Adjustment for nonignorable nonresponse using latent homogeneous response groups

Caren Hasler, *University of Neuchâtel, Switzerland*, caren.hasler@unine.ch
Alina Matei, *University of Neuchâtel and IRDP Neuchâtel, Switzerland*, alina.matei@unine.ch

Abstract. estimated response probabilities are used to compute a two-phase estimator of the population total. Simulations are performed in order to compare the proposed estimators with other estimators currently used. The advantages in terms of bias and variance of the proposed approaches are confirmed through these simulations. We consider a setup in which nonignorable nonresponse is present in the survey. In such a case, the unit response probabilities depend on the variable of interest. When the variable of interest follows a mixture distribution (a typical example of such a variable is income), it is possible to highlight latent homogeneous response groups based on the variable of interest and auxiliary information. Two approaches are discussed. In both approaches, response probabilities are estimated through logistic regression. The estimated response probabilities are then used to compute a two-phase estimator of the population total. Simulations are performed in order to compare the performance of the proposed estimators with that of other estimators currently used. The advantages in terms of reduction of nonresponse bias and variance of the proposed approaches are confirmed through these simulations.

Keywords. Survey sampling, Unit response probability, Two-phase estimation

1 Introduction

Reweighting procedures are commonly used to compensate for unit nonresponse in surveys. The main idea is to increase the sampling weights of each respondent in order to compensate for the nonrespondents. One refers to such procedures as nonresponse weighting adjustment (NWA) methods. Nonresponse can be viewed as a second phase of the survey. Theory of two-phase sampling hence suggests a two-phase estimator which extends the usual Horvitz-Thompson estimator by multiplying the sampling weights of the respondents by the inverse of their response probabilities. As the response probabilities are unknown, a preliminary step consists of estimating them. The sampling weights of the respondents are then multiplied by

the inverse of their estimated response probabilities and a two-phase estimator adjusted for nonresponse is obtained. In the literature, several approaches have been used to estimate the response probabilities, as for example response homogeneity groups, calibration, or parametric modelling as in [2] and [7]. Auxiliary information available at the sample or population level plays a central role in the estimation process. It can simultaneously decrease variance and nonresponse bias of estimators if it is adequately used in the response probabilities estimation. The reader may refer to [11] for an overview of NWA methods.

Nonignorable nonresponse refers to a nonresponse mechanism which depends on the variable of interest itself (see [9] for a formal definition). It is particularly difficult to handle as the process that leads to nonresponse is defined through characteristics of interest which are partially or completely missing. Sophisticated techniques must therefore be used to control for nonresponse bias and variance in this framework. The problem of nonignorable nonresponse in surveys has already been addressed as for instance in [6], [10], [1], and [4].

We propose two NWA procedures for handling nonignorable nonresponse, when the variable of interest follows a mixture distribution with different components. The goal is to reduce nonresponse bias and variance of estimators. Latent homogeneous response groups based on both auxiliary information and the variable of interest are highlighted for respondents and are imputed using auxiliary information for nonrespondents. In the presented procedures, the response probabilities are modelled through logistic regression including information about the groups (observed or imputed). The estimated response probabilities are then used in a two-phase estimator for the total of the variable of interest. The inclusion of information about the groups in the estimation of the response probabilities allows to control simultaneously nonresponse bias and variance of the two-phase estimator.

A typical example of application where the proposed methods can be used is a survey whose variable of interest is the income. Indeed, it is customary and sensible to suppose that the willingness to answer questions related to income depends on the income itself. On the other hand, income data typically shows heterogeneity and mixture distributions represent a powerful tool to model such data (see [5]). It follows that a natural assumption is the existence of homogeneous response groups depending on the underling income mixture groups and auxiliary information.

The paper is organized as follows. Section 2 introduces the framework and notation. Section 3 discusses the response probabilities estimation for nonignorable nonresponse using logistic regression. The proposed procedures are presented in Section 4. Next, in Section 5, the performance of the proposed procedures is tested and compared to that of other NWA procedures through a simulation study. Finally, Section 6 closes the paper with brief concluding remarks.

2 Framework

Consider a finite population U of size N, indexed by i from 1 to N. Let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iq})^\top$ be a vector of q auxiliary variables attached to unit i and suppose that the parameter of interest is the population total $Y = \sum_{i \in U} y_i$, for some continuous or categorical variable of interest y. In a first phase, a sample s of size n is selected from the population U using a sampling design $p(s)$. Let $\pi_i = \sum_{s; s \ni i} p(s)$ denote the first-order inclusion probability of unit i and suppose thereafter that $\pi_i > 0$ for all $i \in U$. The vector of auxiliary variables $\mathbf{x}_i$ is assumed to be available for each population unit $i \in U$ or at least for each sampled unit $i \in s$. In the presence of unit

nonresponse, some selected units do not respond to the survey. This results in two subsets which form a partition of s: the survey *respondents* (the set r) and the survey *nonrespondents* (the set $\overline{r}$). The value y_i of the variable of interest is then observed for each respondent $i \in r$ but is missing for each nonrespondent $i \in \overline{r}$. For $i \in s$, let R_i be the response indicator of y_i which takes value 1 if unit i is a respondent (i.e. if $i \in r$) and 0 if unit i is a nonrespondent (i.e. if $i \in \overline{r}$). Let p_i be the response propensity of unit i, that is $p_i = \Pr(i \in r | s; i \in s)$. It is supposed that units respond independently from each other. The response indicator R_i is therefore generated from a Bernoulli random variable with parameter p_i. Moreover, it is thereafter assumed that $p_i > 0$ for all $i \in U$. In the ideal case of complete response, the Horvitz-Thompson estimator

$$\widehat{Y}_\pi = \sum_{i \in s} \frac{1}{\pi_i} y_i, \tag{1}$$

is a design unbiased estimator for Y. In the presence of nonresponse, however, this latter is intractable as the values y_i of the variable of interest are missing for nonrespondents $i \in \overline{r}$. Nonresponse can be viewed as a second phase of the survey. A subsample r of s is indeed selected according to a Poisson sampling design $q(r|s) = \prod_{i \in r} p_i \prod_{i \in \overline{r}} (1 - p_i)$. Theory of two-phase sampling proposes, in this case, the double expansion estimator $\widehat{Y}_{\pi,p} = \sum_{i \in r} \frac{1}{\pi_i} \frac{1}{p_i} y_i$, which extends the estimator in Expression (1). This estimator would be unbiased for Y if the response probabilities p_i were known. Unfortunately, this is never the case. A preliminary step therefore consists of estimating the response probabilities. Those are then replaced by the estimated response probabilities $\widehat{p}_i$ in the previous estimator and the two-phase estimator adjusted for nonresponse

$$\widehat{Y}_{\pi,\widehat{p}} = \sum_{i \in r} \frac{1}{\pi_i} \frac{1}{\widehat{p}_i} y_i, \tag{2}$$

is obtained. If the response probabilities are parametrically modeled, then it is shown in [7] that estimator $\widehat{Y}_{\pi,\widehat{p}}$ is more efficient than estimator $\widehat{Y}_{\pi,p}$ when maximum likelihood is used to estimate the parameters. In Section 3, the question of the response probabilities estimation for nonignorable nonresponse is discussed.

3 Estimating response probabilities

Under nonignorable nonresponse, a solution to estimate the response probabilities consists of modelling them with logistic regression in which the variable of interest plays the role of a covariate. Hence, the following two models can be considered:

$$p_i = \mathbb{E}(R_i | y_i) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 y_i)]}, \tag{3}$$

$$p_i = \mathbb{E}(R_i | y_i, \mathbf{x}_i) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \beta_1 y_i + \mathbf{x}_i^\top \boldsymbol{\alpha}\right)\right]}, \tag{4}$$

where β_0, β_1, and $\boldsymbol{\alpha}$ are parameters. In the presence of nonresponse, however, these parameters can not be estimated as the values y_i of the variable of interest are missing for the nonrespondents.

A solution is proposed in [2] and is presented below. It consists of considering only the auxiliary variables as covariates. This results in the following model

$$p_i = \mathbb{E}(R_i | \mathbf{x}_i) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\alpha}\right)\right]}, \tag{5}$$

where β_0 and $\boldsymbol{\alpha}$ are parameters. As the values $\mathbf{x}_i$ of the auxiliary variables are known for each sampled unit $i \in s$, the parameters can now be estimated considering $(R_i, \mathbf{x}_i)$ for $i \in s$. Consider $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\alpha}}$ the maximum likelihood estimates of parameters β_0 and $\boldsymbol{\alpha}$, and estimate the response probabilities by replacing the parameters by their estimates in Expression (5), that is $\widehat{p}_i = 1/\left\{1 + \exp\left[-\left(\widehat{\beta}_0 + \mathbf{x}_i^\top \widehat{\boldsymbol{\alpha}}\right)\right]\right\}$. If the auxiliary variables are good predictors for the variable of interest or for the response probabilities, then this procedure provides protection against nonresponse bias (see [2]).

4 Latent homogeneous response groups

We assume that the variable of interest y follows a mixture distribution with t components $y_i \sim \sum_{\ell=1}^{t} \lambda_\ell f_\ell(y_i|\mathbf{x}_i, \theta_\ell)$, $\lambda_\ell \geq 0$, $\sum_{\ell=1}^{t} \lambda_\ell = 1$, where λ_ℓ is the prior probability of component ℓ (y_i is drawn from a mixture of densities of underlying groups or clusters or subpopulations in unknown proportions $\lambda_1, \ldots, \lambda_t$) and θ_ℓ is the specific parameter vector for the density function f_ℓ in the ℓth component. If f_ℓ is a univariate normal density and $\theta_\ell = (\mu_\ell, \sigma_\ell^2)'$, one describes a mixture of standard linear regression models, also called latent class regression or cluster-wise regression (see [3]). Other f_ℓ densities can also be used.

