Regularized Multiple Correspondence Analysis

Yoshio Takane and Heungsun Hwang

1 Introduction

Multiple correspondence analysis (MCA) is a useful technique for the structural analysis of multivariate categorical data (Greenacre, 1984; Lebart, Morineau, & Warwick, 1984; Nishisato, 1980). MCA assigns scores to rows (representing the subjects) and columns (representing the response categories) of a data matrix, yielding a graphical display of the rows and the columns of the data matrix. The graphical display facilitates our intuitive understanding of the relationships among the categories of the variables.

MCA, however, has remained largely descriptive, although there have been some attempts to make it more inferential (Gifi, 1990; Greenacre, 1984, 1993; Markus, 1994). These attempts mostly focussed on the assessment of stability of solutions using a bootstrap resampling technique (Efron, 1979). However, inferential data analysis is not limited to assessing stability, but is also intended in estimating population characteristics using sample data. The quality of solutions should be assessed in terms of how close the estimates are to the corresponding population quantities.

An interesting question arises from this perspective: Is the conventional method of MCA the best method for obtaining estimates of parameters? In MCA, the conventional method is well established, is computationally simple, and has been in use almost exclusively in data analytic situations involving MCA. However, does it provide estimates that are on average closest to the population parameters? The answer is “not always”. In this chapter, we propose an alternative estimation procedure for MCA, called regularized MCA (RMCA), and demonstrate that in some cases it provides estimates that are on average closer to population parameters than the conventional estimation method. This method is easy to apply and is also computationally simple, in fact almost as simple as the conventional method.

The basic idea of RMCA comes from ridge regression, proven useful to mitigate multi-collinearity problems often encountered in multiple regression analysis (Hoerl & Kennard, 1970). Let $\mathbf{X}$ and $\mathbf{y}$ denote a matrix of predictor variables and an observed criterion vector, respectively, and let $\mathbf{b}$ denote a vector of regression coefficients. Then, the ordinary least squares (LS) estimates of regression coefficients are obtained by

$$\hat{\mathbf{b}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{1}$$

In ridge regression, on the other hand, regression coefficients are estimated by

$$\mathbf{b}^* = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}, \tag{2}$$

where the additional quantity, λ, is a regularization (or ridge) parameter, which typically takes a small positive value. The LS estimation provides the best (minimum variance) estimates among all linear unbiased estimates (BLUE) under mild distributional assumptions on errors. However, it may provide poor estimates of regression coefficients (associated with large variances) when the matrix $\mathbf{X}^\top\mathbf{X}$ is ill-conditioned (nearly singular) due to multi-collinearity (high correlations among the predictor variables). The ridge estimator, on the other hand, is biased but is more robust against multi-collinearity. A small positive number added to the diagonals of $\mathbf{X}^\top\mathbf{X}$ works almost magically to provide estimates that are more stable than the ordinary LS estimates.

The quality of parameter estimates is measured by the squared Euclidean distance between the estimates and parameters. If we take the expected value of the squared distance over replicated samples of data, we obtain mean squared error (MSE). MSE can be decomposed into two distinct parts. One is the squared bias (the squared distance between the population parameters and the mean of the estimates over replications), and the other is the variance (the average distance between individual estimates and the mean of the estimates). The LS squares estimates have zero bias, but they may have large variances (as is typically the case in the presence of multi-collinearity). The ridge estimates, on the other hand, although often biased, are also typically associated with a smaller variance, and most importantly, if the variance is small enough, ridge estimates may well have a smaller MSE than their LS counterparts. That is, in spite of their bias, they are on average closer to the population values. Indeed, for a certain range of values of λ, it is known that ridge estimators always have a smaller MSE than the ordinary LS estimates (Hoerl & Kennard, 1970) regardless of the existence of multi-collinearity problems. We exploit this fact to obtain better estimates in MCA.

To see the effect of regularization in MCA, let us look at Figure 1 as an example. This figure displays a two-dimensional configuration of response categories obtained by the usual (non-regularized) MCA of Nishisato's (1994) small survey data. In this data set, 23 subjects responded to four multiple-choice items each having three response categories. The questionnaire items, the associated response categories, and the data are given in Appendix (A). In this figure, each response category is labelled using an alphabet/number pairing; the alphabet indicates an item, and the number indicates a category number within the item. Ellipses surrounding the points are 95% confidence regions obtained by the bootstrap procedure (Efron, 1979). The smaller the ellipse, the more reliably the point is estimated. The ellipses are fairly large in all cases, indicating that the category points are not very reliably estimated. This is understandable from the fact that the sample size is rather small in this data set. Now let us look at Figure 2 for comparison. This figure is essentially the same as Figure 1, except that it was derived from the regularized MCA (RMCA). As can be seen, ellipses are almost uniformly smaller compared to those in Figure 1, indicating that the point locations are more reliably estimated by RMCA. This exemplifies the kind of benefit we might expect to get as a result of regularization. (More will be said about this example later).

***** Insert Figures 1 and 2 about here. *****

This chapter is organized as follows. In the next section, we present the proposed method of regularized MCA (RMCA) in some detail. We first (section 2.1) briefly discuss ordinary (non-regularized) MCA to provide a context for introducing regularization (section 2.2). We then discuss how to choose an optimal value of the regularization parameter (section 2.3). In section 3, we give some examples that illustrate the use of RMCA. We first (section 3.1) follow up on Nishisato's data and complete our discussion given above, and then (section 3.2) consider a large data set from Greenacre (1993; pp. 124-125) to demonstrate the conditions under which RMCA is most effective. Appendices give descriptions of the data used for illustration.

2 The Method

2.1 Multiple Correspondence Analysis (MCA)

In this section we briefly discuss ordinary MCA as an introduction to regularized MCA (RMCA), which we develop in the next section.

Let $\mathbf{Z}_k$ ($k = 1, \cdots, K$) denote an n (cases) by p_k (categories) matrix of raw indicator variables for the k^{th} categorical variable. We use $\mathbf{Z}$ to denote a block matrix formed by arranging $\mathbf{Z}_k$ side by side. That is, $\mathbf{Z} = [\mathbf{Z}_1, \cdots, \mathbf{Z}_K]$. Define a block diagonal matrix $\tilde{\mathbf{D}}$ consisting of $\tilde{\mathbf{D}}_k = \mathbf{Z}_k^\top \mathbf{Z}_k$ as the k^{th} diagonal block. Let $\mathbf{X}$ denote a column-wise centered data matrix obtained from the raw data matrix by $\mathbf{X} = \mathbf{Q}_n \mathbf{Z} = \mathbf{Z}\mathbf{Q}_{1_p/\tilde{D}}$, where $\mathbf{Q}_n$ is the centering matrix of order n (i.e., $\mathbf{Q}_n = \mathbf{I}_n - \mathbf{1}_n \mathbf{1}_n^\top / n$, where $\mathbf{1}_n$ is the n-element vector of ones), and $\mathbf{Q}_{1_p/\tilde{D}}$ is the block diagonal matrix with $\mathbf{Q}_{1_{p_k}/\tilde{D}_k} = \mathbf{I}_{p_k} - \mathbf{1}_{p_k}\mathbf{1}'_{p_k}\tilde{\mathbf{D}}_k/n$ as the k^{th} diagonal block where $\mathbf{1}_{p_k}$ is the p_k-element vector of ones. We assume $\mathbf{X}$ is partitioned in the same way as $\mathbf{Z}$ (i.e., $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_k]$). Define $\mathbf{D} = \tilde{\mathbf{D}}\mathbf{Q}_{1_p/\tilde{D}}$, which is the block diagonal matrix with $\mathbf{D}_k = \mathbf{X}_k^\top \mathbf{X}_k$ as the k^{th} diagonal block.

In MCA, we find the matrix of column scores (weights) $\mathbf{W} = [\mathbf{W}_1, \cdots, \mathbf{W}_K]$ partitioned in the same way as $\mathbf{X}$, that maximizes

$$\phi(\mathbf{W}) = \mathrm{tr}(\mathbf{W}^\top \mathbf{X}^\top \mathbf{X}\mathbf{W}), \tag{3}$$

subject to the ortho-normalization restriction that $\mathbf{W}^\top \mathbf{D}\mathbf{W} = \mathbf{I}_A$, where A is the dimensionality of the representation space. This leads to the following generalized eigen-equation,

$$\mathbf{X}^\top \mathbf{X}\mathbf{W} = \mathbf{D}\mathbf{W}\Delta^2, \tag{4}$$

where Δ^2 is the diagonal matrix of generalized eigenvalues in descending order of magnitude, and $\mathbf{W}$ is the matrix of generalized eigenvectors of $\mathbf{X}^\top \mathbf{X}$ with respect to $\mathbf{D}$. Once $\mathbf{W}$ and Δ^2 are obtained by solving the above eigen-equation, the matrix of row scores, $\mathbf{F}$, can be obtained by $\mathbf{F} = \mathbf{X}\mathbf{W}\Delta^{-1}$.

