Inverse correspondence analysis

Groenen, PJF; van de Velden, M

Published in:
Linear Algebra and Its Applications

DOI:
10.1016/j.laa.2003.10.016

Available online at www.sciencedirect.com
SCIENCE DIRECT®
LINEAR ALGEBRA AND ITS APPLICATIONS
ELSEVIER Linear Algebra and its Applications 388 (2004) 221–238
www.elsevier.com/locate/laa

Inverse correspondence analysis

Patrick J.F. Groenen [a,*], Michel van de Velden [b]

[a] *Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands*
[b] *Department of Marketing and Marketing Research, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands*

Received 12 September 2002; accepted 21 October 2003

Submitted by H.J. Werner

Abstract

In correspondence analysis (CA), rows and columns of a data matrix are depicted as points in low-dimensional space. The row and column profiles are approximated by minimizing the so-called weighted chi-squared distance between the original profiles and their approximations, see for example, [Theory and applications of correspondence analysis, Academic Press, New York, 1984]. In this paper, we will study the *inverse CA* problem, that is, the possibilities for retrieving one or more data matrices from a low-dimensional CA solution. We will show that there exists a nonempty closed and bounded polyhedron of such matrices. We also present two algorithms to find the vertices of the polyhedron: an exact algorithm that finds all vertices and a heuristic approach for larger sized problems that will find some of the vertices. A proof that the maximum of the Pearson chi-squared statistic is attained at one of the vertices is given. In addition, it is discussed how extra equality constraints on some elements of the data matrix can be imposed on the inverse CA problem. As a special case, we present a method for imposing integer restrictions on the data matrix as well. The approach to inverse CA followed here is similar to the one employed by De Leeuw and Groenen [J. Classification 14 (1997) 3] in their inverse multidimensional scaling problem.

Keywords: Correspondence analysis; Inverse problems; Maximum chi-square

* Corresponding author.
E-mail address: groenen@few.eur.nl (P.J.F. Groenen).

0024-3795/$ - see front matter
doi:10.1016/j.laa.2003.10.016

1. Introduction

In correspondence analysis (CA), the rows and columns of a data matrix **F** are depicted as points in low-dimensional space. Most often, **F** is a contingency matrix, but this need not be the case. The only restriction on **F** is that its elements are nonnegative. A CA solution is obtained by simultaneously approximating the row and column profiles through minimization of the so-called chi-squared distance. It is well known that the CA solution for both the rows and columns can be obtained immediately from the singular value decomposition of the scaled data matrix.

Much is known about the properties of CA (see, for example [4,5,10]). In this paper, we concentrate on a problem that has not been treated before. Given a low dimensional CA solution, which matrices **F** would have produced the current solution as a CA solution? We call this problem *the inverse CA problem*.

There are several reasons to investigate the inverse CA problem. First of all, the size of the set of matrices **F** may reveal information about the uniqueness of the original solution. If this set is large, then there are many nonnegative matrices **F** that yield the same CA solution. Thus, even though the data have lead to a perfectly normal CA solution, it is good to realize that there are many other data sets that would have led to exactly the same solution. On the other hand, if the set is small, there are far fewer nonnegative matrices **F** yielding the solution of the original problem. In particular, if the set only consists of the original data, then we know that there is a unique relation between the CA solution and the data. Second, when CA solutions are reported in the literature, the original data are not always presented. The solution of the inverse CA problem enables us to generate data that has the original CA solution as its CA solution. These generated data can then be used in simulation studies, for example, to investigate stability of the solution or to apply novel extensions to the technique of CA. Thirdly, we believe that the study of inverse CA deepens our understanding of CA. Finally, through inverse CA, we are able to derive the upper bound of the Pearson chi-square given marginal frequencies but unknown data.

To study the inverse CA problem, we will follow a similar approach to the one proposed by De Leeuw and Groenen [2], in their treatment of the inverse multidimensional scaling problem (see also [6]).

This paper is organized as follows. First, we introduce notation for CA. Then we formalize the inverse CA problem. Next, we present a computational method for computing the inverse CA solution. Then, we discuss where the upper bound of the Pearson chi-square statistic is attained. The next section discusses how additional equality and integer constraints can be imposed. We illustrate our method by an example. This paper is ended with some concluding remarks.

2. The correspondence analysis problem

Before we start with the inverse CA problem, let us introduce notation needed for CA. Let **F** denote an $n_r \times n_c$ matrix of nonnegative elements on which CA is performed. Let $\mathbf{r}$ be the vector of row sums of $\mathbf{F}$, that is, $\mathbf{r} = \mathbf{F1}$ and $\mathbf{c}$ the vector of column sums, $\mathbf{c} = \mathbf{F}'\mathbf{1}$, where $\mathbf{1}$ denotes a vector of ones of appropriate length. Furthermore, define n as the sum of all elements of $\mathbf{F}$, that is, $n = \mathbf{1}'\mathbf{F1}$.

Define the scaled data matrix $\widetilde{\mathbf{F}}$ as $\widetilde{\mathbf{F}} = \mathbf{D}_r^{-1/2}\mathbf{F}\mathbf{D}_c^{-1/2}$, where $\mathbf{D}_r$ and $\mathbf{D}_c$ are diagonal scaling matrices with, respectively, the elements of $\mathbf{r}$ and $\mathbf{c}$ on their diagonal. The task of CA is to find k-dimensional coordinates matrices $\mathbf{R}_k$ and $\mathbf{C}_k$ for row and column points such that the loss function

$$\phi(\mathbf{R}_k, \mathbf{C}_k) = \|\widetilde{\mathbf{F}} - \mathbf{D}_r^{1/2}\mathbf{R}_k\mathbf{C}_k'\mathbf{D}_c^{1/2}\|^2 \tag{1}$$

is minimized, where $\|\mathbf{A}\|^2$ denotes the sum of squared elements of $\mathbf{A}$. Consider the (complete) singular value decomposition

$$\widetilde{\mathbf{F}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}', \quad \text{where } \mathbf{U}'\mathbf{U} = \mathbf{I}_{n_r}, \mathbf{V}'\mathbf{V} = \mathbf{I}_{n_c}, \tag{2}$$

where $\mathbf{I}_i$ denotes the $i \times i$ identity matrix. Then, by Eckart and Young [3] we can minimize $\phi(\mathbf{R}_k, \mathbf{C}_k)$ by

$$\mathbf{R}_k = \mathbf{D}_r^{-1/2}\mathbf{U}_k\mathbf{\Lambda}_k^{\alpha} \quad \text{and} \quad \mathbf{C}_k = \mathbf{D}_c^{-1/2}\mathbf{V}_k\mathbf{\Lambda}_k^{1-\alpha},$$

where $\mathbf{U}_k$ and $\mathbf{V}_k$ are respectively the $n_r \times k$ and $n_c \times k$ matrices of singular vectors corresponding to the k largest singular values gathered, in decreasing order, in the $k \times k$ diagonal matrix $\mathbf{\Lambda}_k$, and α is a nonnegative scalar. Clearly,

$$\mathbf{R}_k'\mathbf{D}_r\mathbf{R}_k = \mathbf{\Lambda}_k^{2\alpha} \quad \text{and} \quad \mathbf{C}_k'\mathbf{D}_c\mathbf{C}_k = \mathbf{\Lambda}_k^{2(1-\alpha)}.$$

For $\alpha = 1$, we obtain *row principal* coordinates and for $\alpha = 0$ *column principal* coordinates.

