2003-07-07/10 IMPS-2003, Chia Laguna Hotel, Cagliari, Italy
The 13th International Meeting and the 68th Annual American meeting of the Psychometric Society

Ability of Correspondence Analysis to Find Latent Structures Comparison with Association Model

Tatsuo OTSU
Research and Development Division
The National Center for University Entrance Examinations
otsu@rd.dnc.ac.jp

0. Outline of TalK

Partly based on "A Practical Introduction to Survey Data Analysis" in Frontiers in Statistical Sciences, 10 Takemura,A. et al. eds. (2003) Tokyo: Iwanami (in Japanese).

1. Correspondence Analysis (CA)
   Pitfalls in dimension reduction

2. The Structre of Multivariate Normal Distributions (MVN)

3. Goodman's Association Model
   Compatibility with MVN structure

4. Tame Data and Wild Data
   Difficulties of Association Model in Large Tables

1. Correspondence Analysis (CA): Pitfalls in Dimension Reduction

Guttman (1950) showed CA Solutions for Artificial Data.

Some Japanese authors also showed solutions on several artificial data.
Kusunoki,T.(1986), Iwatsubo,S. (1987), Otsu,T. (1990), Okamoto,M.(1993)

An Artificial Example (Triangular Case)

The Data has Essentially One dimensional Structure.

	A	B	C	D	E
a	1	1	1	1	1
b	0	1	1	1	1
c	0	0	1	1	1
d	0	0	0	1	1
e	0	0	0	0	1

The Solutions are High Order Polynomials.

Singular Values of CA

$$\rho_i = \frac{1}{i+1}, \quad (i = 0, 1, 2, ..., N-1),$$

where N is the size of the square matrix.

The scores for rows (and columns)

The ith solution is an ith order polynomial. The score for the k-th column is given by

$$p_i^N(k) = \sum_{j=0}^{i} (-1)^j \frac{1}{j+1} \begin{pmatrix} i \\ j \end{pmatrix} \begin{pmatrix} i+j+1 \\ j \end{pmatrix} \frac{(k-1)^{(j)}}{(N)^{(j)}},$$

$$\text{for} \quad (i = 0, ..., N-1; k = 1, ..., N),$$

where $(x)^{(y)}$ shows $x \times (x-1) \cdots (x-y+1)$. (Otsu,1990).

Case of N=5 (first scores are set to be 1)

Dim 1	1.000	0.6250	0.2500	-0.1250	-0.5000
Dim 2	1.000	0.0000	-0.4444	-0.3333	0.3333
Dim 3	1.000	-0.8750	-0.2500	0.6875	-0.2500
Dim 4	1.000	-2.0000	2.0000	-1.0000	0.2000

Another Example (Diagonal Band matrix)

	A	B	C	D	E	F
a	1	1	0	0	0	0
b	0	1	1	0	0	0
c	0	0	1	1	0	0
d	0	0	0	1	1	0
e	0	0	0	0	1	1

$N \times (N+1)$ table.

Singular Values $\rho_i = \cos\left(\frac{i\pi}{2N}\right)$, for $i = 0, 1, ..., N-1$, where N shows the row size.

Row Scores

$$p_i^N(k) = \cos\left(\frac{i\pi(2k-1)}{2N}\right), \text{ for } k = 1, ..., N; i = 0, 1, ..., N-1$$

(c.f. $\cos(i\theta)$ is a ith order polynomial of $\cos\theta$.)

Clear one dimension structure brings slowly decreasing singular values. We term them Resonances.

Question:

Do they have practical meaning in data analysis?

Guttman termed them *Intensity* (quadratic) and *Closure* (cubic), and suggested their supplemental use for data analysis.

Clear One Dimensional Latent Structure $\Rightarrow$ Polynomial (like) solutions

Is the **reverse** true ?

Answer: No. If there were vague latent factors, they might not be detected by CA. Resonances may disturb detection of less prominent latent factors.

2. The Structure of Multivariate Normal Distributions (MVN)

The structure of CA is more clearly explaind in MVN cases.

Lancaster,H.O.(1957) pointed out an important property of two variate normal distribution and Hermite polynomials.

Hermite Polynomials

$$H_m(x) = \frac{(-1)^m}{N(x)} \frac{d^m}{dx^m} N(x),$$

where $N(x)$ shows standared normal distribution. (c.f. Usual notation of H is slightly different in scale.)

$$\begin{aligned} H_0(x) &= 1 \\ H_1(x) &= x \\ H_2(x) &= x^2 - 1 \\ H_3(x) &= x^3 - 3x \end{aligned}$$

Properties of Hermite Polynomials

$$E[H_m(X)] = 0, (m = 1, 2, ...)$$

and

$$E[H_m(X)H_s(X)] = m!\delta_{ms}, (m, s = 0, 1, 2, ...) \ ,$$

where δ shows Kronecker delta.

