# Ordinal Data Analysis Based on Kullback-Leibler Distance

Wei Gao

School of Mathematics and Statistics, Northeast Normal University

Email: gaow@nenu.edu.cn

Taipei, Dec. 18th, 2011

# Outline

Since precise measurements may be expensive or cannot be obtained for latent variables or latent variables are ordinal categorial variable, ordinal categorical data often appears in data analysis.

In clinic diagnosed results for some disease : **Having diseases, probably having, probably not having, and not having**;

In social surveys, income may be classified into **high , middle and low income.**

**”The benefits include the potential for considerable improvements in power over methods that ignore the ordinal information and greater parsimony in model-building from a reduction in the number of parameters needed to describe association structure. In particular, ordinal models can represent predictions about monotone or linear trends in effects and associations.”**, stated by Agresti and Coull(2005).

An ordinal variable $X$ with possible $k$ values and the following problems are often considered

$$p_j \leq h_j \text{ for } j = 1, \cdots, k-1;$$

or

$$\sum_{j=1}^{l} p_j \leq \sum_{j=1}^{l} h_j \text{ for } j = 1, \cdots, k-1$$

where $p_i = P\{X = i\}$ and $h$ is a known probability vector. ( See Robertson, Wright and Dykstra(1988), and Agresti and Coull(2005)).

A series $2 \times 2$ contingency tables, one may be interested in odds ratios $\psi_l = (p_{11l}p_{22l}/(p_{12l}p_{21l})$, having some trends, for example,

$$\psi_1 \leq \psi_2 \leq \cdots \leq \psi_K;$$

$$\psi_1 \geq \psi_2 \geq \cdots \geq \psi_K;$$

or

$$\psi_1 \leq \psi_2 \leq \cdots \leq \psi_m \geq \psi_{m+1} \geq \cdots \geq \psi_K.$$

Robertson, Wright and Dykstra(1988), Silvapulle and Sen(2005), Agresti(2010).

# Outline

## Kullback-Leibler distance

$$I(p|q) = \sum_{l=1}^{k} p_i \log(p_i/q_i)$$

where $p$ and $q$ are probability vectors.

## Minimize the Kullback-Leibler distance

$$\min \sum_{i=1}^{k} p_i \log \frac{p_i}{\pi_i} \quad \text{subject to} \quad p \in D, \tag{1}$$

$$D = \{p \in P : \quad h - A^{'} p \in C^{*}\},$$

$P$ is the probability space, $h$ is a given vector,

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1s} \\ a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \ldots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{ks} \end{pmatrix},$$

and $C^{*}$ is the dual cone of an isotonic cone of $C$ in $R^{s}$.

**Isotonic cones:** Let $\preceq$ be a quasi order defined on $\{1, 2, \cdots, s\}$. Then,

$$C = \{x \in R^s : x_i \leq x_j \text{ if } \; i \preceq j\}$$

is called the **isotonic cone** induced by the quasi order $\preceq$.

**Dual Cones:** $C^*$ is defined to be the **dual cone** (or **polar cone**) of $C$ if

$$C^* = \{y \in R^s : \sum_{j=1}^{s} y_j x_j \leq 0, \quad \forall x \in C\}.$$

1. **Linear equalities** For $C = R^s$, $C^* = \{(0, \cdots, 0)'\}$;

2.**Stochastic orderings:** for
$C = \{x \in R^s : \; x_1 \leq x_2 \leq \cdots \leq x_s\};$,

$$C^* = \{y \in R^s : \sum_{j=1}^{l} y_j \geq 0 \text{ for } l = 1, \cdots, s-1 \\ \text{and } \sum_{j=1}^{s} y_j = 0\}.$$

3. **Strict controls:** For
$C = \{x \in R^s : \ \ x_j \leq x_s \ \text{ for } \ j = 1, \cdots, s-1\}$,

$$C^* = \{y \in R^s : \ \ y_1 \geq 0, \cdots, y_{s-1} \geq 0 \ \text{ and } \ \sum_{j=1}^{s} y_j = 0\}.$$

