Inverse Correspondence Analysis

Patrick J.F. Groenen* Michel van de Velden†

September 10, 2002

Econometric Institute Report EI 2002-31

*Econometric Institute, Erasmus University Rotterdam, The Netherlands, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands (e-mail: groenen@few.eur.nl).

†Universitat Pompeu Fabra, Barcelona, Spain

Abstract

In correspondence analysis, 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 or example, Greenacre (1984). In this paper, we will study the *inverse correspondence analysis* problem, that is, the possibilities of retrieving one or more data matrices from a low dimensional correspondence analysis solution. We will show that there exists a nonempty closed and bounded polyhedron of such matrices. We also present an algorithm to find the vertices of the polyhedron. 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 correspondence analysis problem. As a special case, we present a method for imposing integer restrictions on the data matrix as well. The approach to inverse correspondence analysis followed here is similar to the one employed by De Leeuw and Groenen (1997) in their inverse multidimensional scaling problem.

Keywords: Correspondence Analysis, Inverse Problems, Maximum Chi-Square.

1 Introduction

In correspondence analysis (CA), the rows and columns of a data matrix $\mathbf{F}$ are depicted as points in low-dimensional space. Most often, $\mathbf{F}$ is a contingency matrix, but this need not be the case. The only restriction on $\mathbf{F}$ 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, Greenacre (1984), Gifi (1990), and Van de Velden (2000). In this paper, we concentrate on a problem that has not been treated before. Given a low dimensional CA solution, which matrices $\mathbf{F}$ would have produced the current solution as a CA solution? We call this problem *the inverse correspondence analysis problem.*

There are several reasons to investigate the inverse CA problem. First of all, the size of the set of matrices $\mathbf{F}$ may reveal information about the uniqueness of the original solution. If this set is large, then there are many nonnegative matrices $\mathbf{F}$ that yield the same CA solution. Thus, even though the data have lead to a perfectly normal correspondence analysis 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 much less nonnegative matrices $\mathbf{F}$ 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 correspondence analysis 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. Thirdly, we believe that the study of inverse CA deepens our understanding of CA. Finally, through inverse CA, we are able to prove 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 (1997), in their treatment of the inverse multidimensional scaling problem (see also, Groenen, De Leeuw, & Mathar, 1996).

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 $\mathbf{F}$ 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 columns 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 $\tilde{\mathbf{F}}$ as $\tilde{\mathbf{F}} = \mathbf{D}_r^{-1/2}\mathbf{F}\mathbf{D}_c^{-1/2}$, 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 correspondence analysis is to find k-dimensional coordinates matrices $\mathbf{R}$ and $\mathbf{C}$ for row and column points such that the loss function

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

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

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

where $\mathbf{I}_i$ denotes the $i \times i$ identity matrix. Then, by Eckart and Young (1936) we can minimize $\phi(\mathbf{R}_k, \mathbf{C}_k)$ by

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

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 the $k \times k$ diagonal matrix $\mathbf{\Lambda}_k$, and α is a nonnegative scalar. Clearly,

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

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 correspondence analysis 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 $n_r \geq n_c$, 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 correspondence analysis 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 correspondence analysis. Not only do they provide the proper scaling for the coordinates, they are also referred to as the so-called trivial solution, see, e.g., Greenacre (1984). Typically, one ignores this trivial solution, which can be done by simply discarding the solution, or by considering the singular value decomposition of $\mathbf{D}_r^{-1/2}\left(\mathbf{F} - n^{-1}\mathbf{r}\mathbf{c}'\right)\mathbf{D}_c^{-1/2}$ 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 $\tilde{\mathbf{F}}$ for $1 \leq k \leq n_c$.