Now, suppose that the marginals $\mathbf{r}$ and $\mathbf{c}$ and the coordinates $\mathbf{R}_k$ and $\mathbf{C}_k$ are given. Then, *the inverse CA problem* is concerned with the question what matrix $\mathbf{F}$ could have produced $\mathbf{R}_k$ and $\mathbf{C}_k$ as its CA solution. In other words, given a CA solution, can we find one or more matrices $\mathbf{F}$ that have the given CA solution as its CA solution?

In the next section, we shall investigate the properties of the set F satisfying the requirements for inverse CA. Necessarily, F must contain the original data matrix $\mathbf{F}$ as an element. We assume, without loss of generality, that $n_r \geqslant n_c$, so that the rank of $\mathbf{F}$ equals n_c or less. If $k = n_c$, the inverse CA problem is trivial and set F only contains $\mathbf{F}$. For $k < n_c$, however, the problem is not trivial and is discussed below.

3. Formalizing the inverse correspondence analysis problem

Suppose that we have a CA solution $\mathbf{R}_k$ and $\mathbf{C}_k$ in k dimensions. In addition, we will assume throughout this paper that the row and column sums of $\mathbf{F}$ are known, so that the scaling matrices $\mathbf{D}_r$ and $\mathbf{D}_c$ are known. Note that these vectors of row and column totals are of great importance in CA. Not only do they provide the proper scaling for the coordinates, they also define the so-called trivial solution that corresponds to the independence model for two categorical variables (see, e.g., [5]).

This trivial solution corresponds to the rank one approximation of $\widetilde{\mathbf{F}}$ and is equal to $n^{-1}\mathbf{D}_r^{1/2}\mathbf{1}\mathbf{1}'\mathbf{D}_c^{1/2}$ that can be obtained by choosing $\mathbf{R}_1 = n^{-1/2}\mathbf{D}_r^{1/2}\mathbf{1}$ and $\mathbf{C}_1 = n^{-1/2}\mathbf{D}_c^{1/2}\mathbf{1}$ where $\mathbf{R}_1$ is $n_r \times 1$ and $\mathbf{C}_1$ is $n_c \times 1$. Typically, one ignores this trivial solution, which can be done by simply discarding the solution, or by considering the singular value decomposition of $\mathbf{D}_r^{-1/2}(\mathbf{F} - n^{-1}\mathbf{rc}')\mathbf{D}_c^{-1/2}$ rather than that of $\mathbf{D}_r^{-1/2}\mathbf{F}\mathbf{D}_c^{-1/2}$. In the following, we will assume that the trivial solution is contained in the coordinate matrices $\mathbf{R}_k$ and $\mathbf{C}_k$. Hence, we will consider the singular value decomposition of $\widetilde{\mathbf{F}}$ for $1 \leqslant k \leqslant n_c$.

In the *inverse CA* problem, we look for set F of all $\mathbf{F}$ that have

1. column sum $\mathbf{c}$ and row sum $\mathbf{r}$, that is, $\mathbf{F1} = \mathbf{c}$ and $\mathbf{1}'\mathbf{F} = \mathbf{r}$,
2. $\mathbf{R}_k$ and $\mathbf{C}_k$ in its CA solution, and
3. only nonnegative elements.

Note that condition 2 does not imply that CA on a particular $\mathbf{F}$ yields $\mathbf{R}_k$ and $\mathbf{C}_k$ as the *first k* dimensions. Condition 2 only tells us that $\mathbf{R}_k$ and $\mathbf{C}_k$ will be among the CA dimensions. In the *strict inverse* CA problem with set F_{strict}, the additional condition imposed is that $\mathbf{R}_k$ and $\mathbf{C}_k$ must be the first k dimensions. In the remainder of this section, we investigate properties of the (strict) inverse CA problem.

Recall the complete singular value decomposition

$$\widetilde{\mathbf{F}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}', \quad \text{where } \mathbf{U}'\mathbf{U} = \mathbf{I}_{n_r}, \mathbf{V}'\mathbf{V} = \mathbf{I}_{n_c}. \tag{3}$$

Let

$$\mathbf{U} = [\mathbf{U}_k \mid \mathbf{U}_c], \quad \mathbf{V} = [\mathbf{V}_k \mid \mathbf{V}_c] \quad \text{and} \quad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_k & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_c \end{bmatrix},$$

where $\mathbf{U}_c$ is $n_r \times (n_r - k)$, $\mathbf{V}_c$ is $n_c \times (n_c - k)$ and $\boldsymbol{\Lambda}_c$ is an $(n_r - k) \times (n_c - k)$ matrix that can be partitioned as $\boldsymbol{\Lambda}_c = [\tilde{\boldsymbol{\Lambda}}_c\ \mathbf{0}]'$ where $\tilde{\boldsymbol{\Lambda}}_c$ is diagonal of order $(n_c - k) \times (n_c - k)$ and, generically, $\mathbf{0}$ denotes a matrix of zeros of appropriate order. Furthermore, as $\mathbf{U}'\mathbf{U} = \mathbf{I}_{n_r}$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}_{n_c}$ it follows that

$$\mathbf{U}_k'\mathbf{U}_c = \mathbf{0} \quad \text{and} \quad \mathbf{V}_k'\mathbf{V}_c = \mathbf{0}. \tag{4}$$

Assuming for the moment that $\mathbf{F}$ is known, then the complete singular value decomposition for the scaled matrix $\widetilde{\mathbf{F}} = \mathbf{D}_r^{-1/2}\mathbf{F}\mathbf{D}_c^{-1/2}$ can be expressed in the following way

$$\widetilde{\mathbf{F}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}' = \mathbf{U}_k\boldsymbol{\Lambda}_k\mathbf{V}_k' + \mathbf{U}_c\boldsymbol{\Lambda}_c\mathbf{V}_c'.$$

Now assume that $\mathbf{F}$ and thus $\widetilde{\mathbf{F}}$ are unknown, but $\mathbf{R}_k, \mathbf{C}_k, \mathbf{D}_r, \mathbf{D}_c$ and thus $\mathbf{U}_k\boldsymbol{\Lambda}_k\mathbf{V}_k'$ are known. From the orthogonality restrictions (4) we can obtain matrices $\widetilde{\mathbf{U}}_c = \mathbf{U}_c\mathbf{T}$ and $\widetilde{\mathbf{V}}_c = \mathbf{V}_c\mathbf{Q}$, where $\mathbf{T}$ and $\mathbf{Q}$ are unknown orthogonal matrices of the appropriate orders. Then, $\widetilde{\mathbf{F}}$ is decomposed into two orthogonal parts

$$\widetilde{\mathbf{F}} = \mathbf{U}_k\boldsymbol{\Lambda}_k\mathbf{V}_k' + \widetilde{\mathbf{U}}_c\mathbf{G}\widetilde{\mathbf{V}}_c', \tag{5}$$

where $\mathbf{G} = \mathbf{T}'\mathbf{\Lambda}_c\mathbf{Q}$. From (5) it can easily be derived that $\mathbf{F}$ can be reconstructed as

$$\mathbf{F} = \mathbf{D}_r^{1/2}(\mathbf{U}_k\mathbf{\Lambda}_k\mathbf{V}_k' + \widetilde{\mathbf{U}}_c\mathbf{G}\widetilde{\mathbf{V}}_c')\mathbf{D}_c^{1/2}. \tag{6}$$

Therefore, in the inverse CA problem, we search for those matrices $\mathbf{G}$ for which $\mathbf{F}$ reconstructed by (6) satisfies the three earlier mentioned conditions.