Its norm standardized version is

$$G_m(x) = \frac{1}{\sqrt{m!}} H_m(x). \tag{1}$$

For MVN,

$$\mathrm{Corr}[G_m(X_i), G_s(X_j)] = \delta_{ms}\rho_{ij}^m, \ (m = 0, 1, ...) \tag{2}$$

holds.

Suppose the observed variables are the polynomial transformations of the latent MVN variables.

$$Y_j = \sum_{m=1}^{M} a_{jm} G_m(X_j), \ (j = 1, 2) \tag{3}$$

$$E[Y_j] = 0, \ \mathrm{Var}[Y_j] = \sum_m a_{jm}^2 = 1$$

$$\mathrm{Corr}[Y_1, Y_2] = r_{12} = \sum_m a_{1m} a_{2m} \rho^m$$

If $0 < \rho < 1$ holds, then the maximum correlation is given by $Y_1 = X_1$, $Y_2 = X_2$, and $r = \rho$.

CA is approximately optimal nonlinear inverse function search (restricted to discrete values).

One dimensional CA is a process for inquiring $g_1^{-1} = h_1$ and $g_2^{-1} = h_2$ that lead to the maximum correlation between

$$g_1^{-1}(Y_1) = X_1 \text{ and } g_2^{-1}(Y_2) = X_2.$$

If the latent structre is MVN, X_1 and X_2 are approximately recoverd by CA.

Resonances in MVN

$$\text{Corr}[G_1(X_i), G_1(X_j)] = \rho_{ij},\ \text{Corr}[G_2(X_i), G_2(X_j)] = \rho_{ij}^2,\ \ \cdots$$

$\{G_1(X_i), G_1(X_j)\}, \{G_2(X_i), G_2(X_j)\}, \cdots$ are approximately CA solutions, and their correlations are $\rho_{ij}, \rho_{ij}^2, \cdots$.

If ρ is high, the resonances decrease slowly.

3. Goodman's Association Model

Goodman proposed association model (and correlation model) for two-way tables. (Goodman, 1985,1991).

Association model :

1. Log-linear based model with restricted interactions.
2. ML estimation based on multinomial distribution.
3. Row and column scores are subject to marginal weighted standardized constraints. They are the same as the constraints in CA.

$$\sum_i x_i P_{i+} = 0, \ \sum_i x_i^2 P_{i+} = 1, \ \sum_j y_i P_{+j} = 0, \ \sum_i y_j^2 P_{+j} = 1$$

4. Theoretically preferable properties, especially in latent MVN cases.

Association Model:

Cell Probabilities are modeled as

$$\pi_{ij} = \alpha_i \beta_j \exp\left(\sum_{m=1}^{M} \phi_m x_{im} y_{jm}\right), \text{ for } (i = 1, ..., I;\ j = 1, ..., J), \quad (4)$$

where x_{im} is the score for the ith row in the mth dimension, and y_{jm} is for the jth column.

Standard constraints that are similar to CA are proposed on (x_i) and (y_j).

If $\log \alpha_i \propto -x_{im}^2$ and $\log \beta_j \propto -y_{jm}^2$ hold,
π_{ij} are approximately two variate MVN density.

In this case, $\phi_m \sim \rho_m/(1 - \rho_m^2)$ holds.

Questions:

1. Usually, one dimensional scores are similar to CA solutions.
2. Is it really effective in practical data anlaysis? When does ciritical difference between the methods appear?

One Answer:

Resonaces in latent MVN cases are different. Multidimensional solutions may lead to different interpretation.

Example1 : Two variable case

The correlation in MVN is $\rho = 0.6$. Values are breaked into 10 categories in balanced frequencies.

The table size is 10×10. Cell probabilities are obtained by a numerical integration software (mvndstpack) (Genz, 1992).

