Analysis of Association on Non-Product Spaces

Anna Klimova

IST Austria

September 11, 2013

Joint work with Tamás Rudas

Outline

- **Motivation.**
- **Relational models without the overall effect.**
- **Properties of the MLE.**
- **Computation of the MLE.**

Non-product sample spaces

- **A sample space is a proper subset of the Cartesian product of the ranges of the variables of interest (structural zeros - combinations that do not exist logically or in a particular population)**

- **Patterns of participation in waves of a panel study**
- **Lists of traffic violations**
- **Market basket analysis**
- **Congenital malformations**

Data: Indicators of Features

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
1	1	0	1	1	1
2	1	1	1	0	1
3	1	1	1	0	0
4	1	1	0	0	0
5	1	0	1	1	1
6	1	0	0	0	0
7	0	1	1	1	0
8	0	1	0	1	0
9	0	1	1	0	1
10	0	1	1	0	1

At least one feature is present.

Independence of Malformations

- Do malformations X_1 and X_2 occur independently of each other?

X_2	X_1 No	X_1 Yes
No	-	p_{01}
Yes	p_{10}	p_{11}

The model of independence: $\quad p_{11} = p_{01}p_{10}.$

A.Klimova, T.Rudas, A.Dobra (2012).
Relational Models for Contingency Tables.
J. Multivariate Anal., 104, 159–173.

Observed Data

Patient	X_1	X_2
1	0	1
2	1	1
3	1	0
4	1	0
5	1	1
6	1	0
7	1	0
...	...	

X_2	X_1 No	X_1 Yes
No	-	47
Yes	42	13

Relational Model

The model of independence: $\mathbf{p}_{11} = \mathbf{p}_{01}\mathbf{p}_{10}$.

Generating subsets: X_1 is present; X_2 is present.

Model matrix: $\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$.

Log-linear representation: $\log \mathbf{p} = \mathbf{A}'\boldsymbol{\beta}$, where $\boldsymbol{\beta} = exp(\boldsymbol{\theta})$.

Multiplicative representation:
$p_{01} = \theta_1, p_{10} = \theta_2, p_{11} = \theta_1\theta_2$.

Under such a model, there is no parameter that is common to every cell in the table. This is a model without the overall effect.

MLE in Curved Families

- **Assume that the observed distribution $\mathbf{q}$ is positive.The MLE $\hat{\mathbf{p}}$ exists, and it is the unique solution of the system:**

$$\begin{aligned} \mathbf{A}\hat{\mathbf{p}} &= \gamma \mathbf{A}\mathbf{q}, \\ p_{01}p_{10} &= p_{11}, \\ p_{11} + p_{01} + p_{10} &= 1. \end{aligned}$$

- **Here $\gamma = \gamma(\mathbf{q})$ is an adjustment factor.**

- **The mean-value parameters of the MLE are proportional to those of the observed distribution. (For regular exponential families, they are equal!).**

How IPF works

- Starts with a distribution $\mathbf{p}^{(0)}$ in the model: $p_{01}^{(0)} p_{10}^{(0)} = p_{11}^{(0)}$.
- Rescales the components of $\mathbf{p}^{(n)}$ according to the values of $A_j\mathbf{q}$, where A_j are the rows of $\mathbf{A}$:

$$p_{ij}^{(n)} = p_{ij}^{(n-1)} \left(\frac{A_j \mathbf{q}}{A_j \mathbf{p}^{(n-1)}} \right)^{a_{ji}}.$$

- The sequence $\mathbf{p}^{(n)}$ converges to a $\mathbf{p}^*$ that satisfies

$$\mathbf{A}\mathbf{p}^* = \mathbf{A}\mathbf{q}, \;\; p_{01}^* p_{10}^* = p_{11}^*.$$

- If $p_{01}^* + p_{10}^* + p_{11}^* = 1$, then $\mathbf{p}^* = \hat{\mathbf{p}}$ is the MLE.

▶ Can this procedure be modified to include the adjustment factor: $\mathbf{A}\hat{\mathbf{p}} = \gamma \mathbf{A}\mathbf{q}$? Is it implied that $p_{01}^* + p_{10}^* + p_{11}^* = 1$?

G-IPF Algorithm (Klimova and Rudas, 2013)

- Select a value $\tilde{\gamma}$ of the adjustment factor.
- Choose a $\mathbf{p}^{(0)}$ in the model: $p_{01}^{(0)} p_{10}^{(0)} = p_{11}^{(0)}$.
- Rescale the components of $\mathbf{p}^{(n)}$:

$$p_{ij}^{(n)} = p_{ij}^{(n-1)} \left(\tilde{\gamma} \frac{A_j \mathbf{q}}{A_j \mathbf{p}^{(n-1)}} \right)^{a_{ji}}.$$

- Then $\mathbf{p}^{(n)} \to \tilde{\mathbf{p}}$ that satisfies $\mathbf{A}\tilde{\mathbf{p}} = \tilde{\gamma}\mathbf{A}\mathbf{q}, \;\; \tilde{p}_{01}\tilde{p}_{10} = \tilde{p}_{11}$.
- If $\tilde{p}_{01} + \tilde{p}_{10} + \tilde{p}_{11} = 1$, then $\tilde{\mathbf{p}}$ is the MLE.
- Otherwise, choose a smaller or a larger $\tilde{\gamma}$ depending on whether $\tilde{p}_{01} + \tilde{p}_{10} + \tilde{p}_{11} > 1$ or $\tilde{p}_{01} + \tilde{p}_{10} + \tilde{p}_{11} < 1$.

G-IPF (Klimova and Rudas, 2013)

Iterative Scaling in Curved Exponential Families.
arXiv: 1307.3282

- **The algorithm converges to the MLE.**
- **The proof of convergence is based on Bregman divergence (generalization of Kullback-Leibler divergence).**
- **R-package gIPFrm.**

X_2	X_1: **No**	X_1: **Yes**
No	-	47(44.872)
Yes	42(39.679)	13(17.456)

The adjustment factor = 1.039. **P-value = 0.24.**