A typical example of such a variable y is income. Models based on mixed distributions better explain the income heterogeneity in different subpopulations. When nonresponse treatment is added, latent homogeneous response groups can be highlighted based on these subpopulations. These response groups depend on the variable of interest and the auxiliary information. An important gain in terms of reduction of nonresponse bias and variance can be derived from including information about these groups in the estimation of the response probabilities. In the presence of nonresponse, however, these groups are not fully observed as the values y_i of the variable of interest are unknown for nonrespondents. In the current section, a procedure to reconstruct these latent homogeneous response groups is presented. Then, two solutions to include them in the response probabilities estimation are proposed.

As stated above, homogeneous response groups are observed for respondents only. A procedure to reconstruct the group membership of the nonrespondents is provided here. The main idea is to impute the missing groups by nearest neighbor imputation. Suppose that k homogeneous groups are observed for the respondents. Moreover, let $c_i \in \{1, 2, \ldots, k\}$ be the observed group membership value of respondent $i \in r$ and consider $c_i^* \in \{1, 2, \ldots, k\}$ the reconstructed membership group value of a unit $i \in s$. As the membership group value is observed for each respondent, we set $c_i^* = c_i$ for $i \in r$. For a nonrespondent, however, the membership group value is unobserved and that one is reconstructed by nearest neighbor imputation using auxiliary information. Hence, for $i \in \overline{r}$, consider $c_i^* = c_{j(i)}$ where $j(i)$ satisfies $d(\mathbf{x}_i, \mathbf{x}_{j(i)}) = \min_{j \in r} d(\mathbf{x}_i, \mathbf{x}_j)$ for some distance measure $d(\cdot, \cdot)$. Therefore, observed group membership values are combined with imputed group membership values. This leads to a reconstructed group membership variable whose values c_i^* are available for every sampled unit $i \in s$.

Two different models can be constructed. In the first one, the reconstructed group membership variable (observed or imputed) is added as a categorical covariate. This results in the following model

$$p_i = I\!E\left(R_i|\mathbf{x}_i, c_i^*\right) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}_1 + \beta_2 c_i^* + \mathbf{x}_i^\top \boldsymbol{\beta}_3 c_i^*\right)\right]}, \tag{6}$$

where β_0, $\boldsymbol{\beta}_1$, β_2, and $\boldsymbol{\beta}_3$ are parameters. The maximum likelihood estimation is then applied to fit this model considering $(R_i, \mathbf{x}_i, c_i^*)$ for $i \in s$. This leads to estimates $\widehat{\beta}_0$, $\widehat{\boldsymbol{\beta}}_1$, $\widehat{\beta}_2$, $\widehat{\boldsymbol{\beta}}_3$, and $\widehat{p_i} = 1/\left\{1 + \exp\left[-\left(\widehat{\beta}_0 + \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_1 + \widehat{\beta}_2 c_i^* + \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_3 c_i^*\right)\right]\right\}$. If the auxiliary variables are good predictors of the variable of interest or good predictors of the response probabilities and moreover if the reconstructed groups are homogeneous with respect to the variable of interest or with respect to the response probabilities, then this procedure provides additional protection against nonresponse bias and variance compared to Model (5).

In the second proposed procedure, the missing values of the variable of interest are imputed in each reconstructed group. The response probabilities are estimated using logistic regression and the variable of interest (observed or imputed); see also [8]. Hence, let y_i^* denote a reconstructed value of the variable of interest of a unit $i \in s$. For a respondent $i \in r$, this value corresponds to the observed value of the variable of interest, that is $y_i^* = y_i$. Then, for the nonrespondents, the missing y_i's are reconstructed by using regression imputation independently in each reconstructed group. Hence, for each nonrespondent $i \in \overline{r}$ we set $y_i^* = \left(\sum_{j \in r | c_j^* = c_i^*} \frac{1}{\pi_j} \mathbf{x}_j \mathbf{x}_j^\top\right)^{-1} \left(\sum_{\ell \in r | c_\ell^* = c_i^*} \frac{1}{\pi_\ell} \mathbf{x}_\ell y_\ell\right) \mathbf{x}_i$. Therefore, observed values of the variable of interest are combined with imputed values. This leads to a reconstructed variable of interest whose values y_i^* are available for every sampled unit $i \in s$. This variable then plays the role of covariate in the logistic regression used to estimate the response probabilities. Hence, the parameters δ_0 and δ_1 of the logistic regression model

$$p_i = I\!E\left(R_i | y_i^*\right) = \frac{1}{1 + \exp\left[-\left(\delta_0 + \delta_1 y_i^*\right)\right]}, \tag{7}$$

are estimated by maximum likelihood considering (R_i, y_i^*) for $i \in s$. This leads to estimates $\widehat{\delta}_0$, $\widehat{\delta}_1$, and $\widehat{p_i} = 1/\left\{1 + \exp\left[-\left(\widehat{\delta}_0 + \widehat{\delta}_1 y_i^*\right)\right]\right\}$. If the auxiliary variables are good predictors of the variable of interest within the reconstructed groups but not necessarily within the whole population, then this procedure provides additional protection against nonresponse bias and variance compared to Model (5). Even though y_i^* is essentially a linear combination of the outer product of the auxiliary variables and the (imputed) latent groups, Model (7) is different from the model including $\mathbf{x}_i$ and c_i^* as covariates as in Expression (6), because it uses the original y_i for the respondents and performs closer to the assumed response model.

5 Simulations

A simulation study was conducted to evaluate the performance of the procedures proposed in Section 4. Two different settings were considered. In each setting, a population of size $N = 1000$ divided into two groups of equal size, a variable of interest y generated from a mixture distribution, and an auxiliary variable x were considered. A census was considered in both cases, which implies that we set $U = s$ and $\pi_i = 1$ for each $i \in s$. Ten thousand simulations were conducted.

For each setting, the simulations were conducted as follows. First, for each unit i, the response probabilities were obtained from the logistic function $p_i = 1/\left\{1 + \exp\left[-\left(\beta_0 + \beta_1 y_i\right)\right]\right\}$, where β_0 and β_1 were fixed to obtain a mean response rate close to 65%. Then, 10000 response sets were created by generating 10000 response indicator vectors R. Each component $R_i, i \in U$ of R was generated from a Bernoulli distribution with parameter p_i. For each response set generated, the population total for the variable of interest was estimated through the two-phase

estimator adjusted for nonresponse of Expression (2) by considering different choices for the estimated response probabilities $\widehat{p}_i$ as follows:

1. $\widehat{Y}_{\widehat{p}(x)}$: estimator proposed in [2], i.e. response probabilities estimated through logistic regression with the auxiliary variables as covariates as in Model (5),

2. $\widehat{Y}_{\widehat{p}(x,c^*)}$: first proposed procedure, i.e. response probabilities estimated through logistic regression with the auxiliary variables and the reconstructed membership groups variable as covariates as in Model (6),

3. $\widehat{Y}_{\widehat{p}(y^*)}$: second proposed procedure, i.e. response probabilities estimated through logistic regression with the values of the variable of interest (observed or imputed through regression imputation in the reconstructed groups) as covariates as in Model (7),

4. $\widehat{Y}_{\widehat{p}(y^{nn})}$: response probabilities estimated through logistic regression with the vector of observed and imputed by nearest neighbor values of the variable of interest y^{nn} as covariate. The coefficients of y^{nn} are thus defined as $y_i^{nn} = y_i$ if $i \in r$ and $y_i^{nn} = y_{j(i)}$ where $|x_i - x_{j(i)}| = \min_{j \in r} |x_i - x_j|$ if $i \in \overline{r}$,

5. $\widehat{Y}_p$: true response probabilities considered in the two-phase estimator.

The following comparison measures were considered for these five estimators, here generically denoted by $\widehat{Y}$:

- The Monte Carlo relative bias: $\text{RB} = B/Y$, where $B = I\!E_{sim}(\widehat{Y}) - Y$, $I\!E_{sim}(\widehat{Y}) = \sum_{i=1}^{M} \widehat{Y}_i/M$, $\widehat{Y}_i$ is the estimate of $\widehat{Y}$ obtained at the i-th simulation, and M is the number of simulations,

- The Monte Carlo variance: $\text{VAR} = \frac{1}{M-1} \sum_{i=1}^{M} \left[\widehat{Y}_i - I\!E_{sim}\left(\widehat{Y}\right)\right]^2$,

- The Monte Carlo mean square error: $\text{MSE} = \text{B}^2 + \text{VAR}$.

Details and results from the two considered settings are presented below.

Setting 1: A single auxiliary variable $x = (x_i)_{i=1}^N$ was considered. Its coefficients were generated by independent draws of a uniform distribution with parameters 0 and 1 for units that belong to the first group, and by independent draws of a uniform random variable with parameters 2 and 3 for units that belong to the second group. Next, the variable of interest $y = (y_i)_{i=1}^N$ was generated as follows: $y_i = 5 + 5x_i + 3\varepsilon_i$ if i belongs to the first group and $y_i = 40 - (x_i - 5)^2 + 3\varepsilon_i$ if i belongs to the second group, where ε_i are independent draws of a normal random variable with mean 0 and variance 1. Simulations were then conducted according to the scheme described above. The results are presented in Table 1.

