Regularized Nonsymmetric Correspondence Analysis

Yoshio Takane *, Sunho Jung

McGill University, Department of Psychology, 1205 Dr. Penfield Avenue, Montreal, QC, H3A 1B1, Canada

Abstract

Nonsymmetric correspondence analysis (NSCA) is designed to analyze two-way contingency tables in which rows and columns assume an asymmetric role, e.g., columns depend on rows, but not vice versa. A ridge type of regularization was incorporated into a variety of NSCA: Ordinary NSCA, and Partial and/or Constrained NSCA. The regularization has proven useful in obtaining estimates of parameters, which are on average closer to the true population values. An optimal value of the regularization parameter is found by a G-fold cross validation method, and the best dimensionality of the solution space is determined by permutation tests. A bootstrap method is used to evaluate the stability of the solution. A small Monte Carlo study and an illustrative example demonstrate the usefulness of the proposed procedures.

Key words: Reduced rank approximation, Covariates, Linear constraints, Ridge regularization method, Generalized singular value decomposition (GSVD), Permutation tests, G-fold cross validation, Bootstrap method

1 Introduction

In two-way contingency tables, rows and columns often assume an asymmetric role. Table 1 (Haberman, 1978, p.113) shows a typical example. This table was constructed by classifying 1,441 psychiatric patients by diagnostic group and the type of therapy. It is clear that the type of therapy is affected by the diagnostic group, while the reverse is not necessarily true. Such a situation

* Corresponding author (tel: 514-398-6125; fax: 514-398-4896). Matlab programs that carried out the analyses reported in this paper are available upon request.

Email addresses: takane@psych.mcgill.ca (Yoshio Takane), sunho.jung@mail.mcgill.ca (Sunho Jung).

Table 1
Number of psychotic patients in specified diagnostic groups receiving a principal type of psychiatric therapy (from Haberman, 1978, p. 113).

	Type of therapy			
Diagnostic group	Psychotherapy	Organic therapy	Custodial care	Total
Affective	30	102	28	160
Alcoholic	48	23	20	91
Organic	19	80	75	174
Schizophrenic	121	344	382	847
Senile	18	11	141	170
Total	236	560	646	1422

raises two interesting questions (Kroonenberg and Lombardo, 1999): 1) How strongly does a category in the diagnostic group predict a category in the type of treatment? 2) What is the best possible way of visualizing the predictive role of diagnostics on treatments? Nonsymmetric correspondence analysis (NSCA; D'Ambra and Lauro, 1992; Lauro and D'Ambra, 1984) is designed to answer such questions.

How should the predictive power of row i on column j be measured? Consider the size of the conditional probability of column j given row i relative to the size of the "average" (= unconditional) probability of column j. The larger the difference between the two, the larger the predictive power of row i on column j. (Throughout this paper, it will be assumed that rows represent the predictive categories, and columns the criterion categories.) Let p_{ij} denote the joint probability of row i and column j, and let $p_{i.} = \sum_{j=1}^{C} p_{ij}$ and $p_{.j} = \sum_{i=1}^{R} p_{ij}$ represent the marginal probabilities of row i and column j, respectively. Then, the conditional probability of column j given row i is given by $p_{j|i} = p_{ij}/p_{i.}$, and the unconditional probability of column j by the marginal probability $p_{.j}$ of column j. The predictive power of row i on column j is calculated by $a_{ij} = p_{ij}/p_{i.} - p_{.j}$. The generalized singular value decomposition (GSVD) is then used for visualizing the set of a_{ij} in a low dimensional space. Let $\mathbf{A}$ denote the $R \times C$ matrix with a_{ij} as the ij^{th} element, and let $\mathbf{P}$ denote the matrix of p_{ij} arranged in the same way. Let $\mathbf{P}_R$ and $\mathbf{P}_C$ represent the diagonal matrices of row and column marginal probabilities $p_{i.}$ and $p_{.j}$, respectively. Then,

$$\mathbf{A} = \mathbf{P}_R^{-1}\mathbf{P} - \mathbf{1}_R\mathbf{1}_C^{\top}\mathbf{P}_C, \tag{1}$$

where $\mathbf{1}_R$ and $\mathbf{1}_C$ are, respectively, R- and C-component vectors of ones. The GSVD of $\mathbf{A}$ is calculated with the row metric $\mathbf{P}_R$ and the identity column metric. The non-identity row metric $\mathbf{P}_R$ is used to reflect the size of row

marginal probabilities in a representation of rows and columns, as rows with larger probabilities should have more influence in the representation. In a resultant map constructed from the results of GSVD, the predictive power of row i on column j is indicated by the magnitude of the inner product of the two vectors representing the row and the column.

Contingency tables are often accompanied by some auxiliary information. For example, the diagnostic groups in Table 1 may be characterized by sets of scores on MMPI (Minnesota Multiphasic Personality Inventory) subscales. Such additional information can be incorporated as linear constraints on the rows of the table (Böckenholt and Böckenholt, 1990; Böckenholt and Takane, 1994; Hwang and Takane, 2002; Takane, Yanai, and Mayekawa, 1991; ter Braak, 1986). By imposing linear constraints on predictor categories, a variant of NSCA is obtained, called constrained NSCA. Constrained NSCA may be viewed as a nonsymmetric version of canonical correspondence analysis (CCA; ter Braak, 1986), which is based on symmetric CA. (While CCA analyzes a mutual relationship between rows and columns of a contingency table with constraints on either rows or columns, constrained NSCA analyzes a predictive relationship between the two with constraints on predictive categories.) Constraints may have an added benefit of stabilizing the estimates of parameters, provided that they are consistent with the predictive relationship in the data.

The predictive relationship between rows and columns of a contingency table is often mediated by variables other than those that define the rows and the columns of the table. For example, the psychiatric patients in Table 1 may differ among themselves in various aspects other than the diagnostic group and the type of treatment. They may be of different gender, of different age, with different socio-economic backgrounds, and so on. Part of the predictive relationship between the diagnostic group and the treatment may be due to these extraneous variables. The effects of these extraneous variables can be eliminated (Yanai, 1988) to capture the more intrinsic aspects of the predictive relationship between rows and columns. The resultant procedure may be called partial NSCA. Note, however, that partial NSCA requires patient level data; it is not feasible if the data are provided only in the form of a contingency table, which has no facet corresponding to individual patients.

Partial NSCA and constrained NSCA may be combined into a unified procedure, which may be called partial and constrained NSCA. In partial and constrained NSCA, not only are the effects of extraneous variables eliminated in predicting the criterion variable (columns), but linear restrictions are also imposed on the predictor variable (rows).

D'Ambra and Lauro (1989) proposed similar techniques for constrained and partial NSCA's in the specific context of three-way contingency tables. How-

ever, their proposal was rather limited in scope. Their procedures allow only a special kind of constraints and variables to be partialled out. (See sections 4.1 and 5 for more information about their procedures.) Our methods, on the other hand, are much more general. Any kind of linear constraints can be imposed, and the effects of any kind of variables can be eliminated.

So far, NSCA has been used to a large extent as a descriptive tool, although some effort has been expended to make it more inferential. These efforts have mostly been directed toward assessing the stability of parameter estimates in NSCA (Balbi, 1992; 1994) using the bootstrap method (Efron and Tibshirani, 1993). While this is an important development, there remains another important issue to be addressed. Solutions of NSCA are almost always obtained by the GSVD of $\mathbf{A}$ in (1) with the row metric $\mathbf{P}_R$ and an identity column metric. Does this method produce the best estimates of parameters in the sense that they are on average closest to the true population values? It will be shown in this paper that this is not always the case, and a better estimation method for NSCA is proposed. This method is easy to implement, almost as simple as the original method, and yet is capable of obtaining better quality estimates.

Our method is based on the ridge regularization method, which was originally developed by Hoerl and Kennard (1970) for multiple regression analysis. Let $\mathbf{X}$ and $\mathbf{y}$ represent a matrix of predictor variables and a vector of the criterion variable, respectively. In ridge regression, an estimate of the vector of regression coefficients $\mathbf{b}$ is obtained by $\hat{\mathbf{b}}(\lambda) = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$, where λ is a ridge parameter, and $\mathbf{I}$ is an identity matrix of appropriate order. A small positive value of λ often provides an estimate of $\mathbf{b}$ that is on average closer to true parameter values than their least squares (LS) counterpart, particularly when the sample size is small, and/or predictor variables are highly collinear (Hoerl and Kennard, 1970). A similar idea can be exploited to develop a better estimation procedure for NSCA.

The ridge type of regularization has recently been incorporated in a variety of multivariate data analysis techniques with considerable success. Takane and Hwang (2006) developed regularized multiple correspondence analysis (Greenacre, 1984; Nishisato, 1980), which subsumes symmetric CA as a special case. Takane and Hwang (2007) also developed a similar regularization procedure for redundancy analysis (RA), and Takane and Jung (2006) further extended the regularized RA to partial and/or constrained RA. The latter developments are particularly important for NSCA, since, as will be shown, NSCA turns out to be a special case of RA, in which both predictor and criterion variables are dummy coded categorical variables. The results given in Takane and Hwang (2007) and Takane and Jung (2006) on RA are readily extensible to NSCA with relatively minor modifications.