Essentially the same results can also be obtained by the generalized singular value decomposition (GSVD) of $\mathbf{XD}^-$ with row metric $\mathbf{D}$. (This is based on the well-known relationship between the generalized eigenvalue decomposition and the GSVD. See, for example, Takane (2002).) This is written as GSVD$(\mathbf{XD}^-)_{I_n,D}$, where $\mathbf{D}^-$ indicates a generalized inverse (g-inverse) of $\mathbf{D}$. (For definition and computation of GSVD, see Greenacre (1984) or Takane & Hunter (2001).) Let GSVD$(\mathbf{XD}^-)_{I_n,D}$ be denoted by $\mathbf{XD}^- = \mathbf{F}^*\Delta^*\mathbf{W}^{*\top}$, where $\mathbf{F}^*$ is the matrix of left singular vectors such that $\mathbf{F}^{*\top}\mathbf{F}^* = \mathbf{I}_r$, $\mathbf{W}^*$ is the matrix of right generalized singular vectors such that $\mathbf{W}^{*\top}\mathbf{DW}^* = \mathbf{I}_r$, and Δ^* is the positive-definite (*pd*) diagonal matrix of generalized singular values arranged in descending order of magnitude. Here, r is the rank of the column-wise centered data matrix, $\mathbf{X}$. Matrix $\mathbf{W}$ in (4) is obtained from $\mathbf{W}^*$ by retaining only the first A columns of $\mathbf{W}^*$ corresponding to the A largest generalized singular values, and matrix Δ is obtained from Δ^* by retaining only the first A rows and columns. The matrix of row scores $\mathbf{F}$ can be similarly obtained by retaining only the first A columns of $\mathbf{F}^*$.

Whether we use (4) or GSVD to obtain MCA solutions, we can replace $\mathbf{D}$ in the formulae by $\tilde{\mathbf{D}}$ (and $\mathbf{D}^-$ by $\tilde{\mathbf{D}}^{-1}$) without affecting the computational results. This simplifies the computation considerably because $\tilde{\mathbf{D}}$ is diagonal and can be directly calculated from the raw data, whereas the computation of $\mathbf{D}$ requires an additional step. Also, $\tilde{\mathbf{D}}^{-1}$ is much easier to compute than $\mathbf{D}^-$, which is a g-inverse of $\mathbf{D}$ of a specific kind ($\mathbf{D}$ is necessarily of rank deficient because each diagonal block of $\mathbf{D}$ is of rank deficient) to obtain $\mathbf{W}_k$ that satisfies $\mathbf{1}_{p_k}^\top \mathbf{D}_k \mathbf{W}_k = \mathbf{0}^\top$ for every k, which is a standard requirement in MCA solutions. The use of $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{D}}^{-1}$ automatically assures this requirement.

2.2 Regularized MCA

We now introduce a regularization procedure. Define

$$\tilde{\mathbf{D}}(\lambda) = \tilde{\mathbf{D}} + \lambda \mathbf{J}_p, \tag{5}$$

where λ is a regularization parameter (whose value is determined by some cross validation method, as will be discussed in the next section), and $\mathbf{J}_p$ is a block diagonal matrix with $\mathbf{J}_{p_k} = \mathbf{X}_k^\top (\mathbf{X}_k \mathbf{X}_k^\top)^- \mathbf{X}_k$ as the k^{th} diagonal block. (Matrix $\mathbf{J}_{p_k}$ is the orthogonal projector onto the row space of $\mathbf{X}_k$.) Also, define

$$\mathbf{D}(\lambda) = \tilde{\mathbf{D}}(\lambda)\mathbf{Q}_{1_p/\tilde{D}} = \tilde{\mathbf{D}}\mathbf{Q}_{1_p/\tilde{D}} + \lambda \mathbf{J}_p \mathbf{Q}_{1_p/\tilde{D}} = \mathbf{D} + \lambda \mathbf{J}_p. \tag{6}$$

Note that $\mathbf{J}_p \mathbf{Q}_{1_p/\tilde{D}} = \mathbf{J}_p$. In regularized MCA (RMCA), we maximize

$$\phi_\lambda(\mathbf{W}) = \mathrm{tr}(\mathbf{W}^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{J}_p)\mathbf{W}) \tag{7}$$

with respect to $\mathbf{W}$, subject to the ortho-normalization restriction that $\mathbf{W}^\top \mathbf{D}(\lambda)\mathbf{W} = \mathbf{I}_A$. This criterion is an extension of (3). The above criterion leads to the following generalized eigen equation analogous to (4):

$$(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{J}_p)\mathbf{W} = \mathbf{D}(\lambda)\mathbf{W}\Delta^2. \tag{8}$$

In order to derive the GSVD equivalent to the above generalized eigen-equation, we need to define a special metric matrix, $\mathbf{M}(\lambda)$, as follows:

$$\mathbf{M}(\lambda) = \mathbf{I}_n + \lambda(\mathbf{X}\mathbf{X}^\top)^+, \tag{9}$$

where $(\mathbf{X}\mathbf{X}^\top)^+$ indicates the Moore-Penrose inverse of $\mathbf{X}\mathbf{X}^\top$. Note that using $\mathbf{M}(\lambda)$, $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{J}_p$ can be rewritten as,

$$\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{J}_p = \mathbf{X}^\top \mathbf{M}(\lambda)\mathbf{X}, \tag{10}$$

assuming that $\mathbf{X}_k$'s are disjoint (i.e., $\text{rank}(\mathbf{X}) = \sum_{k=1}^{K} \text{rank}(\mathbf{X}_k)$). This condition is usually met in practical data analysis situations. See Takane & Hwang (2004) for more details of the derivation. The equivalent GSVD problem can now be stated as $\text{GSVD}(\mathbf{X}\mathbf{D}(\lambda)^-)_{M(\lambda),D(\lambda)}$.

As in the non-regularized case, $\mathbf{D}(\lambda)$ and $\mathbf{D}(\lambda)^-$ in (8) or in the above GSVD problem can be replaced by $\tilde{\mathbf{D}}(\lambda)$ and $\tilde{\mathbf{D}}(\lambda)^{-1}$, respectively. Again, this simplifies the computation. The exact rationale that allows this replacement can again be found in Takane & Hwang (2004).

The criterion, (7), can be further generalized into:

$$\phi_\lambda^{(L)}(\mathbf{W}) = \text{tr}(\mathbf{W}^\top(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{L})\mathbf{W}), \tag{11}$$

which is maximized with respect to $\mathbf{W}$, subject to the restriction that $\mathbf{W}^\top(\mathbf{D} + \lambda\mathbf{L})\mathbf{W} = \mathbf{I}_A$, where $\mathbf{L}$ is a block diagonal matrix with $\mathbf{L}_k$ as the k^{th} diagonal block. Matrix $\mathbf{L}_k$ could be any symmetric *nnd* (non-negative definite) matrix such that $\mathrm{Sp}(\mathbf{L}_k) = \mathrm{Sp}(\mathbf{X}_k')$, where $\mathrm{Sp}(\mathbf{Y})$ indicates the space spanned by the column vectors of $\mathbf{Y}$. This criterion leads to a solution of the following generalized eigen equation:

$$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{L})\mathbf{W} = \mathbf{D}(\mathbf{L})\mathbf{W}\Delta^2, \tag{12}$$

where $\mathbf{D}(\mathbf{L}) = \mathbf{D} + \lambda\mathbf{L}$. This generalization is often useful when we need a regularization term more complicated than $\lambda\mathbf{J}_p$. Such cases arise, for example, when we want to incorporate, by way of regularization, certain degrees of smoothness in the function to be approximated (e.g., Ramsay & Silverman, 1997). Adachi (2002) used this form of regularization to modulate the degree of smoothness of the trajectory describing changes in responses to a categorical variable over a period of time. In this case, $\mathbf{M}(\lambda)$ defined in (9) should also be generalized into:

$$\mathbf{M}^{(L)}(\lambda) = \mathbf{I}_n + \lambda(\mathbf{X}\mathbf{L}^+\mathbf{X}^\top)^+. \tag{13}$$

Properties similar to those for $\mathbf{M}(\lambda)$ hold for $\mathbf{M}^{(L)}(\lambda)$ as well.

2.3 The Choice of λ

In this section, we first discuss a cross validation procedure for selecting an optimal value of the regularization (ridge) parameter, λ. We then briefly discuss other nonparametric procedures that help make MCA more inferential.

