L_1-regression for multivariate clustered data

Jaakko Nevalainen and Denis Larocque

Abstract In this chapter, we are considering L_1-type estimation for multivariate clustered data. Although valid, using the direct L_1 estimation of the regression coefficients in the clustered data setting is likely to lack efficiency since it does not use the intracluster correlation structure. A transformation-retransformation method is proposed to overcome this problem. This method first transforms the original model in an attempt to eliminate the intracluster correlation. Secondly, the L_1 estimates are obtained with the transformed data, which are then transformed back to the original scale. One particular implementation of this method is investigated in a simulation study which shows that it is more efficient than using the direct L_1 estimators.

1 Introduction

In this chapter, we consider L_1-regression models for clustered data with a multivariate response. For a univariate response, Jung and Ying (2003) proposed a generalization of the Wilcoxon-Mann-Whitney statistic for analyzing repeated measurements data. The estimating function is based on the unweighed ranks of the residuals, which is equivalent to the method proposed by Jurečková (1969, 1971) for independent observations. In order to recover some of the information present in the clustering structure, Wang and Zhu (2006) generalized this approach by partitioning the ranks into between- and within-subject ranks. Two estimators are obtained and then combined in an optimal way. However, to get the combined estimator, an estimation of the covariance matrix of the two estimating functions is required. To

Jaakko Nevalainen
School of Health Sciences, University of Tampere, 33014 Tampere, Finland, e-mail: jaakko.nevalainen@uta.fi

Denis Larocque
Department of Decision Sciences, HEC Montréal, 3000 chemin de la Côte-Sainte-Catherine, Montréal, Canada e-mail: denis.larocque@hec.ca

achieve this, a resampling method is used in Wang and Zhu (2006). Fu, Wang and Bai (2010) proposed a smoothing method to avoid this computationally intensive approach. In another attempt to use the clustering structure, Wang and Zhao (2008) proposed a weighted version of the loss function, where the weights are functions of the cluster sizes. Their approach is related to the one proposed in Datta and Satten (2005) for the two-sample problem. Fu and Wang (2012) argue that the Wang and Zhao (2008) approach performs well for cluster-level covariates but not necessarily for within-cluster covariates. They derive a new optimal rank-based estimating functions in terms of asymptotic variance of regression parameter estimators. Finally, Kloke, McKean and Rashid (2009) study R-estimators of the fixed effects in an experiment done over clusters, blocks, groups, or subjects, including for example, repeated measure designs, split plot designs, multicenter clinical trials, randomized block designs, and two-stage cluster samples.

All the articles above are aimed at the univariate response case. Nonparametric methods for multivariate data, and especially, methods based on spatial signs and ranks have been developed extensively in the last 20 years (Oja, 2010). They are also available for the user through the R package `MNM` (Nordhausen and Oja, 2011). Moreover, specialized methods for multivariate responses and clustered data have also been developed; see Nevalainen, Larocque, Oja and Pörsti (2010) and the references therein. However, for clustered data, the available methods are limited so far to the one, two and several samples cases. In this chapter we propose an L_1-type (spatial sign) estimation method for a regression setting with multivariate clustered data and investigate it in a simulation study.

2 A multivariate multiple linear regression model for clustered data

Let $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^\top$ be a sample of p-variate ($p > 1$) random response vectors with sample size n. The data are assumed to be clustered with a total of d clusters. The cluster memberships are given by the $n \times d$ matrix $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\top$:

$$(\mathbf{Z})_{ij} = \begin{cases} 1, \text{ if the } i\text{th observation is from cluster } j; \\ 0, \text{ otherwise.} \end{cases}$$

It is useful to note that

$$\left(\mathbf{Z}\mathbf{Z}^\top\right)_{ij} = \begin{cases} 1, \text{ if the } i\text{th and the } j\text{th observation are from the same cluster;} \\ 0, \text{ otherwise,} \end{cases}$$

and that $\mathbf{Z}^\top\mathbf{Z}$ is a $d \times d$ diagonal matrix with the cluster sizes on the diagonal, say, $m_1, \ldots, m_d$. We also write $\mathbf{1}_n$ for a column n-vector of ones, $\mathrm{vec}(\mathbf{Y})$ for the vector obtained by stacking the columns of $\mathbf{Y}$, and $\otimes$ for the Kronecker product.

Consider the multivariate multiple linear regression model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{E},$$

where

- $\mathbf{X}$ is an $n \times q$ design matrix for explanatory variables with the first column consisting of 1's;
- β is the $q \times p$ matrix of regression coefficients;
- $\mathbf{E} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is an $n \times p$ matrix of random errors,

stating that the responses are linearly related to the explanatory variables, and $\mathbf{E}$ is a matrix of random errors with

$$\mathbf{COV}\left(\mathrm{vec}(\mathbf{E}^\top)\right) = \Omega \otimes \Sigma.$$

Here $\Sigma = E(\varepsilon_i \varepsilon_i^\top)$ and $\Omega = \{\rho_{ij}\}$ is a matrix consisting of intracluster correlations, with unit entries on the diagonal. We thus assume that $E(\varepsilon_i \varepsilon_j^\top) = \rho_{ij}\Sigma$, where $\rho_{ij} \neq 0$ if $(\mathbf{Z}\mathbf{Z}^\top)_{ij} = 1$, and $\rho_{ij} = 0$, otherwise.