Lemma 1. *For any* $\mathbf{G}$, *the matrix* $\widetilde{\mathbf{F}}$ *reconstructed by* (5) *has singular values* $\mathbf{\Lambda}_k$ *and corresponding matrices of singular vectors* $\mathbf{U}_k$ *and* $\mathbf{V}_k$.

Proof. The matrices of singular vectors $\mathbf{U}_k$ and $\mathbf{V}_k$, are matrices of eigenvectors of $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}'$ and $\widetilde{\mathbf{F}}'\widetilde{\mathbf{F}}$ respectively. From (4) it follows immediately that for any $\widetilde{\mathbf{F}}$ reconstructed using (5) we have $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}'\mathbf{U}_k = \mathbf{U}_k\mathbf{\Lambda}_k^2$ and $\widetilde{\mathbf{F}}'\widetilde{\mathbf{F}}\mathbf{V}_k = \mathbf{V}_k\mathbf{\Lambda}_k^2$. □

Lemma 2. *For any* $\mathbf{G}$, *the matrix* $\mathbf{F}$ *reconstructed by* (6) *has row sums equal to* $\mathbf{r}$ *and column sums equal to* $\mathbf{c}$.

Proof. It is easily verified that $\mathbf{u} = n^{-1/2}\mathbf{D}_r^{1/2}\mathbf{1}$ and $\mathbf{v} = n^{-1/2}\mathbf{D}_c^{1/2}\mathbf{1}$ are standardized eigenvectors of $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}'$ and $\widetilde{\mathbf{F}}'\widetilde{\mathbf{F}}$ respectively, both corresponding to the eigenvalue $\lambda = 1$. As $\lambda = 1$ is the largest eigenvalue (for a proof of this property, see [11]), the upper left element of $\mathbf{\Lambda}_k$ equals one, and $\mathbf{u}$ and $\mathbf{v}$ are the first columns of $\mathbf{U}_k$ and $\mathbf{V}_k$ respectively. (Note that this first dimension is the trivial solution.) From the orthonormality of $\mathbf{V}$ it follows that $\widetilde{\mathbf{V}}_c'\mathbf{v} = \mathbf{0}$ and $\mathbf{V}_k'\mathbf{v} = (1, 0, \ldots, 0)'$. Hence, for the row sums of $\mathbf{F}$ reconstructed by (6) we have

$$\begin{aligned}
\mathbf{F1} &= \mathbf{D}_r^{1/2}(\mathbf{U}_k\mathbf{\Lambda}_k\mathbf{V}_k' + \widetilde{\mathbf{U}}_c\mathbf{G}\widetilde{\mathbf{V}}_c')\mathbf{D}_c^{1/2}\mathbf{1} \\
&= \mathbf{D}_r^{1/2}(\mathbf{U}_k\mathbf{\Lambda}_k\mathbf{V}_k' + \widetilde{\mathbf{U}}_c\mathbf{G}\widetilde{\mathbf{V}}_c')n^{1/2}\mathbf{v} \\
&= n^{1/2}\mathbf{D}_r^{1/2}\mathbf{u} \\
&= \mathbf{D}_r^{1/2}\mathbf{D}_r^{1/2}\mathbf{1} \\
&= \mathbf{r}.
\end{aligned}$$

In a similar fashion it can be shown that the vector with the column sums of $\mathbf{F}$ reconstructed by (6) is $\mathbf{c}$. □

Lemma 1 tells us that any $\mathbf{G}$ inserted in (6) gives a CA decomposition that includes the original $\mathbf{R}_k$ and $\mathbf{C}_k$. However, without any additional constraints on $\mathbf{G}$ some of the elements of $\mathbf{F}$ may become negative. Thus, we have additional restrictions on $\mathbf{G}$ to make the elements of $\mathbf{F}$ nonnegative. Note that if $\mathbf{G}$ is constrained so that all elements of $\widetilde{\mathbf{F}}$ are nonnegative, then $\mathbf{F}$ must have nonnegative elements as well, since $\mathbf{F} = \mathbf{D}_r^{-1/2}\widetilde{\mathbf{F}}\mathbf{D}_c^{-1/2}$ and $\mathbf{D}_r$ and $\mathbf{D}_c$ have nonnegative elements only. To meet these extra constraints all elements of $\widetilde{\mathbf{U}}_c\mathbf{G}\widetilde{\mathbf{V}}_c'$ must be larger than (or equal to) the corresponding elements of $-\mathbf{U}_k\mathbf{\Lambda}_k\mathbf{V}_k'$.

Let $\mathbf{g} = \mathrm{vec}(\mathbf{G})$, where the vec operator stacks the columns of $\mathbf{G}$ below each other. Using the relationship

$$\mathrm{vec}(\widetilde{\mathbf{U}}_c\mathbf{G}\widetilde{\mathbf{V}}_c') = (\widetilde{\mathbf{V}}_c \otimes \widetilde{\mathbf{U}}_c)\ \mathrm{vec}(\mathbf{G}) \tag{7}$$

between the vec operator and the Kronecker product, we can express the nonnegativity restrictions as

$$\mathbf{Cg} \geqslant -\mathbf{d}, \tag{8}$$

where $\mathbf{C} = \widetilde{\mathbf{V}}_c \otimes \widetilde{\mathbf{U}}_c$ and $\mathbf{d} = \mathrm{vec}(\mathbf{U}_k\mathbf{\Lambda}_k\mathbf{V}_k')$.

Lemma 3. *The system of inequalities* (8) *is consistent.*

Proof. Choosing $\mathbf{G} = \mathbf{T}'\mathbf{\Lambda}_c\mathbf{Q}$ reconstructs the original $\mathbf{F}$. Therefore, the set of matrices $\mathbf{G}$ or vectors $\mathbf{g}$ satisfying (8) is nonempty. Thus, the system of inequalities (8) is consistent. □

Theorem 4. *The solution set F of the inverse CA problem is a convex set.*

Proof. Each inequality in (8) defines a convex half space. The intersection of convex sets is convex, so that F is convex, too. □

Theorem 5. *The set F is a bounded closed polyhedron.*

Proof. The fact that F is a closed polyhedron follows immediately since it is an intersection of half spaces defined by the system of inequalities (8). Boundedness can be established if it can be proved that F does not contain a ray. If F contains a ray, then there exists a $\mathbf{G}_1$ in F such that $\beta\mathbf{G}_1 \in F$ for $\beta > 0$. Let $\mathbf{F}_\mathrm{t}$ denote the trivial solution, that is, $\mathbf{F}_\mathrm{t} = n^{-1}\mathbf{D}_r^{1/2}\mathbf{1}\mathbf{1}'\mathbf{D}_c^{1/2}$, and let $\mathbf{F}_c = \widetilde{\mathbf{U}}_c\mathbf{G}\widetilde{\mathbf{V}}_c'$. From (4) it follows that $\mathbf{F}_\mathrm{t}'\mathbf{F}_c = \mathbf{0}_{(n_c\times n_c)}$ and $\mathbf{F}_\mathrm{t}\mathbf{F}_c' = \mathbf{0}_{(n_r\times n_r)}$. As $\mathbf{F}_\mathrm{t}$ is strictly positive, that is, all its elements are greater than zero, it follows immediately that each row and column of $\mathbf{F}_c$ must contain at least one negative element. Multiplying $\mathbf{F}_c = \mathbf{U}_c\mathbf{G}\mathbf{V}_c'$ with a sufficiently large β will make $\mathbf{F}$ contain one or more negative values so that $\mathbf{F}$ falls outside the polyhedron. Therefore, F does not contain a ray and is consequently bounded. □

Let the convex hull of a polyhedron be defined here as those points in F for which at least one of the inequalities of (8) is an equality and the remaining inequalities of (8) hold. Note that each inequality in (8) defines a half space where the inequality is satisfied, a half space where it is not satisfied, and an $(n_r - k)(n_c - k) - 1$ dimensional hyperplane that separates these two half spaces. Therefore, each face of a polyhedron is a part of such a hyperplane where at least one inequality is an exact equality.

