Umezu *et al. Pacific Journal of Mathematics for Industry* (2016) 8:5
DOI 10.1186/s40736-016-0024-x

Pacific Journal of Mathematics for Industry

ORIGINAL ARTICLE Open Access

Ridge-type regularization method for questionnaire data analysis

Yuta Umezu[1*], Hidetoshi Matsuoka[2], Hiroshi Ikeda[2] and Yoshiyuki Ninomiya[3]

Abstract

In questionnaire studies for evaluating objects such as manufacturing products, evaluators are required to respond to several evaluation items for the objects. When the number of objects is large, a part of the objects is often assigned randomly to each evaluator, and the response becomes a matrix with missing components. To handle this kind of data, we consider a model by using a dummy matrix representing the existence of the missing components, which can be interpreted as an extension of the GMANOVA model. In addition, to cope with the case where the numbers of the object and evaluation items are large, we consider a ridge-type estimator peculiar to our model to avoid instability in estimation. Moreover, we derive a C_p criterion in order to select the tuning parameters included in our estimator. Finally, we check the validity of the proposed method through simulation studies and real data analysis.

Keywords: Array data, C_p criterion, GMANOVA, Missing value, Ridge estimator

1 Introduction

In questionnaire studies, N evaluators are often required to respond K evaluation items for M objects selected randomly from J objects. For instance, they evaluate on a scale 1 to 5 for each of the evaluation items. Because the evaluator responds only for M objects, evaluations for the rest of $J - M$ objects are missing, that is, we can not observe them. Therefore, we have a three-dimensional array data of size $J \times K \times N$ consisting of the MKN observations and $(J - M)KN$ missing values.

For such data, we are often interested in predicting the missing values based on the observations (for example, recommendation systems of Amazon or Netflix). Nowadays methods such as collaborative filtering or matrix completion are developed to predict the missing part. To predict it, it is indispensable to assume some conditions for the data structure in general. For example, Candès & Recht [2], and Koltchinskii et al. [5] reconstruct the matrix by assuming that the data structure is low-rank, and this is useful since it enables us to use a popular method of convex optimization. However, it is difficult to select a tuning parameter which is included in the method because we have no reasonable information criterion. In addition, it is difficult to evaluate the prediction accuracy because we have no evaluation formula for the variance of the predicted value.

On the other hand, correspondence analysis has been used in questionnaire data analysis in order to extract features from the data (e.g., Benzécri [1]). Correspondence analysis, however, is an exploratory method just like principal component analysis and is not applicable to the data including missing values. So we can not use it to analyze our data. As described in Section 2, it is possible to construct a parametric model by using a dummy matrix representing the existence of the missing values. The model we will consider can be interpreted as an extension of the generalized multivariate analysis of variance (GMANOVA) model in Potthoff & Roy [9] to for three-dimensional array data. Usually, a noise in the GMANOVA model is assumed to be distributed some Gaussian distribution. Unfortunately, since the data obtained from questionnaire study are discrete in general, it is unnatural to assume the normality for noise. Even so, we can express the ordinal least squares estimator explicitly and moreover evaluate the average or variance of the estimator. However, we encounter a problem that the estimator becomes unstable when M or K is large or when the multicollinearity is present in the data.

The ridge-type estimator is often used in order to assure the stability of the estimator (e.g., Hoerl & Kennard [4]).

*Correspondence: umezu.yuta@nitech.ac.jp
[1] Nagoya Institute of Technology, Gokiso-cho, Showa-ku, 466-8555 Nagoya, Japan
Full list of author information is available at the end of the article

We then need to choose appropriate tuning parameters included in the estimator. Computational methods such as cross validation (CV; Stone [10]) are usually used for this choice although they come at a considerable computation cost. Information criteria such as C_p (Mallows [6, 7]) may also be used to choose it. For example, Nagai [8] derived an unbiased estimator of the standardized mean squared error for the ridge-type estimator in the GMANOVA model. However, the objective variable in our data is an $(M \times K)$-dimensional matrix, and so we can not apply his result since it is only for the usually GMANOVA model, that is, they assumed the normality for noise, and he does not considered the missing values.

Although it is sometimes important to predict the missing part in questionnaire studies, our goal in this paper is to construct an appropriate model. To do this, we derive an unbiased estimator of the standardized mean squared error for the model that is defined in Section 2. Moreover, in Section 3, making good use of matrix calculations, we develop a C_p-type information criterion in order to select tuning parameters included in the estimator. The proposed method is shown to be valid through a simulation study in Section 4, and then the result in which the method is applied to real data is reported in Section 5. Some concluding remarks are presented in Section 6. Several matrix algebras used in this paper and some proofs are relegated to Appendix.

2 Setting and assumptions

In the following sections, we will denote $\mathbf{0}_d$, $\mathbf{1}_d$ and I_d by a d-dimensional zero-vector, one-vector and $(d \times d)$-dimensional identity matrix for a positive integer d.

Let $\mathcal{J} = \{1, 2, , \ldots, J\}$ be an index set of objects and $\mathcal{J}_i = \{j_{i1}, j_{i2}, \ldots, j_{iM}\}$ be a subset of $\mathcal{J}$ arranged in ascending order. In addition, let $y_{ij_{im}k}$ be a response of the k-th evaluation item of the j_{im}-th object for the i-th evaluator, and we denote the data for the i-th evaluator by an $(M \times K)$-dimensional matrix $Y_i = (y_{ij_{im}k})_{m=1,2,\ldots,M;k=1,2,\ldots,K}$. For these data, we consider the model

$$y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk},$$

where μ is a general mean, α_j and β_k are main effects, γ_{jk} is an interaction effect between the j-th object and the k-th item, and ε_{ijk} is noise. Note that we can not fully observe the response y_{ijk}'s , more specifically speaking, y_{ijk} is missing whenever $j \notin \mathcal{J}_i$. Let $\tilde{X}_i$ be an $(M \times J)$-dimensional matrix whose (m, j)-th element is 1 when $j = j_{im}$ and 0 otherwise. Then, we can rewrite this model as

$$Y_i = \bar{X}_i \bar{B} \bar{A} + E_i, \tag{1}$$

where $\bar{X}_i = (\mathbf{1}_M, \tilde{X}_i) \in \mathbb{R}^{M \times (J+1)}$, $\bar{A} = (\mathbf{1}_K, I_K)' \in \mathbb{R}^{(K+1) \times K}$, and

$$\bar{B} = \begin{pmatrix} \mu & \beta_1 & \cdots & \beta_K \\ \alpha_1 & \gamma_{11} & \cdots & \gamma_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_J & \gamma_{J1} & \cdots & \gamma_{JK} \end{pmatrix} \in \mathbb{R}^{(J+1) \times (K+1)}.$$

