. These 81 cell frequency counts are listed according to the three levels: "too little", "about right" and "too much", defined with the variables: environment (E), health (H), big cities (C), and law enforcement (L). The basic analysis of this data by log-linear modeling leads to the accepted model: deletion of the four-way interaction and all the three-way effects, that is, fitting the data to the summary of the four main effects and all the six two-way effects (Agresti, 2002, Table 8.20). It is so fitted with the log-linear modeling because the deviance between this fitted model and the full model is evaluated to be 31.67, which approximately equals to the 35th percentile of the chi-square distribution with 48 d.f. Readers may refer to subsequent estimation of odds ratios parameters to the log-linear model fitting with all two-way effects.

The information modeling begins with finding the most relevant variable to be the CV, according to Step 1 of the selection scheme as illustrated in Section 3.2. It is found that the variable health (H) yields the greatest significant sum of two-way effects, which is 61.87 to the chi-square distribution with 12 d.f. It then easily follows by Steps 2 and 3 to find that the only significant terms in the information equation (3.2), using $Y = H$ and $X = E$, are the two-way effects, the deviances $2I(C;H) \cong 28.74$ and $2I(E;H) \cong 24.18$. By Step 4, it can be easily checked that the deviance, the sum of the residual insignificant CMI's, is 71.59, which is close to the 76th percentile of the chi-square distribution with 64 d.f. A complete selection scheme in accordance with equation (3.2) for the four-way table is summarized in Table 1 below. Thus, a prototype of information model selection by equation (3.2) is tentatively concluded with the information model: "four main effects plus the two-way effects $\{CH, EH\}$", with the total fitted d.f.= 16, that is only half of 32, the fitted log-linear model of all two-way effects. This exhibits a basic advantage of information modeling that it usually selects a more parsimonious model with more concise and simpler explanation compared to the classical log-linear modeling.

In case another variable is of interest, say, Environment (E), which may be closely related to budget spending and worth an investigation, then, the variable (E) can be taken as the CV. It also follows by equation (3.2) and the same selection scheme that a similar model is selected: "the main effects, plus the two-way effects $\{CE, EH\}$", for which the remainder deviance is 81.98, close to the 93.5 percentile of the chi-square distribution with 64 d.f. This provides a similar CV model selection and interpretation. It can be easily checked from this four-way data that the other two categories, cities (C) and law enforcement (L), may not be considered as useful CV, because the selected models will include at least one three-way CMI term, in addition to the two-way terms and main effects.

In case no particular variable is fixed as CV, it is checked by equation (3.3) and the same selection scheme (omitting Step 1) would lead to a non-CV model: "$\{CE, HL\}$, plus the four main effects", for which the remainder deviance is 77.41, close to the 88 percentile of the chi-square (64 d.f.) distribution. This provides another equally parsimonious information model in which all the four variables share the two-way effects together in a pair of two-way terms, instead of all the six two-way terms as used in the selected log-linear model (Agresti, 2002).

The (conditional) odds ratio parameters and the deviances of the above three selected information models, including interval estimates, can be computed along with the selection scheme. For brevity, these calculations are not discussed here for each selected information models, and the readers may contact the authors for details.

Table 1: Mutual Information (3.2)

MI, CMI \ values	chi-squares	d.f.	p-values
$I(C;L\|E,H)$	34.90	36	0.55
$I(C;E\|H)$	19.77	12	0.07
$I(E;L\|H)$	7.99	12	0.78
$I(H;L)$	8.95	4	0.07
$I(E;H)$	24.18	4	0.001
$I(C;H)$	28.74	4	0.001

Example 4.2: A five-way contingency table

A data of cross-classification of individuals according to five dichotomized factors had been studied in six publications prior to Goodman (1978, p.112, Table 1). The purpose was to understand the association relationship between the knowledge (good or poor, denoted by K) of cancer, and the presence or absence of the other four qualitative attributes: $L(=\text{lecture})$, $R(=\text{radio})$, $N(=\text{newspaper})$, and $S(=\text{solid reading})$. The factor of primary interest is the "knowledge K", and hypotheses about the logits of K, plus estimates of the hypothesized effects were examined by Goodman (1978, Tables 5 to 7) using hierarchical log-linear models. However, it is so far unknown in the literature whether there are specific criteria or schemes of selecting a definite, or tentatively entertained, parsimonious log-linear model, for the present five-way data that had been much discussed prior to Goodman (1978). It is thus worth investigating the proposed selection scheme of linear information models, in contrast to the hierarchical log-linear modeling, for the current five-way data.

