Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Statistical Planning and Inference 134 (2005) 26–35
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Influence function and correspondence analysis

Emmanuel Nowak[a, b], Avner Bar-Hen[c, d, *]

[a] *USTL, Stat & Prob., FRE-CNRS 2222, 59655 Villeneuve d'Ascq Cedex, France*
[b] *ISA, 41, rue du Port, 59046 Lille Cedex, France*
[c] *UMR INAPG-INRA BIA-ENGREF 518, 16, rue Claude Bernard, 75231 Paris, France*
[d] *Université Aix-Marseille III, IMEP case 451, 13397 Marseille Cedex 20, France*

Received 4 February 2003; accepted 23 February 2004
Available online 23 July 2004

Abstract

Correspondence analysis appears to be sensitive to modalities with low margins leading to outliers. We use the notions of influence function and Hadamard differentiability to provide a criterion for deciding when such an outlier can be considered as too influential.

MSC: 62H25; 62F35

Keywords: Correspondence analysis; Influence function; Hadamard differentiability

1. Introduction

Correspondence analysis consists primarily of techniques that display the rows and columns of a two-way contingency table on a sub-space. The idea is to summarize the data, losing in the process as little information as possible. Many computational techniques and mathematical structures of correspondence analysis are similar to those of principal component analysis. In principal component analysis, the distance between observations corresponds to a Euclidean distance and in correspondence analysis the distance between modalities corresponds to a chi-square distance. An innate property within the construction of a chi-square distance is a sensitivity to modalities with low marginals leading to outliers. An outlier can by itself determine an entire component.

* Corresponding author.
E-mail address: avner@bar-hen.net (A. Bar-Hen).

0378-3758/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2004.02.013

The influence function (see Hampel, 1974), in Statistics, has two main uses. First, it allows to assess the relative influence of individual observations toward the value of an estimate. In correspondence analysis, Kim (1992) has calculated the influence function for eigenvalues. Pack and Jolliffe (1992) measure the influence for eigenvectors and propose various synthetical indices. These results allow to describe the impact of a modality but does not provide a decision rule: when can we say that a low margin modality is too influential? Also, the influence curve allows assessment of asymptotic properties of an estimate: under standard conditions, this one is asymptotically normal and the asymptotic variance can be expressed in terms of the influence function (Huber, 1981).

The purpose of this paper is to use this aspect and the delta-method established by Gill (1989), to obtain a central limit theorem (CLT) for eigenvalues in correspondence analysis. This will lead to a decision criterion for influential modalities detection.

2. Influence function

Let $X_1, \ldots, X_n$ be random variables with common distribution function (df) F on $\mathbb{R}^d$ ($d \geqslant 1$). To simplify notations, we will confuse distribution function and probability measure: F is either one or the other. Suppose that we are interested in a parameter that can be expressed, as often in statistics, as a functional $T(F)$ of the generating df, T being defined at least on the space $\mathscr{F}$ of df's. The natural estimator is $T(F_n)$ where F_n is the empirical df, defined by

$$F_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

with δ_{X_i} the point mass 1 at X_i.

To evaluate the importance of an additional observation $x \in \mathbb{R}^d$, we can define, under condition of existence, the quantity

$$IC_{T,F}(x) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)}{\varepsilon} \tag{1}$$

which measures the influence of an infinitesimal perturbation along the direction δ_x (see Hampel, 1974). The influence function (or influence curve) $IC_{T,F}(\cdot)$ is defined pointwise by (1), if the limit exists for every x.

There is strong connection between influence function and jackknife (Miller, 1974): let $F_{n-1}^{(i)} = 1/(n-1)\sum_{j,j\neq i}\delta_{x_j}$, then $F_n = (n-1)/n F_{n-1}^{(i)} + \frac{1}{n}\delta_{x_i}$. If $\varepsilon = -1/(n-1)$, we have

$$\begin{aligned} IC_{T,F_n}(x_i) &\approx \frac{T((1-\varepsilon)F_n + \varepsilon\delta x_i) - T(F_n)}{\varepsilon} \qquad (2)\\ &= (n-1)(T(F_n) - T(F_{n-1}^{(i)})) \\ &= T_{n,i}^* - T(F_n). \qquad (3) \end{aligned}$$

$T_{n,i}^*$ are the pseudo-values of the jackknife (i.e. values computed on $n-1$ observations), (Miller, 1974).