The two proposed estimators ($\widehat{Y}_{\widehat{p}(x,c^*)}$ and $\widehat{Y}_{\widehat{p}(y^*)}$) display a decrease in relative bias compared to estimators $\widehat{Y}_{\widehat{p}(x)}$ and $\widehat{Y}_{\widehat{p}(y^{nn})}$. The gap between the relative bias of $\widehat{Y}_{\widehat{p}(x,c^*)}$ and that of $\widehat{Y}_{\widehat{p}(y^{nn})}$ is not large and makes it difficult to clearly rank these two estimators. The proposed estimators, however, imply a clear decrease in variance compared to estimators $\widehat{Y}_{\widehat{p}(x)}$ and $\widehat{Y}_{\widehat{p}(y^{nn})}$. Estimator $\widehat{Y}_p$ is clearly the best in terms of bias, which is not surprising. Indeed, it uses the true response probabilities and is therefore unbiased for the total (the small relative bias is due to the simulation process). Finally, the four estimators with estimated probabilities imply a huge decrease in

Table 1: Comparison measures for five estimators in setting 1.

Estimator	RB ($\times 10^{-3}$)	Var ($\times 10^{3}$)	MSE ($\times 10^{4}$)
$\widehat{Y}_{\widehat{p}(x)}$	6.96	7.59	2.83
$\widehat{Y}_{\widehat{p}(x,c^*)}$	4.63	5.10	1.43
$\widehat{Y}_{\widehat{p}(y^*)}$	3.25	5.17	0.97
$\widehat{Y}_{\widehat{p}(y^{nn})}$	5.56	9.34	2.25
$\widehat{Y}_p$	−0.13	226.90	22.69

variance compared to the estimator with the true probabilities ($\widehat{Y}_p$), which confirms the result in [7].

Setting 2: The values y_i of the variable of interest $y = (y_i)_{i=1}^N$ were generated independently from a gamma distribution with parameters 10 and 1 for units that belong to the first group and from a gamma distribution with parameters 40 and 1 for units that belong to the second group. Next, values of an auxiliary variable $x = (x_i)_{i=1}^N$ were generated as follows. We set $x_i = 5 + \rho_1 y_i + \varepsilon_i$, where $\rho_1 = 0.7$ and where ε_i was drawn from a normal random variable with mean 0 and variance $10(1-\rho_1^2)$ if i belongs to the first group. Moreover, we set $x_i = 5+\rho_2 y_i+\varepsilon_i$, where $\rho_2 = 0.93$, and where ε_i was drawn from a normal random variable with mean 0 and variance $40(1-\rho_2^2)$ if unit i belongs to the second group. Simulations were then conducted according to the scheme described above. The results are presented in Table 2. These results

Table 2: Comparison measures for five estimators in setting 2.

Estimator	RB ($\times 10^{-3}$)	Var ($\times 10^{3}$)	MSE ($\times 10^{4}$)
$\widehat{Y}_{\widehat{p}(x)}$	11.04	13.61	9.01
$\widehat{Y}_{\widehat{p}(x,c^*)}$	6.69	10.02	3.82
$\widehat{Y}_{\widehat{p}(y^*)}$	5.44	9.83	2.84
$\widehat{Y}_{\widehat{p}(y^{nn})}$	6.94	14.93	4.52
$\widehat{Y}_p$	0.02	267.78	26.78

follow a fairly similar pattern to those of setting 1. The two proposed estimators ($\widehat{Y}_{\widehat{p}(x,c^*)}$ and $\widehat{Y}_{\widehat{p}(y^*)}$) display a decrease in relative bias compared to estimators $\widehat{Y}_{\widehat{p}(x)}$ and $\widehat{Y}_{\widehat{p}(y^{nn})}$. However, the gap between the relative bias of $\widehat{Y}_{\widehat{p}(x,c^*)}$ and that of $\widehat{Y}_{\widehat{p}(y^{nn})}$ is very small and does not allow us to rank these two estimators. The proposed estimators again imply a clear decrease in variance compared to estimators $\widehat{Y}_{\widehat{p}(x)}$ and $\widehat{Y}_{\widehat{p}(y^{nn})}$. Finally, estimator $\widehat{Y}_p$ also displays by far the smallest relative bias and the largest variance.

6 Conclusion

We have proposed two NWA procedures for handling nonignorable nonresponse when the variable of interest follows a mixture distribution. Homogeneous response groups can be constructed

based on the hidden structure of the variable of interest; they include information about the variable of interest and the auxiliary information. Benefits in terms of reduction of nonresponse bias and variance of the total estimator can be obtained if these groups are taken into account in the response probability estimation. Our results are confirmed through a simulation study. We have not considered the problem of variance estimation of the total estimator when the proposed methods are applied. This problem is currently under investigation.

Acknowledgement

The authors are grateful to a reviewer for his constructive comments and suggestions.

Bibliography

[1] Beaumont, J.-F. (2000). An estimation method for nonignorable nonresponse. *Survey Methodology*, 26:131–136.

[2] Cassel, C. M., Särndal, C.-E., and Wretman, J. H. (1983). Some uses of statistical models in connexion with the nonresponse problem. In Madow, W. G. and Olkin, I., editors, *Incomplete Data in Sample Surveys*, volume 3, pages 143–160. Academic Press, New York.

[3] DeSarbo, W. S., and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5, 249–282.

[4] Fang, F., Hong, Q., and Shao, J. (2010). Empirical likelihood estimation for samples with nonignorable nonresponse. *Statistica Sinica*, 20:263–280.

[5] Flachaire, E. and Nuñez, O. (2007). Estimation of the income distribution and detection of subpopulations: An explanatory model. *Computational Statistics & Data Analysis*, 51:3368–3380.

[6] Greenlees, J. S. and Reece, W. S., and Zieschang, K. (1982) Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 378:251–261.

[7] Kim, J. K. and Kim, J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 35(4):501–514.

[8] Laaksonen, S. and Chambers, R. (2006). Survey estimation under informative nonresponse with follow-up. *Journal of Official Statistics*, 22(1):81–95.

[9] Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77(378):237–250.

[10] Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, Canada.

[11] Särndal, C.-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Wiley, New York.

Bartlett adjustment of deviance statistic for three types of binary response models

Nobuhiro Taneichi, *Kagoshima University*, taneichi@sci.kagoshima-u.ac.jp
Yuri Sekiya, *Hokkaido University of Education*, sekiya.yuri@k.hokkyodai.ac.jp
Jun Toyama, *The Institute for the Practical Application of Mathematics*, mandheling@nifty.com

Abstract. A logistic regression model, complementary log-log model and probit model are frequently used for a generalised linear model of binary data. We consider deviance (log likelihood ratio statistic) as a goodness-of-fit statistic. In this paper, using the continuous term of asymptotic expansion for the deviance under the null hypothesis that each model is correct, we obtain the Bartlett adjusted deviance statistic for each model that improves the speed of convergence to chi-square limiting distribution of deviance. Performance of each adjusted deviance statistic is also investigated numerically.

Keywords. Asymptotic expansion, Bartlett adjustment, Complementary log-log model, Deviance, Generalized linear model, Logistic regression model, Probit model

1 Introduction

We consider generalized linear models (Nelder and Wedderburn [5]) in which the response variables are measured on a binary scale. Let random variables $Y_\alpha,\ \alpha = 1, \ldots, S$ be the number of successes in S different subgroups, which are independent distributed according to binomial distributions $B(n_\alpha, \pi_\alpha)$, $\alpha = 1, \ldots, S$. If we use a monotone and differentiable function $g(\cdot)$ as a link function, we obtain a generalized linear model for binary data as

$$g(\pi_\alpha) = \boldsymbol{x}'_\alpha \boldsymbol{\beta}, \qquad \alpha = 1, \ldots, S, \tag{1}$$

where $\boldsymbol{x}_\alpha = (x_{\alpha 1}, \ldots, x_{\alpha p})', \alpha = 1, \ldots, S$, are covariate vectors and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is an unknown parameter vector and $(p < S)$. When $g(t)$ is a canonical link function, that is,

$$g(t) = \log\left(\frac{t}{1-t}\right),$$

model (1) is a logistic regression model. When

$$g(t) = g_P(t) = \Phi^{-1}(t),$$

where

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left(-\frac{x^2}{2}\right) dx,$$

model (1) is a probit model. When

$$g(t) = \log\{-\log(1-t)\},$$

model (1) is a complementary log-log model.

We consider the null hypothesis

$$H_0 : \pi_\alpha = \pi_\alpha(\boldsymbol{\beta}) = g^{-1}(\boldsymbol{x}'_\alpha \boldsymbol{\beta}), \qquad \alpha = 1, \ldots, S. \tag{2}$$

The deviance (log likelihood ratio statistic) is

$$D = 2 \sum_{\alpha=1}^{S} n_\alpha \left\{ \frac{Y_\alpha}{n_\alpha} \log\left(\frac{Y_\alpha}{n_\alpha \hat{\pi}_\alpha}\right) + \left(1 - \frac{Y_\alpha}{n_\alpha}\right) \log\left(\frac{1 - \frac{Y_\alpha}{n_\alpha}}{1 - \hat{\pi}_\alpha}\right) \right\},$$

where $\hat{\pi}_\alpha = \pi_\alpha(\hat{\boldsymbol{\beta}}), \alpha = 1, \ldots, S$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ under H_0 given by (2). Under the null hypothesis H_0, it is known that the deviance D has a χ^2_{S-p} limiting distribution if

$$n_\alpha / n \to \mu_\alpha \ (0 < \mu_\alpha < 1) \text{ for each } \alpha, \quad \text{as} \quad n \to \infty, \tag{3}$$

where $n = \sum_{\alpha=1}^{S} n_\alpha$ and $\sum_{\alpha=1}^{S} \mu_\alpha = 1$. Usually, using large sample results, we test H_0 by using the statistic D for a goodness-of-fit test statistic of each model.