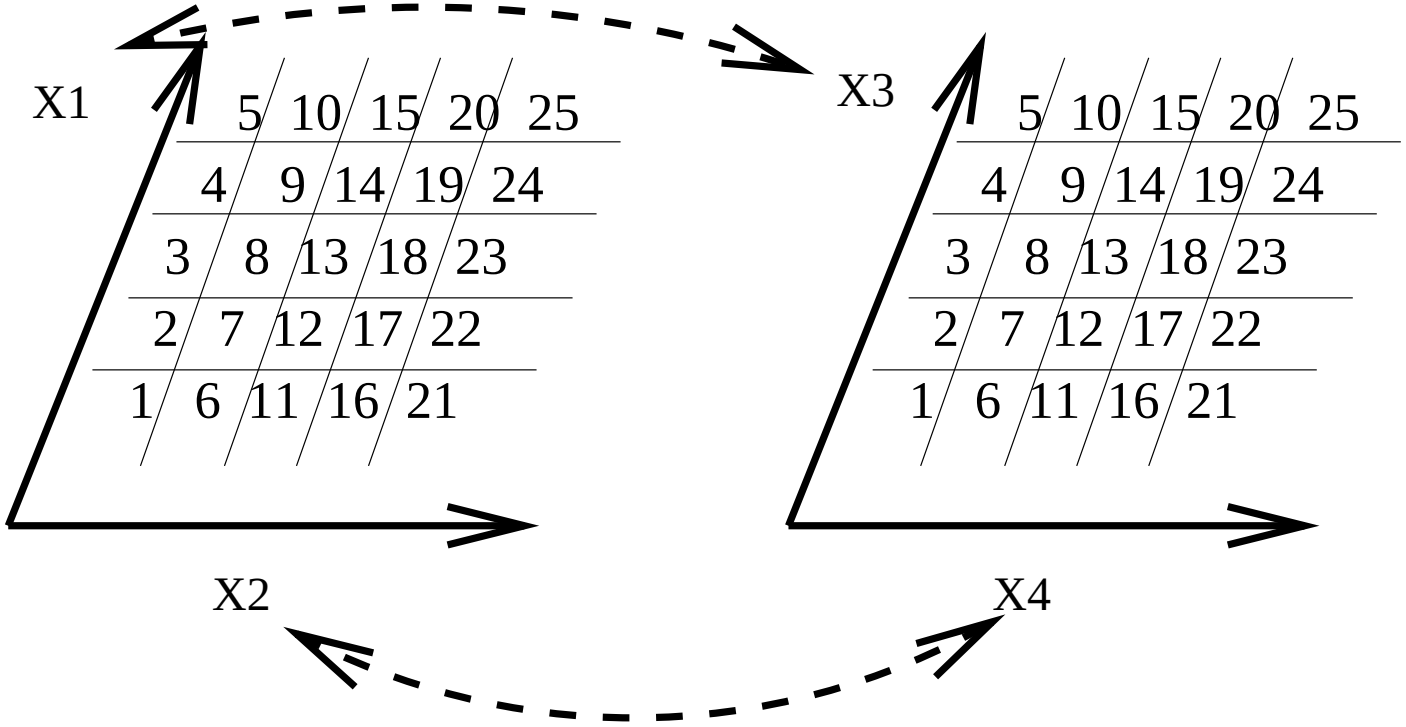
Association model estimation was computed by a a program on R that uses quasi-Newton multiplier method (a kind of nonlinear constrained optimisation).

Solutions of CA and Association Model(AssocM) ($M = 3$)
two variate normal case(10×10, $\rho = 0.6$)

		Dim		
Model		1	2	3
CA	r_k	0.580	$\underline{0.277}$	0.088
	$r_k/(1 - r_k^2)$	0.875	$\underline{0.300}$	0.089
AssocM	ϕ_k	0.827	0.024	0.011

Underline shows *resonance*. Association model does not show resonances.

Example 2: A four variable MVN case



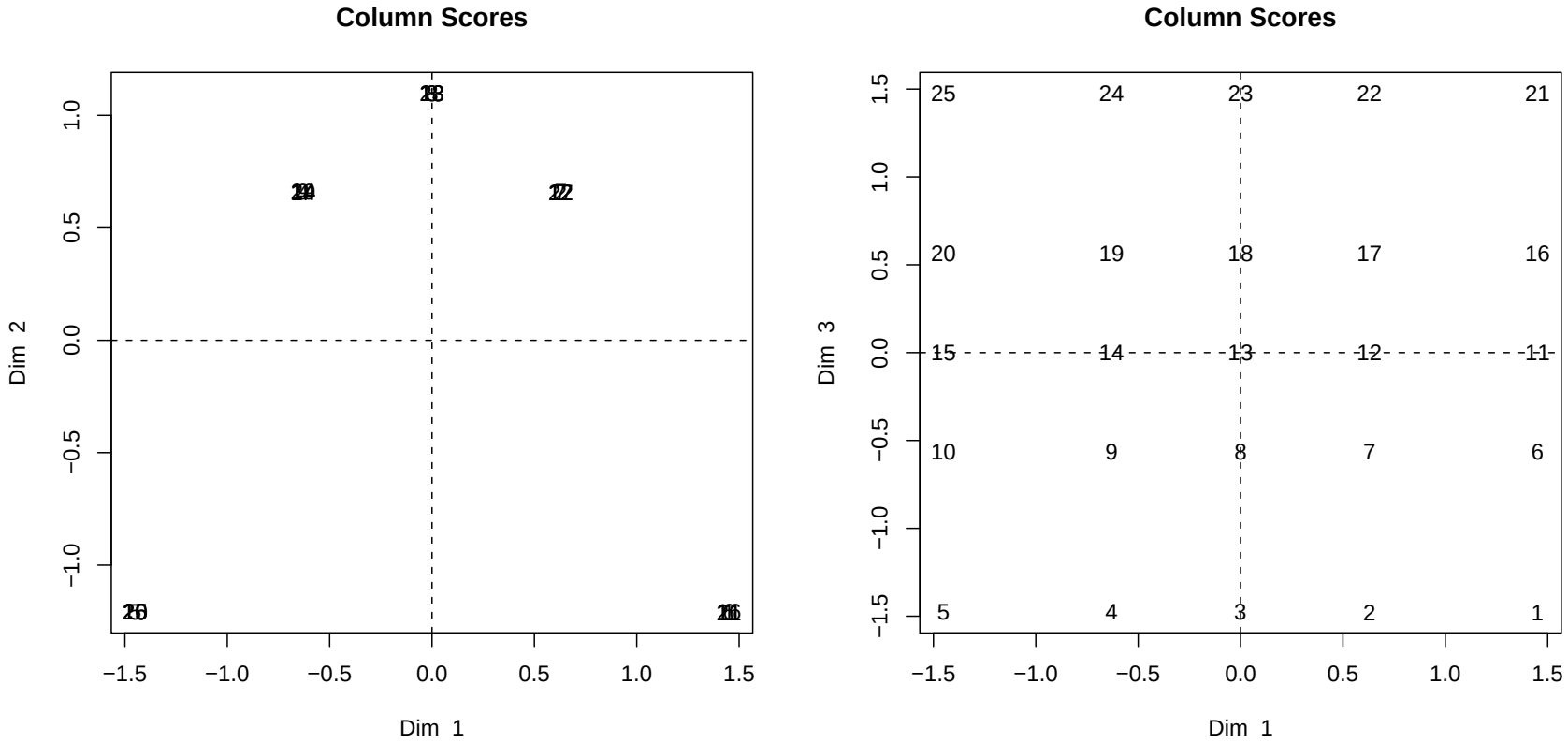
The correlation between X_1 and X_3 is ρ_1, between X_2 and X_4 is ρ_2. Others are independent.