The rest of this paper is organized as follows. In section 2, first it is shown

that NSCA is indeed a special case of RA, and then how to incorporate a regularization procedure into ordinary NSCA is described. In the remainder of section 2, regularized partial and/or constrained NSCA is discussed in a similar fashion. Section 3 describes how to choose an optimal value of regularization parameter by cross validation. Permutation tests are used to determine the best dimensionality of the solution space. The bootstrap method is then described for evaluating the stability of the estimates of parameters. In section 4, two illustrative examples are given to demonstrate the usefulness of the proposed procedures. The final section gives concluding remarks.

2 The Method

In this section we develop methods of parameter estimation for regularized NCSA. We first discuss the case of regularized ordinary NSCA in some detail, and then extend it to the other varieties of NSCA discussed in the introduction section.

2.1 Ordinary Nonsymmetric Correspondence Analysis

For the sake of generality, we start with two dummy coded (indicator) data matrices, although in some cases only a contingency table calculated from these matrices is required for computation. Let $\mathbf{Z}_Y$ and $\mathbf{Z}_X$ denote n by C and n by R indicator matrices of criterion and predictor variables, respectively, where n is the number of subjects, C is the number of categories in the criterion variable, and R is the number of categories in the predictor variable. Note that $\mathbf{Z}_X^\top \mathbf{Z}_Y = n\mathbf{P}$, $\mathbf{D}_R \equiv \mathbf{Z}_X^\top \mathbf{Z}_X = n\mathbf{P}_R$, and $\mathbf{D}_C \equiv \mathbf{Z}_Y^\top \mathbf{Z}_Y = n\mathbf{P}_C$.

In NSCA, $\mathbf{Z}_Y$ and $\mathbf{Z}_X$ are usually columnwise centered to eliminate an intercept term in regression analysis. This is done by $\mathbf{Y} = \mathbf{Q}_n\mathbf{Z}_Y$ and $\mathbf{X} = \mathbf{Q}_n\mathbf{Z}_X$, where $\mathbf{Q}_n = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^\top/n$ is the centering matrix of order n, and $\mathbf{I}_n$ is the identity matrix of order n, and $\mathbf{1}_n$ is the n-component vector of ones. Note that the centering operation applied to an indicator matrix reduces the rank of the resultant matrix by 1. We attempt to predict as much as possible of variations in $\mathbf{Y}$ based on as small a number of components of $\mathbf{X}$ as possible. Let

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{2}$$

denote the multivariate regression model, where $\mathbf{B}$ is the matrix of regression coefficients, and $\mathbf{E}$ is the matrix of disturbance terms. We would like to find

$\mathbf{B}$ that minimizes

$$\phi(\mathbf{B}) = \text{SS}(\mathbf{E}) = \text{SS}(\mathbf{Y} - \mathbf{XB}), \tag{3}$$

where $\text{SS}(\mathbf{E}) = \text{tr}(\mathbf{E}^\top\mathbf{E})$, subject to $\text{rank}(\mathbf{B}) = r \leq \text{rank}(\mathbf{X}^\top\mathbf{Y}) \leq \min(\text{rank}(\mathbf{Y}), \text{rank}(\mathbf{X})) \leq \min(R-1, C-1)$. This is called redundancy analysis (RA; Van den Wollenberg, 1977). RA is a general-purpose technique to analyze predictive relationships between two sets of multivariate data (Lambert, Wildt, and Durand, 1988).

Let

$$\hat{\mathbf{B}} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y} \tag{4}$$

be a rank free least squares (LS) estimate of $\mathbf{B}$, where $(\mathbf{X}^\top\mathbf{X})^-$ indicates a generalized inverse (g-inverse) of $\mathbf{X}^\top\mathbf{X}$. Matrix $\mathbf{X}$ is necessarily a rank deficient matrix, and so a g-inverse of $\mathbf{X}^\top\mathbf{X}$ is always required. Then, (3) can be rewritten as

$$\phi(\mathbf{B}) = \text{SS}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) + \text{SS}(\hat{\mathbf{B}} - \mathbf{B})_{X^\top X}, \tag{5}$$

where $\text{SS}(\hat{\mathbf{B}} - \mathbf{B})_{X^\top X} = \text{tr}(\hat{\mathbf{B}} - \mathbf{B})'\mathbf{X}^\top\mathbf{X}(\hat{\mathbf{B}} - \mathbf{B})$ (e.g., Takane and Hwang, 2007). Since the first term in (5) is unrelated to $\mathbf{B}$, a reduced rank estimate of $\mathbf{B}$ can be obtained by minimizing the second term, which is accomplished by the generalized singular value decomposition (GSVD) of $\hat{\mathbf{B}}$ with the row metric matrix $\mathbf{X}^\top\mathbf{X}$ and the column metric matrix $\mathbf{I}$. This GSVD problem is written as $\text{GSVD}((\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{Y})_{X^\top X, I}$.

How is this solution related to the NSCA solution discussed in the introduction section? Note that $\mathbf{X}^\top\mathbf{Y} = \mathbf{Z}_X^\top\mathbf{Z}_Y - \mathbf{Z}_X^\top\mathbf{1}_n\mathbf{1}_n^\top\mathbf{Z}_Y/n = n(\mathbf{P} - \mathbf{P}_R\mathbf{1}_R\mathbf{1}_C^\top\mathbf{P}_C)$. However, $\mathbf{X}^\top\mathbf{X} \neq \mathbf{D}_R = \mathbf{Z}_X^\top\mathbf{Z}_X$. Fortunately, $\mathbf{X}^\top\mathbf{X}$ in the above GSVD problem can be replaced by $\mathbf{D}_R$ (Takane, Hwang, and Abdi, in press; Theorem 1 and Corollary 1 in Appendix (B)). When $\mathbf{D}_R$ is nonsingular as assumed in the introduction section, $\mathbf{D}_R^{-1}\mathbf{X}^\top\mathbf{Y} = (n\mathbf{P}_R)^{-1}n(\mathbf{P} - \mathbf{P}_R\mathbf{1}_R\mathbf{1}_C^\top\mathbf{P}_C) = \mathbf{A}$. The multiplicative factor n in the row metric $\mathbf{D}_R = n\mathbf{P}_R$ has no effect on the solution. Thus, the two solutions are equivalent. This shows that NSCA is a special case of RA when both sets of variables consist of indicator variables. A variety of useful extensions have been proposed for RA (Takane and Jung, 2006). Similar extensions may be useful for NSCA, including regularized NSCA, and partial and/or constrained NSCA. When $\mathbf{D}_R$ is not necessarily nonsingular, as will be assumed in the rest of this paper, the Moore-Penrose inverse of $\mathbf{D}_R$ (denoted as $\mathbf{D}_R^+$) may be used as a g-inverse of $\mathbf{X}^\top\mathbf{X}$ (Takane, et al., in press; Theorem 1).

We solve GSVD$(\mathbf{D}_R^+\mathbf{X}^\top\mathbf{Y})_{D_R,I}$. Let this GSVD be denoted by $\mathbf{D}_R^+\mathbf{X}^\top\mathbf{Y} = \mathbf{U}\Delta\mathbf{V}^\top$. This can be solved as follows. We first premultiply $\mathbf{D}_R^+\mathbf{X}^\top\mathbf{Y}$ by $\mathbf{D}_R^{1/2}$ to obtain $(\mathbf{D}_R^+)^{1/2}\mathbf{X}^\top\mathbf{Y}$, whose ordinary SVD is calculated. Let this SVD be denoted by $(\mathbf{D}_R^+)^{1/2}\mathbf{X}^\top\mathbf{Y} = \mathbf{U}^*\Delta^*\mathbf{V}^{*\top}$. Then, the above GSVD is obtained by $\mathbf{U} = (\mathbf{D}_R^+)^{-1/2}\mathbf{U}^*$, $\Delta = \Delta^*$, and $\mathbf{V} = \mathbf{V}^*$. To obtain the reduced rank approximation we retain only those portions of $\mathbf{U}$, Δ, and $\mathbf{V}$ pertaining to the r largest singular values. Let these portions of $\mathbf{U}$, Δ, and $\mathbf{V}$ be denoted by $\tilde{\mathbf{U}}$, $\tilde{\Delta}$, and $\tilde{\mathbf{V}}$, respectively. Then, the reduced rank estimate of $\mathbf{B}$ is obtained by $\tilde{\mathbf{B}} = \tilde{\mathbf{U}}\tilde{\Delta}\tilde{\mathbf{V}}'$. In the r-dimensional solution space, the standard coordinates of vectors representing criterion categories (columns) are defined by $\tilde{\mathbf{V}}$, and the principal coordinates of vectors representing predictive categories (rows) by $\tilde{\mathbf{U}}\tilde{\Delta}$. All subsequent GSVD problems in this paper are solved in essentially the same way.