We should note at the outset that a wide range of values of λ exists for which the regularization method works reasonably well, so that we do not have to be overly concerned about its choice. As will be shown in section 3.2, any value between 2 and 20 works substantially better than $\lambda = 0$, and they all give similar results. In this sense, the proposed regularization method is robust. Having said that, we may still want to have an objective procedure that can determine a near optimal value of λ. We may use some kind of cross validation method such as the bootstrap or the G-fold cross validation method. We use the latter without any good reason to favor one over the other. In this method the data set at hand are randomly divided into G sub-samples. One of the sub-samples is set aside, and estimates of parameters are obtained from the remaining data. These estimates are then used to predict the cases in the sample set aside to assess the amount of prediction error. We repeat this process G times, each time setting aside one of the G sub-samples in turn.

Let $\mathbf{X}^{(-g)}$ denote the data matrix with data in sample g eliminated from $\mathbf{X}$. We denote the data in sample g (that are eliminated) by $\mathbf{X}^{(g)}$. We apply RMCA to $\mathbf{X}^{(-g)}$ to obtain $\mathbf{W}^{(-g)}$, from which we calculate $\mathbf{X}^{(g)}\mathbf{W}^{(-g)}\mathbf{W}^{(-g)\top}$. This gives the cross validation prediction of $\mathbf{X}^{(g)}\mathbf{D}(\lambda)^{-}$. We repeat this for all G sub-samples, and collect all cross validated predictions, $\mathbf{X}^{(g)}\mathbf{W}^{(-g)}\mathbf{W}^{(-g)\top}$, in matrix $\widehat{\mathbf{X}\mathbf{D}}(\lambda)^{-}$. We then calculate

$$\varepsilon(\lambda) = \mathrm{SS}(\mathbf{XD}(\lambda)^{-} - \widehat{\mathbf{XD}}(\lambda)^{-})_{M(\lambda),D(\lambda)} \tag{14}$$

as an index of prediction error, where $\mathrm{SS}(\mathbf{Y})_{M(\lambda),D(\lambda)} = \mathrm{tr}(\mathbf{Y}^{\top}\mathbf{M}(\lambda)\mathbf{YD}(\lambda))$. We compare the value of $\varepsilon(\lambda)$ for different values of λ (e.g., $\lambda = 0, 1, 2, 5, 10, 20, 30$), and choose the value of λ associated with the smallest value of $\varepsilon(\lambda)$.

When G is taken equal to the size of the original data set, this procedure is called leaving-one-out (LOO) or jackknife method. The LOO method may be avoided for large data sets because it requires $G = N$ solutions of RMCA and could be quite time consuming, although for smaller sized data, it may still be used. In most cases, however, the G-fold cross validation method with $G < N$ gives similar results to the LOO method.

The above procedure for determining an optimal value of λ presupposes that we already know the number of dimensions in the RMCA solution. For dimensionality selection, we may use permutation tests similar to the one used by Takane & Hwang (2002) in generalized constrained canonical correlation analysis. Although the permutation tests may also be affected by the value of λ, our experience indicates that the best dimensionality is rarely, if ever, affected by the value of the regularization parameter. Thus, the permutation tests may be applied initially with $\lambda = 0$, by which a tentative dimensionality is selected, and the G-fold cross validation method is applied to select an optimal value of λ. We may then reapply the permutation tests with the selected optimal value of λ to make sure that the best dimensionality remains the same. General descriptions of the permutation tests for dimensionality selection can be found in Legendre & Legendre (1998), and ter Braak (1990).

We may also use a bootstrap method (Efron, 1979) to assess the reliability of parameter estimates derived by RMCA. In this procedure, random samples (called bootstrap samples) of the same size as the original sample are repeatedly sampled from the original sample with replacement. RMCA is applied to each bootstrap sample to obtain estimates of parameters each time. We then calculate the mean and the variance-covariance of the estimates across the bootstrap samples, from which we calculate estimates of standard errors of the parameter estimates, or draw confidence regions to indicate how reliably parameters are estimated. The latter is done under the assumption of asymptotic multivariate normality of the parameter estimates. Figures 1 and 2 presented earlier for Nishisato's data were obtained in this way.

When the assumption of asymptotic normality is suspect, we may use a nonparametric method for constructing confidence regions (e.g., Markus, 1994). Alternatively, we may simply plot as many point estimates (as we obtained by the bootstrap procedure) in the configuration of category points (Greenacre, 1984; 1993). This is usually good enough to give a rough indication of how tightly or loosely category points are estimated. See Figures 8 and 9 in section 3.2.

Significance tests of the elements of $\mathbf{W}$, may also be performed as byproducts of the bootstrap method described above. We simply count the number of times bootstrap estimates "cross" the value of zero (i.e., if the original estimate obtained from the original sample is positive, we count the number of times the corresponding bootstrap estimates turn out to be negative, and vice versa). If the relative frequency (the p-value) of the cross-over is smaller than a prescribed α level, we conclude that the corresponding parameter is significantly different from zero.

3 Examples

We report two sets of numerical results in this section. One involves Nishisato's (1994) small survey data we have discussed previously in the introduction section. The other pertains to Greenacre's (1993) car purchase data. This is a huge data set (over half a million cases) representing the total population of car purchases made in some country during a certain period of time. Because it is a population data set, we can sample data of varying sizes from the population, and directly estimate MSE as well as bias and variance of the estimates that result from a certain estimation procedure. We can directly compare these quantities across different values of λ by systematically varying its value.

3.1 Analysis of Nishisato's Data Continued

As mentioned earlier, these data consist of 23 subjects responding to four items each having three response categories. Information regarding this data set is given in Appendix (A).

Permutation tests indicated that the first dimension (corresponding to the singular value of 2.59) was highly significant ($p = 0$), while the second dimension (corresponding to the singular value of 1.79) was only marginally so ($p = .07$). The p-values were reported only for $\lambda = 2$ found to be optimal for this data set. However, a similar pattern of significance (p-values) was observed for other values of λ. All subsequent analyses on this data set assumed $A = 2$.

The LOO method was then applied to find an optimal value of the regularization parameter. (The LOO method was feasible because this was a small data set.) The estimate of prediction error (ε) was found to be .391 for $\lambda = 0$, .383 for $\lambda = 1$, .382 for $\lambda = 2$, and .391 for $\lambda = 5$. Thus, an optimal value of λ was found to be 2.

A bootstrap method was used to assess the reliability of the parameter estimates. This was done for both $\lambda = 0$ (non-regularized MCA) and the optimal value of $\lambda = 2$ (RMCA) for comparison. One thousand bootstrap samples were generated, parameter estimates were obtained for each sample, and the mean and the variance-covariance estimates of the estimated parameters were calculated, from which the 95%

confidence regions depicted in Figures 1 and 2 were drawn. As we have seen already, confidence regions are almost uniformly smaller for the parameter estimates obtained by RMCA (with the optimal value of $\lambda = 2$) than those obtained by non-regularized MCA (under $\lambda = 0$), indicating that the parameters are more reliably estimated in the former. The regularization tends to shrink the estimates. To avoid getting smaller confidence regions just because of the shrinking effects, the configuration obtained by RMCA was sized up to match the size of the configuration obtained by the ordinary MCA, and the variance-covariance estimates of the former were adjusted accordingly.

Confidence regions were also drawn for subject points (row scores). This was done as follows. The column scores ($\mathbf{W}$) derived from each bootstrap sample were applied to the original data to derive estimates of the coordinates of subject points. The mean and the variance-covariance estimates of the subject points were calculated over the bootstrap samples, and the confidence regions were drawn in the same way as before. Figures 3 and 4 display the configurations of the subject points along with the 95% confidence regions obtained by MCA and RMCA, respectively. Again, we find that the subject points were more reliably estimated by RMCA. Confidence regions are almost uniformly smaller in Figure 4 than in Figure 3. This corroborates our finding in Figures 1 and 2.

***** Insert Figures 3 and 4 about here *****

3.2 Analysis of Greenacre's Car Purchase Data

The second data set we analyze comes from Greenacre (1993, pp. 124-125). The total number of 581,515 cars purchased in the USA during the last quarter of year 1988 were cross classified in terms of 14 size classes of car and purchaser profiles, and are reported in the form of a two-way contingency table. The purchaser profiles were defined by the age of the oldest person in the household of the purchaser, and the income of the household. The age variable was classified into 7 groups, and the income variable into 9 levels, which are factorially combined to create 63 categories. More detailed descriptions of the categories in the three variables (size classes, age and income) can be found in Appendix (B).

Because the data are population data, there is no inferential problem arising. On the other hand, this provides a golden opportunity to perform various sampling experiments on the data to examine the quality of estimates against population parameters. In particular, we can directly estimate MSE as well as its breakdown into squared bias and variance by applying RMCA to data sampled from this population. We may compare MSE across different values of the regularization parameter including the non-regularized case ($\lambda = 0$) to assess the effect of regularization. We may also systematically vary the sample size.