A vertex of a polyhedron is defined here as a point where exactly $(n_r - k)(n_c - k)$ equalities in (8) hold and the remaining inequalities are satisfied. Geometrically, a vertex may be seen as a corner point of the polyhedron where $(n_r - k)(n_c - k)$ faces meet. A vertex is also an extremal point because it cannot be expressed as a convex combination of any two other points in the polyhedron [8, p. 162].

Lemma 6. *Each* **F** *at the convex hull of the polyhedron has at least one element equal to zero and each vertex has* $(n_r - k)(n_c - k)$ *values equal to zero.*

Proof. The system of inequalities (8) is derived from the nonnegativity restrictions on the elements of **F**. Since **G** is an $(n_r - k) \times (n_c - k)$ matrix, there are $(n_r - k)(n_c - k)$ independent elements in **g**. Any **F** at the convex hull of the polyhedron corresponds to a **g** for which at least one of the inequalities is an equality. Since an equality in (8) corresponds to a zero element in **F**, any **g** on the convex hull corresponds to a zero element in **F**.

Because a vertex is the meeting point of $(n_r - k)(n_c - k)$ faces of the convex hull, there are at least $(n_r - k)(n_c - k)$ elements of **F** equal to zero. □

Theorem 7. *The set* F_{strict} *defined by strict inverse CA is a bounded convex set.*

Proof. Set F_{strict} is an intersection between F and the set G of matrices **G** with singular values smaller than or equal to λ_k. To prove that the latter set is convex, we use a result of Magnus and Neudecker [7, p. 205] stating that the largest eigenvalue $\lambda^2_{\max}$ of $\mathbf{G}'\mathbf{G}$ defines a convex function. Therefore, the set G of matrices **G** with $\lambda^2_{\max} \leqslant \lambda^2_k$ is convex. This property also holds for strict monotone functions of $\lambda^2_{\max}$ such as the square root. Therefore, the set G of **G**'s with $\lambda_{\max} \leqslant \lambda_k$ is convex as well. The intersection of two convex sets is also convex, so that the intersection of F and G is convex. Since F is bounded, F_{strict} must also be bounded. □

4. Computing the inverse map

In De Leeuw and Groenen [2], a similar problem was investigated, the so-called inverse multidimensional scaling problem. Here, we take a similar computational approach.

To reconstruct F is equivalent to specifying the polyhedron. We shall do this by searching for all of the vertices of the polyhedron. Any convex combination of the vertices will produce a **g** that is in the polyhedron and from which a valid **F** can be reconstructed through (6) that is a solution to the inverse CA problem.

The basic idea is to check all potential vertices of the system of inequalities defined by $\mathbf{Cg} \geqslant -\mathbf{d}$. We make use of Lemma 6 that states that any vertex must have at least $m = (n_r - k)(n_c - k)$ rows for which the inequalities in $\mathbf{Cg} \geqslant -\mathbf{d}$ are equalities. For an arbitrary selection ψ of m rows, we solve the system $\mathbf{C}_\psi \mathbf{g}_\psi \geqslant -\mathbf{d}_\psi$ by $\mathbf{g}_\psi = -\mathbf{C}_\psi^{-1}\mathbf{d}$. To see whether the candidate vertex $\mathbf{g}_\psi$ is valid, the remaining inequalities are checked. If $\mathbf{Cg}_\psi \geqslant -\mathbf{d}$ holds for all rows, then $\mathbf{g}_\psi$ is a valid vertex of the polyhedron, otherwise it is discarded. To find all vertices, we check for all $\binom{n_r n_c}{m}$ combinations of m rows whether the combination defines a valid vertex. This approach amounts to complete enumeration of all potential vertices of the polyhedron and it can be summarized by the following algorithm.

The inverse CA algorithm:

1. Let the set of vertices V of the polyhedron be empty.
2. Do for all $\binom{n_r n_c}{m}$ combinations ψ:
3. Let $\mathbf{C}_\psi$ and $\mathbf{d}_\psi$ be the m rows of $\mathbf{C}$ and $\mathbf{d}$ respectively defined by ψ.
   - Let $\mathbf{g}_\psi$ be the solution of the system $\mathbf{C}_\psi \mathbf{g} = \mathbf{d}_\psi$.
   - Check if $\mathbf{C}\mathbf{g}_\psi \geqslant \mathbf{d}$. If so, then add $\mathbf{g}_\psi$ to the set of vertices V.
4. End do.
5. Compute for each $\mathbf{g}_\psi$ in V an $\mathbf{F}_\psi$ by (6). Any convex combination of the $\mathbf{F}_\psi$'s satisfies the requirements for inverse CA.

Note that if $\mathbf{C}_\psi$ is not of full rank, then there is no unique solution of $\mathbf{C}_\psi \mathbf{g} = \mathbf{d}_\psi$. Because a vertex must be an extremal point, it is by definition unique. Therefore, if $\mathbf{C}_\psi$ is not of full rank, then ψ cannot define a vertex, so it is simply discarded.

If CA is applied to any $\mathbf{F}_\psi$ found by the inverse CA algorithm or their convex combination, it will have $\mathbf{R}_k$ and $\mathbf{C}_k$ as its row and column coordinates. Note that the inverse CA algorithm gives the solution to the inverse CA problem, not to the strict inverse CA problem. Consequently, the algorithm above cannot guarantee that $\mathbf{R}_k$ and $\mathbf{C}_k$ appear as the first k dimensions in the CA solution.

The complete enumeration approach in the inverse CA algorithm may require a large computational effort, even for small sized CA problems. We discuss this issue in more detail in Section 8 where also a heuristic approach is presented for inverse CA.

5. A strict upper bound for the Pearson chi-squared statistic

Let χ^2 denote the Pearson chi-squared statistic for testing independence. That is, $\chi^2(\mathbf{F}) = \sum_i \sum_j (f_{ij} - e_{ij})^2 / e_{ij}$ with $e_{ij} = r_i c_j / n$. Note that the $\mathbf{r}$ and $\mathbf{c}$ are known in advance. Now we can make use of the results for inverse CA to obtain the upper bound of the chi-squared statistic under the independence model. However, we first consider the general case of the maximum chi-squared statistic in inverse CA.

Theorem 8. *The maximum of χ^2 over the inverse CA set F is attained at one of the vertices of the polyhedron defined by* (8).