However, in the case in which all $n_\alpha, \alpha = 1, \ldots, S$ are not large enough, such an approximation by a χ^2_{S-p} limiting distribution to the distribution of D under H_0 becomes poor. So, there are risks that the hypothesis test based on large sample theory will give results opposite to those of an exact test. In this paper, in order to reduce the risks, we propose a new adjusted statistic $\tilde{D}^B$ of D whose speed of convergence to a chi-square distribution is quicker than that of D. To construct $\tilde{D}^B$, we use the following procedure. First, we formally obtain the asymptotic expansion of the original statistic D assuming a continuous distribution of D. Next, we obtain adjusted statistic $\tilde{D}^B$ by performing Bartlett adjustment to D on the basis of the asymptotic expansion assuming a continuous distribution of D.

2 An asymptotic approximation for the distribution of D under H_0

With regard to evaluation of the lower probability of the deviance D under H_0, we obtain the followig theorem (a special case of Taneichi *et al.* [13]). Here, we consider the following Assumption 2.1 instead of the assumption given by (3).

Assumption 2.1. $n_\alpha \to \infty$, $\alpha = 1, \cdots, S$, as $n \to \infty$, with n_α depending on n in such a way that $n_\alpha / n = \mu_\alpha, \alpha = 1, \ldots, S$, where $0 < \mu_\alpha < 1$ and $\sum_{\alpha=1}^{S} \mu_\alpha = 1$.

Theorem 2.1. When g^{-1} is a fourth time continuously differentiable function, under Assumption 2.1 and assuming that D is continuously distributed, the lower probability of the deviance D under H_0 is evaluated as

$$\Pr\{D \le x|H_0\} = \Pr\{\chi^2_{S-p} \le x\} + \frac{1}{n}\sum_{j=0}^{1} v_j \Pr\{\chi^2_{S-p+2j} \le x\} + O(n^{-2}),$$

where χ^2_f denotes a chi-square random variable with degrees of freedom f,

$$v_0 = -\frac{1}{24}(2A_1 - 6A_2 + 12A_3 - 3A_4 + 4B_1 - 12B_2 + 6B_3 - 3B_4),$$

$$v_1 = -v_0,$$

where

$$A_1 = \sum_{\alpha=1}^{S} \frac{1-\pi_\alpha+\pi_\alpha^2}{\mu_\alpha\pi_\alpha(1-\pi_\alpha)}, \quad A_2 = \sum_{\alpha=1}^{S} \frac{\mu_\alpha(1-3\pi_\alpha+3\pi_\alpha^2)}{\pi_\alpha^3(1-\pi_\alpha)^3} G_1^4(\alpha)\sigma_{\alpha\alpha}^2,$$

$$A_3 = \sum_{\alpha=1}^{S} \frac{\mu_\alpha(1-2\pi_\alpha)}{\pi_\alpha^2(1-\pi_\alpha)^2} G_1^2(\alpha)G_2(\alpha)\sigma_{\alpha\alpha}^2, \quad A_4 = \sum_{\alpha=1}^{S} \frac{\mu_\alpha}{\pi_\alpha(1-\pi_\alpha)} G_2^2(\alpha)\sigma_{\alpha\alpha}^2,$$

$$B_1 = \sum_{\alpha=1}^{S}\sum_{\gamma=1}^{S} \frac{\mu_\alpha(1-2\pi_\alpha)}{\pi_\alpha^2(1-\pi_\alpha)^2} \frac{\mu_\gamma(1-2\pi_\gamma)}{\pi_\gamma^2(1-\pi_\gamma)^2} G_1^3(\alpha)G_1^3(\gamma)\sigma_{\alpha\gamma}^3,$$

$$B_2 = \sum_{\alpha=1}^{S}\sum_{\gamma=1}^{S} \frac{\mu_\alpha}{\pi_\alpha(1-\pi_\alpha)} \frac{\mu_\gamma(1-2\pi_\gamma)}{\pi_\gamma^2(1-\pi_\gamma)^2} G_1(\alpha)G_2(\alpha)G_1^3(\gamma)\sigma_{\alpha\gamma}^3,$$

$$B_3 = \sum_{\alpha=1}^{S}\sum_{\gamma=1}^{S} \frac{\mu_\alpha}{\pi_\alpha(1-\pi_\alpha)} \frac{\mu_\gamma}{\pi_\gamma(1-\pi_\gamma)} G_1(\alpha)G_2(\alpha)G_1(\gamma)G_2(\gamma)\sigma_{\alpha\gamma}^3,$$

$$B_4 = \sum_{\alpha=1}^{S}\sum_{\gamma=1}^{S} \frac{\mu_\alpha}{\pi_\alpha(1-\pi_\alpha)} \frac{\mu_\gamma}{\pi_\gamma(1-\pi_\gamma)} G_1(\alpha)G_2(\alpha)G_1(\gamma)G_2(\gamma)\sigma_{\alpha\alpha}\sigma_{\alpha\gamma}\sigma_{\gamma\gamma},$$

$$G_i(\alpha) = u^{(i)}(\boldsymbol{x}'_\alpha\boldsymbol{\beta}), \qquad \alpha = 1,\dots,S,\ i = 1,2,$$

$$\sigma_{\alpha\gamma} = \boldsymbol{x}'_\alpha K^{-1}\boldsymbol{x}_\gamma,$$

$$K = \sum_{\lambda=1}^{S} \frac{\mu_\lambda}{\pi_\lambda(1-\pi_\lambda)} G_1^2(\lambda)\boldsymbol{x}_\lambda\boldsymbol{x}'_\lambda,$$

where $u^{(i)}$ is the ith derivative of $u(x) = g^{-1}(x)$.

Evaluation for the logistic regression model is given by applying

$$g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)},$$

$$G_1(\alpha) = \pi_\alpha(1-\pi_\alpha), \qquad \alpha = 1,\dots,S,$$

and

$$G_2(\alpha) = \pi_\alpha(1-\pi_\alpha)(1-2\pi_\alpha), \qquad \alpha = 1,\dots,S$$

to Theorem 2.1. Similarly, evaluation for the probit model is given by applying

$$g^{-1}(x) = \Phi(x),$$
$$G_1(\alpha) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{\left\{\Phi^{-1}(\pi_\alpha)\right\}^2}{2}\right], \qquad \alpha = 1, \ldots, S,$$

and

$$G_2(\alpha) = -\frac{1}{\sqrt{2\pi}}\Phi^{-1}(\pi_\alpha) \exp\left[-\frac{\left\{\Phi^{-1}(\pi_\alpha)\right\}^2}{2}\right], \qquad \alpha = 1, \ldots, S$$

and evaluation for the complementary log-log model is given by applying

$$g^{-1}(x) = 1 - \exp\{-\exp(x)\},$$
$$G_1(\alpha) = -(1-\pi_\alpha)\log(1-\pi_\alpha), \qquad \alpha = 1, \ldots, S,$$

and

$$G_2(\alpha) = -(1-\pi_\alpha)\{\log(1-\pi_\alpha)\}\{1+\log(1-\pi_\alpha)\}, \qquad \alpha = 1, \ldots, S$$

to Theorem 2.1, respectively.

We consider the appropriateness of using the Edgeworth approximation assuming a continuous distribution like Theorem 2.1. Yarnold [14] obtained an asymptotic expansion for the null distribution of X^2 (Pearson's chi-square statistic). The expansion consists of continuous and discontinuous terms. Yarnold [14] numerically examined the accuracy of approximations based on the expansion, χ^2 approximation, and Edgeworth approximation assuming a continuous distribution for the null distribution of X^2 and concluded that the Edgeworth approximation assuming a continuous distribution should never be used when random variable has a lattice distribution. In a similar fashion to X^2 statistic, approximations based on asymptotic expansions for null distributions of the log likelihood ratio test statistic and the Freeman-Tukey statistic were obtained by Siotani and Fujikoshi [9], that of the power-divergence statistics was obtained by Read [6] and that of the ϕ-divergence statistics was obtained by Menéndez et al. [4]. The numerical accuracy of the approximation was shown by Yarnold [14] for X^2 statistic and by Read [7] for power-divergence statistics. When the discontinuous term in the asymptotic expansion can be expressed in a simple form as the discontinuous term for the null distribution of above statistics, we must respect Yarnold's recommendation.

On the other hand, from the numerical results obtained by Yarnold [14], we notice that the χ^2 approximation rarely performs better than the Edgeworth approximation assuming a continuous distribution. Thus, the Edgeworth approximation assuming a continuous distribution appears to be an effective approximation when the discontinuous term in the asymptotic expansion cannot be expressed in a simple form. Unlike in the case of the null distribution of above statistics, it is very difficult to represent the discontinuous term in a simple form in the case of the distribution of statistics under alternative hypothesis and in the case of that for more general multinomial models such as contingency tables. The reason for the results are shown in Taneichi *et al.* [11] and Taneichi and Sekiya [12], mathematically. Edgeworth approximations of the distributions of some kinds of multinomial goodness-of-fit statistics under alternative hypotheses have been investigated Taneichi *et al.* [11, 10] and Sekiya and Taneichi [8]. Taneichi and Sekiya [12] discussed approximations for the distribution of statistics for the test of independence in $r \times s$ contingency tables. Based on numerical investigations, we found that an omission of the discountinuous term does not lead to a serious error.