In order to apply MCA, the data were first rearranged into a multiple-choice data format. We took the three variables (size classes, age and income) all constituting column categories defining profiles of purchases. Rows of this data represent 882 (= 14 × 63) distinct profiles of purchases indicated by patterns of size classes of cars purchased, age groups and income levels of the household of purchasers. Figure 5 presents the two-dimensional population configuration of the 30 (= 14 + 7 + 9) category points obtained by applying the ordinary MCA to this data set. Size classes of car are indicated by upper case alphabetic characters (A through N), age groups by small "a" followed by the number indicating the age group (a1 through a7), and income levels by small "i" followed by the number indicating the income level (i1 through i9). Age groups are connected by dotted line segments, and so are income levels.

The direction that goes from the bottom right corner to the top left corner roughly corresponds with the size dimension with the bottom right representing larger cars and the top left smaller cars. The direction perpendicular to this dimension (going from the bottom left corner to the top right corner) roughly corresponds with the price dimension with the left bottom corner representing more expensive cars as opposed to the top right corner representing more economical cars. The age variable is more or less linearly related to the size dimension with older people tending to prefer larger cars. Its relation to the price dimension, however, is quadratic with middle-aged people preferring more expensive cars, while younger and older people more economical cars. The higher end of the income variable is linearly related to the price dimension, while the lower end is more in line with the smaller side of the size dimension. These results are similar to those of Greenacre's (1993) obtained by an analysis of data in contingency table form, except that the current configuration is rotated about 45^o.

***** Insert Figure 5 about here *****

We then obtained 100 samples each of varying sizes (N = 200, 500, 1000, 2000, and 5000) from this data set, applied RMCA to those sampled data with the value of regularization parameter systematically varied ($\lambda = 0$, 5, 10, 20, 30, 40, and 50), and calculated MSE, squared bias and variance. Since this is a multi-parameter situation, these quantities have to be aggregated across parameters. The simple sum of squared discrepancies was taken as an aggregate measure of discrepancies. These aggregated measures of discrepancies are then averaged across the samples. Let θ denote the vector of parameters, $\hat{\theta}_i$ their estimate from sample i, and $\bar{\theta}$ the mean of $\hat{\theta}_i$ across samples. Then,

$$\text{Squared Bias} = (\bar{\theta} - \theta)^\top (\bar{\theta} - \theta),$$

$$\text{Variance} = (1/I) \sum_{i=1}^{I} (\hat{\theta}_i - \bar{\theta})^\top (\hat{\theta}_i - \bar{\theta}),$$

and

$$\text{MSE} = (1/I) \sum_{i=1}^{I} (\hat{\theta}_i - \theta)^\top (\hat{\theta}_i - \theta),$$

where I is the number of sampled data sets within each condition. In the present case, $I = 100$ in all conditions.

Figure 6 displays average MSE's plotted as a function of the sample size and the value of regularization parameter. Average MSE goes down dramatically from the non-regularized case ($\lambda = 0$) to regularized cases in all sample sizes. MSE takes the smallest value for λ between 10 and 20 in all cases. A few remarks are in order. First of all, if we do not regularize, we need a much larger sample size to achieve the same degree of MSE than when we regularize with a near optimal value of the regularization parameter. For example, to achieve without regularization the level of MSE (with $\lambda = 20$) achieved for a sample size of $N = 500$, we roughly need four times as many observations ($N = 2000$). This ratio diminishes as the sample size increases. However, it can still be substantially larger than 1 for the sample size as large as $N = 5000$. Secondly, MSE does not go up drastically even if we overshoot its value, that is, if we happen to choose too large a value of λ by mistake. This indicates that we do not have to be overly concerned about the choice of the value of the regularization parameter. This tendency holds across different sample sizes. The computation of MSE in Figure 6 presupposed a two-dimensional configuration of category points. However, essentially the same results hold for other dimensionalities, including the single dimensional case.

***** Insert Figure 6 about here *****

Figure 7 breaks down the MSE presented in Figure 6 into squared bias and variance components for $N = 500$. The squared bias tends to go up, while the variance goes down, as the value of λ increases. The sum of these quantities, MSE, takes the smallest value somewhere in the mid range. This is the characteristic of the MSE function that Hoerl and Kennard (1970) theoretically derived in the context of ridge regression. It is reassuring to find a similar tendency in RMSA, although the curves depicted in Figure 6 were derived empirically, and not theoretically. Although only the results for $N = 500$ are presented, essentially the same tendency was observed for other sample sizes. Also, these curves were derived in all cases from two-dimensional solutions. As before, a similar tendency holds for other dimensionalities.

***** Insert Figure 7 about here *****

To compare the quality of estimates obtained from RMCA with those obtained from the non-regularized MCA, two categories of the size class variable were picked out and subjected to further examination. Those two categories are Class C and Class A cars. The former has a relatively small marginal frequency (9,626 purchases out of 581,515), the latter a relatively large observed frequency (68,977). These translate into less than 2% and nearly 12% of the total purchases, respectively. It is anticipated that the sampling error is much larger in the small frequency category.

Figure 8(a) plots estimates of the point representing Class C cars obtained by the non-regularized MCA from 100 samples of size $N = 500$ from the population. A small circle indicates the population location, while the × indicates the mean of the one hundred estimates. (The plus symbol indicates the origin of the representation space.) It can be seen that the estimates are quite widely scattered, indicating that their reliability is relatively low. The difference between o and × indicates the bias in the estimation. Figure 8(b), on the other hand, displays the RMCA solutions with a near optimal value of the regularization parameter ($\lambda = 20$). Estimates of the point are much less scattered than in Figure 8, although the bias is slightly larger. In total, we get much smaller MSE for the estimates.

***** Insert Figure 8 about here *****

Figure 9 presents essentially the same information as the previous figure for Class A cars, which received a much larger observed frequency. A similar tendency as in the previous figure can also be observed in this figure, but on a much smaller scale. Estimates obtained from the non-regularized MCA are not as scattered as those for Class C cars. (Compare Figures 8(a) and 9(a).) However, RMCA is still an improvement over the non-regularized case. It is associated with a slightly larger bias, but also with a smaller variance and a smaller MSE than the non-regularized case. This indicates that regularization is most effective when we have categories with small observed frequencies, while it is not harmful to regularize (it may still improve the conventional estimation method) even for categories with relatively large frequencies.

***** Insert Figure 9 about here *****

4 Concluding Remarks

Regularization techniques similar to the one developed in this chapter have been investigated in many other statistical methods including regression analysis (Hoerl & Kennard, 1970; see also Groß, 2003), canonical correlation analysis (Vinod, 1976; Ramsay & Silverman, 1999), discriminant analysis (Dipillo, 1976; Friedman, 1989; see also Hastie, Tibshirani, & Friedman, 2001), PCA (Principal Component Analysis), etc. This chapter is the first demonstration of its usefulness in MCA. Its usefulness was demonstrated through numerical experiments involving actual data sets. A similar regularization technique may be incorporated into many other multivariate data analysis techniques, for example, generalized canonical correlation analysis (Takane & Hwang, 2004), redundancy analysis (Takane & Yanai, 2003), hierarchical linear models (HLM), logistic regression and discrimination, generalized linear models, log-linear models, and structural equation models (SEM), etc.

Incorporating prior knowledge is essential in many data analysis situations. Information obtained from the data is never sufficient, and must be supplemented by prior information. In regression analysis, for example, the regression curves or surfaces (the conditional expectation of the criterion variable, $\mathbf{y}$, as a function of predictor variables, $\mathbf{X}$) are estimated for the entire range of $\mathbf{X}$ based on a finite number of observations. In linear regression analysis, this is enabled by the prior knowledge (assumption) that the regression curves and surfaces are linear within a certain range of $\mathbf{X}$. Regularization may be viewed as a way of incorporating prior knowledge in data analysis. In the ridge type of regularization, the prior knowledge takes the form that any parameters in the model (category points in MCA) should not be too far away from 0 (the origin). That is, the effect of λ is to shrink the estimates toward zero in the light of this prior knowledge.

In a broader perspective, the results presented in this chapter cast serious doubts about the adequacy of the conventional estimation methods (such as the maximum likelihood estimation method) that rely on asymptotic rationale. For problems with small to moderate sample sizes, there are better estimation procedures in the sense of achieving smaller MSE. It is not easy to theoretically prove the superiority of the regularization method, and little theoretical work has been done outside regression analysis. However, the kind of numerical experiments used in the present chapter is easy to implement in a variety of contexts other than MCA, and it is expected that similar results will be obtained.

5 Software Notes

MATLAB programs used to obtain the results reported in this paper can be obtained by sending a request to takane@takane2.psych.mcgill.ca

6 Acknowledgments

The work reported in this paper is supported by Grant A6394 to the first author by Natural Sciences and Engineering Research Council of Canada. Thanks are due to Michael Hunter of the University of Victoria and Ringo Ho of McGill University for their valuable comments on an earlier draft of this paper.