Each variable is breaked into 5 categories. A two-way 25×25 table.

CA and Association Model (AssocM) ($M = 3$) Four Variate Normal Case (25×25, $\rho_1 = 0.7$, $\rho_2 = 0.2$)

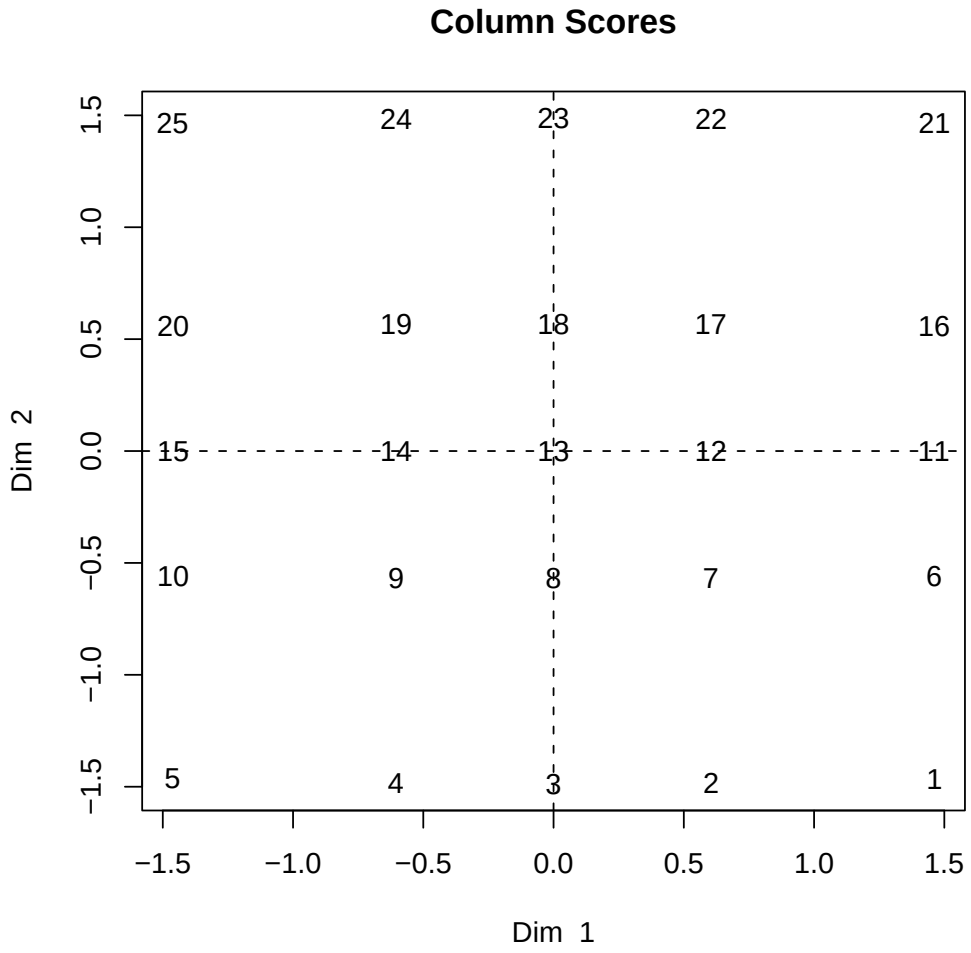
		Dim		
Model		1	2	3
CA	r_k	0.648	<u>0.296</u>	0.180
	$r_k/(1 - r_k^2)$	1.116	<u>0.325</u>	0.186
AssocM	ϕ_k	0.974	0.185	0.058

Underline shows *resonance*.