3 Bartlett adjusted deviance statistic

In this section, we propose the Bartlett adjusted deviance statistic for improving small sample accuracy of χ^2 approximation of the distribution of a random variable.

Suppose that a nonnegative random variable T has an asymptotic expansion such that

$$\Pr\{T \leq x\} = \Pr\{\chi_f^2 \leq x\} + \frac{1}{n}\sum_{j=0}^{1} a_j \Pr\{\chi_{f+2j}^2 \leq x\} + O(n^{-2}).$$

Also suppose that the coefficients a_j, $(j = 0, 1)$ do not depend on the parameter $n(> 0)$ and must satisfy the relation $a_0 + a_1 = 0$.

In order to increase the accuracy of χ^2 approximation of a random variable T, we consider Bartlett adjustment of random variable T defined by T_B.

$$T_B = \left(1 + \frac{2a_0}{fn}\right) T. \tag{4}$$

Then, it holds that

$$\Pr\{T_B \leq x\} = \Pr\{\chi_f^2 \leq x\} + O(n^{-2}).$$

Lawley [3], Barndorff-Nielsen and Cox [1], and Barndorff-Nielsen and Hall [2] discussed Bartlett adjustment for the log likelihood ratio statistic. Applying Theorem 2.1 to T_B given by (4), we obtain the Bartlett adjusted deviance statistic D^B.

$$D^B = \left\{1 + \frac{2v_0}{n(S-p)}\right\} D.$$

Practically, we must use estimate $\hat{v}_0$ obtained by substituting the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ for true value $\boldsymbol{\beta}$ in v_0. Therefore, we propose the statistic $\tilde{D}^B$ that is obtained by substitutig $\hat{v}_0$ for v_0 in D^B.

4 Numerical studies

In this section, we compare the performance of the Bartlett adjusted deviance statistic $\tilde{D}^B$ with that of the original deviance D using the Monte Carlo procedure.

We consider a generalized linear model given by (1) with $p = 2$ and $x_{\alpha,1} = 1$ and $x_{\alpha,2} = x_\alpha, \alpha = 1, \ldots, S$. Let the true values of parameters β_1 and β_2 be β_1^* and β_2^*, respectively. Then, the true value of π_α is

$$\pi_\alpha^* = g^{-1}(\beta_1^* + \beta_2^* x_\alpha), \qquad \alpha = 1, \ldots, S.$$

As a link fuction $g(\cdot)$, we consider the logit link, complementary log-log link and probit link. We give a design matrix

$$\boldsymbol{X} = (\mathbf{1}, \mathrm{vec}\{\boldsymbol{x}\})$$

and execute the following procedure.

For each α, we generate n_α, $\alpha = 1, \ldots, S$ binomial random numbers that are distributed according to $B(1, \pi_\alpha^*)$. From them, we calculate the number of successes $Y_\alpha, \alpha = 1, \ldots, S$ and the maximum likelihood estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ for the parameters β_1 and β_2 by Fisher scoring

method. Using the estimates, we calculate the values $\pi_\alpha(\hat{\boldsymbol{\beta}})$, $\alpha = 1, \ldots, S$, where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)'$, and the observed values of the statistics D and $\tilde{D}^B$. This process is repeated J times.

Among the J times, let V be the number of times that the observed values of the statistic exceed the upper ε point of the χ^2 distribution with degrees of freedom $S - p$, that is, $\chi^2_{S-p}(\varepsilon)$. The performance of χ^2 approximation for the distribution of each statistic can be evaluated on the basis of the index

$$I = \frac{V}{J} - \varepsilon.$$

We consider the following two true parameters

(i) $\beta_1^* = -0.1, \beta_2^* = 0.1$,

(ii) $\beta_1^* = 0.1, \beta_2^* = -0.1$,

and investigate the performance of the following four cases of design matrix when $S = 8$.

(I) $\mathrm{vec}\{\boldsymbol{x}\} = (2.7, 3.0, 3.3, 3.6, 3.9, 4.2, 4.5, 4.8)'$.

(II) $\mathrm{vec}\{\boldsymbol{x}\} = (2.85, 3.05, 3.25, 3.45, 3.65, 3.85, 4.05, 4.25)'$.

(III) $\mathrm{vec}\{\boldsymbol{x}\} = (\log(2.7), \log(3.0), \log(3.3), \log(3.6), \log(3.9), \log(4.2), \log(4.5), \log(4.8))'$.

(IV) $\mathrm{vec}\{\boldsymbol{x}\} = (\log(2.85), \log(3.05), \log(3.25), \log(3.45), \log(3.65), \log(3.85), \log(4.05), \log(4.25))'$.

For each case, we consider the following two sample designs

(A) $n_1 = \cdots = n_8 = n_A$,

(B) $n_1 = \cdots = n_4 = n_B,\ n_5 = \cdots = n_8 = 2n_B$.

We investigate the performance for all combinations of two true parameters (i) and (ii), four design matrices (I), (II), (III), and (IV), and sample design (A), where $n_A = 10$, 20, and 30, and sample design (B), where $n_B =$10, 20, and 30. In the investigation, the number of repetitions is $J = 10^6$. Figure 1 shows the absolute values of index I in the cases of true parameters (i) and (ii), design matrices (I)–(IV) and significance level $\varepsilon = 0.01$, 0.05, and 0.10 when the model is given by the complementary log-log link , sample design is (A) and $n_A = 10, 20$, and 30. Figure 2 shows those for the model that is given by the probit link in the same situation as that in Figure 1. When models are given by complementary log-log link and probit link with sample design (B) and when the model is given by logit link with sample designs (A) and (B), results of simulation are almost the same as those in Figure 1 and Figure 2.

From the results of our simulation, we find that the performance of the Bartlett adjusted deviance statistic $\tilde{D}^B$ is better than that of the original deviance statistic D.

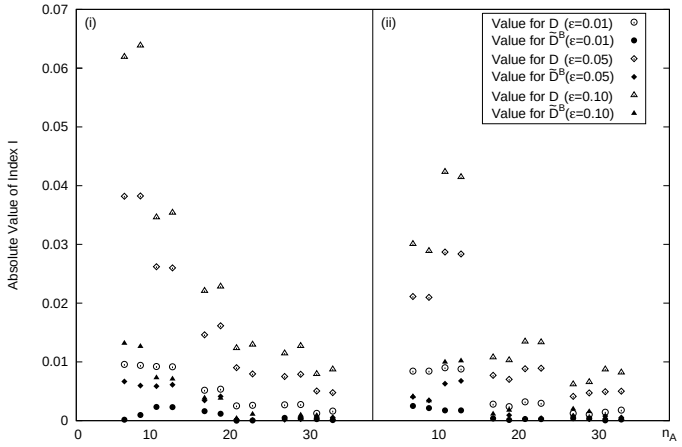

Figure 1: Absolute value of index I when the model is given by the complementary log-log link function for true parameters (i) and (ii) and sample design (A) with $n_A = 10$, 20, 30: $\circ$, $\Diamond$ and $\triangle$ are the values for D when $\varepsilon = 0.01$, 0.05 and 0.10, respectively, and $\bullet$,$\blacklozenge$ and $\blacktriangle$ are the values for $\tilde{D}^B$ when ε =0.01, 0.05 and 0.10, respectively. The 1st column is for design matrix (I), the 2nd column is for design matrix (II), the 3rd column is for design matrix (III), and the 4th column is for design matrix (IV).

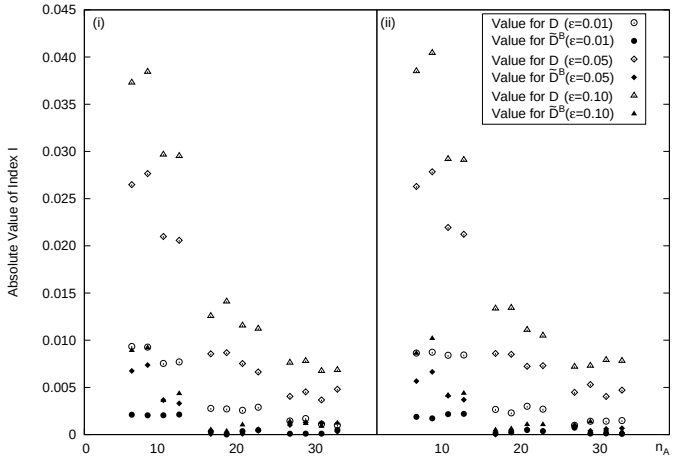

Figure 2: Absolute value of I when the model is given by the probit link function for true parameters (i) and (ii) and sample design (A) with $n_A = 10$, 20, 30: $\circ$, $\Diamond$ and $\triangle$ are the values for D when ε =0.01, 0.05 and 0.10, respectively, and $\bullet$,$\blacklozenge$ and $\blacktriangle$ are the values for $\tilde{D}^B$ when ε =0.01, 0.05 and 0.10, respectively. The 1st column is for design matrix (I), the 2nd column is for design matrix (II), the 3rd column is for design matrix (III), and the 4th column is for design matrix (IV).

Bibliography

[1] Barndorff-Nielsen, O. E. and Cox, D. R. (1984) *Bartlett adjustments to the likelihood ratio statistic and the distribution of maximum likelihood estimator.* J. R. Statist. Soc., B, **46**, 483–495.

[2] Barndorff-Nielsen, O. E. and Hall, P. (1988) *On the level-error after Bartlett adjustment of the likelihood ratio statistic.* Biometrika, **75**, 374–378.