Let us suppose that $E_i = (\boldsymbol{\varepsilon}_{i1}, \boldsymbol{\varepsilon}_{i2}, \ldots, \boldsymbol{\varepsilon}_{iK}) = (\varepsilon_{ij_{im}k})_{m=1,2,\ldots,M;k=1,2,\ldots,K}$ are independent random matrices with mean $\mathrm{E}[E_i] = \mathbf{0}_M \mathbf{0}_K'$ and covariance $\mathrm{V}[\mathrm{vec}(E_i)] = \Sigma \otimes \Xi$, where Σ and Ξ are an unknown $(K \times K)$-dimensional matrix and a known $(M \times M)$-dimensional matrix, respectively. This means that the k-th column $\boldsymbol{\varepsilon}_{ik}$ and the ℓ-th column $\boldsymbol{\varepsilon}_{i\ell}$ of E_i have a covariance matrix $\mathrm{E}[\boldsymbol{\varepsilon}_{ik}\boldsymbol{\varepsilon}_{i\ell}'] = \sigma_{k\ell}\Xi$ for $k, \ell = 1, 2, \ldots, K$. Although j_m's are assigned randomly, we consider X_i deterministic for simplicity. Note that this model includes the so-called GMANOVA model of Potthoff & Roy [9] in a special case when $M = 1$ and E_i is distributed according to some Gaussian distribution.

To avoid redundancy of the model, we impose

$$\sum_{j=1}^{J} \alpha_j = \sum_{k=1}^{K} \beta_k = \sum_{j=1}^{J} \gamma_{jk} = \sum_{k=1}^{K} \gamma_{jk} = 0$$

on the parameter as is often used in the ANOVA model. Since

$$\alpha_J = -\sum_{j=1}^{J-1} \alpha_j, \quad \beta_K = -\sum_{k=1}^{K-1} \beta_k,$$
$$\gamma_{Jk} = -\sum_{j=1}^{J-1} \gamma_{jk}, \quad \text{and} \quad \gamma_{jK} = -\sum_{k=1}^{K-1} \gamma_{jk},$$

we can remove this restriction. In fact, by defining $C = (I_{J-1}, -\mathbf{1}_{J-1})'$, $D = (I_{K-1}, -\mathbf{1}_{K-1})'$, $\bar{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{J-1})'$, $\bar{\boldsymbol{\beta}} = (\beta_1, \beta_2, \ldots, \beta_{K-1})'$ and $\bar{\Gamma} = (\gamma_{jk})_{j=1,2,\ldots,J-1;k=1,2,\ldots,K-1}$, $\bar{B}$ can be rewritten as

$$\bar{B} = \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & C \end{pmatrix} \begin{pmatrix} \mu & \bar{\boldsymbol{\beta}}' \\ \bar{\alpha} & \tilde{\Gamma} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & D' \end{pmatrix},$$

and thus we can define

$$\bar{X}_i \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & C \end{pmatrix} \in \mathbb{R}^{M \times J}, \quad \begin{pmatrix} \mu & \bar{\boldsymbol{\beta}}' \\ \bar{\boldsymbol{\alpha}} & \tilde{\Gamma} \end{pmatrix} \in \mathbb{R}^{J \times K},$$

and

$$\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & D' \end{pmatrix} \bar{A} = \begin{pmatrix} 1'_{K-1} & 1 \\ I_{K-1} & -\mathbf{1}_{K-1} \end{pmatrix} \in \mathbb{R}^{K \times K}$$

by X_i, B and A, respectively.

In the following, let us denote $\sum_{i=1}^{N} X_i' X_i$ and $\sum_{i=1}^{N} X_i' Y_i$ by $X'X$ and $X'Y$, respectively. Then an ordinary least

square estimator of the model (1), that is, a minimizer of $\sum_{i=1}^{N} \|Y_i - X_i BA\|_F^2$, is given by

$$\tilde{B} = (X'X)^{-1}X'YA'(AA')^{-1},$$

where $\|\cdot\|_F$ denotes a Frobenius norm, i.e, $\|T\|_F = (\mathrm{tr}(T'T))^{1/2}$ for a matrix T. It is easy to see that $\tilde{B}$ is an unbiased estimator of B, and that, if $X'X/N$ converges to some positive definite matrix, $\tilde{B}$ is a consistent estimator of B from the Chebyshev's inequality. In addition, an unbiased estimator of Σ is given by

$$\hat{\Sigma} = \frac{1}{n\mathrm{tr}(\Xi) - S}\sum_{i=1}^{N}(Y_i - X_i\tilde{B}A)'(Y_i - X_i\tilde{B}A), \tag{2}$$

where $S = \sum_{i=1}^{N} \mathrm{tr}\{X_i(X'X)^{-1}X_i\Xi\}$. The details for deriving the unbiasedness of (2) are given in Appendix 2.

However, when J or K are large, the inverse of $X'X$ or AA' may not exist or the variance of the estimator may become unstable, and so we consider the ridge-type estimator given by

$$\hat{B}_{\lambda,\mu} = (X'X + \lambda I_J)^{-1}X'YA'(AA' + \mu I_K)^{-1}, \tag{3}$$

where λ and μ are positive constants, which are also known as tuning parameters (see, e.g., Hoerl & Kennard [4], Nagai [8]). Then we can obtain the predictor

$$\hat{Y}_i = X_i\hat{B}_{\lambda,\mu}A. \tag{4}$$

3 Deriving the C_p criterion

3.1 Preparation

Nagai [8] derived a C_p criterion for a ridge-type estimator in the GMANOVA model. His result and ours are different because there are missing values in the data and the observation is an $(M \times K)$-dimensional matrix in our case. Moreover, we do not assume the normality of E_i.