In the *inverse CA* problem, we look for 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, 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

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

Let

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

where $\mathbf{U}_c$ is $n_r \times (n_r - k)$, $\mathbf{V}_c$ is $n_c \times (n_c - k)$ and $\mathbf{\Lambda}_c$ is an $(n_r - k) \times (n_c - k)$ matrix that can be partitioned as $\mathbf{\Lambda}_c = [\tilde{\mathbf{\Lambda}}_c \ \ \mathbf{0}]'$ where $\tilde{\mathbf{\Lambda}}_c$ 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

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

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

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

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

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

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

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

Therefore, in the inverse correspondence analysis 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* $\tilde{\mathbf{F}}$ *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 $\tilde{\mathbf{F}}\tilde{\mathbf{F}}'$ and $\tilde{\mathbf{F}}'\tilde{\mathbf{F}}$ respectively. From (4) it follows immediately that for any $\tilde{\mathbf{F}}$ reconstructed using (5) we have $\tilde{\mathbf{F}}\tilde{\mathbf{F}}'\mathbf{U}_k = \mathbf{U}_k\mathbf{\Lambda}_k^2$ and $\tilde{\mathbf{F}}'\tilde{\mathbf{F}}\mathbf{V}_k = \mathbf{V}_k\mathbf{\Lambda}_k^2$. ■

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. This follows immediately from Lemma 1 and the fact that the trivial solution in correspondence analysis (that is, the first dimension) is equal to $\lambda_1\mathbf{u}_1\mathbf{v}_1' = n^{-1}\mathbf{D}_r^{1/2}\mathbf{1}\mathbf{1}'\mathbf{D}_c^{1/2}$. Pre multiplying by $\mathbf{D}_r^{1/2}$ and post multiplying by $\mathbf{D}_r^{1/2}$ gives $n^{-1}\mathbf{D}_r\mathbf{1}\mathbf{1}'\mathbf{D}_c = n^{-1}\mathbf{r}\mathbf{c}'$, so that the row sums equal $n^{-1}\mathbf{r}\mathbf{c}'\mathbf{1} = \mathbf{r}$ and the column sums equal $n^{-1}\mathbf{1}'\mathbf{r}\mathbf{c}' = \mathbf{c}'$. ■

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 $\tilde{\mathbf{F}}$ are nonnegative, then $\mathbf{F}$ must have nonnegative elements as well, since $\mathbf{F} = \mathbf{D}_r^{-1/2}\tilde{\mathbf{F}}\mathbf{D}_c^{-1/2}$ and $\mathbf{D}_r$ and $\mathbf{D}_c$ have nonnegative elements only. To meet these extra constraints all elements of $\tilde{\mathbf{U}}_c\mathbf{G}\tilde{\mathbf{V}}_c'$ must be larger than (or equal to) the elements of $-\mathbf{U}_k\mathbf{\Lambda}_k\mathbf{V}_k'$.

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

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A})\ \text{vec}(\mathbf{B}) \tag{7}$$

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

$$\mathbf{Cg} \geq -\mathbf{d} \tag{8}$$

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

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 correspondence analysis 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 $\mathbf{F}_t$ denote the trivial solution, that is, $\mathbf{F}_t = n^{-1}\mathbf{D}_r^{1/2}\mathbf{1}\mathbf{1}'\mathbf{D}_c^{1/2}$, and let $\mathbf{F}_c = \tilde{\mathbf{U}}_c\mathbf{G}\tilde{\mathbf{V}}_c'$. From (4) it follows that $\mathbf{F}_t'\mathbf{F}_c = \mathbf{0}_{(n_c \times n_c)}$ and $\mathbf{F}_t\mathbf{F}_c' = \mathbf{0}_{(n_r \times n_r)}$. As $\mathbf{F}_t$ 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. ■

Lemma 6 *Each* $\mathbf{F}$ *at the hull of the polyhedron has at least* $(n_r - k)(n_c - k)$ *values equal to zero.*

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

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 $\mathbf{G}$ 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 (1988, 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 $\mathbf{G}$ with $\lambda^2_{\max} \leq \lambda^2_k$ is convex. This property also holds for strict monotone functions of $\lambda^2_{\max}$ such as the square root. Therefore, the set G of $\mathbf{G}$'s with $\lambda_{\max} \leq \lambda_k$ 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 (1997), a similar problem was investigated, the so-called inverse multidimensional scaling problem. Here, we take a similar computational approach.

The basic idea is to check all possible vertices of the system of inequalities defined by $\mathbf{Cg} \geq -\mathbf{d}$. Let $m = (n_r - k)(n_c - k)$ be the length of vector $\mathbf{g}$. Then, check for all $\binom{n_r n_c}{m}$ combinations of rows whether the combination defines a valid vertex.