[3] Lawley, D. N. (1956) *A general method for approximating to the distribution of the likelihood ratio criteria.* Biometrika, **43**, 295–303.

[4] Menéndez, M. L., Pardo, J. A., Pardo, L. and Pardo, M. C. (1997) *Asymptotic approximations for the distributions of the (h, ϕ)-divergence goodness-of-fit statistics: application to Renyi's statistic.* Kybernetes, **26**(4), 442–452.

[5] Nelder, J. A. and Wedderburn, R. W. M. (1972) *Generalized linear models.* J. R. Statist. Soc. A, **135**, 370–384.

[6] Read, T. R. C. (1984) *Closer asymptotic approximations for the distributions of the power divergence goodness-of-fit statistics.* Ann. Inst. Statist. Math., **36**, 59–69.

[7] Read, T. R. C. (1984) *Small-sample comparisons for the power divergence goodness-of-fit statistics.* J. Am. Statist. Assoc., **79**, 929–935.

[8] Sekiya, Y. and Taneichi, N. (2004) *Improvement of approximations for the distributions of multinomial goodness-of-fit statistics under nonlocal alternatives.* J. Multivariate Anal., **91**, 199–223.

[9] Siotani, M. and Fujikoshi Y. (1984) *Asymptotic approximations for the distributions of multinomial goodness-of-fit statistics.* Hiroshima Math. J., **14**, 115–124.

[10] Taneichi, N., Sekiya, Y. and Suzukawa, A. (2001) *An asymptotic approximation for the distribution of ϕ-divergence multinomial goodness-of-fit statistic under local alternatives.* J. Japan Statist. Soc., **31**(2), 207–224.

[11] Taneichi, N., Sekiya, Y. and Suzukawa, A. (2002) *Asymptotic approximations for the distributions of the multinomial goodness-of-fit statistics under local alternatives.* J. Multivariate Anal., **81**, 335–359.

[12] Taneichi, N. and Sekiya, Y. (2007) *Improved transformed statistics for the test of independence in $r \times s$ contingency tables.* J. Multivariate Anal., **98**, 1630–1657.

[13] Taneichi, N., Sekiya, Y. and Toyama, J. (2014) *Transformed goodness-of-fit statistics for a generalized linear model of binary data.* J. Multivariate Anal., **123**, 311–329.

[14] Yarnold, J. K. (1972) *Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set.* Ann. Math. Statist., **43**, 1566–1580.

Performance of acceleration of ALS algorithm in nonlinear PCA

Yuichi Mori, *Okayama University of Science*, mori@soci.ous.ac.jp
Masahiro Kuroda, *Okayama University of Science*, kuroda@soci.ous.ac.jp
Masaya Iizuka, *Okayama University*, iizuka@okayama-u.ac.jp
Michio Sakakihara, *Okayama University of Science*, sakaki@mis.ous.ac.jp

Abstract. Nonlinear principal components analysis with optimal scaling (NLPCA-OS) is useful for analyzing mixed measurement level data. The algorithm in NLPCA-OS is based on the alternating least squares (ALS) algorithm, where optimal transformation and low-rank matrix approximation are alternated until convergence. We have proposed an accelerated ALS algorithm using the vector ε algorithm (vε-ALS) which increases the speed of convergence, and have observed that computational costs by vε-ALS are less expensive than those by ordinary ALS in small examples in which all variables are categorical. In this paper, we try to evaluate the performance of proposed vε-ALS by simulation, in which NLPCA with vε-ALS is applied to several simulated datasets which have large numbers of variables with a variety of mixing rates of numerical and categorical variables. The simulation study indicates that the performance of approximation by vε-ALS is improved for all simulated datasets and that the larger the number of categorical variables is and the higher the mixing rate is, the more the vε-ALS reduces the computational costs.

Keywords. Vector ε algorithm, Acceleration of convergence, Alternating least squares, Mixed measurement level data, Simulation study.

1 Introduction

Nonlinear principal components analysis with optimal scaling (NLPCA-OS) is useful for analyzing mixed measurement level (nominal, ordinal and numerical) data. The algorithm in NLPCA-OS is based on the alternating least squares (ALS) algorithm, where optimal transformation and low-rank matrix approximation are alternated until convergence, that is, the algorithm alternates between optimal scaling for quantifying nominal and ordinal data and ordinary PCA for the optimally scaled data. PRINCIPALS [6] and PRINCALS [1] are the typical ALS algorithms for NLPCA.

Kuroda et al. [2] have proposed an accelerated ALS algorithm for NLPCA using the vector ε (vε) algorithm of Wynn [7] which increases the speed of convergence. We have applied the method to some numerical examples (e.g., [2] and [3]) and have proposed some more accelerated methods (e.g., two-step algorithm in [5] and re-starting method in [4]), and observed that computational costs of NLPCA with the vε alternating least squares (vε-ALS) are less expensive than those of NLPCA with ordinary ALS.

In the previous studies, we applied the proposed methods to datasets with small number of variables (the number of variables is 20 at most) and all datasets we used consist of only categorical (nominal) variables but not a mixture of numerical and categorical ones. In this paper, we try to evaluate the performance of vε-ALS in further detail to clarify how well the algorithm performs for large data and mixed measurement level data. To do this, we conduct some simulations in which NLPCA with vε-ALS is applied to several artificial datasets which have large numbers of variables with a variety of mixing rates of numerical and categorical variables.

We give an overview of NLPCA-OS and its acceleration by vε-ALS in Section 2 and illustrate numerical experiments on sixteen different types of datasets generated artificially (four different sizes of datasets with four different mixing rates of categorical variables) in Section 3. We discuss the performance of NLPCA with vε-ALS in Section 4.

2 Nonlinear PCA and its acceleration by vector ε ALS

Let $\mathbf{X} = (\mathbf{X}_1\ \mathbf{X}_2\ \cdots\ \mathbf{X}_p)$ be an $n \times p$ standardized matrix of observations on n objects and p numerical variables. PCA postulates that $\mathbf{X}$ is approximated by the bilinear form

$$\hat{\mathbf{X}} = \mathbf{Z}\mathbf{A}^\top, \tag{1}$$

where $\mathbf{Z}$ is an $n \times r$ matrix of n component scores on r $(1 \leq r \leq p)$ components and $\mathbf{A}$ is a $p \times r$ matrix of p component loadings on r components.

In order to handle any categorical data or mixture of numerical and categorical data, NLPCA requires the optimal scaled data $\mathbf{X}^*$, in addition to estimating $\mathbf{Z}$ and $\mathbf{A}$, in which categorical variables in $\mathbf{X}$ are optimally scaled and satisfies restrictions

$$\mathbf{X}^{*\top}\mathbf{1}_n = \mathbf{0}_p \qquad \text{and} \qquad \operatorname{diag}\left[\frac{\mathbf{X}^{*\top}\mathbf{X}^*}{n}\right] = \mathbf{I}_p, \tag{2}$$

where $\mathbf{1}_n$ and $\mathbf{0}_p$ are vectors of ones and zeros of length n and p, respectively. Thus NLPCA is a least square problem to estimate optimal scaling parameter $\mathbf{X}^*$ and model parameters $\mathbf{Z}$ and $\mathbf{A}$ simultaneously, which minimize

$$\theta = \operatorname{tr}(\mathbf{X}^* - \hat{\mathbf{X}})^\top(\mathbf{X}^* - \hat{\mathbf{X}}) = \operatorname{tr}(\mathbf{X}^* - \mathbf{Z}\mathbf{A}^\top)^\top(\mathbf{X}^* - \mathbf{Z}\mathbf{A}^\top). \tag{3}$$

The ALS algorithm can be used in NLPCA-OS. It alternates between ordinary PCA and optimal scaling, and minimizes θ^* in (3) under restriction (2). For given initial data $\mathbf{X}^{*(0)}$, the procedure based on PRINCIPALS [6] is to iterate the following two steps until convergence:

Step 1 *Model parameter estimation step*: Obtain $\mathbf{A}^{(t)}$ by solving an eigenvalue problem

$$\left[\frac{\mathbf{X}^{*(t)\top}\mathbf{X}^{*(t)}}{n}\right]\mathbf{A} = \mathbf{A}\mathbf{D}_r, \tag{4}$$

where $\mathbf{A}^\top\mathbf{A} = \mathbf{I}_r$ and $\mathbf{D}_r$ is an $r \times r$ diagonal matrix of eigenvalues. Compute $\mathbf{Z}^{(t)}$ from $\mathbf{Z}^{(t)} = \mathbf{X}^{*(t)}\mathbf{A}^{(t)}$.

Step 2 *Optimal scaling step*: Calculate $\hat{\mathbf{X}}^{(t+1)} = \mathbf{Z}^{(t)}\mathbf{A}^{(t)\top}$ from Equation (1). Find $\mathbf{X}^{*(t+1)}$ such that

$$\mathbf{X}^{*(t+1)} = \arg\min_{\mathbf{X}^*} \mathrm{tr}(\mathbf{X}^* - \hat{\mathbf{X}}^{(t+1)})^\top(\mathbf{X}^* - \hat{\mathbf{X}}^{(t+1)})$$

for fixed $\hat{\mathbf{X}}^{(t+1)}$ under measurement restrictions on each of the variables. Since $\mathbf{X}^{*(t+1)}$ is obtained by separately estimating $\mathbf{X}_j^*$ for each j $(j = 1, \ldots, p)$, scale $\mathbf{X}^{*(t+1)}$ by columnwise centering and normalizing. Re-compute $\mathbf{X}_j^{(t+1)}$ by an additional transformation to keep the monotonicity restriction for ordinal variables and skip this computation for numerical variables.