Scores by CA

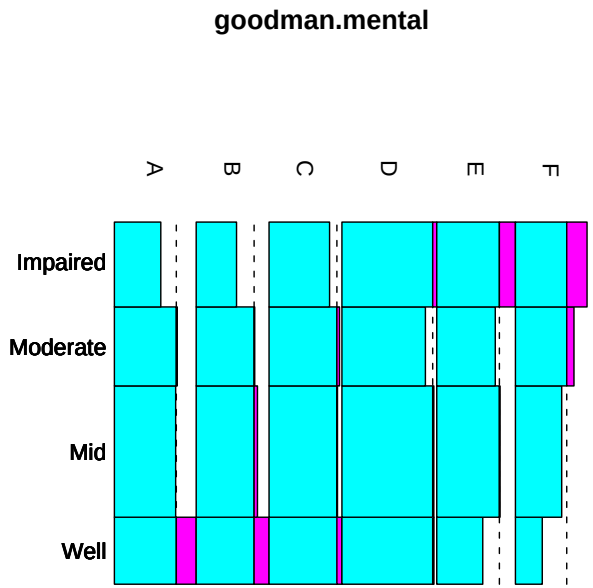
The quadracic polynomial hides the true second dimension.



Scores by Association model

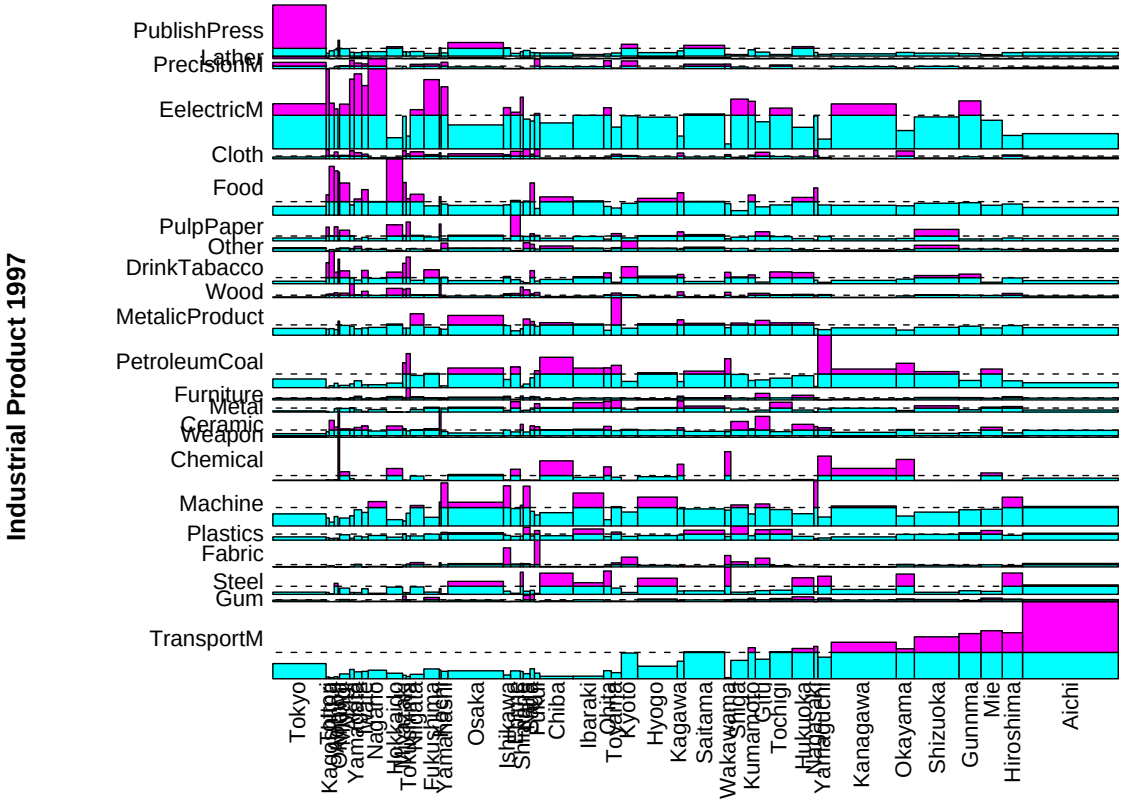
4. Tame Data and Wild Data

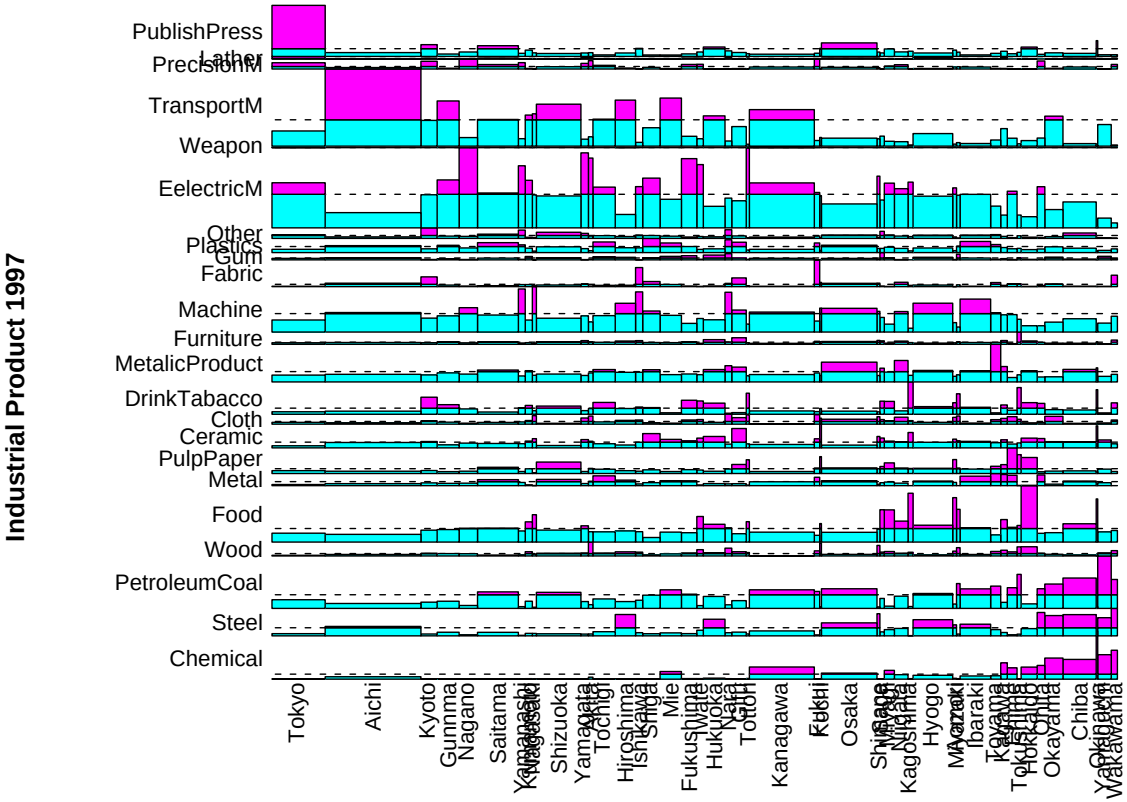
Examples in Goodman's papers are *tame*, i.e. well modeled by CA or AssocM. (Table2, Goodman 1985). Association model ($M = 1$) fits well ($\chi^2 = 3.56$ for $df = 8$).