To derive a C_p criterion, we need some preparation with matrix calculation. Let us define

$$H_\mu = A'(AA' + \mu I_K)^{-1}A$$

and

$$G_\lambda = (X'X + \lambda I_J)^{-1}.$$

Note that by definition of A,

$$AA' = \begin{pmatrix} K & \mathbf{0}' \\ \mathbf{0} & I_{K-1} + \mathbf{1}_{K-1}\mathbf{1}'_{K-1} \end{pmatrix}.$$

Because the inverse matrix of $(1+\mu)I_{K-1} + \mathbf{1}_{K-1}\mathbf{1}'_{K-1}$ is given by

$$\frac{1}{1+\mu}\left(I_{K-1} - \frac{1}{K+\mu}\mathbf{1}_{K-1}\mathbf{1}'_{K-1}\right)$$

from (20), it follows that

$$H_\mu = \frac{1}{(1+\mu)(K+\mu)}\begin{pmatrix} \tilde{H}_\mu & \mathbf{0} \\ \mathbf{0}' & K(1+\mu) \end{pmatrix}, \tag{5}$$

where $\tilde{H}_\mu = (K+\mu)I_{K-1} + \mu\mathbf{1}_{K-1}\mathbf{1}'_{K-1}$.

Next, we see that

$$X_i = (\mathbf{1}_M, \tilde{X}_i)\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & C \end{pmatrix} = (\mathbf{1}_M, \tilde{X}_i C)$$

and then $X'X$ can be expressed as

$$\sum_{i=1}^{N}\begin{pmatrix} M & \mathbf{1}'_M\tilde{X}_i C \\ C'\tilde{X}'_i\mathbf{1}_M & C'\tilde{X}'_i\tilde{X}_i C \end{pmatrix}.$$

Let us define

$$\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_J)' = \sum_{i=1}^{N}\tilde{X}'_i\mathbf{1}_M. \tag{6}$$

Note that δ_j represents the number of times such that the j-th object is assigned and $\Delta = \sum_{i=1}^{N}\tilde{X}'_i\tilde{X}_i$ is a diagonal matrix whose (j,j)-th element is δ_j. From (18) in Appendix 1, G_λ can be expressed as

$$G_\lambda = \begin{pmatrix} \frac{(NM+\lambda)+\boldsymbol{\delta}'C\tilde{G}_\lambda^{-1}C'\boldsymbol{\delta}}{(NM+\lambda)^2} & \frac{\boldsymbol{\delta}'C\tilde{G}_\lambda^{-1}}{NM+\lambda} \\ \frac{\tilde{G}_\lambda^{-1}C'\boldsymbol{\delta}}{NM+\lambda} & \tilde{G}_\lambda^{-1} \end{pmatrix}, \tag{7}$$

where

$$\tilde{G}_\lambda = C'\left(\Delta - \frac{1}{NM+\lambda}\boldsymbol{\delta}\boldsymbol{\delta}'\right)C + \lambda I_{J-1}.$$

Let us define $\tilde{\Delta} = \Delta - (NM+\lambda)^{-1}\boldsymbol{\delta}\boldsymbol{\delta}'$. Then, from (19), we see that

$$\tilde{\Delta}^{-1} = \Delta^{-1} + \frac{1}{\lambda}\mathbf{1}_J\mathbf{1}'_J, \tag{8}$$

since $\Delta^{-1}\delta = \mathbf{1}_J$ and $\boldsymbol{\delta}'\mathbf{1}_J = NM$. Moreover, by using (19) again, we have

$$\tilde{G}_\lambda^{-1} = \frac{1}{\lambda}I_{J-1} - \frac{1}{\lambda^2}C'\left(\tilde{\Delta}^{-1} + \frac{1}{\lambda}CC'\right)^{-1}C.$$

Let $\Delta_{-J} \in \mathbb{R}^{(J-1)\times(J-1)}$ be the sub-matrix of Δ made by removing J-th column and row of Δ, and $\Delta^\dagger = \Delta_{-J}^{-1} + \lambda^{-1}I_{J-1}$. Note that the (j,j)-th element of $\Delta^\dagger$ is given by $\delta_j^{-1} + \lambda^{-1}$ for $j = 1, 2, \ldots, J-1$. Then $\tilde{\Delta}^{-1} + \lambda^{-1}CC'$ can be expressed as

$$\begin{pmatrix} \Delta^\dagger + \lambda^{-1}\mathbf{1}_{J-1}\mathbf{1}'_{J-1} & \mathbf{0} \\ \mathbf{0}' & \delta_J^{-1} + \lambda^{-1}J \end{pmatrix}$$

from (8). Let us define

$$P = \Delta_{-J}(\Delta_{-J} + \lambda I_{J-1})^{-1}. \tag{9}$$

Note that $\Delta^{\dagger-1} = \lambda P$. Then the inverse matrix of $\Delta^\dagger + \lambda^{-1}\mathbf{1}_{J-1}\mathbf{1}'_{J-1}$ can be expressed as

$$\begin{aligned} &\Delta^{\dagger-1} - \frac{\Delta^{\dagger-1}\mathbf{1}_{J-1}\mathbf{1}'_{J-1}\Delta^{\dagger-1}}{\lambda + \mathbf{1}'_{J-1}\Delta^{\dagger-1}\mathbf{1}_{J-1}} \\ =&\lambda\left(P - \frac{P1_{J-1}1'_{J-1}P}{1+\mathrm{tr}(P)}\right). \end{aligned}$$

In this equality, we just use $\mathbf{1}'_{J-1}\Delta^{\dagger-1}\mathbf{1}_{J-1} = \lambda\mathrm{tr}(P)$. Finally, we obtain that

$$\tilde{G}_\lambda^{-1} = \frac{I_{J-1}-P}{\lambda} + \frac{P\mathbf{1}_{J-1}\mathbf{1}'_{J-1}P}{\lambda(1+\mathrm{tr}(P))} - \frac{\delta_J\mathbf{1}_{J-1}\mathbf{1}'_{J-1}}{\lambda(\lambda+J\delta_J)}. \quad (10)$$