We follow a similar line of development to obtain a reduced rank ridge LS (RLS) estimate of $\mathbf{B}$ (Takane and Hwang, 2007). In regularized ordinary NSCA, we minimize

$$\phi_\lambda(\mathbf{B}) = \mathrm{SS}(\mathbf{Y} - \mathbf{X}\mathbf{B}) + \lambda\mathrm{SS}(\mathbf{B})_{P_{X^\top}}, \tag{6}$$

where λ is the ridge parameter, $\mathbf{P}_{X^\top} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^-\mathbf{X}$ is the orthogonal projection operator onto the row space of $\mathbf{X}$. Let

$$\hat{\mathbf{B}}(\lambda) = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{P}_{X^\top})^-\mathbf{X}^\top\mathbf{Y} \tag{7}$$

represent a rank free RLS estimate of $\mathbf{B}$ that minimizes the above criterion. Then (6) can be rewritten as

$$\phi_\lambda(\mathbf{B}) = \mathrm{SS}(\mathbf{Y})_{Q_X(\lambda)} + \mathrm{SS}(\hat{\mathbf{B}}(\lambda) - \mathbf{B})_{X^\top X + \lambda P_{X^\top}}, \tag{8}$$

where

$$\mathbf{Q}_X(\lambda) = \mathbf{I} - \mathbf{P}_X(\lambda) = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{P}_{X^\top})^-\mathbf{X}^\top \tag{9}$$

(Takane and Hwang, 2007). Since the first term in (8) is unrelated to $\mathbf{B}$, (8) can be minimized by minimizing the second term, and the reduced rank RLS estimate of $\mathbf{B}$ is obtained by GSVD$(\hat{\mathbf{B}}(\lambda))_{X^\top X + \lambda P_{X^\top}, I}$. By Theorem 2 of Takane et al. (in press), $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{P}_{X^\top}$ in this GSVD can be replaced by $\mathbf{D}_R + \lambda\mathbf{P}_{X^\top}$, and its g-inverse by the Moore-Penrose inverse of the latter.

In both the LS and RLS estimations of ordinary NSCA, reduced rank estimates of $\mathbf{B}$ could be obtained by first obtaining a rank free estimate of $\mathbf{B}$ followed

by a GSVD. This follows for all the other variants of NSCA we discuss in this paper.

2.2 Partial Nonsymmetric Correspondence Analysis

Suppose that there are two sets of predictor variables. One is an indicator matrix $\mathbf{Z}_{X_1}$ just like $\mathbf{Z}_X$ in the previous section, and the other $\mathbf{Z}_{X_2}$, to be treated as covariates, could be a matrix of continuous variables, another matrix of indicator variables, or a mixture of the two types. We assume that all three data matrices concerned ($\mathbf{Y}$, $\mathbf{X}_1$, and $\mathbf{X}_2$) are columnwise centered. We write the model as

$$\mathbf{Y} = \mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2\mathbf{B}_2 + \mathbf{E}. \tag{10}$$

We impose a rank restriction on $\mathbf{B}_1$, but not on $\mathbf{B}_2$, since the latter pertains to the effects of extraneous variables in which we have no vested interest. For convenience, we orthogonalize the two predictor sets (Reinsel and Velu, 1998), and rewrite the model as

$$\mathbf{Y} = \mathbf{Q}_{X_2}\mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2\mathbf{B}_2^* + \mathbf{E}, \tag{11}$$

where

$$\mathbf{Q}_{X_2} = \mathbf{I} - \mathbf{P}_{X_2} = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2^\top\mathbf{X}_2)^-\mathbf{X}_2^\top \tag{12}$$

and

$$\mathbf{B}_2^* = \mathbf{B}_2 + (\mathbf{X}_2^\top\mathbf{X}_2)^-\mathbf{X}_2^\top\mathbf{X}_1\mathbf{B}_1. \tag{13}$$

In the LS estimation, we minimize

$$\phi(\mathbf{B}_1, \mathbf{B}_2^*) = \mathrm{SS}(\mathbf{Y} - \mathbf{Q}_{X_2}\mathbf{X}_1\mathbf{B}_1 - \mathbf{X}_2\mathbf{B}_2^*). \tag{14}$$

Let

$$\hat{\mathbf{B}}_1 = (\mathbf{X}_1^\top\mathbf{Q}_{X_2}\mathbf{X}_1)^-\mathbf{X}_1^\top\mathbf{Q}_{X_2}\mathbf{Y}, \tag{15}$$

and

$$\hat{\mathbf{B}}_2^* = (\mathbf{X}_2^\top\mathbf{X}_2)^-\mathbf{X}_2^\top\mathbf{Y} \tag{16}$$

be rank free LS estimates of $\mathbf{B}_1$ and $\mathbf{B}_2^*$ that minimize (14). Then (14) can be rewritten as

$$\phi(\mathbf{B}_1, \mathbf{B}_2^*) = \mathrm{SS}(\mathbf{Q}_X \mathbf{Y}) + \mathrm{SS}(\hat{\mathbf{B}}_1 - \mathbf{B}_1)_{X_1' Q_{X_2} X_1} + \mathrm{SS}(\hat{\mathbf{B}}_2^* - \mathbf{B}_2^*)_{X_2' X_2}, \quad (17)$$

where $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ (Takane and Jung, 2006). Since the first and the third terms in (17) are unrelated to $\mathbf{B}_1$, a reduced rank LS estimate of $\mathbf{B}_1$ can be obtained by minimizing the second term, which is achieved by $\mathrm{GSVD}(\hat{\mathbf{B}}_1)_{X_1' Q_{X_2} X_1, I}$. Matrix $\mathbf{X}_1^\top \mathbf{Q}_{X_2} \mathbf{X}_1 = \mathbf{X}_1^\top \mathbf{X}_1 - \mathbf{X}_1^\top \mathbf{P}_{X_2} \mathbf{X}_1$ in this GSVD may be replaced by $\mathbf{D}_1 - \mathbf{D}_{12} \mathbf{D}_2^+ \mathbf{D}_{12}^\top$, provided that $\mathrm{Sp}(\mathbf{1}_n) \subset \mathrm{Sp}(\mathbf{Z}_{X_2})$, and its generalized inverse by the Moore-Penrose inverse of the latter, where $\mathbf{D}_1 = \mathbf{Z}_{X_1}^\top \mathbf{Z}_{X_1}$, $\mathbf{D}_2 = \mathbf{Z}_{X_2}^\top \mathbf{Z}_{X_2}$, and $\mathbf{D}_{12} = \mathbf{Z}_{X_1}^\top \mathbf{Z}_{X_2}$. This follows from $\mathbf{Q}_{Z_{X_2}} = \mathbf{Q}_{Z_{X_2}} \mathbf{Q}_n = \mathbf{Q}_{X_2} \mathbf{Q}_n$, which in turn follows from $\mathbf{Q}_{Z_{X_2}} = \mathbf{Q}_{X_2} + \mathbf{P}_n$, where $\mathbf{P}_n = \mathbf{1}_n \mathbf{1}_n^\top / n$, and $\mathbf{Q}_n = \mathbf{I}_n - \mathbf{P}_n$ as defined earlier. Also, see Yanai and Puntanen (1993; Lemma 1(b)). The above replacement has a slight computational advantage when $\mathbf{D}_2$ is diagonal.

To incorporate ridge regularization in partial NSCA, we minimize

$$\phi_\lambda(\mathbf{B}_1, \mathbf{B}_2) = \mathrm{SS}(\mathbf{Y} - \mathbf{X}_1 \mathbf{B}_1 - \mathbf{X}_2 \mathbf{B}_2) + \lambda \mathrm{SS}(\begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix})_{P_{X^\top}}. \quad (18)$$

To minimize this criterion, we first rewrite the model (10) as

$$\mathbf{Y} = \mathbf{Q}_{X_2}(\lambda) \mathbf{X}_1 \mathbf{B}_1 + \mathbf{X}_2 \mathbf{B}_2^* + \mathbf{E}, \quad (19)$$

where

$$\mathbf{Q}_{X_2}(\lambda) = \mathbf{I} - \mathbf{P}_{X_2}(\lambda) = \mathbf{I} - \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{X}_2 + \lambda \mathbf{P}_{X_2^\top})^- \mathbf{X}_2^\top, \quad (20)$$

and

$$\mathbf{B}_2^* = \mathbf{B}_2 + (\mathbf{X}_2^\top \mathbf{X}_2 + \lambda \mathbf{P}_{X_2^\top})^- \mathbf{X}_2^\top \mathbf{X}_1 \mathbf{B}_1. \quad (21)$$

Let

$$\hat{\mathbf{B}}_1(\lambda) = (\mathbf{X}_1^\top \mathbf{Q}_{X_2}(\lambda) \mathbf{X}_1 + \lambda \mathbf{P}_{X_1^\top})^- \mathbf{X}_1^\top \mathbf{Q}_{X_2}(\lambda) \mathbf{Y}, \quad (22)$$

and

$$\hat{\mathbf{B}}_2^*(\lambda) = (\mathbf{X}_2^\top \mathbf{X}_2 + \lambda \mathbf{P}_{X_2^\top})^- \mathbf{X}_2^\top \mathbf{Y} \quad (23)$$

be rank free estimates of $\mathbf{B}_1$ and $\mathbf{B}_2^*$. Then (18) can be rewritten as

$$\begin{aligned}\phi_\lambda(\mathbf{B}_1, \mathbf{B}_2^*) &= \mathrm{SS}(\mathbf{Y})_{Q_X(\lambda)} \\ &+\mathrm{SS}(\hat{\mathbf{B}}_1(\lambda) - \mathbf{B}_1)_{X_1^\top Q_{X_2}(\lambda) X_1 + \lambda P_{X_1^\top}} + \mathrm{SS}(\hat{\mathbf{B}}_2^*(\lambda) - \mathbf{B}_2^*)_{X_2^\top X_2 + \lambda P_{X_2^\top}}, \end{aligned} \tag{24}$$

where, as before, $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ (Takane and Jung, 2006). Since the first and the third terms in (24) are unrelated to $\mathbf{B}_1$, a reduced rank RLS estimate of $\mathbf{B}_1$ can be obtained by minimizing the second term, which is achieved by $\mathrm{GSVD}(\hat{\mathbf{B}}_1(\lambda))_{X_1^\top Q_{X_2}(\lambda) X_1 + \lambda P_{X_1^\top}, I}$.