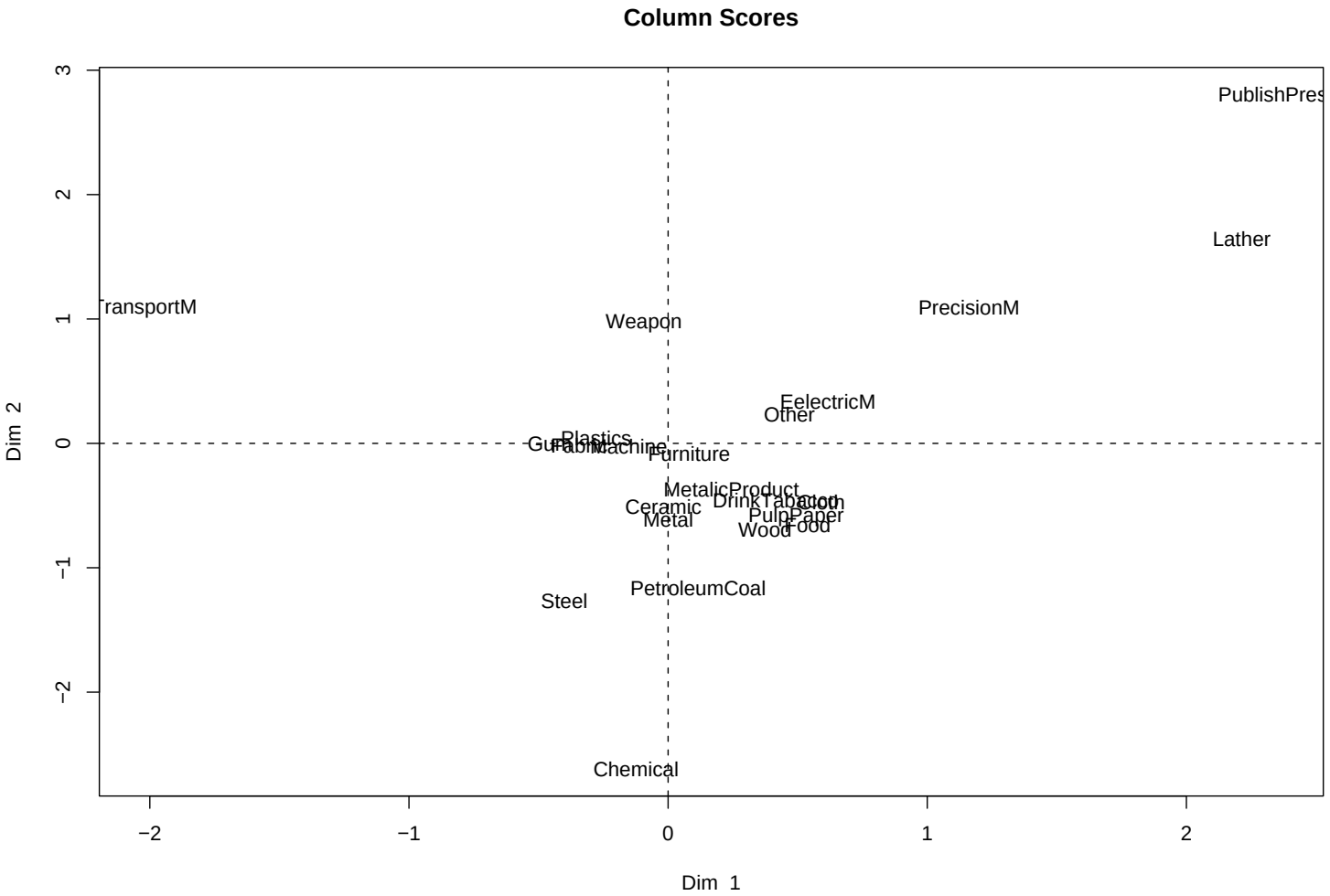
Graphical method based on Jacques Bertin (1984 Englsh trs.).