The superscript (t) indicates the t-th iteration. From the above iteration, we obtain a convergence sequence $\{\mathbf{X}^{*(t)}\}_{t\geq 0}$. Although the true limit points are theoretically obtained at $t = \infty$, the solutions by NLPCA-OS are parameters based on $\mathbf{X}^{*(t)}$ obtained when the iteration converges by the criterion θ.

Here we accelerate the above NLPCA with ALS using the vε algorithm of Wynn [7] which is very effective to accelerate the slow convergence of a linearly convergent vector sequence. Let $\{\dot{\mathbf{X}}^{(t)}\}_{t\geq 0} = \{\dot{\mathbf{X}}^{(0)}, \dot{\mathbf{X}}^{(1)}, \dot{\mathbf{X}}^{(2)}, \ldots\}$ be the accelerated sequence of $\{\mathbf{X}^{(t)}\}_{t\geq 0}$. We define the inverse of vector $\mathbf{X}$ by $[\mathbf{X}]^{-1} = \mathbf{X}/\langle \mathbf{X}, \mathbf{X}\rangle$, where $\langle\cdot,\cdot\rangle$ is the inner product of vectors. Then, the vε algorithm generates $\{\dot{\mathbf{X}}^{(t)}\}_{t\geq 0}$ by using

$$\mathrm{vec}\dot{\mathbf{X}}^{*(t-1)} = \mathrm{vec}\mathbf{X}^{*(t)} + \left[\left[\mathrm{vec}(\mathbf{X}^{*(t-1)} - \mathbf{X}^{*(t)})\right]^{-1} + \left[\mathrm{vec}(\mathbf{X}^{*(t+1)} - \mathbf{X}^{*(t)})\right]^{-1}\right]^{-1}, \quad (5)$$

where $\mathrm{vec}\mathbf{X}^* = (\mathbf{X}_1^{*\top}\ \mathbf{X}_2^{*\top}\cdots\ \mathbf{X}_p^{*\top})^\top$. It is expected that this new sequence $\{\dot{\mathbf{X}}^{(t)}\}_{t\geq 0}$ converges to a limit point $\mathbf{X}^{(\infty)}$ of $\{\mathbf{X}^{(t)}\}_{t\geq 0}$ faster than $\{\mathbf{X}^{(t)}\}_{t\geq 0}$. Our previous numerical experiments (e.g., [2], [3], [4] and [5]) demonstrated that its speed of convergence is significantly higher than that of the ordinary ALS algorithm .

The procedure to accelerate the ALS algorithm in PRINCIPALS described above iterates the following two steps:

Step 1 *PRINCIPALS step*: Compute model parameters $\mathbf{A}^{(t)}$ and $\mathbf{Z}^{(t)}$ and determine optimal scaling parameter $\mathbf{X}^{*(t+1)}$.

Step 2 *Acceleration step*: Calculate $\dot{\mathbf{X}}^{*(t-1)}$ using $\{\mathbf{X}^{*(t-1)}, \mathbf{X}^{*(t)}, \mathbf{X}^{*(t+1)}\}$ from Equation (5) and check the convergence by

$$\left\|\mathrm{vec}(\dot{\mathbf{X}}^{*(t-1)} - \dot{\mathbf{X}}^{*(t-2)})\right\|^2 < \delta, \quad (6)$$

where δ is a desired accuracy.

3 Numerical experiments

We examine the performance of the proposed acceleration for PRINCIPALS using vε-ALS by employing simulated data generated as below, and demonstrate the advantage of vε accelerated PRINCIPALS (vε-ALS in NLPCA) over ordinary PRINCIPALS (ordinary ALS in NLPCA) in terms of the number of iterations and CPU time (in second) required for convergence.

Data generation

Since we are interested in the performance of the proposed vε accelerated PRINCIPALS when it is performed for data which have large numbers of variables and include both numerical and categorical variables, we generate random data matrices with the following four types of number of observations (n) and variables (p): (A) n=100, p=20, (B) n=100, p=50, (C) n=500, p=100 and (D) n=200, p=150. All categorical variables have 10 levels (10 categories). To each of the datasets we further set four kinds of mixing rates of categorical variables: 0.25, 0.50, 0.75 and 1.00. The mixing rate 0.25 means that 25% of variables (rounded) are processed as categorical data and 75% as numerical data, and so on. The number of components (r) is two for all datasets.

We apply ordinary PRINCIPALS and vε accelerated PRINCIPALS to the above datasets. Consequently we execute thirty-two types of experiments ({four data types} $\times$ {four mixing rates of categorical variables} $\times$ {ALS and vε-ALS}).

Results of experiments

For all experiments, δ for convergence is set to 10^{-12}, and PRINCIPALS terminates when $|\theta^{(t+1)} - \theta^{(t)}| < 10^{-12}$, where $\theta^{(t)}$ is the t-th update of θ calculated from Equation (3). Each algorithm also stops when the number of iterations exceeds 10,000. The procedure is replicated 100 times. All computations are performed with the statistical package R executing on Intel Core i5 3.3 GHz with 4 GB RAM. CPU times taken are measured by the function `proc.time`.

Table 1 is summary statistics of the numbers of iterations from thirty-two 100 simulations. Figure 1 shows the same thirty-two simulations in boxplots. The first graph from the left in Figure 1 displays boxplots of eight simulations ({four mixing rates of categorical variables} $\times$ {ALS and vε-ALS}) for data type (A), the second for (B), the third for (C) and the last for (D). The CPU times are similarly summarized in Table 2 and Figure 2.

From these tables and figures, as the data size is increasing, a greater number of iterations and more CPU time are required, but vε-ALS greatly reduces the number of iterations and CPU time. In case of 0.25 mixing rate, for example, ordinary ALS needs 178 iterations with 1.9 seconds for dataset (A) (p=20) but 639 iterations and 140 seconds for dataset (D) (p=150). On the other hand, vε-ALS needs 49 iterations and 0.7 seconds for (A) but 188 iterations and 45 seconds for (D). We can observe similar results as the mixing rate of categorical variables increases. The increase of the number of categorical variables requires computational cost and the acceleration by vε-ALS is therefore effective. In case of dataset (D) (p=150), for example, ordinary ALS needs 639 iterations with 140 seconds for 0.25 mixing rate but 1697 iterations and 805 seconds for 1.00 mixing rate. On the other hand, vε-ALS needs 188 iterations and 45 seconds for 0.25 mixing rate but 644 iterations and 314 seconds for 1.00 mixing rate.

It can be observed that the vε-ALS converges almost 3 times faster than ordinary ALS in all simulations. The tables also show the average speed-up rates in [SpeedUp] row of each data type, which is computed by dividing the number of iterations (CPU time) required for ordinary ALS divided by the number of iterations (CPU time) required for vε-ALS. In Figure 3, we illustrate the speed-up rates of 100 simulations only for dataset (D) in boxplot. The similar boxplots of the speed-up rate can be obtained for other datasets. Regardless of the data size and the mixing rate of categorical variables, vε-ALS is smaller 2.62 – 3.61 times of iterations and shorter 2.50 – 3.13 times of CPU time than those of ordinary ALS, although the speed-up rates slightly decrease according to the increase of mixing rate.

		0.25		0.50		0.75		1.00	
Data type	Stats	ALS	vε-ALS	ALS	vε-ALS	ALS	vε-ALS	ALS	vε-ALS
(A) n=100	Min.	51	17	89	29	125	40	180	65
p=20	1st Qu.	88.5	27	150.8	47.75	225.8	72.75	332.5	103.8
	Median	123.5	35	210	60	323	96	474.5	154.5
	Mean	178.2	49.33	302	93.81	448	133.69	605.7	193.8
	3rd Qu.	182.2	50.5	308.8	86	532.5	147.5	737.5	241
	Max.	2687	623	2464	1344	2752	935	3578	820
	[SpeedUp]		[3.61]		[3.22]		[3.35]		[3.13]
(B) n=100	Min.	85	26	170	58	254	101	296	94
p=50	1st Qu.	150	48.75	290.5	97.5	476.8	147	586.5	194
	Median	218	68.5	435	133	674	224.5	798	285
	Mean	294.4	90.22	505.7	170.4	780.5	266.7	1032	372.3
	3rd Qu.	334.5	95.25	580.8	208.5	922.5	337.5	1200.5	410
	Max.	1799	409	1777	1267	3717	783	4894	1959
	[SpeedUp]		[3.26]		[2.97]		[2.93]		[2.77]
(C) n=500	Min.	83	27	150	53	228	85	307	105
p=100	1st Qu.	181.8	61	308.8	100.2	468.5	152	615.8	227
	Median	261	83	414.5	135	593	207.5	792.5	304.5
	Mean	345.4	103.8	548.2	179.5	697.3	245.8	1147.2	437.6
	3rd Qu.	357.5	111	672.8	216.8	875.2	297.5	1178.5	418.8
	Max.	2187	499	3474	793	1708	1051	10000	2752
	[SpeedUp]		[3.33]		[3.05]		[2.84]		[2.62]
(D) n=200	Min.	190	63	378	116	418	179	501	202
p=150	1st Qu.	341	107.8	619.8	224	898	329.8	1058	397
	Median	468.5	143.5	867	329.5	1194	431	1436	569.5
	Mean	639	187.7	1090.3	388.1	1411	527.3	1697	644.2
	3rd Qu.	670.8	212	1297	448.2	1638	601.8	1987	785.5
	Max.	8329	1263	3667	1534	4406	2354	7416	2192
	[SpeedUp]		[3.40]		[2.81]		[2.68]		[2.63]

Table 1: Summary of statistics of the numbers of iterations of ordinary ALS and vε accelerated ALS for four data types and four mixing rates of categorical variables.