2.3 Constrained Nonsymmetric Correspondence Analysis

The model of constrained NSCA remains the same as (2). Additional information about the rows of a contingency table is incorporated in the form $\mathbf{B} = \mathbf{T}\mathbf{B}^*$, where $\mathbf{T}$ is a known constraint matrix. We may assume without loss of generality that $\mathbf{T}'\mathbf{P}_{X^\top}\mathbf{T} = \mathbf{I}$. (If a given $\mathbf{T}$ does not satisfy this condition, it can always be turned into one that satisfies the condition. Let $\mathbf{P}_{X^\top}\mathbf{T} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ represent the SVD of $\mathbf{P}_{X^\top}\mathbf{T}$. Then, we may redefine $\mathbf{T}$ by $\mathbf{U}$ and $\mathbf{B}^*$ by $\mathbf{D}\mathbf{V}^\top\mathbf{B}^*$.) In the LS estimation, we minimize

$$\phi(\mathbf{B}) = \mathrm{SS}(\mathbf{Y} - \mathbf{X}\mathbf{B}) = \mathrm{SS}(\mathbf{Y} - \mathbf{X}\mathbf{T}\mathbf{B}^*) = \phi(\mathbf{B}^*). \tag{25}$$

A rank free LS estimate of $\mathbf{B}^*$ can be obtained by $\hat{\mathbf{B}}^* = (\mathbf{T}^\top\mathbf{X}^\top\mathbf{X}\mathbf{T})^-\mathbf{T}^\top\mathbf{X}^\top\mathbf{Y}$ from which a rank free estimate of $\mathbf{B}$ can be obtained by

$$\hat{\mathbf{B}}^{(c)} = \mathbf{T}\hat{\mathbf{B}}^* = \mathbf{T}(\mathbf{T}^\top\mathbf{X}^\top\mathbf{X}\mathbf{T})^-\mathbf{T}^\top\mathbf{X}^\top\mathbf{Y}. \tag{26}$$

Then, (25) can be rewritten as

$$\phi(\mathbf{B}) = \mathrm{SS}(\mathbf{Q}_{XT}\mathbf{Y}) + \mathrm{SS}(\hat{\mathbf{B}}^{(c)} - \mathbf{B})_{X'X}, \tag{27}$$

where

$$\mathbf{Q}_{XT} = \mathbf{I} - \mathbf{X}\mathbf{T}(\mathbf{T}^\top\mathbf{X}^\top\mathbf{X}\mathbf{T})^-\mathbf{T}^\top\mathbf{X}^\top. \tag{28}$$

Since the first term in (27) is unrelated to $\mathbf{B}$, a reduced rank estimate of $\mathbf{B}$ can be obtained by minimizing the second term, which is achieved by $\mathrm{GSVD}(\hat{\mathbf{B}}^{(c)})_{X'X, I}$. As in ordinary NSCA, $\mathbf{X}'\mathbf{X}$ in this GSVD can be replaced by $\mathbf{D}_R$, and its g-inverse by the Moore-Penrose inverse of the latter.

In regularized constrained NSCA, we minimize (6) under the constraint that $\mathbf{B} = \mathbf{T}\mathbf{B}^*$. Let

$$\hat{\mathbf{B}}^{(c)}(\lambda) = \mathbf{T}(\mathbf{T}^\top\mathbf{X}^\top\mathbf{X}\mathbf{T} + \lambda\mathbf{P}_{X^\top})^-\mathbf{T}^\top\mathbf{X}^\top\mathbf{Y} \tag{29}$$

be a rank free RLS estimate of $\mathbf{B}$. Then, (6) can be rewritten as

$$\phi_\lambda(\mathbf{B}) = \mathrm{SS}(\mathbf{Y})_{Q_{XT}(\lambda)} + \mathrm{SS}(\hat{\mathbf{B}}^{(c)}(\lambda) - \mathbf{B})_{X'X+\lambda P_{X^\top}}, \tag{30}$$

where

$$\mathbf{Q}_{XT}(\lambda) = \mathbf{I} - \mathbf{X}\mathbf{T}(\mathbf{T}^\top\mathbf{X}^\top\mathbf{X}\mathbf{T} + \lambda\mathbf{P}_{X^\top})^-\mathbf{T}^\top\mathbf{X}^\top. \tag{31}$$

A reduced rank RLS estimate of $\mathbf{B}$ is obtained by $\mathrm{GSVD}(\hat{\mathbf{B}}^{(c)}(\lambda))_{X'X+\lambda P_{X^\top}, I}$. Again, $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{P}_{X^\top}$ in this GSVD can be replaced by $\mathbf{D}_R + \lambda\mathbf{P}_{X^\top}$, and its g-inverse by the Moore-Penrose inverse of the latter.

2.4 Partial and Constrained Nonsymmetric Correspondence Analysis

Partial and constrained NSCA follows essentially the same line of development as partial NSCA. The major distinction between them is that the predictor variables in the former are subject to both rank restriction and linear constraints, while in the latter only the rank restriction is imposed. Let

$$\hat{\mathbf{B}}_1^{(c)} = \mathbf{T}(\mathbf{T}^\top\mathbf{X}_1^\top\mathbf{Q}_{X_2}\mathbf{X}_1\mathbf{T})^-\mathbf{T}^\top\mathbf{X}_1^\top\mathbf{Q}_{X_2}\mathbf{Y} \tag{32}$$

be a rank free LS estimate of $\mathbf{B}_1$. Then a reduced rank estimate is obtained by $\mathrm{GSVD}(\hat{\mathbf{B}}_1^{(c)})_{X_1^\top Q_{X_2} X_1, I}$. As in partial NSCA, $\mathbf{X}_1^\top\mathbf{Q}_{X_2}\mathbf{X}_1$ can be replaced by $\mathbf{D}_1 - \mathbf{D}_{12}\mathbf{D}_2^+\mathbf{D}_{12}^\top$ and its g-inverse by the Moore-Penrose inverse of the latter, provided that $\mathrm{Sp}(\mathbf{1}_n) \subset \mathrm{Sp}(\mathbf{Z}_{X_2})$.

In the regularized estimation, let

$$\hat{\mathbf{B}}_1^{(c)}(\lambda) = \mathbf{T}(\mathbf{T}^\top\mathbf{X}_1^\top\mathbf{Q}_{X_2}(\lambda)\mathbf{X}_1\mathbf{T} + \lambda\mathbf{P}_{X_1^\top})^-\mathbf{T}^\top\mathbf{X}_1^\top\mathbf{Q}_{X_2}(\lambda)\mathbf{Y} \tag{33}$$

be a rank free RLS estimate of $\mathbf{B}_1$. Then, the reduced rank ridge estimate of $\mathbf{B}_1$ is obtained by $\mathrm{GSVD}(\hat{\mathbf{B}}_1^{(c)}(\lambda))_{X_1^\top Q_{X_2}(\lambda) X_1 + \lambda P_{X_1^\top}, I}$.

3 The choice of dimensionality and λ

There are two important decisions to be made in applying regularized NSCA: 1) How many dimensions should be retained in the solution, and 2) What is the "optimal" value of the ridge parameter?

We use permutation tests to choose the best dimensionality in the solution. They are easy to apply and have been proven useful in similar contexts (e.g., Takane and Hwang, 2002). The permutation test proceeds as follows: First, singular values (sv) are calculated from the original data. Then, the rows of predictor variables are randomly permuted, and sv's are calculated from the permuted data. The largest sv from the permuted data is compared with that from the original data. This is repeated many times, say 1000, and the number of times the sv's from the permuted data exceed the original sv is counted. If this count is smaller than 100α (where α is the prescribed significance level), the largest sv from the original data is considered significantly different from 0 at the α level. Subsequent sv's can be tested in a similar fashion by eliminating the effects of more dominant sv's. The chosen dimensionality is the number of sv's that are significantly different from 0. See Legendre and Legendre (1998), and ter Braak and Šmilauer (1998) for more general discussions on permutation tests in similar contexts.