According to Step 1 of Section 3.2, it is found that factor K has the maximal two-way effect with the other variables, which evidences that it was a useful study design. Let factor K be the CV, a saturated information model can be derived according to Step 2 as follows.

$$\begin{aligned}
&I(K;L;N;R;S)\\
&= I(L;(K,N,R,S)) + I(K,N,R,S)\\
&= I(L;N|K,R,S) + I(L;(K,R,S)) + I(S;(K,N,R)) + I(K,N,R)\\
&= I(L;N|K,R,S) + I(L;S|K,R) + I(L;(K,R)) + I(R;S|K,N)\\
&\quad + I(S;(K,N)) + I(N;(K,R)) + I(K;R)\\
&= I(L;N|K,R,S) + I(L;S|K,R) + I(L;R|K) + I(K;L) + I(R;S|K,N)\\
&\quad + I(K;S) + I(N;S|K) + I(K;N) + I(N;R|K) + I(K;R). \qquad (4.1)
\end{aligned}$$

In the last equation of the saturated model (4.1), it is found that all terms are statistically significant at the nominal level of 0.05, except a few insignificant terms, the five-way CMI term $I(L; N|K, R, S)$, and the four-way CMI $I(R; S|K, N)$. By Step 4, it is checked that "twice the sample sum of $I(R; S|K, N)+I(L; N|K, R, S)$, or 26.6" is insignificantly small, approximately equal to the 65th percentile on the chi-square distriubtion with 12 d.f. This yields the linear information model that exhibits the desired relationship between the CV "Knowledge K" and the other four variables:

$$\begin{aligned} I(K; L; N; R; S) \cong{}& I(L; S|K, R) + I(L; R|K) + I(K; L) + I(K; S) \\ &+ I(N; S|K) + I(K; N) + I(N; R|K) + I(K; R). \end{aligned} \tag{4.2}$$

For the five-way data, it is notable that the selected information model (4.2) treats the variable (Knowledge, K) as a response variable. To summarize the data analysis, Table 2 exhibits the component MI and CMI values of the overall mutual information, in which most are significantly large, except two high-way CMI terms; and, the selection scheme confirms only a slight reduction of two CMI terms by equations (4.1) and (4.2). This presents a case that the variables defined in the study are highly associated, and very little information reduction is possible, although information dissemination by model (4.2) yields Table 2.

Table 2: Mutual Information (4.1) and (4.2)

MI, CMI \ values	chi-squares	d.f.	p-values
$I(L; N\|K, R, S)$	10.58	8	0.230
$I(R; S\|K, N)$	2.72	4	0.610
$I(L; S\|K, R)$	20.48	4	< 0.001
$I(L; R\|K)$	15.56	2	< 0.001
$I(N; S\|K)$	190.89	2	< 0.001
$I(N; R\|K)$	58.96	2	< 0.001
$I(K; L)$	17.16	1	< 0.001
$I(K; N)$	105.78	1	< 0.001
$I(K; R)$	24.25	1	< 0.001
$I(K; S)$	150.45	1	< 0.001

As a supplementary note, the possible choices of five-way information models may be estimated like the previous discussion of four-way models based on equations (3.2) and (3.3). Equation (4.1) allows five ways of separating one variable from the other four, which includes a four-variable model (3.2) as part of the whole model. Thus, the number of saturated five-way CV models is $720 = 5! \times 3 \times 2$, and, the number of all saturated five-way models is $5! \times (4! \times 3) = 25920$. Extensions of equations (3.2), (3.3) and (4.1) to multi-way contingency tables appear to be straightforward.

5. Concluding Remarks

A short summary of the proposed information models and the selection scheme in Section 3 can be illustrated with a few remarks. The primary purpose of developing the linear information models is to recommend the natural factorization of the raw data likelihood, without additional operations on the data. The basic information identities of Section 2 are used to illustrate the advantages of using orthogonal information decomposition, directly using the observed data likelihood, but not through the adapted ANOVA. A basic drawback of the latter, the disadvantage of inevitably crossed and overlapped data information in the summands of the log-linear models, can be especially intricate with high-way tables, for which a three-way case was illustrated in the recent literature (Cheng *et al.*, 2006). Essentially, the classical approach invalidates the development of useful selection schemes among the hierarchical log-linear models.

The important advantages of linear information models are based on direct use of the data likelihood identities as illustrated in Section 3 and exemplified in Section 4. The proposed model selection schemes, either using a CV or not, are naturally born out of the data likelihood. It is the simplest method based on comparing data deviances of any possible remainder terms against appropriate chi-square distributions. Thus, the proposed model identification and selection schemes offers a fundamental likelihood analysis with observed data. While useful information model selections still depend on interpreting the data through the choice of certain CV (or no CV) by the experimenter, it is understood that no best or uniquely optimal model can be defined whichever selection criterion is used. Given a natural selection criterion, such as the current proposal, it may take further study to define optimal model parsimony and selection, together with some additional selection criteria that would be useful in other statistical applications.

References

Agresti, A. (2002). *Categorical Data Analysis, 2nd ed.* Wiley.

Bartlett, M. S. (1935). Contingency table interactions. *J. Roy. Statist. Soc. B* **2**, 248-252.

Berkson, J. (1978). In dispraise of the exact test. *J. Statist. Planning and Inference.* **2**, 27-42.

Birch, M. W. (1964). The detection of partial association. *J. Roy. Statist. Soc. B* **26**, 313-324.