Appendix (A): Nishisato's (1994) Small Survey Data

Nishisato's (1994) data consist of the following four items:

a: Your age? (1=20 to 29; 2=30 to 39; 3=over 40)
b: Children these days are not as well disciplined as children when you were a child. (1=Agree; 2=Can't agree; 3=Can't say which)
c: Children today are not as happy as children when you were a child. (1=Agree; 2=Can't agree; 3=Can't say which)
d. Religion should be taught at school. (1=Agree; 2=Disagree; 3=Don't care)

	item			
respondent	a	b	c	d
1	3	1	2	1
2	2	1	3	2
3	2	1	2	2
4	1	2	2	3
5	3	1	2	2
6	1	3	1	2
7	2	1	2	2
8	2	1	1	2
9	1	2	3	1
10	3	1	2	1
11	1	2	2	3
12	2	1	1	1
13	2	1	3	3
14	3	1	2	1
15	1	1	2	3
16	3	1	2	1
17	3	1	1	1
18	2	3	2	2
19	3	1	2	1
20	2	1	2	2
21	1	3	3	3
22	2	1	2	2
23	1	3	3	3

Appendix (B): Variables and categories in Greenacre's (1993) Car Purchase Data

Greenacre's (1993) data consists of three categorical variables: 1. Size class of car, 2. Age (the age of the oldest person in the household), and 3. Income. The size variable consists of 14 categories (size classes), the age variable 7 categories (age groups), and the income variable 9 categories (income levels) as defined in the table below.

Variable	Symbol	Description
Size Class	A	Full-size Standard
	B	Full-size Luxury
	C	Personal Luxury
	D	Intermediate Regular
	E	Intermediate Specialty
	F	Compact Regular
	G	Compact Specialty
	H	Subcompact Regular
	I	Subcompact Specialty
	J	Passenger Utility
	K	Import Economy
	L	Import Standard
	M	Import Sport
	N	Import Luxury
Age	a1	18 - 24 years of age
	a2	25 - 34
	a3	35 - 44
	a4	45 - 54
	a5	55 - 64
	a6	65 - 74
	a7	75 or older
Income	i1	$75,000 or more
	i2	$50,000 - $74,999
	i3	$35,000 - $49,999
	i4	$25,000 - $34,999
	i5	$20,000 - $24,999
	i6	$15,000 - $19,999
	i7	$10,000 - $14,999
	i8	$8,000 - $9,999
	i9	Less than $8,000

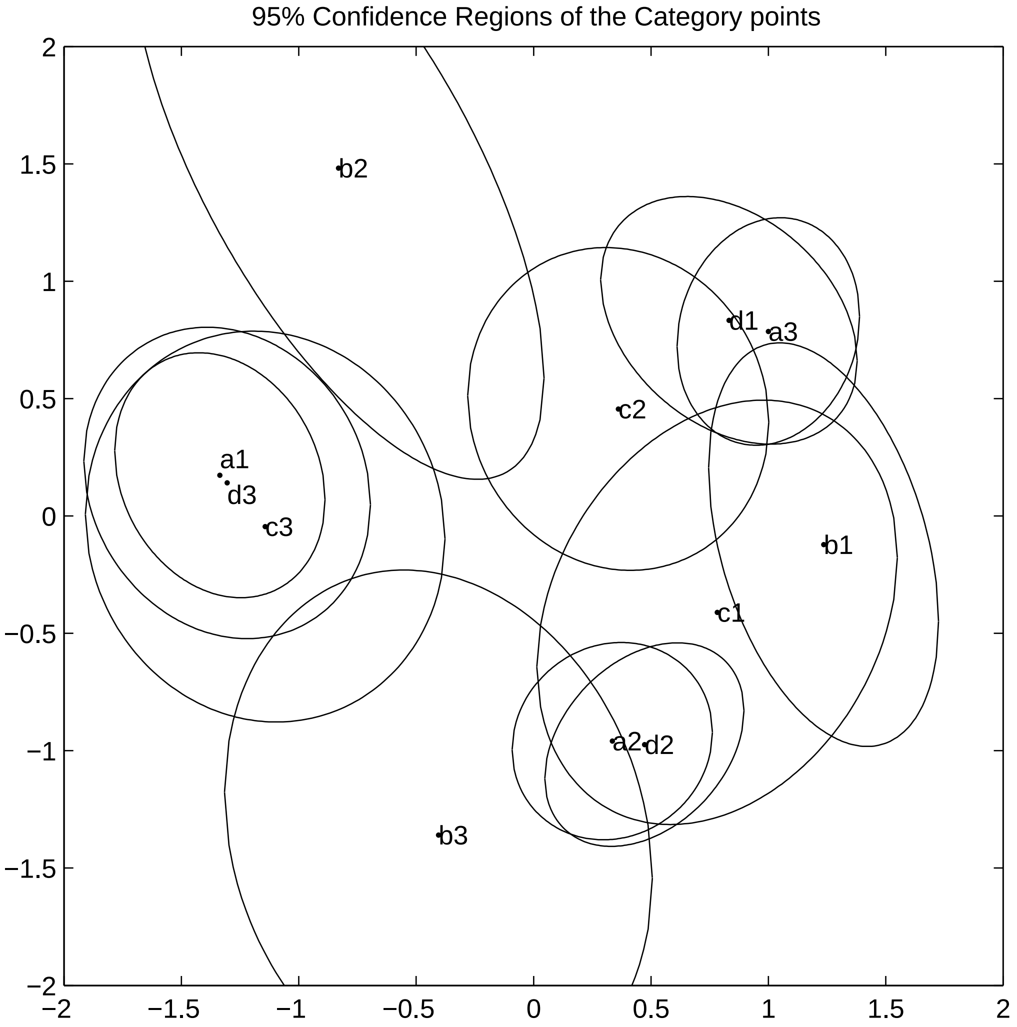

Figure 1: Non-regularized MCA of Nishisato's data: Category points and 95% confidence regions.

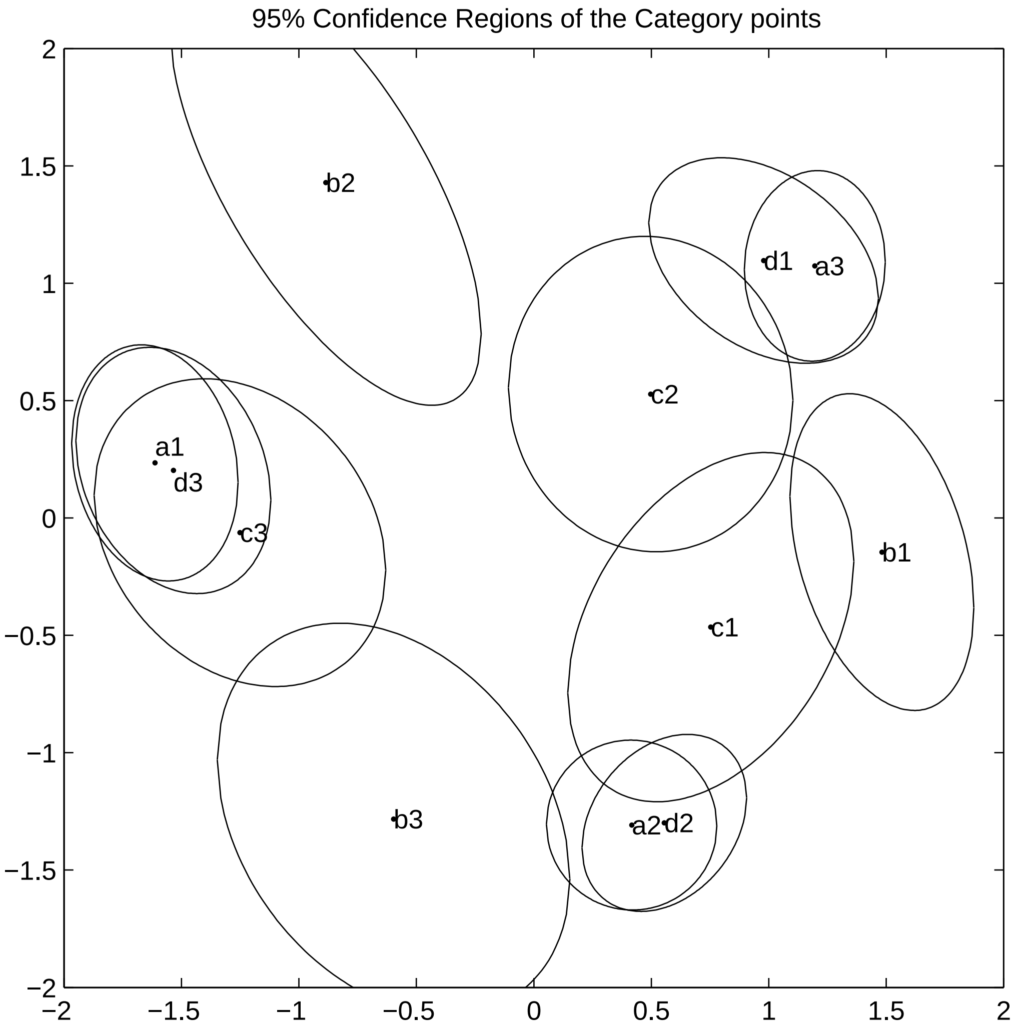

Figure 2: Regularized MCA of Nishisato's data: Category points and 95% confidence regions.