The Inverse Correspondence Analysis Algorithm:

1. Let the set of vertices V 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 $\mathbf{C}\mathbf{g}_\psi \geq \mathbf{d}$. If so, then add $\mathbf{g}_\psi$ to the set of vertices V.

4. End do.

Note that if some $\mathbf{C}_\psi$ is not of full rank, then ψ cannot be a vertex, so it is simply discarded.

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 χ^2 over the inverse CA set F is attained at one of the vertices.*

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 (1970, 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 (Rockafellar, 1970, 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 also be used to obtain the maximum χ^2 under the independence model, where only the the marginal frequencies $\mathbf{r}$ and $\mathbf{c}$ are given and no other CA dimension is known. In the independence case, too, the value 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.

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 $\tilde{\mathbf{F}}$ 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 *The solution of constrained inverse CA is a bounded convex polyhedron that may be empty.*

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 the subspace may fall outside the polyhedron, e.g., by imposing an invalid constraint such as constraining f_{ij} to be larger than either of the corresponding marginals r_i or c_i, the union of the two sets may be empty. ■

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 Correspondence Analysis 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 $\mathbf{C}\mathbf{g}_{\psi^*} \geq \mathbf{d}$. 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 $\mathbf{g} = (\mathbf{C}'_{\phi^*}\mathbf{C}_{\phi^*})^{-1}\mathbf{C}'_{\phi^*}(\tilde{\mathbf{f}}_{\phi^*} - \mathbf{d}_{\phi^*})$. Upon checking the nonnegativity constraints $\mathbf{C}\mathbf{g} \geq \mathbf{d}$, 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 h-th vertex $\mathbf{g}_h$, that is $\mathbf{F}^h = \mathbf{D}_r^{-1/2}(\mathbf{U}_k\mathbf{\Lambda}_k\mathbf{V}_k' + \tilde{\mathbf{U}}_c\mathbf{G}_h\tilde{\mathbf{V}}_c')\mathbf{D}_c^{-1/2}$, where $\text{vec}(\mathbf{G}_h) = \mathbf{g}_h$ and let f_{ij}^h denote the ij-th 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. Also, let $\mathbf{F}_{\min}$ be the matrix with elements $\text{int}_+(\min_h(f_{ij}^h))$ (that is, the smallest ij-th element over all vertices) and $\mathbf{F}_{\max}$ have elements $\text{int}_-(\max_h(f_{ij}^h))$ (that is, the largest ij-th element over all vertices).

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. This follows directly from the convexity of the solution set F in Theorem 4 and the integer constraint for the elements of $\mathbf{F}$. ■

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

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

or in vec notation

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

These additional integer restrictions can be implemented as follows:

The Integer Constrained Inverse Correspondence Analysis Algorithm:

1. Find an initial set of vertices V by the Inverse Correspondence Analysis Algorithm of Section 4.

Table 1: Artificial smoking data of Greenacre (1984).The Pearson Chi-squared statistic for indepence is $\chi^2 = 16.44$.

	Smoking category				Row
Staff group	(1) None	(2) Light	(3) Medium	(4) Heavy	total $\mathbf{r}$
(1) Senior managers	4	2	3	2	11
(2) Junior managers	4	3	7	4	18
(3) Senior employees	25	10	12	4	51
(4) Junior employees	18	24	33	13	88
(5) Secretaries	10	6	7	2	25
Column total $\mathbf{c}'$	61	45	62	25	193

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 $\mathbf{C}_\psi \mathbf{g} = \left(\tilde{\mathbf{f}}_{\max} - \mathbf{d}\right)_\psi$
   - Check if $(\tilde{\mathbf{f}}_{\min} - \mathbf{d}) \leq \mathbf{C}\mathbf{g}_{\psi_1} \leq (\tilde{\mathbf{f}}_{\max} - \mathbf{d})$. 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 $(\tilde{\mathbf{f}}_{\min} - \mathbf{d}) \leq \mathbf{C}\mathbf{g}_{\psi_2} \leq (\tilde{\mathbf{f}}_{\max} - \mathbf{d})$. 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 (1984), see Table 1.