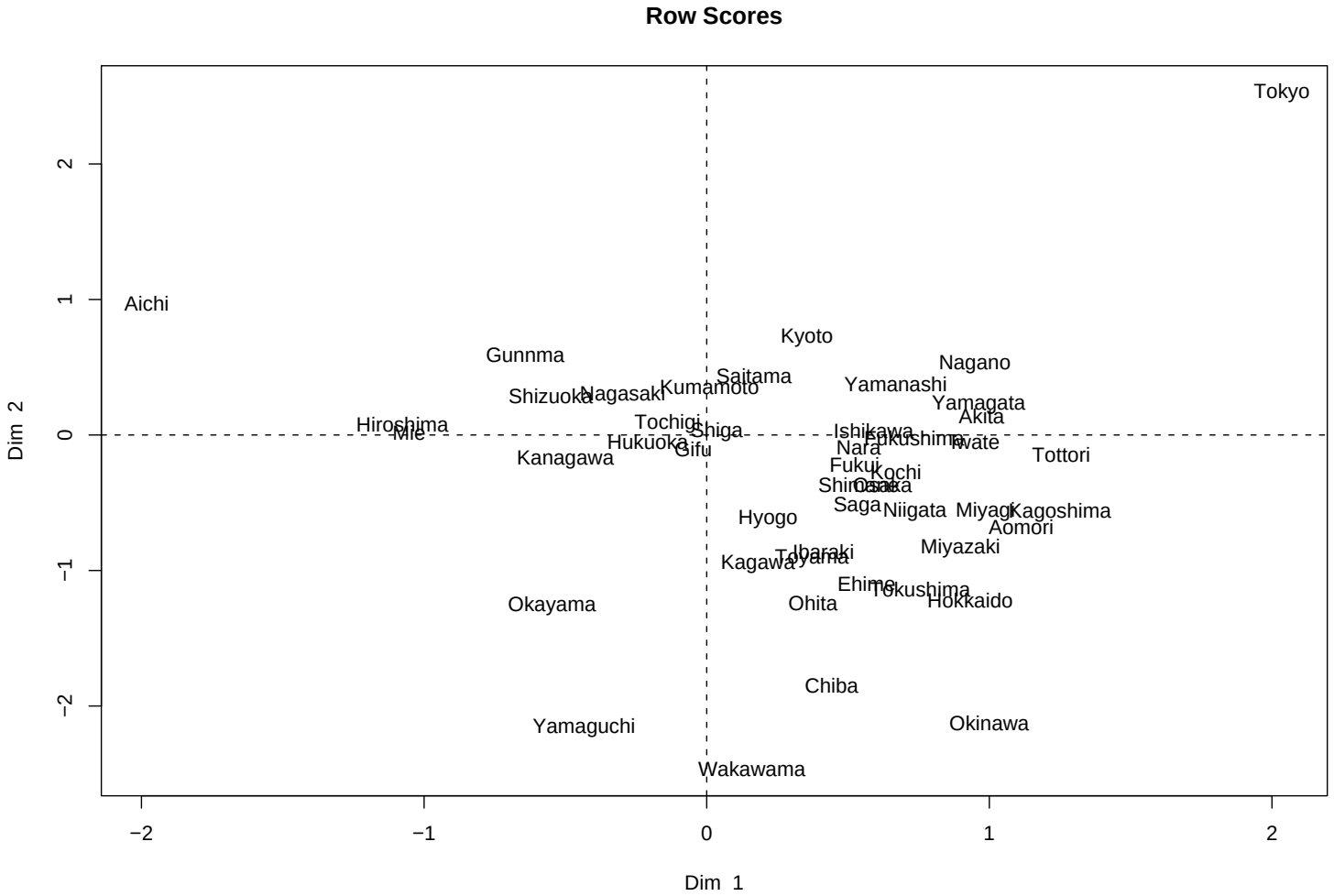
A "Wild" Example: Row and column orders are based on CA.