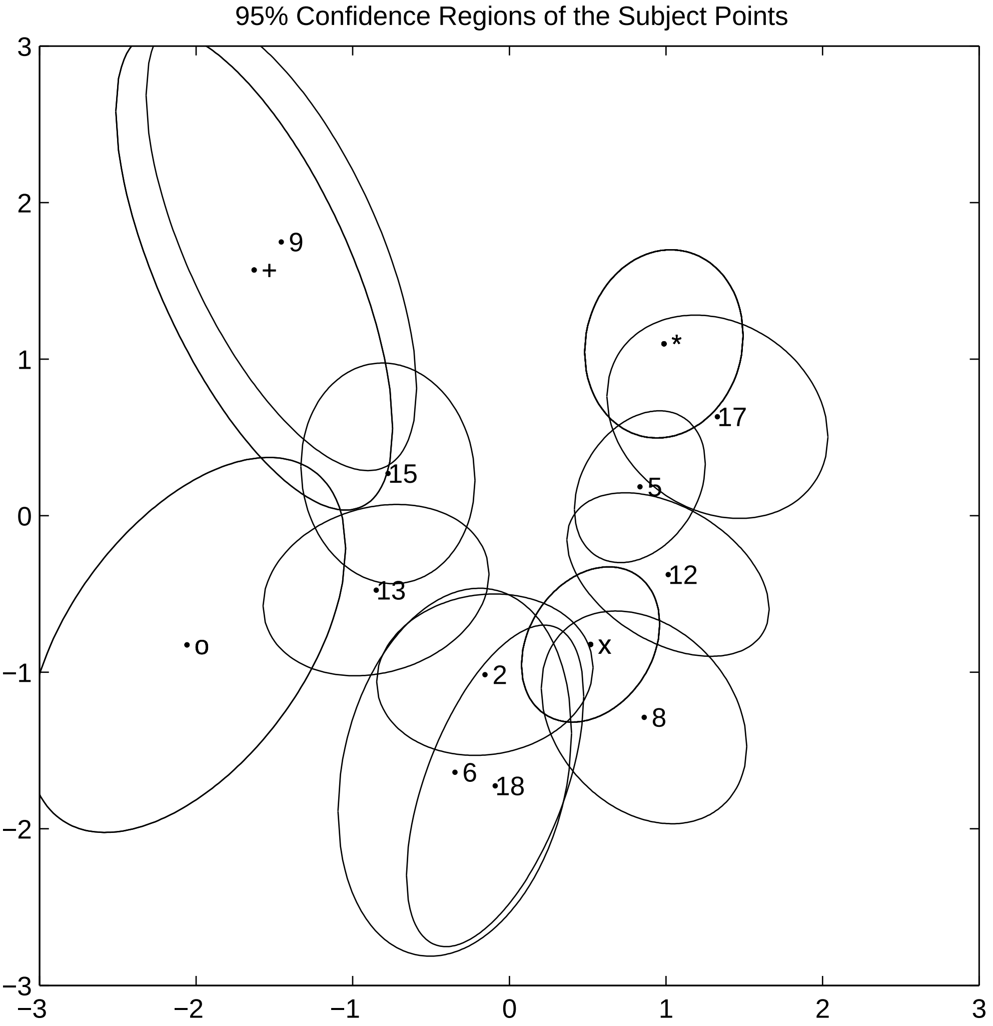

Figure 3: Non-regularized MCA of Nishisato's data: Subject points and 95% confidence regions. Subjects 1, 10, 14, 16 & 19 having an identical response pattern are indicated by "*". Likewise, subjects 3, 7, 20 & 22 are indicated by "x", 4 & 11 by "+", and 21 & 23 by "o".

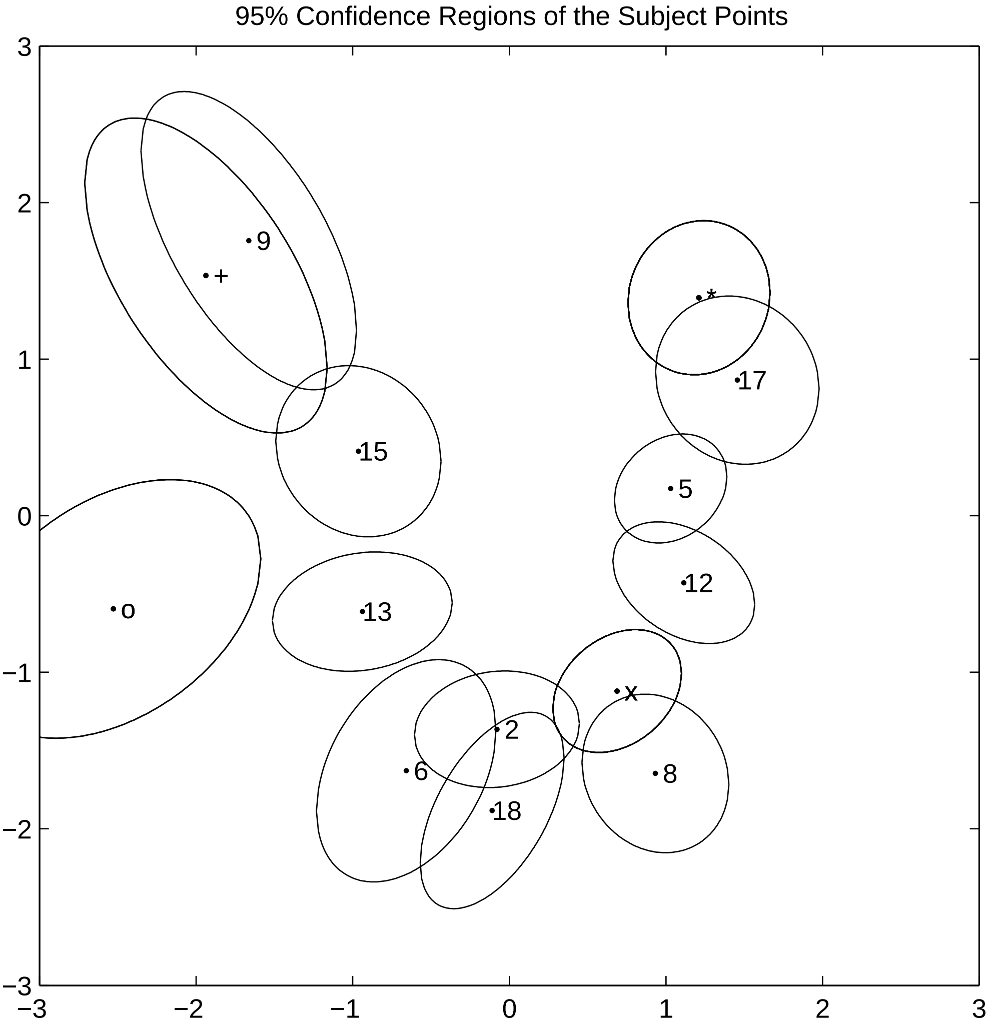

Figure 4: Regularized MCA of Nishisato's data: Subject points and 95% confidence regions. Subjects 1, 10, 14, 16 & 19 having an identical response pattern are indicated by "*". Likewise, subjects 3, 7, 20 & 22 are indicated by "x", 4 & 11 by "+", and 21 & 23 by "o".