3.2 Main result

Now, we can derive the C_p criterion as an unbiased estimator of a standardized mean squared error (MSE) defined by

$$\sum_{i=1}^{N}\mathrm{E}\Big[\mathrm{vec}(\hat{Y}_i - \mathrm{E}[\,Y_i])'(\Sigma\otimes\Xi)^{-1}\mathrm{vec}(\hat{Y}_i - \mathrm{E}[\,Y_i])\Big],$$

where $\hat{Y}_i$ is the predictor defined in (4). From (21), this is equivalent to

$$\sum_{i=1}^{N}\mathrm{E}\Big[\mathrm{tr}\Big\{(\hat{Y}_i - \mathrm{E}[\,Y_i])'\Xi^{-1}(\hat{Y}_i - \mathrm{E}[\,Y_i])\Sigma^{-1}\Big\}\Big].$$

Because $\mathrm{E}[\,Y_i] = Y_i - E_i$ and

$$\begin{aligned}&\mathrm{E}[\,\mathrm{tr}\{\Sigma^{-1}(Y_i-\hat{Y}_i)'E_i\}]\\ =&\mathrm{E}[\,\mathrm{tr}\{\Sigma^{-1}E_i'E_i\}] - \mathrm{E}[\,\mathrm{tr}\{\Sigma^{-1}\hat{Y}_i'E_i\}],\end{aligned}$$

the MSE can be rewritten as

$$\begin{aligned}&\sum_{i=1}^{N}\mathrm{E}[\,\{\mathrm{tr}\{(Y_i-\hat{Y}_i)'\Xi^{-1}(Y_i-\hat{Y}_i)\Sigma^{-1}\}]\\ &-\sum_{i=1}^{N}\mathrm{E}[\,\mathrm{tr}(E_i'\Xi^{-1}E_i\Sigma^{-1})]\\ &+2\sum_{i=1}^{N}\mathrm{E}[\,\mathrm{tr}(\hat{Y}_i'\Xi^{-1}E_i\Sigma^{-1}\}]. \quad (11)\end{aligned}$$

By using (21) and $\mathrm{V}[\,\mathrm{vec}(E_i)] = \Sigma\otimes\Xi$, the second term of the right-hand side in (11) can be reduced to NMK since

$$\begin{aligned}&\mathrm{E}[\,\mathrm{tr}(E_i'\Xi^{-1}E_i\Sigma^{-1})]\\ =&\mathrm{E}[\,\mathrm{vec}(E_i)'(\Sigma^{-1}\otimes\Xi^{-1})\mathrm{vec}(E_i)]\\ =&\mathrm{tr}(I_M\otimes I_K)\\ =&MK.\end{aligned}$$

Next, we evaluate the third term of the right-hand side in (11). From (3) and the definition of the model in (1), we have

$$\begin{aligned}\hat{Y}_i &= \sum_{h=1}^{N}X_iG_\lambda X_h'Y_hH_\mu\\ &= X_iG_\lambda X'XBAH_\mu + \sum_{h=1}^{N}X_iG_\lambda X_h'E_hH_\mu.\end{aligned}$$

Because the first term of the right-hand side in this equality is non-stochastic and E_i's are independent, we see from (22) that

$$\begin{aligned}&\mathrm{E}[\,\mathrm{tr}(\hat{Y}_i'\Xi^{-1}E_i\Sigma^{-1})]\\ =&\sum_{h=1}^{N}\mathrm{E}[\,\mathrm{vec}(E_h)'(H_\mu\Sigma^{-1}\otimes X_hG_\lambda X_i'\Xi^{-1})\mathrm{vec}(E_i)]\\ =&\mathrm{tr}\{(H_\mu\Sigma^{-1}\otimes X_iG_\lambda X_i'\Xi^{-1})(\Sigma\otimes\Xi)\}\\ =&\mathrm{tr}(H_\mu)\mathrm{tr}(G_\lambda X_i'X_i).\end{aligned}$$

Thus the third term of the right-hand side in (11) is reduced to $2\mathrm{tr}(H_\mu)\mathrm{tr}(G_\lambda X'X)$. From (5), we have

$$\begin{aligned}\mathrm{tr}(H_\mu) &= \frac{1}{(1+\mu)(K+\mu)}\{\mathrm{tr}(\tilde{H}_\mu) + K(1+\mu)\}\\ &= \frac{K^2+3K\mu-2\mu}{(1+\mu)(K+\mu)}. \quad (12)\end{aligned}$$

On the other hand, by the definition of G_λ in (7), we have $G_\lambda X'X = I_J - \lambda G_\lambda$ and

$$\mathrm{tr}(G_\lambda) = \frac{1}{NM+\lambda} + \frac{\boldsymbol{\delta}'C\tilde{G}_\lambda^{-1}C'\boldsymbol{\delta}}{(NM+\lambda)^2} + \mathrm{tr}(\tilde{G}_\lambda^{-1})$$

from (7). Since $\mathrm{tr}(P\mathbf{1}_{J-1}\mathbf{1}'_{J-1}P) = \mathrm{tr}(P^2)$, the last term $\mathrm{tr}(\tilde{G}_\lambda^{-1})$ is reduced to

$$\frac{J-1-\mathrm{tr}(P)}{\lambda} + \frac{\mathrm{tr}(P^2)}{\lambda(1+\mathrm{tr}(P))} - \frac{\delta_J(J-1)}{\lambda(\lambda+J\delta_J)}$$

from (10). Moreover, by a simple calculation, we have

$$\begin{aligned}\boldsymbol{\delta}'C\mathbf{1}_{J-1} &= (NM+\lambda) - (\lambda+J\delta_J),\\ \boldsymbol{\delta}'CP\mathbf{1}_{J-1} &= (NM+\lambda) - (\lambda+\delta_J)(1+\mathrm{tr}(P)),\end{aligned}$$

and

$$\begin{aligned}&\boldsymbol{\delta}'C(I_{J-1}-P)C'\boldsymbol{\delta}\\ =&\delta_J(J-1)(\lambda+\delta_J)\\ &-\lambda(\lambda+\delta_J)^2(NM-J\delta_J)\mathrm{tr}(P).\end{aligned}$$