Row and Columns orders are based on the 2nd scores of CA


Column Scores
PublishPres
Lather
TransportM
Weapon
PrecisionM
FelectricM
Other
Plastics
Furniture
MetalicProduct
Ceramic
Metal
Wood
Food
PetroleumCoal
Steel
Chemical
Dim 2
Dim 1
3
2
1
0
−1
−2
−2
−1
0
1
2


Row Scores
Tokyo
Aichi
Kyoto
Gunnma
Nagano
Saitama
Yamanashi
Shizuoka
Nagasaki
Kumamoto
Yamagata
Akita
Hiroshima
Tochigi
Mie
Shiga
Ishikawa
Fukushima
Iwate
Hukuoka
Nara
Tottori
Gifu
Kanagawa
Fuku
Kochi
Saga
Niigata
Miyagi
Kagoshima
Hyogo
Aomori
Miyazaki
Ibaraki
Kagawa
Yama
Ehime
Tokushima
Okayama
Ohita
Hokkaido
Chiba
Yamaguchi
Okinawa
Wakawama
Dim 2
Dim 1
−2
−1
0
1
2

Singular values in the "wild" example

	1	2	3	4	5
ρ_i	0.377	0.348	0.282	0.224	0.208

Serious Necessity for CA/AssocM is in Wild Data Analysis.

Difficulties

Association model is hard to compute with many local maxima on wild data. Simplicity of CA is adequate for wild data, though it might miss latent structures.

Simple linear methods (CA, PCA) have *Incremental Stability*. Model expansion (dimension enlargement) holds solutions.

More sohpisticated methods (Association Model, Factor Analysis) do not (when their fits are not high).

Stable (i.e. orthogonal) coordiate system for representing data is practically strong. Many popular statistical methods have this feature (Balanced Design, FFT, and PCA).

Log-Linear model based method is attractive for multiway data even if the data are not Poisson (or multinomial). Are there Good expantions? Greenacre's method (IFCS-2002). A good candidate?

Conclusion

Nominal Category Quantification : A Deep Problem

It seems to be necessary to use rich sematic constraints on data.

Simple and tough methods that skillfully interact with analysts.

Another direction: More extensive use of meta-data by more intelligent methods that behave like experienced data analysts.

References

Bertin,J. (1984) English translation by Berg and Scott. *Graphics and Grpahics Information Processing*, Walter de Gruyter.

Genz,A. (1992). *Jour. Comutational Graphical Statistics*, **1**, 141-149.

Goodman,L. (1985). *Ann. Stat.*, **13**, 10-69.

Goodman,L. (1986). *Intern. Stat. Rev.*, **54**, 243-309.

Goodman,L. (1991). *J. Amer. Stat. Assoc.* (with discussions), **86**, 1085-1138.

Guttman,L. (1950). in Souffer,S.A. et al. (eds) *Measurement and Prediction*, Princeton Univ. Press

Iwatsubo,S. (1987). *Sūryōkahō no Kiso (Fundamentals of Quantification)* (in Japanese). Tokyo: Asakura.

Kusumoki,T. (1986). *Japanese Jour. Behaviormetrics* (in Japanese), **13(2)**, 8-19.

Lancaster,H.O. (1958). *Ann. Math. Stat.*, **29**, 719–736.

Okamoto,M. (1993). *Mathematica Japonica*, **39**, 523-535.

Otsu,T. (1990). *Behaviormetrika*, **28**, 37-48.