## Relationships between minimizing Kullback-Leibler distance and the MLE of log-linear models

**Poisson regression modeling.** Given the covariates $x_i$, $Y_i$ are independently Poisson distributed with parameters

$$\lambda_i = \exp(\mu + \beta' x_i) \tag{2}$$

where $\mu$ is the intercept and $\beta = (\beta_1, \cdots, \beta_s)'$ are the slope coefficients. Now consider the maximum likelihood estimation of $\beta$ for model (2) with parameters being restricted by

$$\beta \in C \tag{3}$$

where $C$ is an isotonic cone.

**Theorem 1.** $\hat{\lambda} = (\hat{\lambda}_1, \cdots, \hat{\lambda}_N)'$ is the maximum likelihood estimate of $\lambda$ for Models (2) under (3)if and only if

$$\hat{\lambda} = (\sum_{i=1}^{N} y_i) \times \arg\left\{ \min_{p}\{\sum_{j=1}^{N} p_j \log p_j : \quad h - A'p \in C^*\} \right\}; \qquad (4)$$

where

$$h = (\sum_{i=1}^{N} x_{i1} y_i, \cdots, \sum_{i=1}^{N} x_{is} y_i)'$$

and

$$A = (x_1', \cdots, x_N').$$

**Theorem 2.** If there exists a probability vector $p^*$ such that

$$p_i^* = \lambda^* \pi_i \prod_{j=1}^{s} \lambda_j^{*a_{ij}}, \quad (\lambda_1^*, \cdots, \lambda_s^*)' \in C, \tag{5}$$

$$h - A' p^* \in C^*, \text{and} \tag{6}$$

$$\sum_{j=1}^{s} [h_j - \sum_{i=1}^{k} a_{ij} p_i^*] \log \lambda_j^* = 0. \tag{7}$$

then it is the unique optimal solution to (1).

# Unified Generalized Iterative Scaling

Unified Generalized Iterative Scaling(UGIS) is a method to find the the following optimal solution:

$$\min \sum_{i=1}^{k} p_i \log \frac{p_i}{\pi_i} \text{ subject to } h - A' p \in C^*,$$

where

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1s} \\ a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \ldots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{ks} \end{pmatrix},$$

$h = (h_1, \cdots, h_s)'$ are known and $C^*$ is the dual cone of a given isotonic cone $C$.

# Projections on an isotonic cone

For any $x \in R^s$ and weight $w = (w_1, \cdots, w_s)'$, denote the projection of $x$ onto $C$ by $\hat{x} = P_w(x|C)$, which satisfies

$$||x - \hat{x}||_w^2 = \min \sum_{j=1}^{s} (x_j - \mu_j)^2 w_j \quad \text{subject to} \quad \mu \in C. \qquad (8)$$

Such a projection $\hat{x}$ can be easily obtained by the max-min formulas given by Robertson, Wright and Dykstra (1988, page 22).

## Unified Generalized Iterative Scaling(UGIS)

1. For $C = \{\mu \in R^s : \ \mu_1 \leq \mu_2 \leq \cdots \leq \mu_s\}$, $\hat{x} = P_w(x|C)$ is

$$\hat{x}_i = \max_{u \leq i} \min_{v \geq i} \frac{w_u x_u + \cdots + w_v x_v}{w_u + \cdots + w_v};$$

2. For $C = \{\mu \in R^s : \; x_1 \le x_2, \cdots, x_{s-1} \le x_s\}$, $\hat{x} = P_w(x|C)$ is

$$\hat{x}_1 = \min\{x_1, a\},$$

$$\cdots,$$

$$\hat{x}_{s-1} = \min\{x_{s-1}, a\}, \hat{x}_s = a$$

where

$$a = \max_{u \le s} \frac{w_u^* y_u + \cdots + w_s^* y_s}{w_u^* + \cdots + w_s^*},$$

$y = (x_{(1)}, \cdots, x_{(s-1)}, x_s)'$, $x_{(1)}, \cdots, x_{(s-1)}$ is the order statistics of $x_1, \cdots, x_{s-1}$ and $w^*$ is the correspond weight .