Then, it follows that

$$\begin{aligned}&\lambda\boldsymbol{\delta}'C\tilde{G}_\lambda^{-1}C'\boldsymbol{\delta}\\ =&(NM+\lambda)^2\left(\frac{1}{1+\mathrm{tr}(P)} - \frac{\delta_J}{\lambda+J\delta_J}\right)\\ &-\lambda(NM+\lambda),\end{aligned}$$

and thus we have

$$\mathrm{tr}(G_\lambda X'X) = f(P) + \frac{J\delta_J}{\lambda+J\delta_J}, \quad (13)$$

where

$$f(P) = \mathrm{tr}(P) + \frac{\mathrm{tr}(P)-\mathrm{tr}(P^2)}{1+\mathrm{tr}(P)}. \quad (14)$$

Combining all the above, we obtain the following theorem:

Theorem 1 *An unbiased estimator of MSE in* (11) *is given by*

$$\sum_{i=1}^{N} \mathrm{tr}\{(Y_i - \hat{Y}_i)' \Xi^{-1} (Y_i - \hat{Y}_i) \Sigma^{-1}\} - NMK + 2\left(f(P) + \frac{J\delta_J}{\lambda + J\delta_J}\right) \frac{K^2 + 3K\mu - 2\mu}{(1+\mu)(K+\mu)},$$

where δ_j, $\hat{Y}_i$, P *and* $f(P)$ *are defined in* (6), (4), (9) *and* (14), *respectively.*

Our result coincides with Nagai [8] in the special case when $K = 1$. In this case, we can interpret our model (1) as usual multivariate linear regression model except for the missing of the data.

As a result, we propose the following index as a C_p-type information criterion:

$$C_p = \sum_{i=1}^{N} \mathrm{tr}\{(Y_i - \hat{Y}_i)' \Xi^{-1} (Y_i - \hat{Y}_i) \hat{\Sigma}^{-1}\} - NMK + 2\left(f(P) + \frac{J\delta_J}{\lambda + J\delta_J}\right) \frac{K^2 + 3K\mu - 2\mu}{(1+\mu)(K+\mu)}, \quad (15)$$

where $\hat{\Sigma}$ is an unbiased estimator of Σ defined in (2). By minimizing the C_p in (15), we can obtain the optimal values of the tuning parameters (λ, μ).

4 Simulation study

In this section, we conduct some simulation studies to check the performance of the tuning parameter selection based on C_p in (15). The performances for C_p and CV are compared.

Concretely speaking, we assessed the performances in terms of the prediction squared error (PSE), that is,

$$\tilde{\mathrm{E}}\left[(\tilde{Y}_i - X_i \hat{B}_{\hat{\lambda},\hat{\mu}} A)' \Xi^{-1} (\tilde{Y}_i - X_i \hat{B}_{\hat{\lambda},\hat{\mu}} A) \hat{\Sigma}^{-1}\right], \quad (16)$$

where $\tilde{Y}_i$ is the copy of Y_i, $\hat{\lambda}$ and $\hat{\mu}$ are the values of the tuning parameters which minimize each of the criteria, and $\hat{\Sigma}$ is an unbiased estimator of Σ given by (2). In addition, $\tilde{\mathrm{E}}$ denotes the expectation with respect to only $\tilde{Y}_i$. The expectation in PSE is evaluated using an empirical mean of n $(= 1,000)$ tuples of the test data $\{(\tilde{Y}_i, \tilde{X}_i); i = 1, 2, \ldots, n\}$ and we conclude that the criterion giving the small value of the PSE is better. Moreover, we checked the standard deviation for difference between the values of PSE given by two criteria, because the performance of each criterion may almost be the same when it is large, even if the difference between the values of PSE are large. Thus, we conclude that the difference is significant when the value of the standard deviation is small. We also checked the computation time (sec) to compute C_p and CV for each value of the tuning parameters as a secondary index for the assessment.

The simulation settings were as follows. First, we made $(M \times K)$-dimensional matrices X_i $(i = 1, 2, \ldots, N)$ by the sampling uniformly without replacement from $\{1, 2, \ldots, J\}$. We then made Y_i based on the model in (1) for each $i = 1, 2, \ldots, N$, and rounded so that the elements of Y_i were in $\{1, 2, \ldots, 5\}$. Next, we used $\Sigma = (0.5^{|i-j|})_{i,j=1,2,\ldots,K}$ and $\Xi = (1-\rho) I_M + \rho \mathbf{1}_M \mathbf{1}_M'$ with a fixed ρ. This matrix Ξ is known as an intra-class correlation matrix and it represents correlations among the rows of Y_i. We used it as one of the simplest matrices appropriate for representing correlations among objects. Note that while we use the intra-class correlation matrix here, our theory does not depend on any specific structure for Ξ. The true parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{J-1})'$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{J-1})'$, $\Gamma = (\gamma_{jk})_{j=1,2,\ldots,J-1; k=1,2,\ldots,K-1}$ were drawn from

$$\boldsymbol{\alpha} \sim \mathrm{N}(\mathbf{0}, 10^{-3} \mathbb{I}_{J-1}), \qquad \boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, 10^{-3} \mathbb{I}_{K-1}),$$

and

$$\Gamma \sim \mathrm{N}(\mathbf{0}, 10^{-3} \mathbb{I}_{K-1} \otimes \mathbb{I}_{J-1}),$$

where $\mathbb{I}_q = I_q + \mathbf{1}_q \mathbf{1}_q'$ for an positive integer q, and $\mu = 3$. In this case, we can evaluate

$$N\mathrm{tr}(\Xi) - S = NM - J \quad (17)$$

in (2). The details for deriving (17) are given in Appendix 3. Sample size N was set to 500 or 1,000, nine cases were considered for three-tuple (J, M, K), and fifty simulations were conducted.