The ridge parameter λ regulates the strength of the shrinkage effect. Since the optimal level of the shrinkage effect is usually not known in advance, an "optimal" value of λ has to be determined from data. We use cross validation to choose an optimal value of λ. In G-fold cross validation, we randomly divide the data into G groups, one of which is set aside as a test sample, while the remaining $G-1$ groups are used to estimate parameters in the model. The estimates obtained from the calibration sample are used to predict the test sample. We repeat this G times with each of the G test samples used in turn and calculate the mean squared prediction error. The whole procedure is repeated with the value of λ systematically varied (say, 0, 5, 10, 20, 50, 100). We choose the value of λ which gives the smallest mean squared prediction error. When $G = n$ (the number of cases in the original data) is taken, this method reduces to the leaving-one-out (LOO) method (aka the Jackknife method).

A bootstrap method (Efron and Tibshirani, 1993) can be used to assess the stability of parameter estimates. In the bootstrap method, we repeatedly draw random samples of size n (called bootstrap samples) from the original data set with replacement. We apply NSCA to each of the bootstrap samples to obtain parameter estimates. We then calculate means and variances of the estimates, from which we estimate biases and standard errors. The bootstrap method may also be used to test whether the estimated parameters are significantly positive or negative. Suppose that an estimate with the original data set happens to be

positive. We count the number of times the estimate of the same parameter is negative in bootstrap samples. If the relative frequency of the bootstrap estimates crossing over zero is less than a prescribed significance level (e.g., .05 or .01), we conclude that it is significantly positive.

4 Some Numerical Results

We first demonstrate the effectiveness of regularization in ordinary NSCA and constrained NSCA using a Monte Carlo technique, in which we sample data from a population, and the quality of estimation procedures is evaluated in terms of closeness of the estimates they produce to the population parameters. In the second analysis, we apply constrained NSCA to spider abundance data (ter Braak, 1986), where the objective is to predict the abundance of various kinds of spider from the environmental variables characterizing the sites where they are trapped. This data set was previously analyzed by canonical correspondence analysis (CCA; ter Braak, 1986), but because of the directional nature of the relationship between spiders and the sites, constrained NSCA seems to offer a more appropriate way of analyzing the data. We also demonstrate the positive effect of regularization using the bootstrap method. Specifically, we show that estimates obtained by regularized constrained NSCA are more reliable than those derived by non-regularized constrained NSCA.

4.1 A Monte Carlo Study Using Shoplifting Data

The quality of estimates can generally be assessed by how close they are on average to the corresponding true population values. One possible measure of this is the mean squared error (MSE), defined by $\mathrm{E}[\mathrm{SS}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})]$, where $\hat{\boldsymbol{\theta}}$ is the vector of parameter estimates, $\boldsymbol{\theta}$ is the vector of population parameters, and E indicates expectation. The MSE can be decomposed into two parts: the variance, $\mathrm{E}[\mathrm{SS}(\hat{\boldsymbol{\theta}} - \mathrm{E}(\hat{\boldsymbol{\theta}}))]$ (the expected value of squared difference between estimates and their expectations), and the squared bias, $\mathrm{SS}(\boldsymbol{\theta} - \mathrm{E}(\hat{\boldsymbol{\theta}}))$ (the squared difference between the true parameter values and the expected values of the estimates). It is known (e.g., Hoerl and Kennard, 1970) that in regression analysis, the LS estimates of regression coefficients have zero biases but large variances, while the ridge estimates have some biases (usually small), but much smaller variances than their LS counterparts. Consequently, the latter tend to have smaller MSE's.

Of course, the MSE cannot be calculated unless the values of population parameters are known. Fortunately, some contingency tables can be considered as representing populations, exhausting all cases under study. One such exam-

ple is the shoplifting data (Israëls, 1987), in which the total number of 33,101 people suspected of shoplifting in the Netherlands during 1977 and 1978 were cross-classified in terms of gender (male or females), age group (younger than 12, 12 to 14, 15 to 17, 18 to 20, 21 to 29, 30 to 39, 40 to 49, 50 to 64, and 65 and over), and categories of stolen goods (clothing, clothing accessories, provisions, tobacco, writing material, books, records, household goods, sweets, toys, jewelry, perfume, hobby tools, and others). Ordinary NSCA was first applied to the table with the combinations of age group and gender treated as predictor categories. This yields the population parameters. (The solution is not presented here, since Kroonenberg and Lombardo (1999) have already presented it along with a detailed explanation of how to interpret the configuration in NSCA.) One hundred data sets of varying sample sizes (N = 50, 100, 200, 300, and 500) were then sampled from the population contingency table, which were then analyzed by regularized ordinary NSCA with the value of the ridge parameter λ systematically varied (0, 2, 5, 10, 25, and 50), and the MSE evaluated as a function of λ.

Figure 1(A) presents MSEs as functions of the sample size (N) and λ. As can be seen, in all sample sizes the MSE goes down as the value of λ departs from zero, but has an upward trend after a while. This tendency is particularly clear for small sample sizes, although it can still be observed to a lesser extent for larger sample sizes. This means that a moderate value of λ yields the best estimates of parameters associated with the smallest MSE's. Figure 1(B) displays a breakdown of the MSE function for $N = 100$ into variance and squared bias. The variance consistently decreases as the value of λ increases, while the squared bias increases. The MSE attains its smallest value near $\lambda = 5$. Similar behavior of the MSE was first demonstrated in the context of univariate regression by Hoerl and Kennard (1970), and subsequently in many multivariate data analysis contexts by Takane and Hwang (2006; 2007), and Takane and Jung (2006).

Figures 1(C) and 1(D) show the results similar to Figures 1(A) and 1(B) respectively, for constrained NSCA. They are remarkably similar to those obtained for ordinary NSCA. In this analysis, predictor categories of the shoplifting data were created by interactive coding of the gender and age variables. The effects of the predictor categories were then constrained to be an additive function of the main effects of the gender and age variables. Incidentally, this is the kind of constrained NSCA suggested by D'Ambra and Lauro (1989) in the specific context of three-way contingency tables.

In the above, the RLS estimates ($\lambda > 0$) were indeed found to be better estimates than their non-regularized counterparts ($\lambda = 0$). However, we cannot assume that the effect of regularization is uniform across all the estimates. To illustrate the differential effects of regularization, we selected two predictor categories, one (male aged 65 and over) with a relatively small marginal

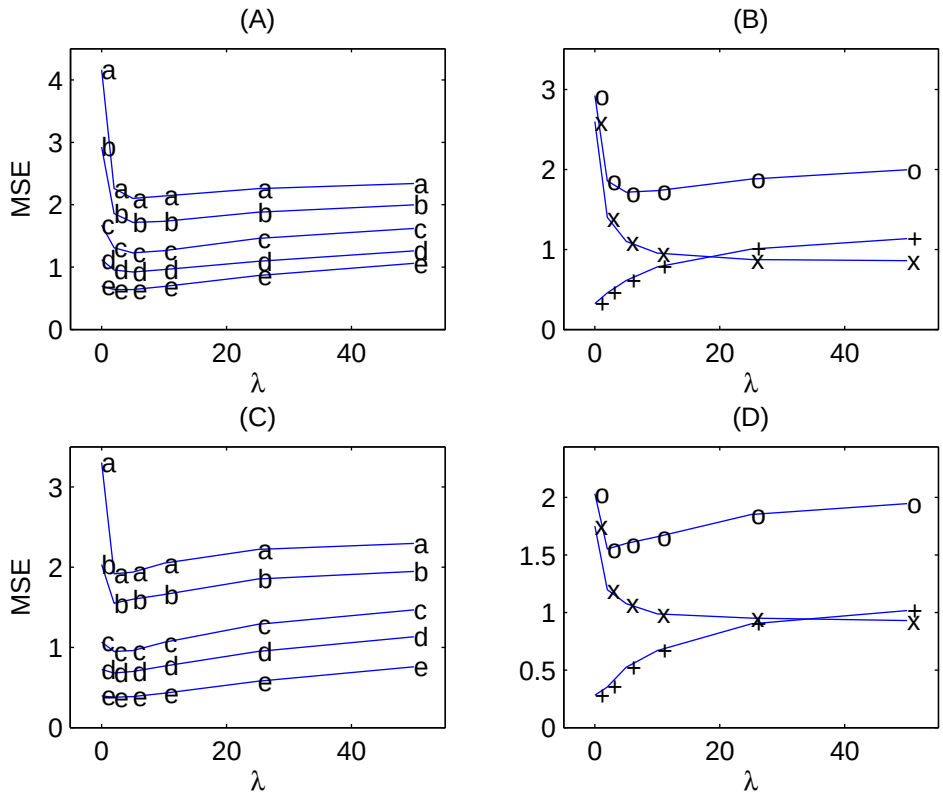

Fig. 1. (A) shows MSE as a function of the ridge parameter (λ) and the sample size, and (B) shows the breakdown of MSE into squared bias and variance for $N = 100$ for ordinary NSCA. (C) and (D) are similar to (A) and (B), but for constrained NSCA with additivity constraints on predictor categories. In (A) and (C), symbols a, b, c, d, and e indicate sample sizes of $N = 50, 100, 200, 500$, and 1000, respectively. In (B) and (D), o indicates MSE, x indicates Variance, and + indicates Squared Bias.

frequency (less than 2%), and the other (male aged between 12 and 14) with a relatively large marginal frequency (nearly 17%). Estimates of parameters were obtained corresponding to these categories by both LS and RLS for 100 random samples each of size $N = 100$ drawn from the original table. Figure 2(A) displays the scatter plot of the 100 LS estimates corresponding to the small marginal frequency category. The asterisk indicates the population parameters, while the circle indicates the mean of the estimates. Although the LS estimates are biased very little (as indicated by closeness of the population values and the mean of the estimates), they scatter widely in the solution space. Figure 2(B), on the other hand, presents the RLS estimates for the same category. The RLS estimates are more biased than the LS estimates. However, they are much more tightly clustered around the mean. This shows that regularization has the strongest effect on small frequency categories. Figure 2(C) and 2(D) show analogous results for the large marginal frequency category. In

this case, even the LS estimates are fairly tightly clustered around their population value. However, the RLS estimates are even more tightly clustered. The effect of regularization is less on large marginal frequency categories, although its effect in reducing the size of MSE can still be observed.