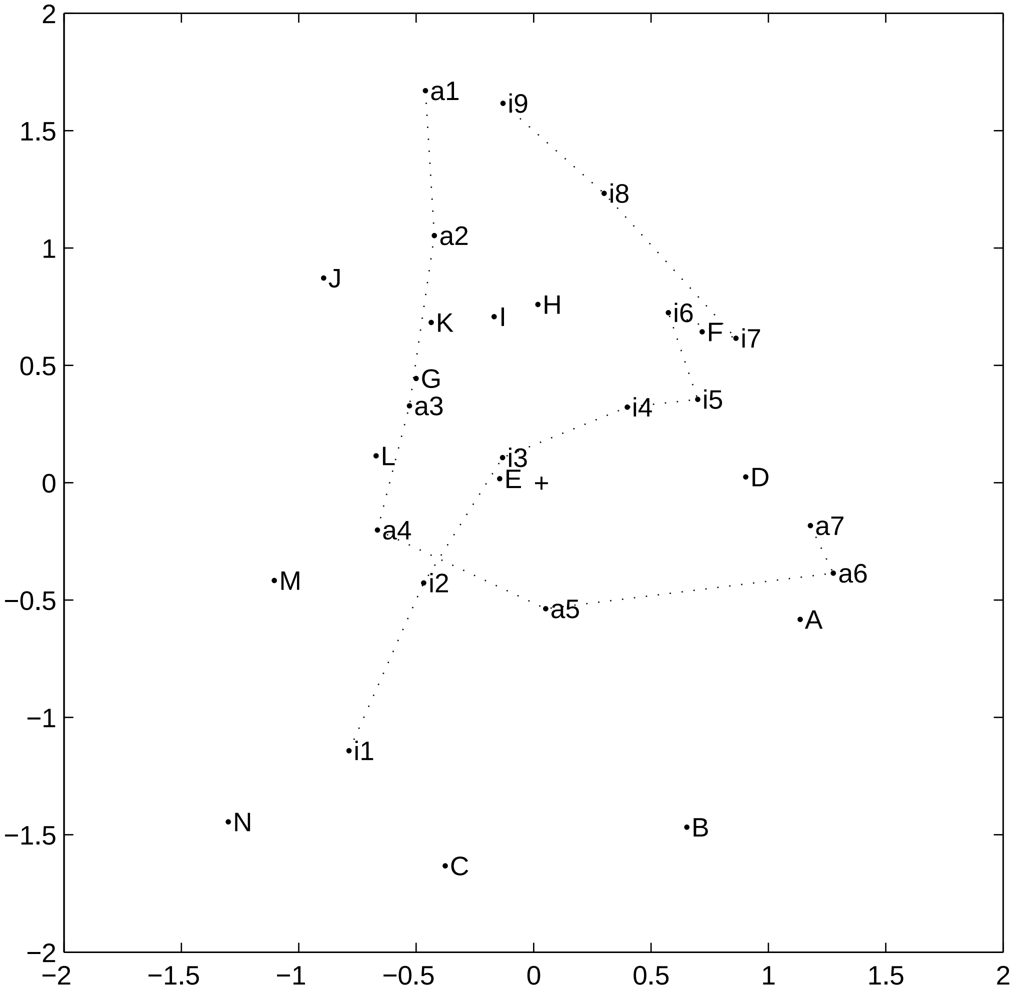

Figure 5: The two-dimensional population configuration for the car data.

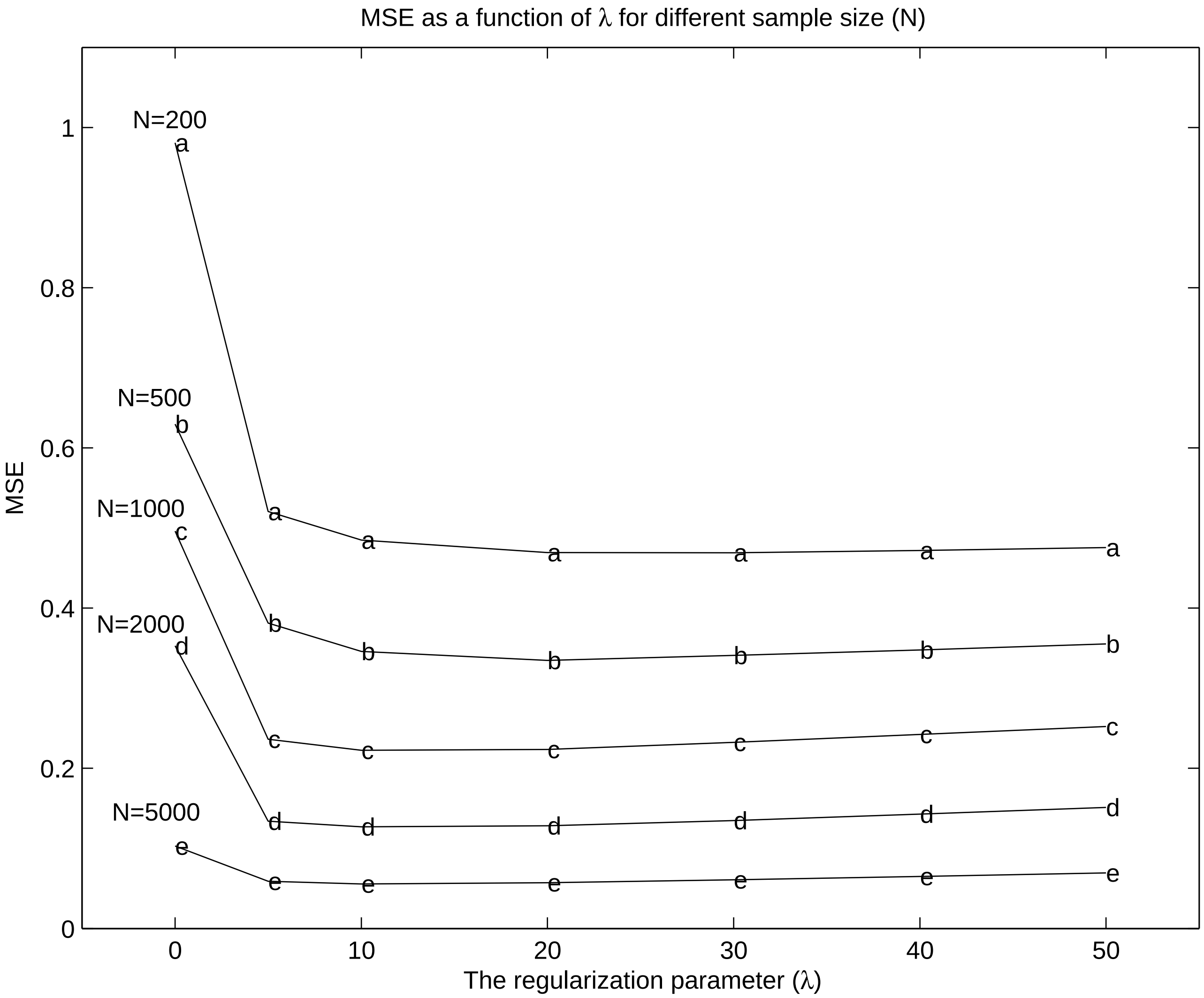

Figure 6: MSE as a function of the regularization parameter (λ) and sample size (N). a: $N = 200$, b: $N = 500$, c: $N = 1000$, d: $N = 2000$, e: $N = 5000$.