Proof. Clearly, $\chi^2(\mathbf{F})$ is quadratic in $\mathbf{F}$ so it is a convex function. Because $\mathbf{F}$ is determined by $\mathbf{G}$ through (6) and $\mathbf{G}$ must be in the convex set F, $\mathbf{F}$ lies in a convex set too. Rockafellar [8, Theorem 32.3, p. 344] states that the maximum of a convex function over a convex set is obtained at an extremal point. An extremal point of a convex set is a point that cannot be expressed as a convex combination of other points in the convex set [8, p. 162]. The extremal points of a polyhedron are the vertices. Because F is a polyhedron, the maximum χ^2 is obtained at a vertex. □

This theorem can be used to find a matrix $\mathbf{F}$ for which the chi-squared statistic for testing independence is a maximum. Hence, if only the marginal frequencies $\mathbf{r}$ and $\mathbf{c}$ are given and no other CA dimension is known, we can construct a matrix $\mathbf{F}$ that yields the upper bound for the chi-squared statistic for testing independence. Again, the value of χ^2 is bounded above and the maximum is attained at one of the vertices. This situation arises in the inverse CA problem when only the trivial dimension is given so that $k = 1$. To obtain the maximum value, the algorithm from Section 4 can be used, although computationally (much) faster methods may exist that make efficient use of the additional structure in the restrictions. Note that the strict lower bound of $\chi^2 = 0$ is obtained at $f_{ij} = e_{ij}$ for all i, j.

6. Additional constraints in inverse CA

We now consider the case where, in addition to the marginals, extra information concerning elements of $\mathbf{F}$ is available. First we discuss the case where one or more elements of $\mathbf{F}$ are known. Then we present an algorithm that can be used to reduce the original set F under the restriction that the elements of the original matrix need to be integers.

6.1. Equality constraints

It may occur that one or more elements of $\mathbf{F}$ are known a priori. For example, if a certain event cannot occur, the corresponding value in $\mathbf{F}$ must be zero. Assume that p values of $\mathbf{F}$, and hence, of $\tilde{\mathbf{F}}$, are known. Let ϕ denote the row indices of $\mathbf{C}$ for which the equality constraints are imposed, so that the rows of the $p \times m$ matrix $\mathbf{C}_\phi$ match the constrained rows of $\mathbf{C}$. Furthermore, let $\tilde{\mathbf{f}}_\phi$ denote the $p \times 1$ vector of corresponding (known) values of $\widetilde{\mathbf{F}}$ and let $\mathbf{d}_\phi$ denote the corresponding rows of $\mathbf{d}$. The new constraints can be expressed as

$$\mathbf{C}_\phi \mathbf{g} = \tilde{\mathbf{f}}_\phi - \mathbf{d}_\phi. \tag{9}$$

Theorem 9. *Assume that the constraints also hold for the original data that were used to compute the CA solution. Then, the solution of constrained inverse CA is a nonempty bounded convex polyhedron.*

Proof. By Theorem 5, the solution of the inverse CA problem defines a bounded convex polyhedron. The equality constraints defined by (9) are linear and thus convex. The union of a bounded polyhedron and a linear subspace is again a bounded polyhedron. Because original data follow the constraints by assumption, there is at least one $\mathbf{F}$ that falls in the polyhedron and satisfies the constraints. □

We distinguish three cases that may occur with respect to the constraints as expressed in (9):

(a) $p < m$: There are fewer constraints than free elements in $\mathbf{g}$. We can implement the restrictions in our algorithm in the following way.

The constrained inverse CA algorithm:
1. Let the set of vertices V be empty.
2. Do for all $\binom{n_r n_c - p}{m - p}$ combinations ψ^*, where each combination contains ϕ, i.e. $\psi^* = \begin{pmatrix}\phi\\ \psi\end{pmatrix}$:
3. Let $\mathbf{C}_{\psi^*} = \begin{pmatrix}\mathbf{C}_\phi\\ \mathbf{C}_\psi\end{pmatrix}$ and $\mathbf{d}_{\psi^*} = \begin{pmatrix}\tilde{\mathbf{f}}_\phi - \mathbf{d}_\phi\\ -\mathbf{d}_\psi\end{pmatrix}$ be the m rows of $\mathbf{C}$ and $\mathbf{d}$, defined by ψ^* and the constrained values $\tilde{\mathbf{f}}_\phi$.
   - Let $\mathbf{g}_{\psi^*}$ be the solution of the system $\mathbf{C}_{\psi^*}\mathbf{g} = \mathbf{d}_{\psi^*}$.
   - Check if $\mathbf{C}\mathbf{g}_{\psi^*} \geqslant \mathbf{d}$. If so, then add $\mathbf{g}_{\psi^*}$ to the set of vertices V.
4. End do.

(b) $p = m$: The number of constraints is equal to the number of unknown elements. Therefore, if the corresponding matrix $\mathbf{C}_\phi$ is nonsingular, that is, if $\mathbf{C}_\phi^{-1}$ exists, we obtain a unique solution for $\mathbf{g}$. Thus, if $\mathbf{F}$ reconstructed using (6) is nonnegative, we have a valid unique solution. Else, if an element of the reconstructed matrix $\mathbf{F}$ is negative, the solution set F is empty.
If $\mathbf{C}_\phi$ is singular, that is, if some constraints are linearly dependent and hence redundant, we cannot uniquely determine $\mathbf{g}$. We thus have a similar situation as in (a). We can obtain vertices satisfying the equality and nonnegativity constraints in the following way: Let p^* denote the rank of $\mathbf{C}_\phi$. Then, consider for all $\binom{n_r n_c - p}{m - p^*}$ combinations of rows of $\mathbf{C}$ that contain $\mathbf{C}_\phi$, the following system of equations:

$$\mathbf{C}_{\phi^*}\mathbf{g} = \tilde{\mathbf{f}}_{\phi^*} - \mathbf{d}_{\phi^*}$$

where $\mathbf{C}_{\phi^*}$ is a $(2p - p^*) \times m$ matrix with as first p rows independent rows of $\mathbf{C}$ corresponding to the equality constraints, $\tilde{\mathbf{f}}_{\phi^*}$ is the vector $\tilde{\mathbf{f}}_\phi$ supplemented with $p - p^*$ zeros and $\mathbf{d}_{\phi^*}$ is the vector of appropriate elements of $\mathbf{d}$. Then, for each $\mathbf{C}_{\phi^*}$ that has rank m, we can calculate $\mathbf{g}$ as $\mathbf{g} = (\mathbf{C}'_{\phi^*}\mathbf{C}_{\phi^*})^{-1}\mathbf{C}'_{\phi^*} \times$

$(\tilde{\mathbf{f}}_{\phi^*} - \mathbf{d}_{\phi^*})$. Upon checking the nonnegativity constraints $\mathbf{Cg} \geqslant \mathbf{d}$, we add or discard the vertices to our solution set. Note that, if $\mathbf{C}_{\phi^*}$ is not of full column rank then ϕ^* cannot be a vertex and we can simply discard it.

(c) $p > m$: There are more constraints than free elements, so that the matrix $\mathbf{C}_\phi$ has more rows than columns. Then, assuming that $\mathbf{C}_\phi$ has full column rank, $\mathbf{g}$ can be calculated as $\mathbf{g} = (\mathbf{C}'_\phi\mathbf{C}_\phi)^{-1}\mathbf{C}'_\phi(\tilde{\mathbf{f}}_\phi - \mathbf{d}_\phi)$. In order for $\mathbf{g}$ to be a valid solution, $\mathbf{F}$ reconstructed using (6) must be nonnegative. Else, the solution set F is empty.
If the rank of $\mathbf{C}_\phi$ is smaller than m, we have essentially the same situation as described under (b) and we can apply the same procedure to obtain vertices.

Note that, by imposing the additional constraints, we have decreased the number of inequalities to be checked. Therefore, with a sufficient number of constraints even large inverse CA problems become computationally feasible.

6.2. Integer constraints

Suppose we know that the elements of the original matrix $\mathbf{F}$ are integers. For example, we may know $\mathbf{F}$ to be a contingency matrix. This information can be used to reduce the solution set F in the following way.