Suppose that in addition to the marginals $\mathbf{r}$ and $\mathbf{c}$, 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

Table 2: Vertices and reconstructed $\mathbf{F}$ by (6) of inverse correspondence analysis of the smoking data using $k = 3$.

Vertex	Reconstructed $\mathbf{F}$	χ^2
$\mathbf{g}_1 = \begin{bmatrix} -.1821 \\ -.2974 \end{bmatrix}$ $\mathbf{H}_1 =$	$\begin{bmatrix} 4.11 & .00 & 6.21 & 0.68 \\ 3.87 & 5.43 & 3.09 & 5.61 \\ 25.11 & 8.00 & 15.22 & 2.67 \\ 18.15 & 21.22 & 37.47 & 11.16 \\ 9.76 & 10.36 & .00 & 4.88 \end{bmatrix}$	$\chi^2_1 = 39.83$
$\mathbf{g}_2 = \begin{bmatrix} -.1821 \\ .1993 \end{bmatrix}$ $\mathbf{H}_2 =$	$\begin{bmatrix} 4.11 & .00 & 6.21 & 0.68 \\ 3.96 & 3.74 & 5.80 & 4.49 \\ 24.83 & 13.14 & 6.95 & 6.08 \\ 17.94 & 25.13 & 31.18 & 13.75 \\ 10.17 & 2.98 & 11.86 & .00 \end{bmatrix}$	$\chi^2_2 = 30.43$
$\mathbf{g}_3 = \begin{bmatrix} .2084 \\ -.2812 \end{bmatrix}$ $\mathbf{H}_3 =$	$\begin{bmatrix} 3.90 & 3.87 & 0 & 3.24 \\ 4.02 & 2.66 & 7.55 & 3.77 \\ 25.22 & 6.00 & 18.44 & 1.35 \\ 18.10 & 22.13 & 36.01 & 11.76 \\ 9.76 & 10.36 & 0 & 4.88 \end{bmatrix}$	$\chi^2_3 = 40.00$
$\mathbf{g}_4 = \begin{bmatrix} .2084 \\ .2155 \end{bmatrix}$ $\mathbf{H}_4 =$	$\begin{bmatrix} 3.90 & 3.87 & .00 & 3.24 \\ 4.11 & 0.97 & 10.26 & 2.66 \\ 24.94 & 11.14 & 10.16 & 4.76 \\ 17.89 & 26.04 & 29.72 & 14.35 \\ 10.17 & 2.98 & 11.86 & .00 \end{bmatrix}$	$\chi^2_4 = 33.71$

trivial solutions. We can derive $\tilde{\mathbf{U}}_c$ and $\tilde{\mathbf{V}}_c$ from $\mathbf{R}_k'\tilde{\mathbf{U}}_c = \mathbf{0}$ and $\mathbf{C}_k'\tilde{\mathbf{V}}_c = \mathbf{0}$. Applying the Inverse Correspondence Analysis Algorithm described in Section 4 with $\mathbf{C} = \tilde{\mathbf{V}}_c \otimes \tilde{\mathbf{U}}_c$ 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 $\mathbf{g}$ are obtained. Tabel 2 contains the four vertices and the corresponding reconstructed $\mathbf{F}$ matrices. Thus, any convex combination of these four vertices yields a CA solution with $\mathbf{R}_k$ and $\mathbf{C}_k$, the marginals are $\mathbf{r}$ and $\mathbf{c}$, and the elements of $\mathbf{F}$ are nonnegative. It may be verified that the convex combination $.1962\mathbf{H}_1 + .2866\mathbf{H}_2 + .2134\mathbf{H}_3 + .3038\mathbf{H}_4$ yields the original contingency matrix in Table 1.