Bishop, Y. M. M., Fienberg, S. E. and Holland. P. W. (1975). *Discrete Multivariate Analysis.* Cambridge, MA: MIT Press.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research, 1. The Analysis of Case-Control Studies.* Lyon.

Cheng, P. E., Liou, M., Aston, J. and Tsai, Arthur C. (2005). Information identities and testing hypotheses: Power analysis for contingency tables. *Statistica Sinica,* in press.

Cheng, P. E., Liou, M. and Aston, J. (2007). Testing association: A two-stage test or Cochran-Mantel-Haenszel test. *Institute of Statistical Science, Academia Sinica, Technical Report 2007-03.*

Cheng, P. E., Liou, J. W., Liou, M. and Aston, J. (2006). Data information in contingency tables: A fallacy of hierarchical log-linear models. *Journal of Data Science.* **4**, 387-398.

Christensen, R. (1997). *Log-linear Models and logistic regression*, Springer-Verlag.

Claringbold, P. J. (1961). The use of orthogonal polynomials in the partition of chi-square. *Aust. J. Statist.* **3**, 48-63.

Cochran, W. G. (1954). Some methods for strengthening the common chi-square tests. *Biometrics* **24**, 315-327.

Darroch, J. N. (1962). Interactions in multifactor contingency tables. *J. Roy. Statist. Soc. B* **24**, 251-263.

Deming, W. E. and Stephan F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* **11**, 427-444.

Fisher, R. A. (1925) (5th ed., 1934). *Statistical Methods for Research Workers*, Oliver & Boyd.

Fisher, R. A. (1962). Confidence limits for a cross-product ratio. *Australian Journal of Statistics*, **4**, 41-41.

Goodman, L. A. (1964). Simple methods for analyzing three-factor interaction in contingency tables. *J Amer. Statist. Assoc.* **59**, 319-352.

Goodman, L. A. (1969). On partitioning and detecting partial association in three-way contingency tables. *J. Roy. Statist. Soc. B* **31**, 486-498.

Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interactions among multiple classifications. *J. Amer. Statist. Assoc.* **65**, 226-256.

Goodman, L. A. (1978). *Analyzing Qualitative/Categorical Data.* Cambridge: Abt Associates Inc.

Hagenaars, J. A. (1993). *Loglinear Models with Latent Variables.* Newbury Park: Sage.

Kullback, S. (1959). *Information Theory and Statistics.* Wiley.

Lancaster, H. O. (1951). Complex contingency tables treated by the partition of chi-square. *J. Roy. Statist. Soc. B* **13**, 242-249.

Lewis, B. N. (1962). On the analysis of interaction in multi-dimensional contingency tables. *J. Roy. Statist. Soc. A* **125**, 88-117.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.* **22**, 719-748.

Mellenberg, G. J. (1982). Contingency table models for assessing item bias. *J. Edu. Statist.* **7**, 105-118.

Mood, A. M. (1950). *Introduction to the Theory of Statistics.* McGraw-Hill.

Norton, H. W. (1945). Calculation of chi-square for complex contingency tables. *J. Amer. Statist. Assoc.* **40**, 251-258.

Pearson, K. (1904). *Mathematical contributions to the theory of evolution XIII: On the theory of contingency and its relation to association and normal correlation.* Draper's Co. Research Memoirs, Biometric Series, no. 1. (Reprinted in Karl Pearson's Early Papers, ed. E. S. Pearson, Cambridge: Cambridge University Press, 1948.)

Plackett, R. L. (1962). A note on interactions in contingency tables. *J. Roy. Statist. Soc. B* **24**, 162-166.

Roy, S. N. and Kastenbaum, M. A. (1956). On the hypothesis of no "interaction" in a multi-way contingency table. *Ann. Math. Statist.* **27**, 749-757.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. B* **13**, 238-241.

Snedecor, G. W. (1958). Chi-squares of Bartlett, Mood, and Lancaster in a contingency table. *Biometrics* **14**, 560-562.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Ann. Human Genetics* **19**, 251-253.

Yates, F. (1934). Contingency tables involving small numbers and the test. *J. Royal Statist. Soc., Suppl.* **1**, 217-235.

Yates, F. (1984). Tests of Significance for contingency tables (with discussion). *J. Royal Statist. Soc., A* **147**, 426-463.

Yule, G. U. (1911). *An Introduction to the Theory of Statistics.* Griffin.

Received April 1, 2007; accepted May 31, 2007.

Philip E. Cheng
Institute of Statistical Science
Academia Sinica
Taipei, 115, Taiwan
pcheng@stat.sinica.edu.tw

Jiun W. Liou
Institute of Statistical Science
Academia Sinica
Taipei, 115, Taiwan
needgem@stat.sinica.edu.tw

Michelle Liou
Institute of Statistical Science
Academia Sinica
Taipei, 115, Taiwan
mliou@stat.sinica.edu.tw

Aston, John A. D.
Institute of Statistical Science
Academia Sinica
Taipei, 115, Taiwan
jaston@stat.sinica.edu.tw