Let $\mathbf{F}^h$ denote the reconstructed matrix for the hth vertex $\mathbf{g}_h$, that is $\mathbf{F}^h = \mathbf{D}_r^{-1/2}(\mathbf{U}_k\mathbf{\Lambda}_k\mathbf{V}'_k + \widetilde{\mathbf{U}}_c\mathbf{G}_h\widetilde{\mathbf{V}}'_c)\mathbf{D}_c^{-1/2}$, where $\text{vec}(\mathbf{G}_h) = \mathbf{g}_h$ and let f_{ij}^h denote the ijth element of $\mathbf{F}^h$. Define $\text{int}_+(x)$ as the first integer larger than x and $\text{int}_-(x)$ as the first integer smaller than x. Furthermore, let $\min_h(f_{ij}^h)$ denote the smallest ijth element over all vertices, and let $\max_h(f_{ij}^h)$ denote the largest ijth element over all vertices. Using these definitions we construct a matrix $\mathbf{F}_{\min}$ which has as its elements $\text{int}_+(\min_h(f_{ij}^h))$, and, similarly, a matrix $\mathbf{F}_{\max}$ with elements $\text{int}_-(\max_h(f_{ij}^h))$.

Theorem 10. *When* $\mathbf{F}$ *is restricted to have elements* f_{ij} *that are integer, then elements of* $\mathbf{F}$ *are bounded below by* $\mathbf{F}_{\min}$ *and bounded above by* $\mathbf{F}_{\max}$.

Proof. As the solution set F is convex, the true matrix $\mathbf{F}$ can be expressed as a convex combination of the vertices. That is, $\mathbf{F} = \sum_h a_h\mathbf{F}^h$, where $\sum_h a_h = 1$, and $0 \leqslant a_h \leqslant 1$. Hence, the ijth element of $\mathbf{F}$ lies between $\min_h(f_{ij}^h)$ and $\max_h(f_{ij}^h)$. Therefore, if the elements of $\mathbf{F}$ are integers, the smallest possible value for the ijth element of $\mathbf{F}$ is the first integer larger than (or equal to) $\min_h(f_{ij}^h)$, and the largest possible value for the ijth element of $\mathbf{F}$ is the first integer smaller than (or equal to) $\max_h(f_{ij}^h)$. □

Define $\widetilde{\mathbf{F}}_{\min} = \mathbf{D}_r^{-1/2}\mathbf{F}_{\min}\mathbf{D}_c^{-1/2}$, $\widetilde{\mathbf{F}}_{\max} = \mathbf{D}_r^{-1/2}\mathbf{F}_{\max}\mathbf{D}_c^{-1/2}$, $\tilde{\mathbf{f}}_{\max} = \text{vec}(\widetilde{\mathbf{F}}_{\max})$, and $\tilde{\mathbf{f}}_{\min} = \text{vec}(\widetilde{\mathbf{F}}_{\min})$. Using (6), we must have

$$\widetilde{\mathbf{F}}_{\min} - \mathbf{U}_k\mathbf{\Lambda}_k\mathbf{V}'_k \leqslant \widetilde{\mathbf{U}}_c\mathbf{G}\widetilde{\mathbf{V}}'_c \leqslant \widetilde{\mathbf{F}}_{\max} - \mathbf{U}_k\mathbf{\Lambda}_k\mathbf{V}'_k,$$

or in vec notation

$$\tilde{\mathbf{f}}_{\min} - \mathbf{d} \leqslant \mathbf{C}\mathbf{g} \leqslant \tilde{\mathbf{f}}_{\max} - \mathbf{d}.$$

These additional integer restrictions can be implemented as follows:

The integer constrained inverse CA algorithm:

1. Find an initial set of vertices V by the Inverse CA Algorithm of Section 4.
2. Repeat until V does not change:
   (a) Compute $\tilde{\mathbf{f}}_{\min}$ and $\tilde{\mathbf{f}}_{\max}$ as described above.
   (b) Do for all $\binom{n_r n_c}{m}$ combinations ψ.
   (c) Let $\mathbf{C}_\psi$ and $\mathbf{d}_\psi$ be the rows of $\mathbf{C}$ and $\mathbf{d}$ defined by ψ.
   - Let $\mathbf{g}_{\psi_1}$ be the solution of the system $\mathbf{C}_\psi \mathbf{g} = (\tilde{\mathbf{f}}_{\max} - \mathbf{d})_\psi$.
   - Check if $(\tilde{\mathbf{f}}_{\min} - \mathbf{d}) \leqslant \mathbf{C}\mathbf{g}_{\psi_1} \leqslant (\tilde{\mathbf{f}}_{\max} - \mathbf{d})$. If so, then add $\mathbf{g}_{\psi_1}$ to the set of vertices V.
   - Let $\mathbf{g}_{\psi_2}$ be the solution of the system $\mathbf{C}_\psi \mathbf{g} = (\tilde{\mathbf{f}}_{\min} - \mathbf{d})_\psi$.
   - Check if $(\tilde{\mathbf{f}}_{\min} - \mathbf{d}) \leqslant \mathbf{C}\mathbf{g}_{\psi_2} \leqslant (\tilde{\mathbf{f}}_{\max} - \mathbf{d})$. If so, then add $\mathbf{g}_{\psi_2}$ to the set of vertices V.

   (d) End do.

As this procedure imposes additional restrictions, the number of vertices may increase. The solution space, however, becomes smaller. Moreover, the matrices $\mathbf{F}_{\min}$ and $\mathbf{F}_{\max}$ provide us with lower and upper bounds for the integer elements of $\mathbf{F}$.

7. An illustrative example

To illustrate our method, consider the artificial smoking data of Greenacre [5] (see Table 1).

Table 1
Artificial smoking data of Greenacre [5]

Staff group	Smoking category				Row total $\mathbf{r}$
	None	Light	Medium	Heavy	
Senior managers	4	2	3	2	11
Junior managers	4	3	7	4	18
Senior employees	25	10	12	4	51
Junior employees	18	24	33	13	88
Secretaries	10	6	7	2	25
Column total $\mathbf{c}'$	61	45	62	25	193

The Pearson chi-squared statistic for independence is $\chi^2 = 16.44$.

Table 2
Vertices and reconstructed **F** by (6) of inverse CA of the smoking data using $k = 3$