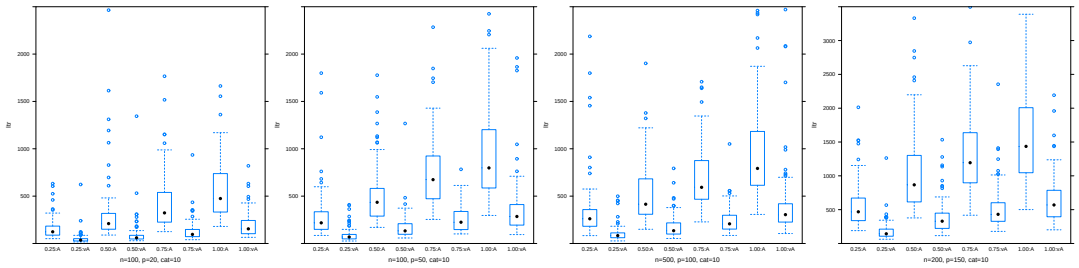

Figure 1: Boxplots of the number of iterations for data type (A) to (D) (from left to right in order).

4 Concluding remarks

In this paper, we examined the performance of the vε-ALS algorithm which accelerates the convergence of the sequence generated from ordinary ALS. To do this, we applied ordinary ALS and vε-ALS to several simulated datasets generated from four different data sizes and four different mixing rates of categorical variables. The numerical experiments for comparing the number of iterations and CPU time by ordinary ALS and vε-ALS demonstrated that the larger the number of categorical variables is and the higher the mixing rate is, the more the vε-ALS reduces the computational costs. They also indicated that the performance of approximation

		0.25		0.50		0.75		1.00	
Data type	Stats	ALS	vε-ALS	ALS	vε-ALS	ALS	vε-ALS	ALS	vε-ALS
(A) n=100	Min.	0.66	0.34	1.21	0.55	1.96	0.77	3.07	1.26
p=20	1st Qu.	1.018	0.4475	1.992	0.7775	3.37	1.258	5.593	1.948
	Median	1.35	0.53	2.695	0.92	4.745	1.6	7.855	2.795
	Mean	1.898	0.6744	3.775	1.3441	6.489	2.141	10.015	3.448
	3rd Qu.	1.925	0.685	3.86	1.2525	7.697	2.345	12.133	4.258
	Max.	26.15	6.43	29.44	16.8	38.88	13.83	58.15	13.93
	[SpeedUp]		[2.81]		[2.81]		[3.03]		[2.90]
(B) n=100	Min.	2.9	1.22	7.09	2.8	13	5.55	17.75	6.05
p=50	1st Qu.	4.76	1.907	11.81	4.402	23.9	7.875	34.78	12.04
	Median	6.675	2.465	17.42	5.785	33.77	11.675	47.29	17.36
	Mean	8.887	3.104	20.22	7.246	39.02	13.812	60.95	22.53
	3rd Qu.	10.053	3.237	23.17	8.797	46.09	17.288	71.05	24.76
	Max.	52.24	12.5	70.13	50.84	184.89	39.74	287.42	116.24
	[SpeedUp]		[2.86]		[2.79]		[2.83]		[2.71]
(C) n=500	Min.	18.68	9.67	39.34	17.85	70.56	30.75	114.2	43.34
p=100	1st Qu.	36.31	15.82	77.09	29.22	142.03	51.48	224	87.26
	Median	50.32	20.01	101.38	37.49	179.84	67.65	285.1	115.35
	Mean	65.14	23.84	132.91	48.49	209.99	79.31	410.4	164.23
	3rd Qu.	66.58	25.62	163.34	57.05	262.47	94.36	420.7	157.34
	Max.	392.5	96.43	825.62	197.59	507.83	323.42	3489.6	991.63
	[SpeedUp]		[2.73]		[2.74]		[2.65]		[2.50]
(D) n=200	Min.	43.41	16.78	115.4	38.93	163.9	74.04	238.9	101
p=150	1st Qu.	75.57	26.95	188	72.22	348.6	133.48	501.9	195.4
	Median	103.07	35.03	261.6	104.67	463.4	173.78	681	278.1
	Mean	139.57	44.66	329	122.92	545.9	211.73	805.1	313.9
	3rd Qu.	146.21	49.99	392.7	141.39	630.3	241.1	942.5	382.1
	Max.	1788.1	284.06	1096.7	476.74	1714.3	932.91	3508.4	1062.2
	[SpeedUp]		[3.13]		[2.68]		[2.58]		[2.56]

Table 2: Summary of statistics of CPU times of ordinary ALS and vε accelerated ALS for four data types and four mixing rates of categorical variables.

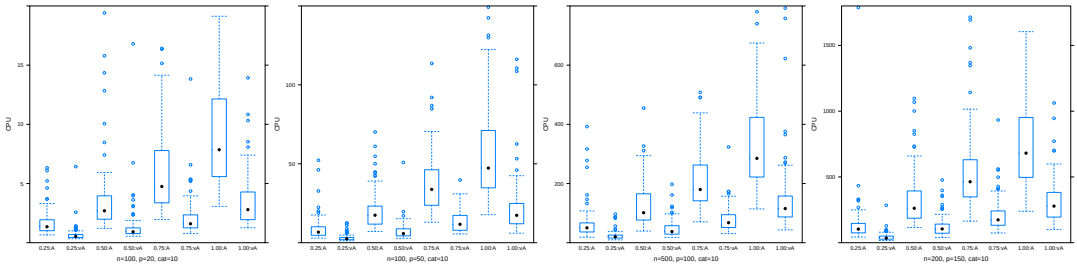

Figure 2: Boxplots of CPU time for data type (A) to (D) (from left to right in order).

by vε-ALS is improved about 3 times of ordinary ALS for any number of categorical variables in data.

For future problems, we have to investigate how much the proposed acceleration improves computational efficiency when it is applied to more complex situations; such as variable selection problem. Since we are developing faster algorithms (e.g., re-starting ALS in [4]), we are trying to evaluate the performances of such algorithms in detail. Furthermore, there exist many other ALS types of algorithms, so we are attempting to speed up the convergence of their ALS algorithms by incorporating the proposed acceleration.

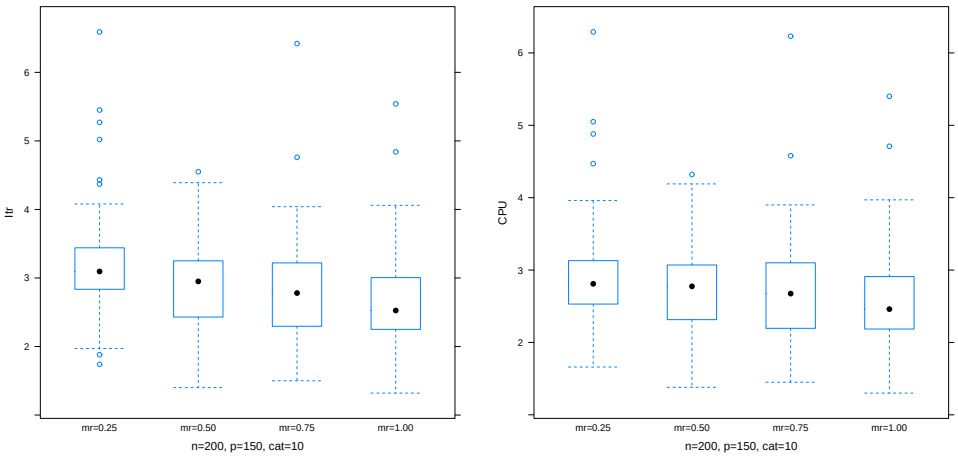

Figure 3: Boxplots of the speed-up rates of 100 simulations for data type (D) (Left: the number of iterations, Right: CPU time).

Acknowledgement

This work is supported by JSPS KAKENHI Grant Numbers 24500353, 26330052.

Bibliography

[1] Gifi, A. (1990). *Nonlinear multivariate analysis.* John Wiley & Sons, Ltd.,

[2] Kuroda, M., Mori, Y., Iizuka, M. and Sakakihara, M. (2011). *Accelerating the convergence of the EM algorithm using the vector epsilon algorithm.* Computational Statistics and Data Analysis, **55**, 143–153.

[3] Kuroda, M., Mori, Y., Iizuka, M. and Sakakihara, M. (2012). *Acceleration of convergence of the alternating least squares algorithm for nonlinear principal components analysis.* In Principal Component Analysis (Sanguansat, P. (Ed.)), InTech Publications, 129-144.

[4] Kuroda, M., Mori, Y., Izuka, M. and Sakakihara, M. (2013). *Accelerating and re-starting the alternating least squares algorithm for non-linear principal components analysis.* Proceedings of the 59th World Statistics Congress, 5426-5431.

[5] Kuroda, M., Sakakihara, M.. Mori, Y. and Izuka, M. (2012). *Two-stage acceleration for non-linear PCA.* Proceedings of COMPSTAT 2012, 461-471.

[6] Young, F.W., Takane, Y., and de Leeuw, J. (1978). *Principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features.* Psychometrika, **43**, 279–281.

[7] Wynn, P. (1962). *Acceleration techniques for iterated vector and matrix problems.* Mathematics of Computation, **16**, 301–322.