Table 1 shows the results for $\rho = 0$ and $\rho = 0.5$, and the average and standard deviation of the PSE. Standard deviation of the differences between the values of the PSE (Diff) are also provided. In each case, we see that both the average and standard deviation of PSE for C_p in (15) are smaller than those of CV. Moreover, comparing $N = 500$ and $N = 1,000$ with the same value of (J, M, K), the value of Diff is small when $N = 1,000$. Thus we can say that the difference between the values given by C_p and CV is significant as N increases.

On the other hand, Fig. 1 shows the comparison of the computation time to compute C_p and CV for each value of the tuning parameters. We set $M = 5$ and $K = 4$. On the left, we can see that the computation time for CV increases although that for C_p is not much changed as N or J increase. An enlarged view of the computation time for C_p is drawn on the right. Since the difference among lines is small, we can say that the computation time for C_p is robust to scale changes. Moreover, the model selection via C_p is easily implemented because C_p in (15) has a simple form. On the whole, we conclude that the C_p in (15) is better than CV.

Table 1 Comparison between C_p and CV for simulated data

ρ	(J, M, K)	N	C_p (sd)	CV (sd)	Diff
0	(30, 5, 2)	500	11.726 (0.432)	12.764 (0.697)	0.516
		1000	11.370 (0.369)	11.661 (0.480)	0.226
	(60, 5, 2)	500	13.311 (0.713)	16.189 (1.547)	1.146
		1000	12.136 (0.408)	13.003 (0.703)	0.476
	(90, 5, 2)	500	14.544 (0.688)	21.219 (2.249)	1.918
		1000	12.819 (0.432)	14.757 (0.835)	0.664
	(30, 5, 4)	500	23.623 (0.931)	25.612 (1.391)	1.068
		1000	22.411 (0.539)	22.870 (0.694)	0.326
	(60, 5, 4)	500	25.642 (1.207)	31.381 (3.331)	2.484
		1000	23.919 (0.715)	26.073 (1.317)	1.050
	(90, 5, 4)	500	26.141 (0.835)	37.661 (4.553)	4.198
		1000	24.376 (0.616)	27.986 (1.371)	1.113
	(30, 10, 4)	500	43.211 (1.101)	43.525 (1.230)	0.336
		1000	41.958 (0.664)	42.084 (0.663)	0.161
	(60, 10, 4)	500	46.600 (1.136)	49.628 (1.929)	1.475
		1000	44.857 (0.781)	45.629 (0.831)	0.408
	(90, 10, 4)	500	48.550 (1.040)	56.096 (3.186)	2.735
		1000	45.542 (0.956)	47.192 (0.991)	0.611
0.5	(30, 5, 2)	500	12.905 (0.670)	14.282 (1.080)	0.892
		1000	12.294 (0.650)	12.609 (0.774)	0.288
	(60, 5, 2)	500	15.197 (1.043)	19.896 (2.325)	1.899
		1000	13.760 (0.659)	14.895 (0.752)	0.481
	(90, 5, 2)	500	16.520 (0.950)	28.466 (4.101)	3.773
		1000	14.553 (0.560)	17.854 (1.386)	1.101
	(30, 5, 4)	500	24.779 (1.289)	26.680 (1.715)	1.126
		1000	22.189 (0.644)	22.972 (0.832)	0.548
	(60, 5, 4)	500	27.027 (1.295)	36.329 (4.517)	3.960
		1000	24.047 (0.714)	27.204 (1.710)	1.322
	(90, 5, 4)	500	28.378 (1.413)	44.593 (5.329)	4.685
		1000	25.965 (0.635)	33.100 (1.840)	1.698
	(30, 10, 4)	500	45.319 (1.232)	46.186 (1.577)	0.725
		1000	42.442 (1.009)	42.490 (0.917)	0.491
	(60, 10, 4)	500	52.135 (3.347)	56.454 (3.215)	3.110
		1000	46.593 (1.320)	48.044 (1.696)	0.841
	(90, 10, 4)	500	55.531 (4.301)	65.578 (4.003)	3.381
		1000	48.748 (1.115)	50.934 (1.566)	1.337

5 Real data analysis

In this section, we compare the methods by applying them to real data. In the data, objects are grouped into three categories and we assume that the data for three categories are independent each other. For the three categories, (N, J)'s are respectively (1884,60), (1364,21), and (1425,44), and $(K, M) = (4, 5)$.

We used 1,200 samples obtained at random as training data for each category, and the rest of the data is used as test data. In addition, we set $\rho = 0$ or $\rho = 0.5$. Table 2 shows the PSE in (16) evaluated from the test data after selecting the tuning parameters based on C_p and CV. Similarly to Section 4, we observe that the criterion giving a smaller value of the PSE is better, and so we can say that

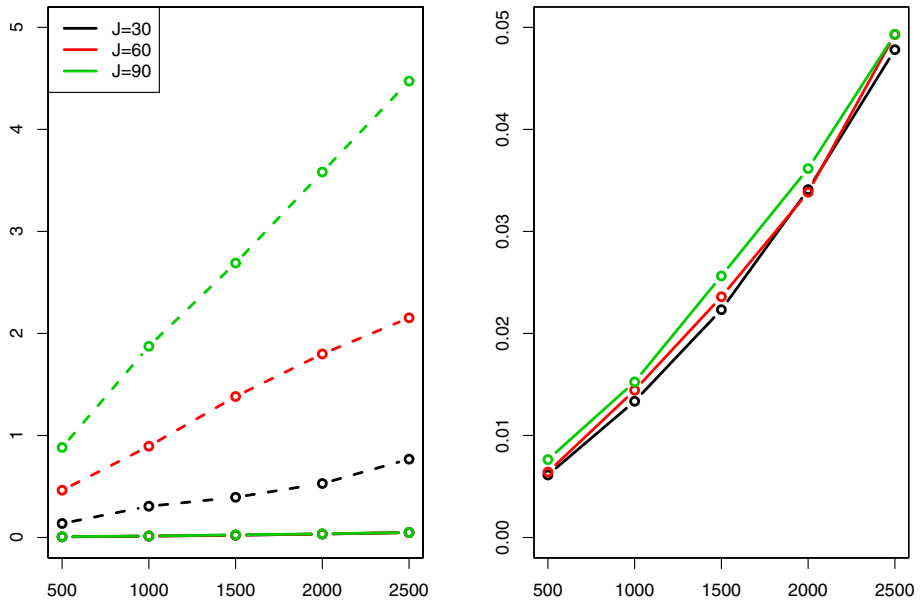

Fig. 1 Transition of computation time. In each figure, the horizontal axis indicates the sample size N and the vertical axis indicates the average of computation time (sec) for each value of the tuning parameters. On the left, the solid and the dashed line represent the average of computation time via C_p and CV, respectively. An extended figure of the average of computation time via C_p is drawn on the right