Vertex	Reconstructed **F**	χ^2
$\mathbf{g}_1 = \begin{bmatrix} -0.1821 \\ -0.2974 \end{bmatrix}$	$\mathbf{H}_1 = \begin{bmatrix} 4.11 & 0.00 & 6.21 & 0.68 \\ 3.87 & 5.43 & 3.09 & 5.61 \\ 25.11 & 8.00 & 15.22 & 2.67 \\ 18.15 & 21.22 & 37.47 & 11.16 \\ 9.76 & 10.36 & 0.00 & 4.88 \end{bmatrix}$	$\chi_1^2 = 39.83$
$\mathbf{g}_2 = \begin{bmatrix} -0.1821 \\ 0.1993 \end{bmatrix}$	$\mathbf{H}_2 = \begin{bmatrix} 4.11 & 0.00 & 6.21 & 0.68 \\ 3.96 & 3.74 & 5.80 & 4.49 \\ 24.83 & 13.14 & 6.95 & 6.08 \\ 17.94 & 25.13 & 31.18 & 13.75 \\ 10.17 & 2.98 & 11.86 & 0.00 \end{bmatrix}$	$\chi_2^2 = 30.43$
$\mathbf{g}_3 = \begin{bmatrix} 0.2084 \\ -0.2812 \end{bmatrix}$	$\mathbf{H}_3 = \begin{bmatrix} 3.90 & 3.87 & 0.00 & 3.24 \\ 4.02 & 2.66 & 7.55 & 3.77 \\ 25.22 & 6.00 & 18.44 & 1.35 \\ 18.10 & 22.13 & 36.01 & 11.76 \\ 9.76 & 10.36 & 0.00 & 4.88 \end{bmatrix}$	$\chi_3^2 = 40.00$
$\mathbf{g}_4 = \begin{bmatrix} 0.2084 \\ 0.2155 \end{bmatrix}$	$\mathbf{H}_4 = \begin{bmatrix} 3.90 & 3.87 & 0.00 & 3.24 \\ 4.11 & 0.97 & 10.26 & 2.66 \\ 24.94 & 11.14 & 10.16 & 4.76 \\ 17.89 & 26.04 & 29.72 & 14.35 \\ 10.17 & 2.98 & 11.86 & 0.00 \end{bmatrix}$	$\chi_4^2 = 33.71$

Suppose that in addition to the marginals **r** and **c**, we have the 2-dimensional CA solution for these data. That is, in our notation, $k = 3$ and $\mathbf{R}_k$ and $\mathbf{C}_k$ are 5×3 and 4×3 matrices with as their first column the trivial solutions. We can derive $\widetilde{\mathbf{U}}_c$ and $\widetilde{\mathbf{V}}_c$ from $\mathbf{R}_k'\widetilde{\mathbf{U}}_c = \mathbf{0}$ and $\mathbf{C}_k'\widetilde{\mathbf{V}}_c = \mathbf{0}$.

Applying the Inverse CA Algorithm described in Section 4 with $\mathbf{C} = \widetilde{\mathbf{V}}_c \otimes \widetilde{\mathbf{U}}_c$ and $\mathbf{d} = \text{vec}(\mathbf{D}_r^{1/2}\mathbf{R}_k\mathbf{C}_k'\mathbf{D}_c^{1/2})$, four valid solutions for **g** are obtained. Table 2 contains the four vertices and the corresponding reconstructed **F** matrices. Thus, any convex combination of these four vertices yields a CA solution with $\mathbf{R}_k$ and $\mathbf{C}_k$, the marginals are **r** and **c**, and the elements of **F** are nonnegative. It may be verified that the convex combination $0.1962\mathbf{H}_1 + 0.2866\mathbf{H}_2 + 0.2134\mathbf{H}_3 + 0.3038\mathbf{H}_4$ yields the original contingency matrix in Table 1.

Because **g** only contains two elements, a visual representation of the inverse CA solution can easily be obtained (see Fig. 1). The axes represent the elements of **g**, that is, g_1 and g_2. For a **g** of this size, the set with $\lambda_{\max} < \lambda_k$ can be graphed as circle.

For the same data, suppose that we want to impose the additional restriction that element $i = 1$ and $j = 4$ is fixed to 2. Clearly, the problem becomes the constrained inverse CA problem. The vertices of the constrained inverse CA solution are presented in Table 3. Again it may be verified that for every convex combination of the two vertices the marginals are **r** and **c**, a CA solution contains $\mathbf{R}_k$ and $\mathbf{C}_k$, the elements of **F** are nonnegative, and element $i = 1$ and $j = 4$ equals 2.

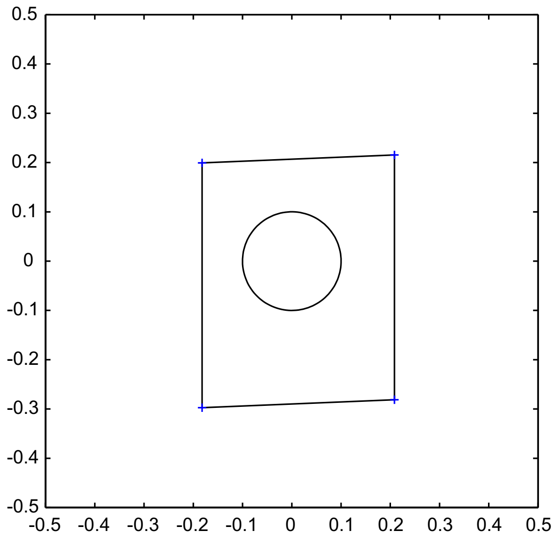

Fig. 1. Polyhedron defined by inverse CA on the smoking data using $k = 3$. The vertices are indicated by crosses. The dimensions are g_1 and g_2. The circle indicates those $\mathbf{g}$ that satisfy the strict inverse CA condition.

Table 3
Vertices and reconstructed $\mathbf{F}$ by (6) of constrained inverse CA of the smoking data using $k = 3$, where element $i = 1$ and $j = 4$ is fixed to 2

Vertex	Reconstructed $\mathbf{F}$	χ^2
$\mathbf{g}_1 = \begin{bmatrix} 0.0199 \\ -0.2890 \end{bmatrix}$	$\mathbf{H}_1 = \begin{bmatrix} 4.00 & 2.00 & 3.00 & 2.00 \\ 3.95 & 4.00 & 5.40 & 4.66 \\ 25.17 & 6.96 & 16.88 & 1.99 \\ 18.13 & 21.69 & 36.72 & 11.47 \\ 9.76 & 10.36 & 0.00 & 4.88 \end{bmatrix}$	$\chi_1^2 = 32.56$
$\mathbf{g}_2 = \begin{bmatrix} 0.0199 \\ 0.2077 \end{bmatrix}$	$\mathbf{H}_2 = \begin{bmatrix} 4.00 & 2.00 & 3.00 & 2.00 \\ 4.04 & 2.31 & 8.11 & 3.54 \\ 24.88 & 12.11 & 8.61 & 5.40 \\ 17.91 & 25.60 & 30.42 & 14.06 \\ 10.17 & 2.98 & 11.86 & 0.00 \end{bmatrix}$	$\chi_2^2 = 24.76$

Finally, suppose it is known that the original matrix is a contingency matrix. Then, using Theorem 10 we can obtain matrices with lower and upper (integer) bounds for the values of $\mathbf{F}$. These matrices, based on the four reconstructed $\mathbf{F}$ matrices from Table 2, are presented in Table 4. Applying the algorithm described in Section 6.2 immediately yields one vertex $\mathbf{g} = [0.0199, 0.0043]'$, with as corresponding $\mathbf{F}$ matrix the original contingency table in Table 1.