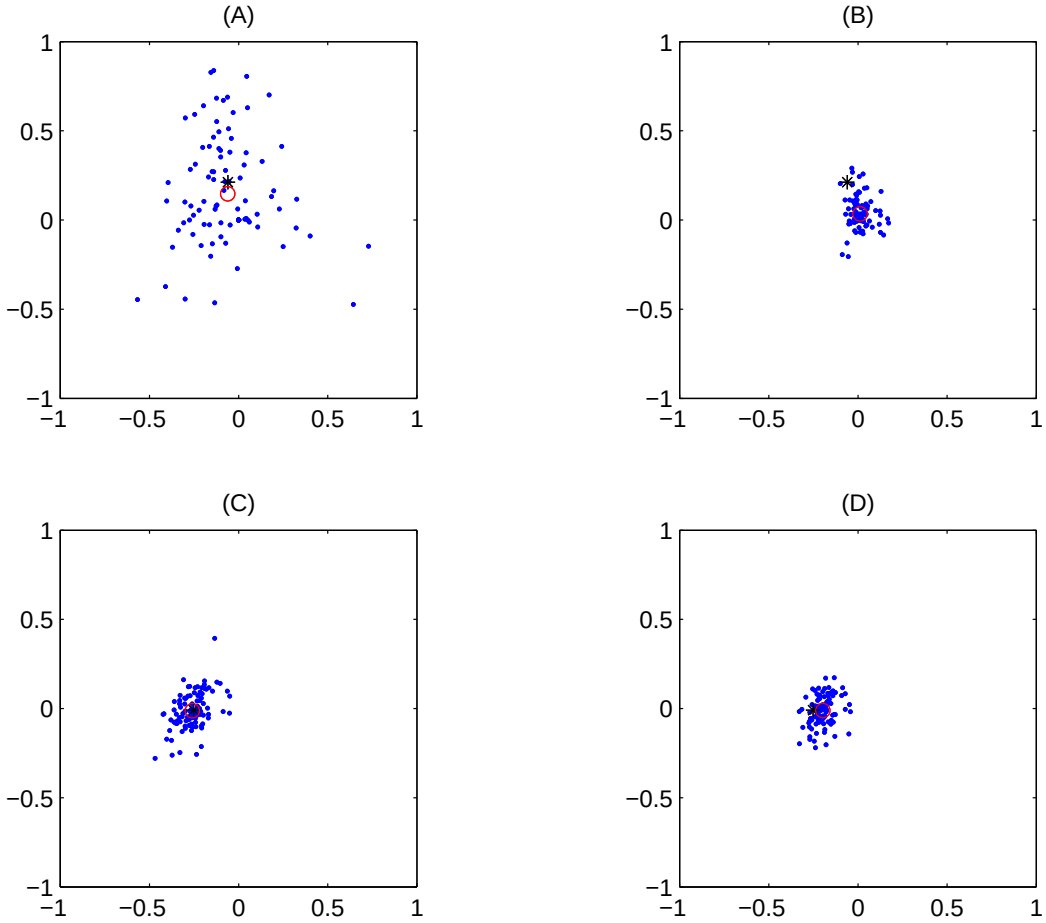

Fig. 2. (A) One hundred estimates (dots) of coordinates of vectors representing a predictor category with a small marginal frequency derived by non-regularized ordinary NSCA. (B) The same as (A), but obtained by regularized ordinary NSCA. (C) The same as (A), but for a predictor category with a large marginal frequency. (D) The same as (C), but obtained by the regularized estimation. In all of these figures, "*" indicates the population parameter, and "o" indicates the mean of the estimates.

4.2 Hunting Spider Abundance Data

The second example concerns a data set pertaining to the abundance of 12 species of hunting spiders at 28 sampled sites originally collected by Van der Aart and Smeek-Enserink (1975). (The data, as analyzed in the present study, were taken from ter Braak (1986), who applied CCA to this data set.) There

are also six environmental variables describing some aspects of environmental conditions of the sites such as: (ws) Water content (percentage of soil dry mass), (bs) Bare sand (percentage cover of bare sand), (cm) Cover moss (percentage cover of moss layer), (lr) Light refl. (reflection of the soil surface with cloudless sky), (ft) Fallen twigs (percentage cover of fallen leaves and twigs), and (ch) Cover herbs (percentage cover of herb layer). In ecology, it is important to understand the influence of environmental factors on species diversity. We thus focus on the predictive relationship between the environmental variables and the distribution of various kinds of hunting spiders. Criterion categories are the 12 species of hunting spider: (al) *Arctosa lutetiana*, (pl) *Pardosa lugubris*, (zs) *Zora spinimana*, (pn) *Pardosa nigriceps*, (pp) *Pardosa pullata*, (aa) *Aulonia albimana*, (tt) *Trochosa terricola*, (ac) *Alopecosa cuneata*, (pm) *Pardosa monticola*, (ae) *Alopecosa accentuana*, (af) *Alopecosa fabrilis*, and (ap) *Arctosa perita*. (Symbols in parentheses are plotting symbols in Figure 3.)

Constrained NSCA was applied with the environmental variables as additional information on predictor categories (sites). Permutation tests were first applied to determine the dimensionality of the solution. They uniformly indicated that the first component was highly significant ($s_1^2 = 18.67, p < .001$ (where s_1^2 indicates the sum of squares in the criterion categories that can be accounted for by the first component), as was the second ($s_2^2 = 8.33, p < .001$), and the third ($s_3^2 = 2.34, p < .001$), while the fourth component was not significant ($s_4^2 = 0.50, p > .670$). (The reported numbers are for $\lambda = 100$, which was later found to be optimal.) Thus, the best dimensionality was three. Although the third dimension is significant, its contribution is relatively small compared to the first two dimensions. A 55-fold cross validation was then applied with the value of λ varied systematically: 0, 10, 25, 50, 100, 150, and 200. The optimal value of λ was found to be 100. The cross validation also found that the three dimensional solution was the best, which was consistent with the result of the permutation tests.

Figure 3 shows the three-dimensional RLS configuration obtained by constrained NSCA. (It was found that the RLS configuration was shrunk slightly toward the origin compared to the LS configuration.) The predictive power of a particular site on a particular kind of spider can be evaluated by the magnitude of the inner product between two vectors representing these entities. For example, cover herbs labeled as (ch) is closest to Pardosa pullata labeled as (pp), suggesting that herbal sites predict the highest occurrences of the spider pp. To help interpret the derived components (dimensions), correlations between the environmental variables and the components were calculated and reported in Table 2. Component 1 seems to contrast dry sites with wet sites. The cm (Cover moss), lr (Light refl.), and bs (Bare sand) are positively correlated with this component, while wc (Water content) and ft (Fallen twigs) are negatively correlated. On this component, spiders ae (*Alopecosa accentuana*)