the tuning parameter selection based on C_p is superior to that of CV. Looking at the result for C_p, the value of the PSE with $\rho = 0.5$ is small in categories 1 and 3, and the value of the PSE with $\rho = 0$ is small in category 2. Thus it is suggested that the correlations among the objects in category 2 are smaller than those of other categories. Note that there is no significant difference of the results for C_p and CV in category 2. It is because the size of test data is small compared to that in categories 1 and 3.

6 Concluding remarks and future work

In this paper, we have considered an appropriate model and estimating method in a questionnaire study and derived the C_p criterion to choose the tuning parameters included in the estimator. More precisely, using a dummy matrix representing the existence of the missing values, we have constructed a model which can be interpreted as an extension of the GMANOVA model of Potthoff & Roy [9] for three-dimensional array data. We have explicitly evaluated the penalty term in the C_p without assuming the normality of the noise and shown that it becomes a simple form. Through the simulation study and real data analysis, we have confirmed the usefulness of the derived C_p. This criterion has a high prediction accuracy and low computational costs compared to CV because it can be expressed by a simple form explicitly.

It is well known that predicting a missing part is important when we construct a recommendation system, which is sometimes required in a recent questionnaire study or WEB survey. However, it is in general difficult to evaluate the prediction accuracy for methods such as collaborative filtering or matrix completion. For this problem, by extending the method in this paper to a model which contains a random effect in the evaluators, it might be possible to draw common statistical inferences, including the evaluation of the prediction accuracy.

In the future, it is expected that similar results will be obtained for more complex models because the model we considered has a particular structure for X_i and A. In addition, it will be necessary to treat the case where X_i is random, Ξ is unknown or there exist correlations among the categories in order to use more flexible models.

Table 2 Comparison between C_p and CV for real data

ρ	category	C_p	CV
0	1	21.433	22.239
	2	20.323	20.364
	3	19.989	20.177
0.5	1	19.861	20.390
	2	20.523	20.566
	3	19.000	19.221

Appendix 1: Matrix algebra

Here, we describe some matrix algebra that we have used in this paper. All the proofs can be found in Harville [3].

First, we describe two matrix inversion formulae. Let $A \in \mathbb{R}^{p\times p}$, $B \in \mathbb{R}^{p\times q}$, and $C \in \mathbb{R}^{q\times q}$, and assume that A is non-singular. Then

$$\begin{aligned}&\begin{pmatrix} A & B \\ B' & C \end{pmatrix}^{-1} \\ =&\begin{pmatrix} A^{-1}+A^{-1}BD^{-1}B'A^{-1} & -A^{-1}BD^{-1} \\ -D^{-1}B'A^{-1} & D^{-1} \end{pmatrix}\end{aligned} \tag{18}$$

if and only if $D = C - B'A^{-1}B$ is non-singular. In addition, assume that C is non-singular. Then

$$\begin{aligned}&(A+BCB')^{-1} \\ =&A^{-1} - A^{-1}B(C^{-1}+B'A^{-1}B)^{-1}B'A^{-1}\end{aligned} \tag{19}$$

if and only if $C^{-1} + B'A^{-1}B$ is non-singular. This is also known as Woodbury's formula, and in the special case where $A = \alpha I_p$, $B = \boldsymbol{b} \in \mathbb{R}^p$, and $C = \beta \in \mathbb{R}$ such that $\alpha \neq 0$ and $\alpha + \beta \boldsymbol{b}'\boldsymbol{b} \neq 0$, we have

$$(\alpha I_p + \beta \boldsymbol{b}\boldsymbol{b}')^{-1} = \frac{1}{\alpha}\left(I_p - \frac{\beta}{\alpha+\beta\boldsymbol{b}'\boldsymbol{b}}\boldsymbol{b}\boldsymbol{b}'\right) \tag{20}$$

Next, we describe the relationship between tr and vec operators. For matrices $A \in \mathbb{R}^{p\times q}$, $B \in \mathbb{R}^{q\times q}$, and $C \in \mathbb{R}^{p\times p}$, we have

$$\mathrm{tr}(A'CAB') = \mathrm{vec}(A)'(B\otimes C)\mathrm{vec}(A). \tag{21}$$

From (21), we can immediately see that

$$\mathrm{tr}(D'A'CAB') = \mathrm{vec}(A)'(DB\otimes C)\mathrm{vec}(A) \tag{22}$$

for matrices A, B, C defined in (21), and $D \in \mathbb{R}^{q\times q}$.

Appendix 2: Unbiasedness of (2)

Noting that $A'(AA')^{-1}A = I_K$ by the definition of A, it follows that

$$Y_i - X_i\tilde{B}A = E_i - \sum_{h=1}^{N} X_i(X'X)^{-1}X_h'E_h. \tag{23}$$

In addition, for an $(M\times M)$-dimensional matrix T, we can see that

$$\mathrm{E}[E_i'TE_i] = \mathrm{tr}(T\Xi)\Sigma. \tag{24}$$

In fact, the (h,l)-th element of $E_i'TE_i$ is given by $\boldsymbol{\varepsilon}_{ih}'T\varepsilon_{il}$ for $h,l = 1,2,\ldots,K$, and thus we have

$$\mathrm{E}[\boldsymbol{\varepsilon}_{ih}'T\boldsymbol{\varepsilon}_{il}] = \mathrm{tr}(T\mathrm{E}[\boldsymbol{\varepsilon}_{il}\boldsymbol{\varepsilon}_{ih}']) = \mathrm{tr}(T\Xi)\sigma_{lh}.$$