Because $\mathbf{g}$ only contains two elements, a visual representation of the inverse CA solution can easily be obtained (see Figure 1). The axes represent the elements of $\mathbf{g}$, that is, g_1 and g_2. The area inside the polyhedron is 0.194. For a $\mathbf{g}$ 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 is presented in Table 3. Again it may be verified that

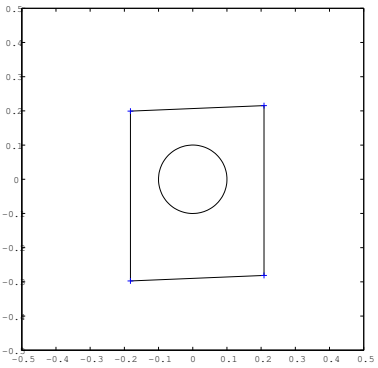

Figure 1: Polyhedron defined by inverse CA on the smoking data using $k = 3$. The vertices are indicated by crosses. The dimensions are g_1 and g_2. The circle indicates those $\mathbf{g}$ that satisfy the strict inverse CA condition.

for every convex combination of the two vertices the marginals are $\mathbf{r}$ and $\mathbf{c}$, a CA solution contains $\mathbf{R}_k$ and $\mathbf{C}_k$, the elements of $\mathbf{F}$ are nonnegative, and element $i = 1$ and $j = 4$ equals 2.

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 $\mathbf{g} = [.0199, \quad .0043]'$, with as corresponding $\mathbf{F}$ matrix the original contingency table in Table 1.

8 Conclusion and Discussion

In this paper, we have specified the set of matrices that all yield a given low dimensional configuration in its correspondence analysis 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. This task is reduced if the number of additional constraints on the elements is sufficiently large. We also specified a strict upper bound for the Pearson chi-squared statistic, not limited to inverse correspondence analysis, 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 origina unknown contingency table can be obtained. In this case, the inverse CA solution set

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

Vertex	Reconstructed $\mathbf{F}$	χ^2
$\mathbf{g}_1 = \begin{bmatrix} .0199 \\ -.2890 \end{bmatrix}$ $\mathbf{H}_1 =$	$\begin{bmatrix} 4.00 & 2.00 & 3.00 & 2.00 \\ 3.95 & 4.00 & 5.40 & 4.66 \\ 25.17 & 6.96 & 16.88 & 1.99 \\ 18.13 & 21.69 & 36.72 & 11.47 \\ 9.76 & 10.36 & .00 & 4.88 \end{bmatrix}$	$\chi^2_1 = 32.56$
$\mathbf{g}_2 = \begin{bmatrix} .0199 \\ .2077 \end{bmatrix}$ $\mathbf{H}_2 =$	$\begin{bmatrix} 4.00 & 2.00 & 3.00 & 2.00 \\ 4.04 & 2.31 & 8.11 & 3.54 \\ 24.88 & 12.11 & 8.61 & 5.40 \\ 17.91 & 25.60 & 30.42 & 14.06 \\ 10.17 & 2.98 & 11.86 & .00 \end{bmatrix}$	$\chi^2_2 = 24.76$

Table 4: Lower and upper bounds for the smoking data.

	Lower Bounds				Upper Bounds			
Staff group	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

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. However, 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 (De Leeuw & Groenen, 1997; Groenen et al., 1996) and principal components analysis (Ten Berge & Kiers, 1999), 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.

References

De Leeuw, J., & Groenen, P. J. F. (1997). Inverse multidimensional scaling. *Journal of Classification*, *14*, 3–21.

Eckart, C., & Young, G. (1936). Approximation of one matrix by another of lower rank. *Psychometrika*, *1*, 211–218.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Chichester: Wiley.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis.* New York: Academic Press.

Groenen, P. J. F., De Leeuw, J., & Mathar, R. (1996). Least squares multidimensional scaling with transformed distances. In W. Gaul & D. Pfeifer (Eds.), *Studies in classification, data analysis, and knowledge organization* (p. 177-185). Berlin: Springer.

Magnus, J. R., & Neudecker, H. (1988). *Matrix differential calculus with appilications in statistics and econometrics.* Chichester: Wiley.

Rockafellar, R. T. (1970). *Convex analysis.* Princeton, NJ: Princeton University Press.

Ten Berge, J. M. F., & Kiers, H. A. L. (1999). Retrieving the correlation matrix from a truncated PCA solution: The inverse principal component solution. *Psychometrika*, *64*, 317–324.

Van de Velden, M. (2000). *Topics in correspondence analysis.* Amsterdam: Tinbergen Institute, University of Amsterdam.