## UGIS Algorithms:

Initially set $p_i^{(0)} = \pi_i \prod_{j=1}^{s} (\lambda_j^{(0)})^{a_{ij}}$ where $(\lambda_1^{(0)}, \cdots, \lambda_s^{(0)})' \in C$.

**Step(n)**

$$p_i^{(n)} = p_i^{(n-1)} \prod_{j=1}^{s} (\lambda_j^{(n)} / \lambda_j^{(n-1)})^{a_{ij}}$$

where $\lambda^{(n)} = P_{w^{(n-1)}}(U^{(n-1)}|C)$, $\lambda_j^{(n)}$ is the $j-$th element of $\lambda^{(n)}$ ,

$$w^{(n-1)} = (\sum_{i=1}^{k} a_{i1} p_i^{(n-1)}, \cdots, \sum_{i=1}^{k} a_{is} p_i^{(n-1)})',$$

and

$$U^{(n-1)} = (\frac{h_1 \lambda_1^{(n-1)}}{w_1^{(n-1)}}, \cdots, \frac{h_s \lambda_s^{(n-1)}}{w_s^{(n-1)}})'.$$

**Theorem 3.** $\{p^{(n)}\}$ given in the above proposed algorithm converges to the optimal solution of (1).

# Remark:

If $C = R^s$, then $p^{(n)}$ given in the UGIS algorithm is

$$p_i^{(n)} = p_i^{(n-1)} \prod_{j=1}^{s} (h_j / \sum_{i=1}^{k} a_{ij} p_i^{(n-1)})^{a_{ij}}$$

which is the Generalized Iterative Scaling(GIS) proposed by Darroch and Ratcliff(1972).

$$\min \sum_{i=1}^{k} p_i \log \frac{p_i}{\pi_i}, \quad \text{suject to} \quad p \in D \tag{9}$$

where $\pi$ is chosen as observed counts,

$$D = \{p : \quad p \in \cap_{r=1}^{l} D_r\}$$

and $D_r = \{p : \quad h^{(r)} - A^{(r)'} p \in C_r^*\}$.

**Theorem 4.** If there exists a probability vector $p^*$ such that

$$p_i^* = \lambda^* \pi_i \prod_{r=1}^{l} \prod_{j=1}^{s_r} {\lambda_j^*}^{(r)\, a_{ij}^{(r)}}, \quad ({\lambda^*}_1^{(r)}, \cdots, {\lambda^*}_{s_r}^{(r)})' \in C_r; \tag{10}$$

$$h^{(r)} - {A'}^{(r)} p^* \in C_r^*; \tag{11}$$

$$\sum_{j=1}^{s_r} [h_j^{(r)} - \sum_{i=1}^{k} a_{ij}^{(r)} p_i^*] \log {\lambda^*}_j^{(r)} = 0 \tag{12}$$

## UGIS Algorithms for intersection

Initially, set $p_i^{(0)} = \pi_i \prod_{r=1}^{l} \prod_{j=1}^{s_r} (\lambda_j^{(0,r)})^{a_{ij}^{(r)}}$ where $\lambda^{(0,r)} \in C_r$ for $r = 1, \cdots, l$.