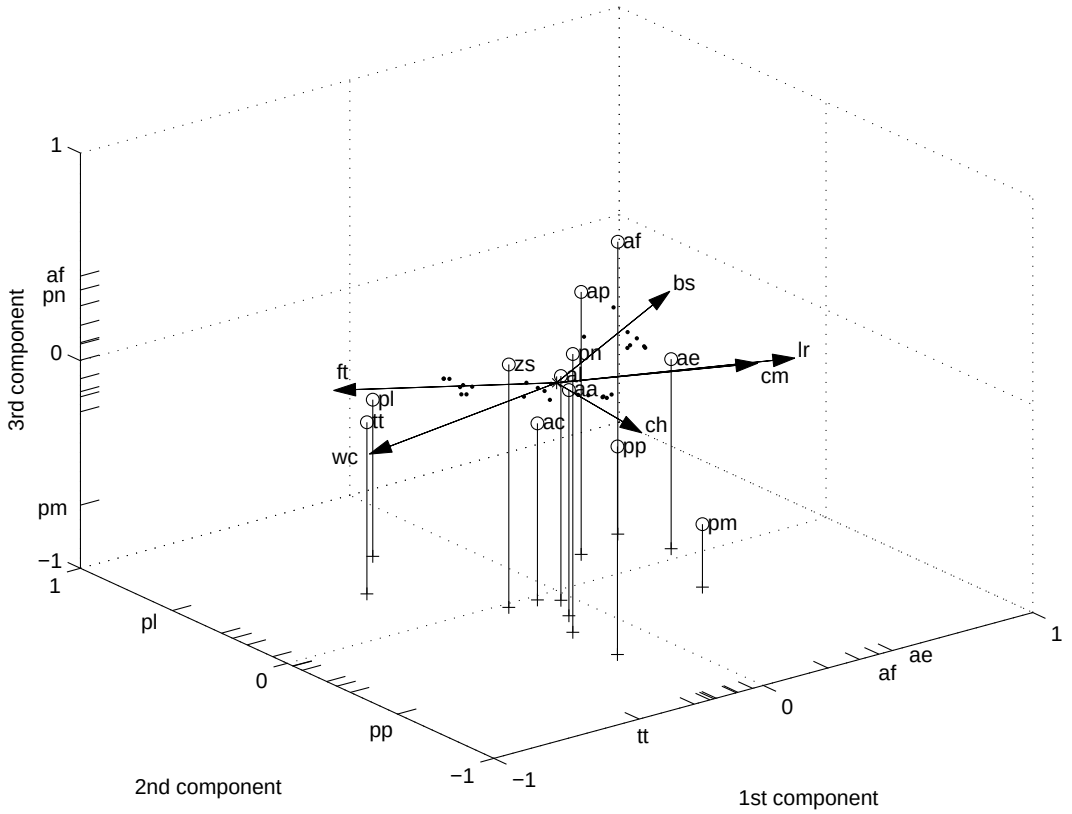

Fig. 3. Three-dimensional configuration of hunting spider abundance data by regularized NSCA. Circles indicate criterion categories (species of hunting spiders) in standard coordinates, dots indicate predictor categories (sampled sites) in principal coordinates, and arrowheads indicate the environmental variables.

and af (*Alopecosa fabrilis*) come at the positive end (preferring dry areas), while tt (*Trochosa terricola*) comes at the negative end (preferring wet sites). The second component separates herbal sites (ch) on the negative side and woody sites (ft) on the positive side. Spider pp (*Pardosa pullata*) seems to prefer herbal sites, while pl (*Pardosa lugubris*) woody areas. The third component contrasts bs (Bare sand) on the positive side and cm (Cover moss) on the negative side. Spiders af (*Alopecosa fabrilis*) and pn (*Pardosa nigriceps*) prefer bare sand, while pm (*Pardosa monticola*) prefers areas covered by moss.

Confidence ellipsoids indicate the degree of stability of parameter estimates. Since it is rather difficult to visualize them in three dimensions, a series of two-dimensional projections of the confidence ellipsoids were drawn. (They are sufficient to get a feel for which of the two estimation methods, regularized or non-regularized, will yield more reliable parameter estimates.) Figures 4 and 5 report the projected 95% confidence regions for the estimates of re-

Table 2
Correlations between the environmental variables and the components derived by the regularized constrained NSCA.

	Component		
E.V.	1	2	3
wc	-.910	-.256	.024
bs	.630	.316	.578
cm	.812	-.063	-.506
lr	.779	-.448	-.026
ft	-.583	.715	-.145
ch	-.097	-.927	.106

gression weights (the top row), predictor categories (the middle row), and criterion categories (the bottom row). Of the three columns, the first depicts the projections onto the first two dimensions, the second column to dimensions 1 and 3, and the last column to dimensions 2 and 3. Figure 4 presents the results of non-regularized LS estimation, and Figure 5 of regularized constrained NSCA. For both the regression weights and the predictor categories, the RLS estimates are consistently more reliable than the corresponding LS estimates. To avoid the impression that the RLS estimates look more reliable simply because they were shrunk toward 0, the RLS configurations were scaled up to match the size of the LS configurations. As can be seen, confidence regions are still smaller for the RLS estimates after this adjustment. (This scaling is presumed to have a bias equalizing effect. Note, however, that it is only for the purpose of the reliability comparison.) There does not seem to be any systematic differences in the reliability of the estimates of criterion categories between the RLS and LS estimations, which are only indirectly affected by the regularization.

5 Concluding Remarks

NSCA is a preferred method of analysis, when rows and columns of a two-way contingency table have a directional dependence structure. NSCA yields a graphical representation of the rows and columns in a low dimensional space. The best dimensionality of the solution can be determined in such a way that the low dimensional representation captures all the important aspects of the predictive relationship in the table. Another potential advantage of NSCA has been pointed out by Gimaret-Carpentier, Chessel, Pascal, and Ramesh (1999) in that NSCA is relatively unaffected by a rare criterion category, which often dominates the symmetric CA solutions. (This is because the symmetric

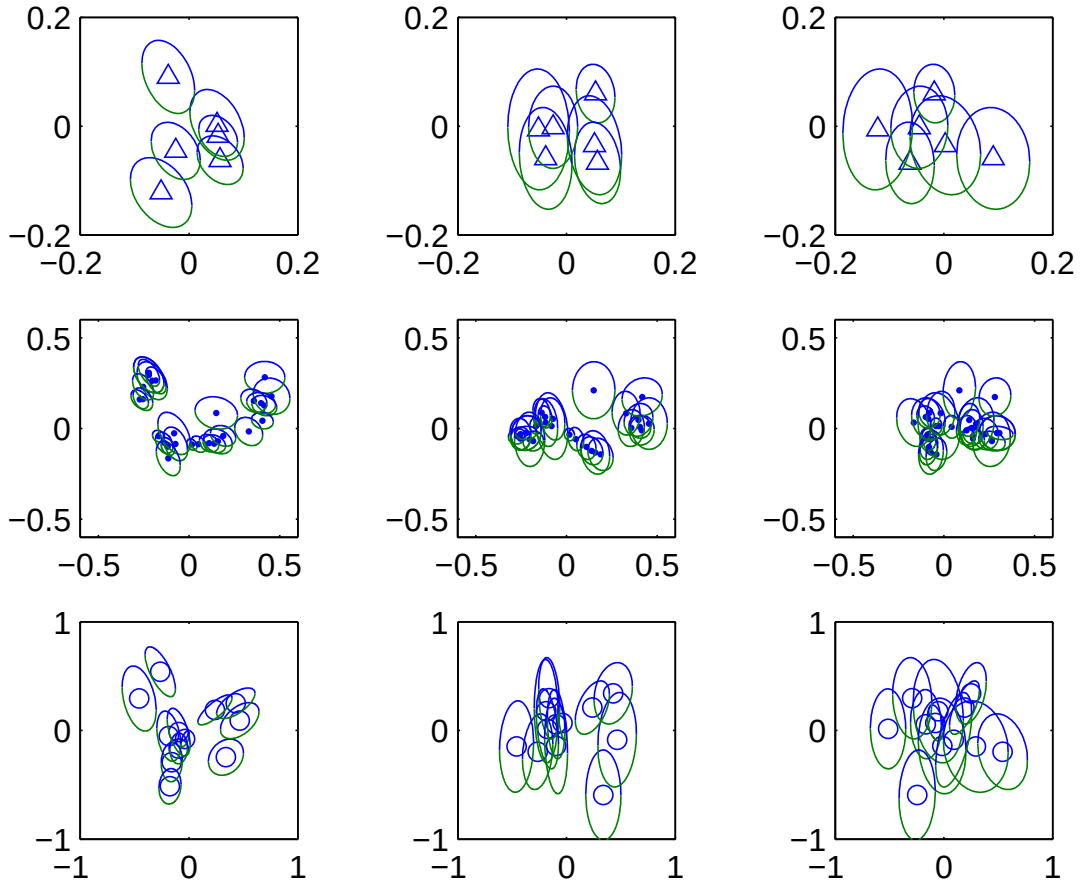

Fig. 4. Bi-dimensional 95% confidence regions for the estimates of weights for the environmental variables and of category vectors by non-regularized NSCA. Top row: Weights for the environmental variables. Middle row: Predictor categories (sampled sites) in principal coordinates. Bottom row: Criterion categories (species of hunting spiders) in standard coordinates. The first column: Dimension 1 vs 2. The second column: Dimension 1 vs 3. The third column: Dimension 2 vs 3.

CA, being a special case of canonical correlation analysis, treats all categories "equally" irrespective of their size.) Partial and/or constrained NSCA were introduced to incorporate external information into NSCA. Subjects under study may have some background information (e.g, demographic information). Partial NSCA eliminates the effects of extraneous variables in analyzing the relationships between rows and columns of a contingency table. Predictor categories may have some interesting structural relationship among themselves. Such information can be incorporated into NSCA as linear constraints, yielding constrained NSCA.