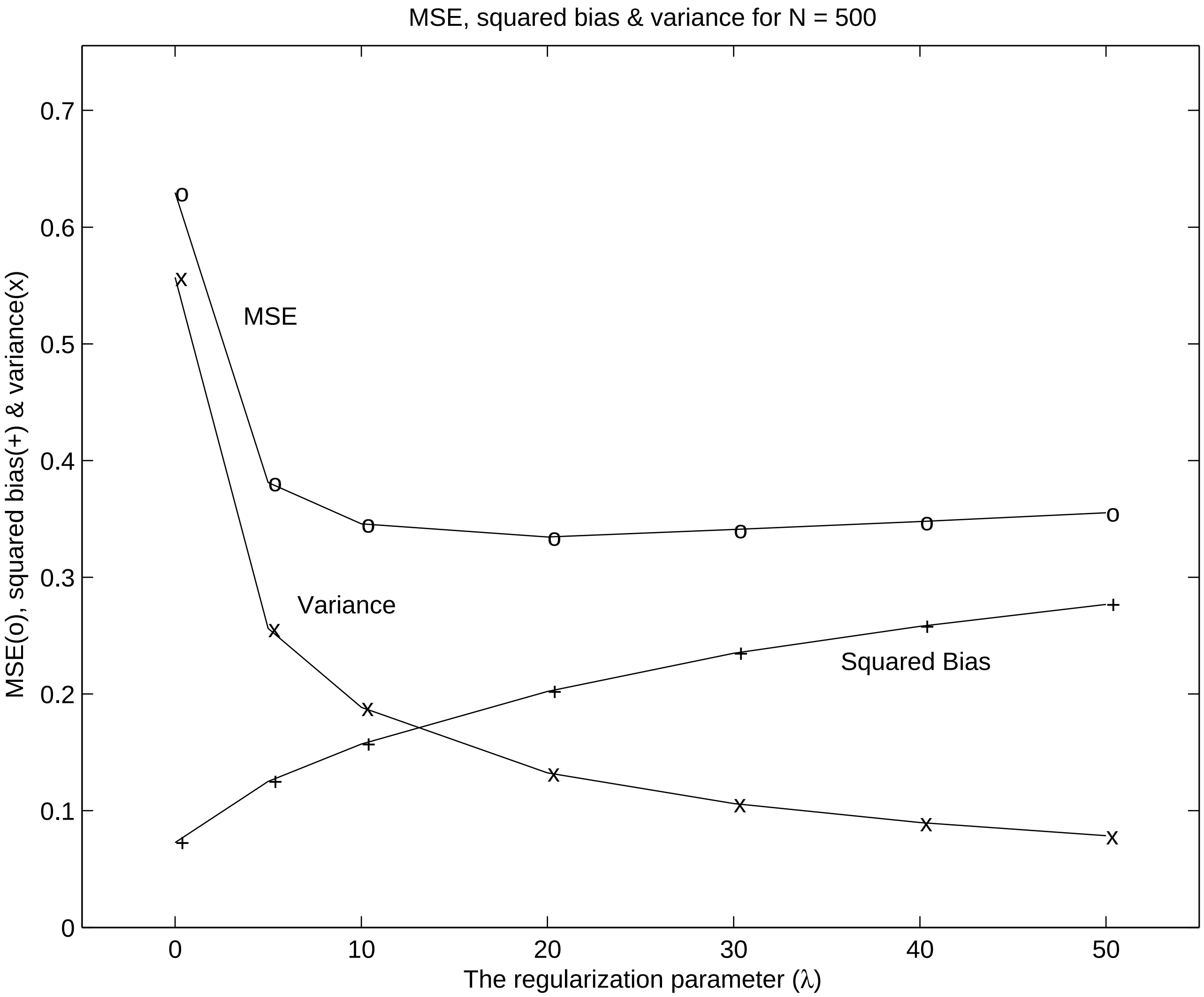

Figure 7: MSE, squared bias and variance as functions of λ for $N = 500$. o: MSE, $+$: squared bias, $\times$: variance.

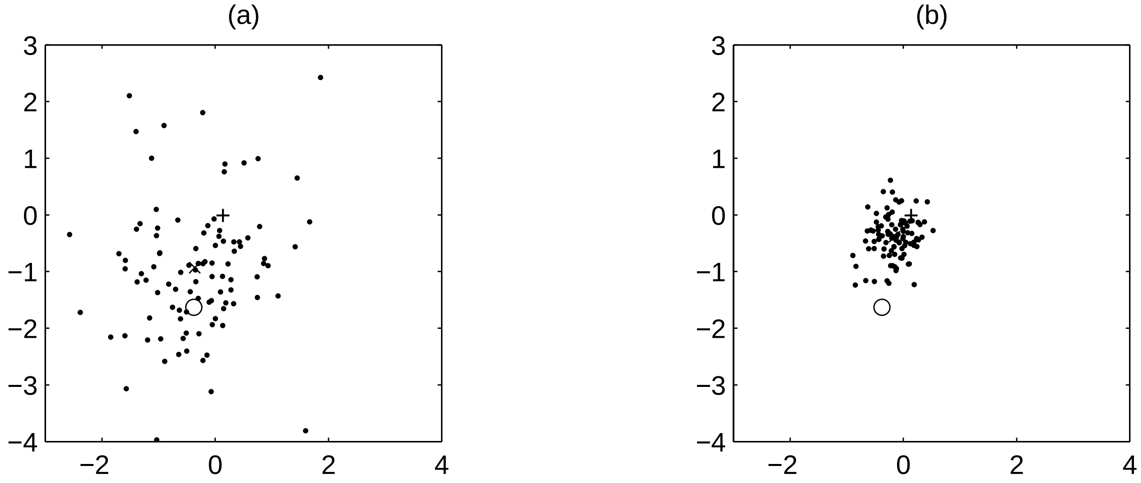

Figure 8: Estimates of point location by (a) non-regularized and (b) regularized MCA for car class C over 100 replicated samples of size 500.

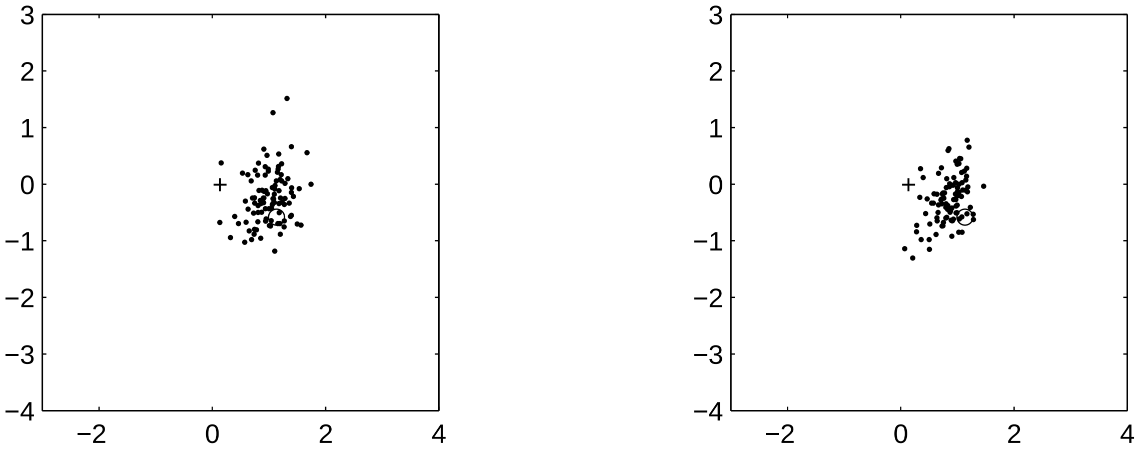

Figure 9: Estimates of point location by (a) non-regularized and (b) regularized MCA for car class A over 100 replicated samples of size 500.

References

Adachi, K. (2002). Homogeneity and smoothness analysis for quantifying a longitudinal categorical variable. In S. Nishisato, Y. Baba, H. Bozdogan & K. Kanefuji (eds), *Measurement and multivariate analysis*, pp. 47-56. Tokyo: Springer.

DiPillo, P.J. (1976). The application of bias to discriminant analysis. *Communications in Statistics - Theory and Methods*, **5**, 843-859.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, **7**, 1-26.

Friedman, J. (1989). Penalized discriminant analysis. *Journal of the American Statistical Association*, **84**, 165-175.

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

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

Greenacre, M.J. (1993). *Correspondence analysis in practice.* London: Academic Press.

Groß, J. (2003). *Linear regression.* Berlin: Springer.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer.

Hoerl, A.F. & Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthgonal problems. *Technometrics*, **12**, 55-67.

Lebart, L., Morineau, A., & Warwick, K.M. (1984). *Multivariate descriptive statistical analysis.* New York: Wiley.

Legendre, P. & Legendre, L. (1998). *Numerical ecology.* Amsterdam: North Holland.

Markus, M.T. (1994). *Bootstrap confidence regions in nonlinear multivariate analysis.* Leiden: DSWO Press.

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

Nishisato, S. (1994). *Elements of dual scaling: An introduction to practical data analysis.* Hillsdale, NJ: Earlbaum Associates.

Ramsay, J.O. & Silverman, B.W. (1997). *Functional data analysis.* New York: Springer.

Takane, Y. (2002). Relationships among various kinds of eigenvalue and singular value decompositions. In Yanai, H., Okada, A., Shigemasu, K., Kano, Y., &

Meulman, J. (Eds.), *New developments in psychometrics*, (pp. 45-56). Tokyo: Springer.

Takane, Y. & Hwang, H. (2002). Generalized constrained canonical correlation analysis. *Multivariate Behavioral Research*, **37**, 163-195.

Takane, Y. & Hwang, H. (2004). Regularized multiple-set canonical correlation analysis. Submitted for publication.

Takane, Y. & Yanai, H. (July, 2003). A simple regularization technique for linear and kernel redundancy analysis. An invited paper presented at IMPS-2003, Sardinia, Italy.

ter Braak, C.J.F. (1990). *Update notes: CANOCO Version 3.10.* Wageningen, The Netherlands: Agricultural Mathematics Group.

Vinod, H.D. (1976). Canonical ridge and econometrics of joint production. *Journal of Econometrics*, **4**, 47-166.

About the author

Yoshio Takane is Professor of Psychology, McGill University, Montréal, Canada. He is a past president of the Psychometric Society. His recent interests are primarily in the development of methods for structured analysis of multivariate data, and artificial neural network simulations.

Address: Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montréal, Québec, H3A 1B1, Canada. E-mail: takane@takane2.psych.mcgill.ca

Heungsun Hwang is Assistant Professor of Marketing, HEC Montréal, Montréal, Canada. His recent interests include generalizations of growth curve models, and correspondence analysis to capture subject heterogeneity.

Address: Department of Marketing, HEC Montréal, 3000 Chemin de la Côte Ste Catherine, Montréal, Québec, H3T 2A7, Canada. E-mail: heungsun.hwang@hec.ca.