Table 4
Lower and upper bounds for the smoking data

Staff group	Lower bounds				Upper bounds			
	None	Light	Medium	Heavy	None	Light	Medium	Heavy
Senior managers	4	0	0	1	4	3	6	3
Junior managers	4	1	4	3	4	5	10	5
Senior employees	25	6	7	2	25	13	18	6
Junior employees	18	22	30	12	18	26	37	14
Secretaries	10	3	0	0	10	10	11	4

8. A heuristic approach to computing the inverse map

Since finding the vertices of a polyhedron is a combinatorial problem, the computational effort may be prohibitive, even for small sized inverse CA problems. Table 5 gives an overview of the number of vertices that need to be checked as a function of the number of rows n_r, the number of columns n_c, and the dimensionality k. As k includes the trivial dimension, the table gives the number of combinations for $k = 2$, 3, and 4. It is clear that even for a reasonably sized CA problem with $n_r = 8$,

Table 5
Number of vertices to be checked by the complete enumeration approach as a function of the dimensionality k, the number of rows n_r, and the number of columns n_c

n_r	n_c					
	3	4	5	6	7	8
$k=2$						
3	9					
4	66	1820				
5	455	38,760	2,042,975			
6	3060	735,471	86,493,225	7,307,872,110		
7	20,349	13,123,110	3,247,943,160	5.1379×10^{11}	6.3205×10^{13}	
8	134,596	225,792,840	1.1338×10^{11}	3.2247×10^{13}	6.6464×10^{15}	1.1188×10^{18}
$k=3$						
4		16				
5		190	12,650			
6		2024	593,775	94,143,280		
7		20,475	23,535,820	1.1058×10^{10}	3.3481×10^{12}	
8		201,376	847,660,528	1.0933×10^{12}	7.8561×10^{14}	4.0104×10^{17}
$k=4$						
5			25			
6			435	58,905		
7			6545	5,245,786	2,054,455,634	
8			91,390	377,348,994	5.5838×10^{11}	4.8853×10^{14}

$n_c = 8$, and $k = 3$, the computational task of enumerating 4.0104×10^{17} vertices is prohibitive. Therefore, we propose a heuristic method that is not guaranteed to find all vertices but it may give a reasonable set of vertices with limited computational effort.

The heuristic approach is based on the idea that we would like to find the vertices by function optimization subject to the inequality constraints. From the proof of Theorem 8 we know that a linear function over a polyhedron reaches its optimum at a vertex. Define an $m = (n_r - k)(n_c - k)$ vector $\mathbf{a}$ of length one to give a random direction. Then, the maximum of the linear program $\mathbf{a}'\mathbf{g}$ subject to $\mathbf{Cg} \geqslant \mathbf{d}$ will be reached at a vertex. Repeatedly doing so for different random $\mathbf{a}$ may yield different vertices. This leads to the heuristic inverse CA algorithm.

The heuristic inverse CA algorithm:
1. Let the set of vertices V of the polyhedron be empty. Set the number of random direction N_{dir} to some large integer.
2. Do for $i = 1$ to N_{dir}:
3. Draw a random direction $\mathbf{a}$ as follows. Let θ be an $m - 1$ vector with random values on the interval $[0, 2\pi]$ for θ_1 and $[-\pi, \pi]$ for θ_i with $2 \leqslant i \leqslant m - 1$. Using polar coordinates, define $a_1 = \sin\theta_1$, $a_i = (\prod_{j=1}^{i-1} \cos\theta_j)\sin\theta_i$ for $2 \leqslant i \leqslant m - 1$, and $a_m = (\prod_{j=1}^{m-1} \cos\theta_j)$ so that $\mathbf{a}$ is randomly distributed on a sphere of length 1.
4. Solve the linear program $\mathbf{a}'\mathbf{g}$ subject to $\mathbf{Cg} \geqslant \mathbf{d}$. Let $\mathbf{g}^*$ be the solution.
5. If $\mathbf{g}^*$ is not in V then add $\mathbf{g}^*$ to V.
6. End do.
7. Compute for each $\mathbf{g}$ in V an $\mathbf{F}$ by (6). Any convex combination of these $\mathbf{F}$'s satisfies the requirements for inverse CA.

It cannot be guaranteed that this heuristic procedure finds all vertices but as N_{dir} increases, the polyhedron reconstructed by this heuristic method will approximate the true polyhedron better. The way of constructing $\mathbf{a}$ in Step 3 ensures that every direction is equally likely. To solve the linear program, we use the standard linear program solver in MatLab 6.1 that is based on the simplex method of Dantzig et al. [1].

To see how well the heuristic approach is able to find the vertices, we applied the approach to the smoking data in Table 1 using $k = 2$. For this choice of k, the inverse CA algorithm took 11 CPU seconds to check 38,760 potential vertices of which 45 were indeed vertices of the polyhedron defined by inverse CA. The heuristic inverse CA algorithm took 7.7 CPU seconds using 300 random directions and found the same 45 vertices.

This example illustrates that the heuristic approach can find all correct vertices within a reasonable time if the inverse CA problem is relatively small. However, for larger inverse CA problems, the number of possible vertices also rapidly increases.

In such cases, the heuristic method proposed here should only be seen as a crude way to obtain an estimate of the polyhedron defined by the inverse CA problem.

9. Conclusion and discussion

In this paper, we have specified the set of matrices that all yield a given low-dimensional configuration in its CA solution. This set is a nonempty bounded closed polyhedron. Computing the vertices of the polyhedron is a computationally very demanding task, even for relatively small CA problems. One way to reduce this task is to impose a number of additional constraints on the elements. However, in practice additional constraints may not be available or there may not be enough constraints to seriously reduce the computational task. We also specified a strict upper bound for the Pearson chi-squared statistic, not limited to inverse CA, but also to the special case of the independence model where only the margins of the data matrix are available. Furthermore, we showed that if the data matrix is known to have integer values (as in a contingency table), then lower and upper integer bounds for the elements of the original unknown contingency table can be obtained. In this case, the inverse CA solution set may be significantly reduced and can be unique.

Throughout this paper, we have assumed that the row and column marginals were known in advance together with the low-dimensional CA solution. This choice can easily be justified by recognizing that the marginals can be directly derived from the trivial CA dimension. An extension of the inverse CA problem to a situation where the marginals are unknown *a priori*, would lead to a much more complicated situation with a set that does not have the nice mathematical properties as in this paper.

The specification of the inverse set is available for some other multivariate analysis techniques such as multidimensional scaling [2,6] and principal components analysis [9], or could be developed in the same spirit as the present paper. We believe that investigation of the inverse set yields better understanding of the original problem.

Acknowledgements

The research of Michel van de Velden was partially funded through EU Grant HPMF-CT-2000-00664. We are greatful for the constructive remarks of an anonymous referee that has helped us to improve this manuscript.

References

[1] G.B. Dantzig, A. Orden, P. Wolfe, Generalized simplex method for minimizing a linear from under linear inequality constraints, Pacific J. Math. 5 (1955) 183–195.

[2] J. De Leeuw, P.J.F. Groenen, Inverse multidimensional scaling, J. Classification 14 (1997) 3–21.
[3] C. Eckart, G. Young, Approximation of one matrix by another of lower rank, Psychometrika 1 (1936) 211–218.
[4] A. Gifi, Nonlinear Multivariate Analysis, Wiley, Chichester, 1990.
[5] M.J. Greenacre, Theory and Applications of Correspondence Analysis, Academic Press, New York, 1984.
[6] P.J.F. Groenen, J. De Leeuw, R. Mathar, Least squares multidimensional scaling with transformed distances, in: W. Gaul, D. Pfeifer (Eds.), Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, 1996, pp. 177–185.
[7] J.R. Magnus, H. Neudecker, Matrix Differential Calculus with Appilications in Statistics and Econometrics, Wiley, Chichester, 1988.
[8] R.T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, NJ, 1970.
[9] J.M.F. Ten Berge, H.A.L. Kiers, Retrieving the correlation matrix from a truncated PCA solution: The inverse principal component solution, Psychometrika 64 (1999) 317–324.
[10] M. Van de Velden, Topics in Correspondence Analysis, Tinbergen Institute, University of Amsterdam, Amsterdam, 2000.
[11] M. Van de Velden, H. Neudecker, On an eigenvalue property relevant in correspondence analysis and related methods, Linear Algebra Appl. 321 (2000) 347–364.