To obtain better estimates of parameters in NSCA, a ridge type of regularization was introduced into NSCA. It was shown through the analysis of two example data sets that this type of estimation was indeed capable of obtaining

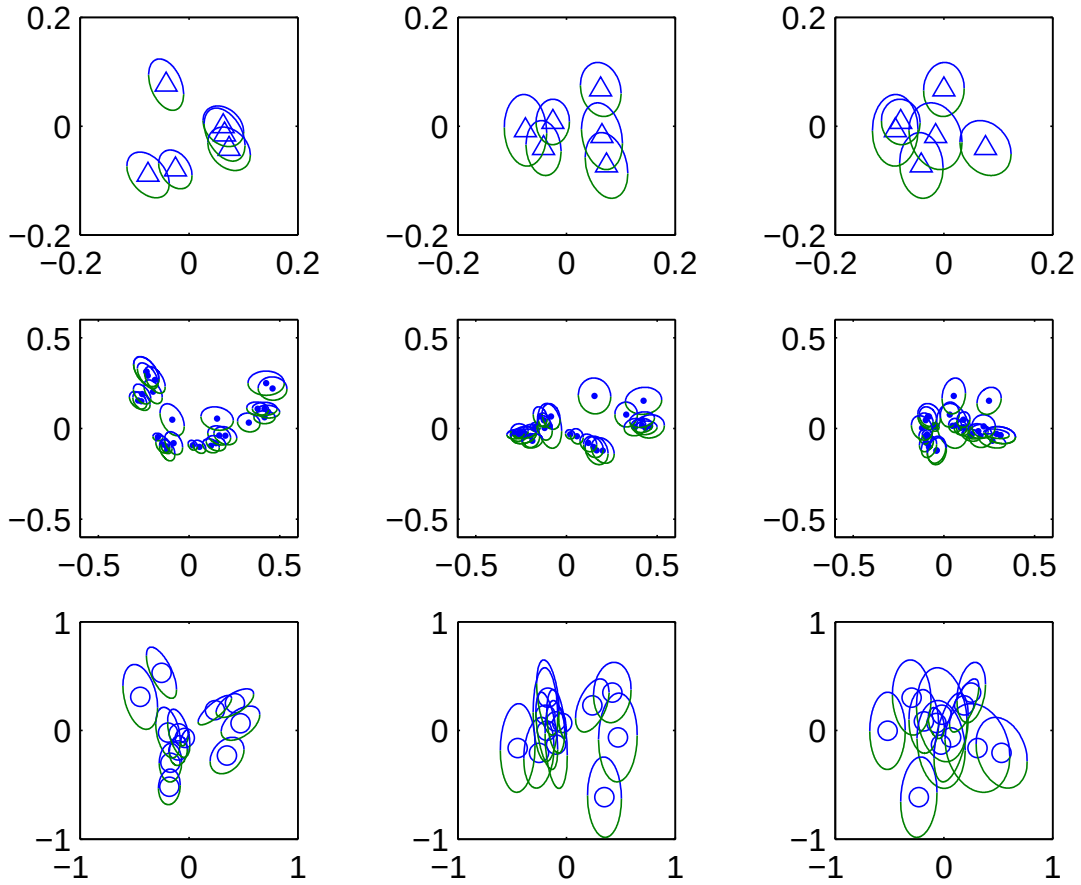

Fig. 5. The same as Figure 4, but obtained by regularized NSCA. Weights for the environmental variables (the first row) and predictor categories (the second row) have been scaled up to the size of the corresponding configurations Figure 4. This is presumed to have a bias equating effect between the two solutions.

estimates of parameters which were more stable, and which were on average closer to the true population values.

No real examples were given for partial NSCA or partial and constrained NSCA in this paper. This is largely due to a lack of suitable and interesting data sets. It is rare to find in literature the subject level indicator matrices with the additional subject level information. Higher order contingency tables are perhaps the only exceptional cases. As mentioned in the introduction section, D'Ambra and Lauro (1989) developed a special kind of partial NSCA for three-way contingency tables. In their procedure, one of the three categorical variables is taken as the criterion variable, while the other two are interactively coded and used as the predictor set. The main effect of one of the two predictor variables (say, A) is partialled out by conditioning on each level (category) of variable A. What remains is the simple main effects of the other predictor variable (say, B) at different levels of A. (The combined effect of these simple

main effects is equivalent to the overall main effect of A plus the interaction between A and B.) This is a special case of our partial NSCA and the kind that can also be dealt with as a special case of constrained NSCA by specifying appropriate constraints (Takane, Yanai, and Mayekawa, 1991). Thus, the kind of partial NSCA that can only be carried out by our procedure still awaits good example data sets.

6 Acknowledgement

We would like to thank Heungsun Hwang for his insightful comments on an earlier draft of this paper. We would also like to thank the editor and two anonymous reviewers for their constructive comments.

References

Balbi, S., 1992. On stability in nonsymmetrical correspondence analysis using Bootstrap. Statistica Applicata, 4, 543-552.

Balbi, S., 1994. Influence and stability in nonsymmetrical correspondence analysis. Metron, 52, 111-128.

Böckenholt, U., and Böckenholt, I., 1990. Canonical analysis of contingency tables with linear constraints. Psychometrika, 55, 633-639.

Böckenholt, U., and Takane, Y., 1994. Linear constraints in correspondence analysis. In M. J. Greenacre and J. Blasius (Eds.), Correspondence analysis in social sciences (pp.112-127). London: Academic Press.

D'Ambra, L., and Lauro, N. C., 1989. Non symmetrical analysis of three-way contingency tables. In R. Coppi and S. Bolasco (Eds.), Multiway data analysis, (pp. 301-315). Amsterdam: Elsevier.

D'Ambra, L., and Lauro, N. C., 1992. Non symmetrical exploratory data analysis. Statistica Applicata, 4, 511-529.

Efron, B., and Tibshirani, R. J., 1993. An introduction to the Bootstrap. Boca Raton, Florida: CRC Press.

Gimaret-Carpentier, C., Chessel, D., Pascal, J.-P., and Ramesh, B. R., 1999. Advantages of non-symmetric correspondence analysis in identifying multi-specific spatial patterns in the rain forest of the western ghats. Data management and modelling using remote sensing and GIS for tropical forest land inventory. Jakarta: Rodeo International Publishers.

Greenacre, M. J., 1984. Theory and applications of correspondence analysis. London: Academic Press.

Haberman, S. J., 1978. Analysis of qualitative data, (Vol. 1). Orlando, FL: Academic Press.

Hoerl, K. E., and Kennard, R. W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12, 55-67.

Hwang, H., Takane, Y., 2002. Generalized constrained multiple correspondence analysis. Psychometrika, 67, 215-228.

Israëls, A., 1987. Eigenvalue techniques for qualitative data. Leiden, The Netherlands: DSWO Press.

Kroonenberg, P., and Lombardo, R., 1999. Nonsymmetric correspondence analysis: A tool for analysing contingency tables with a dependence structure. Multivariate Behavioral Research, 34, 367-396.

Lambert, Z. V., Wildt, A. R., and Durand, R. M., 1988. Redundancy analysis: An alternative to canonical correlation and multivariate multiple regression in exploring interset associations. Psychological Bulletin, 104, 282-289.

Lauro, N. C., and D'Ambra, L., 1984. L'analyse non symétrique des correspondences. In E. Diday et al. (Eds.), Data analysis and informatics. (pp. 433-446). Amsterdam: Elsevier.

Legendre, P., and Legendre, L., 1998. Numerical ecology. The Second English Edition. Oxford: Elsevier.

Nishisato, S., 1980. Analysis of categorical data: Dual scaling and its applications. Toronto: University of Toronto Press.

Takane, Y., and Hwang, H., 2002. Generalized constrained canonical correlation analysis. Multivariate Behavioral Research, 37, 163-195.

Takane, Y., and Hwang, H., 2006. Regularized multiple correspondence analysis. In M. J. Greenacre and J. Blasius (Eds.), Multiple correspondence analysis and related methods, (pp. 259-279). London: Chapman and Hall.

Takane, Y., and Hwang, H., 2007. Regularized linear and kernel redundancy analysis. Computational Statistics and Data Analysis, 52, 394-405.

Takane, Y., Hwang, H., and Abdi, H., in press. Regularized multiple-set canonical correlation analysis. Psychometrika.

Takane, Y., and Jung, S., 2006. Regularized partial and/or constrained redundancy analysis. Submitted for publication.

Takane, Y., Yanai, H., and Mayekawa, S., 1991. Relationships among several methods of linearly constrained correspondence analysis. Psychometrika, 56, 667-684.

ter Braak, C. J. F., 1986. Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. Ecology, 67, 1167-1179.

ter Braak, C. J. F., and Šmilauer, P., 1998. CANOCO Reference Manual and User's Guide to Canoco for Windows. Ithaka, N.Y.: Microcomputer Power.

Van den Wollenberg, A. L., 1977. Redundancy analysis: Alternative for canonical analysis. Psychometrika, 42, 207-219.

Van der Aart, P. J. M., and Semeek-Enserink, N., 1975. Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. Netherlands Journal of Zoology, 25, 1-45.

Yanai, H., 1988. Partial correspondence analysis and its properties. In C. Hayashi, M. Jambu, E. Diday, and N. Osumi (Eds.), Recent developments

in clustering and data analysis, (pp. 259-266). Boston: Academic Press.
Yanai, H., and Puntanen, S., 1993. Partial canonical correlations associated with the inverse and some generalized inverse of a partitioned dispersion matrix. In K. Matsushita, M. L. Puri, and T. Hayakawa, (Eds.), Statistical sciecnces and data analysis (pp. 253-264). Utrecht: VSP International Science Publisher.