**Step(n, 1)**

$$p_i^{(n,1)} = p_i^{(n-1,l)} \prod_{j=1}^{s_1} (\lambda_j^{(n,1)} / \lambda_j^{(n-1,1)})^{a_{ij}^{(1)}},$$

$$\vdots$$

**Step(n, r)**

$$p_i^{(n,r)} = p_i^{(n,r-1)} \prod_{j=1}^{s_r} (\lambda_j^{(n,r)} / \lambda_j^{(n-1,r)})^{a_{ij}^{(r)}},$$

## Step(n, l)

$$p_i^{(n,l)} = p_i^{(n,l-1)} \prod_{j=1}^{s_l} (\lambda_j^{(n,l)} / \lambda_j^{(n-1,l)})^{a_{ij}^{(l)}},$$

where $\lambda^{(n,t)} = P_{w^{(n,t)}}(U^{(n,t)} | C_t)$, $\lambda_j^{(n,t)}$ is the $j$-th element of $\lambda^{(n,t)}$,

$$w^{(n,t)} = (\sum_{i=1}^{k} a_{i1}^{(t)} p_i^{(n,t-1)}, \cdots, \sum_{i=1}^{k} a_{i\,s_t}^{(t)} p_i^{(n,t-1)})',$$

and

$$U^{(n,t)} = (h_1^{(t)} \lambda_1^{(n-1,t)} / w_1^{(n,t)}, \cdots, h_{s_t}^{(t)} \lambda_{s_t}^{(n-1,t)} / w_{s_t}^{(n,t)})'.$$

**Theorem 5.**

$\{p^{(n,r)}\}$ given in the above proposed algorithm converges to the optimal solution to (9).

# Outline

# Marginal stochastic order

Let $pr = (pr_1, \cdots, pr_k)'$ and $pc = (pc_1, \cdots, pc_k)'$ be given probability vectors with $pr_i > 0$ and $pc_i > 0$ for $i = 1, \cdots, k$. Let $D$ be the set of discrete probability measures $\{p_{ij}\}$ defined by

$$\sum_{i=1}^{l} p_{i+} \leq \sum_{i=1}^{l} pr_i, \quad l = 1, \cdots, k-1; \sum_{i=1}^{k} p_{i+} = \sum_{i=1}^{k} pr_i = 1, \quad (13)$$

and

$$\sum_{j=1}^{l} p_{+j} \leq \sum_{j=1}^{l} pc_j, \quad l = 1, \cdots, k-1; \sum_{j=1}^{k} p_{+j} = \sum_{i=1}^{k} pc_i = 1. \quad (14)$$

Now consider the following problem

$$\min \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij} \log p_{ij}/\pi_{ij} \text{ subject to } p \in D. \tag{15}$$

Let $p = (p_{11}, p_{12}, \cdots, p_{1k}, \cdots, p_{k1}, \cdots, p_{kk})'$,

$$A^{(1)} = I \otimes (1, \cdots, 1)' \quad \text{and} \quad A^{(2)} = (1, \cdots, 1)' \otimes I,$$

and then

$$D = \{p : \ pr - A^{(1)'} p \in C^*\} \bigcap \{p : \ pc - A^{(2)'} p \in C^*\}$$

where

$$C^* = \ \{y \in R^s : \sum_{j=1}^{l} y_j \geq 0 \ \text{ for } l = 1, \cdots, k-1 \text{ and } \sum_{j=1}^{k} y_j = 0\}.$$

## Testing

We can also test the marginal probabilities equal to some given probabilities against stochastic marginal order based on Kullback-Leibler information. One can refer to the paper (Gao and Kuriki 2006, JMVA).

# Outline

1.Agresti, A. (2002). *Categorical Data Analysis* . John Wiley & Sons, Inc..
2. Agresti, A. And Coull, B. A.(2002). The analysis of contingency tables under inequality constraints. **J. Statist. Plann. Inference** 107 45-73.
3. Bhattacharya, B. (2006). An Iterative Procedure for General Probability Measures to obtain I-Projection onto Intersection of Convex Sets. **Ann. Statist.** 34, 878-902.
4. Darroch, J. N. and Ratcliff, D. (1972). Generalized iterative scaling for loglinear models. **Ann. Math. Statist.** 43, 1470-1480.
5. Gao, W., Shi, N.-Z. and et.al (2010). Unified generalized iterative scaling and its application. **Computational Statistics and Data Analysis** 54, 1066-1078.
6. Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*, John Wiley and Sons: New York.

Thank you very much!