This and the symmetry of Σ imply (24). From (23), (24), and the independence of E_i, we have

$$\begin{aligned}&\sum_{i=1}^{N}\mathrm{E}[(Y_i - X_i\tilde{B}A)'(Y_i - X_i\tilde{B}A)] \\ =&\sum_{i=1}^{N}\mathrm{E}[E_i'E_i] - 2\sum_{i,h}\mathrm{E}[E_h'X_h(X'X)^{-1}X_i'E_i] \\ &+\sum_{i,h,l}\mathrm{E}[E_h'X_h(X'X)^{-1}X_i'X_i(X'X)^{-1}X_l'E_l] \\ =&N\mathrm{tr}(\Xi)\Sigma - \sum_{h=1}^{N}\mathrm{tr}\{X_k(X'X)^{-1}X_h'\Xi\}\Sigma \\ =&\{N\mathrm{tr}(\Xi) - S\}\Sigma.\end{aligned}$$

This completes the proof.

Appendix 3: Derivation of (17)

By the same argument in Section 3, we see that $(X'X)^{-1}$ can be expressed as

$$\begin{pmatrix} J^{-2}\mathrm{tr}(\Delta^{-1}) & J^{-1}\mathbf{1}_{J-1}'Q \\ J^{-1}Q'\mathbf{1}_{J-1} & R \end{pmatrix},$$

where $Q = \Delta_{-J}^{-1} - J^{-1}\mathrm{tr}(\Delta^{-1})I_{J-1}$ and $R = (C'C)^{-1}C'\Delta^{-1}C(C'C)^{-1}$. Then we have

$$\begin{aligned}&X_i(X'X)^{-1}X_i' \\ =&\frac{1}{J^2}\mathrm{tr}(\Delta^{-1})\mathbf{1}_M\mathbf{1}_M' + \frac{1}{J}\tilde{X}_iCQ'\mathbf{1}_{J-1}\mathbf{1}_M' \\ &+\frac{1}{J}\mathbf{1}_M\mathbf{1}_{J-1}'QC'\tilde{X}_i' + \tilde{X}_iCRC\tilde{X}_i'.\end{aligned}$$

Note that $CQ'1_{J-1} = (\Delta^{-1} - J^{-1}\mathrm{tr}(\Delta^{-1})I_J)\mathbf{1}_J$ and $CRC = (I_J - J^{-1}\mathbf{1}_J\mathbf{1}_J')\Delta^{-1}(I_J - J^{-1}\mathbf{1}_J\mathbf{1}_J')$ by a simple calculation, and that $\tilde{X}_i\mathbf{1}_J = \mathbf{1}_M$. Hence, $X_i(X'X)^{-1}X_i'$ is reduced to $\tilde{X}_i\Delta^{-1}\tilde{X}_i'$ and we see that this is a diagonal matrix. Finally, since $\Xi = (1-\rho)I_M + \rho\mathbf{1}_M\mathbf{1}_M'$ and

$$\begin{aligned}&\sum_{i=1}^{N}\mathrm{tr}\{X_i(X'X)^{-1}X_i'\mathbf{1}_M\mathbf{1}_M'\} \\ =&\sum_{i=1}^{N}\mathrm{tr}\{\tilde{X}_i\Delta^{-1}\tilde{X}_i'\mathbf{1}_M\mathbf{1}_M'\} = \sum_{i=1}^{N}\mathbf{1}_M'\tilde{X}_i\Delta^{-1}\tilde{X}_i'\mathbf{1}_M \\ =&\sum_{i=1}^{N}\mathrm{tr}\{\tilde{X}_i\Delta^{-1}\tilde{X}_i'\} = J,\end{aligned}$$

we obtain

$$\begin{aligned}&\sum_{i=1}^{N}\mathrm{tr}\{X_i(X'X)^{-1}X_i'\Xi\}\\ =&(1-\rho)\sum_{i=1}^{N}\mathrm{tr}\{X_i(X'X)^{-1}X_i'\}\\ &+\rho\sum_{i=1}^{N}\mathrm{tr}\{X_i(X'X)^{-1}X_i'\mathbf{1}_M\mathbf{1}_M'\}\\ =&(1-\rho)J+\rho J=J.\end{aligned}$$

This and $\mathrm{tr}(\Xi)=M$ imply (17).

Acknowledgements
The authors would like to thank the reviewer for his/her valuable commentsand advice to improve the paper. This research was partially supported by a Grant-in-Aid for Scientific Research (23500353) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Author details
[1]Nagoya Institute of Technology, Gokiso-cho, Showa-ku, 466-8555 Nagoya, Japan. [2]Fujitsu Laboratories Ltd, 4-1-1 Kamikodanaka, Nakahara-ku, 211-8588 Kawasaki, Japan. [3]Institute of Mathematics for Industry, 744, Motooka, Nishi-ku, 819-0395 Fukuoka, Japan.

Received: 30 April 2016 Revised: 9 August 2016
Accepted: 18 August 2016
Published online: 30 August 2016

References

1. Benzécri, J-P: Correspondence Analysis Handbook. Marcel Dekker, New York (1992)
2. Candès, EJ, Recht, B: Exact matrix completion via convex optimization. Found. Comput. Math. **9**(6), 717–772 (2009)
3. Harville, DA: Matrix Algebra from a Statistician's Perspective. Springer, New York (1997)
4. Hoerl, AE, Kennard, RW: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. **12**(1), 55–67 (1970)
5. Koltchinskii, V, Lounici, K, Tsybakov, AB: Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Ann. Stat. **39**(5), 2302–2329 (2011)
6. Mallows, CL: Some comments on C_p. Technometrics. **15**(1), 661–675 (1973)
7. Mallows, CL: More comments on C_p. Technometrics. **37**(4), 362–372 (1995)
8. Nagai, I: Modified C_p criterion for optimizing ridge and smooth parameters in the mgr estimator for the nonparametric gmanova model. Open J. Stat. **1**, 1–14 (2011)
9. Potthoff, RF, Roy, S: A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika. **51**(3/4), 313–326 (1964)
10. Stone, M: Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. Ser. B Stat Methodol. **36**(2), 111–147 (1974)