Typically, the regression coefficients of the model are estimated by ordinary least squares based on a minimization of an L_2 criterion function, or by maximum likelihood relying on a multivariate normality assumption on the random errors. These two solutions are in general not the same with clustered data. Under heavy-tailed error distributions these estimates are inefficient, and may be vulnerable to outliers. In such circumstances, a fit based on an L_1-criterion function may be preferable.

3 Estimation based on an L_1-criterion function

The goal is to estimate the unknown β matrix of regression coefficients by minimizing an L_1 norm

$$D_n(\beta) = \sum_{i=1}^{n} \left(|\mathbf{y}_i - \beta^\top \mathbf{x}_i| - |\mathbf{y}_i| \right).$$

This leads to a multivariate *least absolute deviation (LAD) estimate* of β. If the residuals lie in a genuinely p-dimensional space, the resulting estimate $\hat{\beta}$ solves the estimating equation

$$\mathbf{U}(\hat{\beta})^\top \mathbf{X} = \mathbf{0},$$

where

$$\mathbf{U}_i(\beta) = (\mathbf{y}_i - \beta^\top \mathbf{x}_i)/|\mathbf{y}_i - \beta^\top \mathbf{x}_i|,\ i = 1, ..., n$$

is the spatial sign of the residual at β and $\mathbf{U}(\beta) = (\mathbf{U}_1(\beta), ..., \mathbf{U}_n(\beta))^\top$ is the corresponding matrix of residual spatial signs (Oja, 2010). These L_1 estimates of regression coefficients are quite natural and not difficult to compute (Nordhausen and Oja, 2011).

4 Transformation-retransformation L_1 regression estimates

It is possible to use the L_1 norm to directly estimate the parameters also in the clustered data case. Compared to a setting with i.i.d. random errors, the limiting distribution would only require a correction in the variance terms. This permits a valid analysis. However, the estimate as such suffers from one important shortcoming: it makes no use of the underlying and known cluster structure. A reasonable concern is that it may be an inefficient approach; recall that the (optimal) maximum likelihood estimates for linear models in the univariate normal case use the covariance structure as an essential ingredient.

Suppose first that the covariance matrix was known and again has the general form

$$\mathbf{COV}(\mathrm{vec}(\mathbf{E}^\top)) = \Omega \otimes \Sigma.$$

For example, the "compound symmetry" covariance structure,

$$\mathbf{COV}(\mathrm{vec}(\mathbf{E}^\top)) = \mathbf{I}_n \otimes \Sigma + (\mathbf{Z}\mathbf{Z}^\top - \mathbf{I}_n) \otimes \rho\Sigma = (\mathbf{I}_n + \rho(\mathbf{Z}\mathbf{Z}^\top - \mathbf{I}_n)) \otimes \Sigma$$

falls into this class of structures.

Given a pre-specified covariance structure $\Omega \otimes \Sigma$, the original estimation problem can be rewritten as

$$\begin{aligned}
\mathbf{Y} &\to \mathbf{Y}_0 = \Omega^{-1/2}\mathbf{Y}\Sigma^{-1/2} \\
\mathbf{X} &\to \mathbf{X}_0 = \Omega^{-1/2}\mathbf{X} \\
\beta &\to \beta_0 = \beta\Sigma^{-1/2} \\
\mathbf{E} &\to \mathbf{E}_0 = \Omega^{-1/2}\mathbf{E}\Sigma^{-1/2}
\end{aligned}$$

This postulates a new regression model $\mathbf{Y}_0 = \mathbf{X}_0\beta_0 + \mathbf{E}_0$ in which, if the covariance structure is correctly specified, the random errors are uncorrelated. Multiplication from the left attempts to eliminate the intracluster correlation, and multiplication from the right is aimed to standardize the marginal distributions. For the transformed data set on the right-hand side, it is reasonable to conduct ordinary L_1 estimation of regression coefficients as before. Therefore, the estimating equation is

$$\mathbf{U}_0 \left(\beta\Sigma^{-1/2}\right)^\top \Omega^{-1/2}\mathbf{X} = \mathbf{0},$$

where $\mathbf{U}_0$ now consists of the spatial signs of the residuals on the transformed scale, i.e., the rows of $\Omega^{-1/2}(\mathbf{Y} - \mathbf{X}\beta)\Sigma^{-1/2}$. As a final step, the estimates of the regression parameters in the original model are obtained by back-transformation $\hat{\beta}_{trt} = \hat{\beta}\Sigma^{1/2}$.

This procedure has potential for improved efficiency. Similar idea of a working, user-specified correlation structure is met in the framework of generalized estimating equations (Diggle et al., 2002). In that setup, the estimates of the regression coefficients are consistent even if the working correlation structure is misspecified. Simulations will be used to investigate the merits of this method in the next section.

5 Simulation study

We next investigate the finite sample efficiency of the proposed method by simulation studies involving four competing approaches to the problem.

5.1 Practical implementation of the transformation–retransformation L_1 method

We are assuming that the working covariance structure is $E(\varepsilon_i \varepsilon_j^\top) = \rho_{ij}\Sigma$, where $\rho_{ij} = \rho$ if $(\mathbf{Z}\mathbf{Z}^\top)_{ij} = 1$ and $\rho_{ij} = 0$ otherwise. We thus assume a compound symmetry structure with the same correlation (ρ) for each response. In this respect the compound symmetry structure is a special case; other covariance structures would not in general have this property.

A particular implementation of the transformation–retransformation L_1 method for this setting is as follows. To estimate ρ, we take the average of the intracluster correlation estimate obtained from separate linear mixed models to each response. More precisely, let $\hat{\rho}_i$ be the estimated value of the intracluster correlation obtained from fitting a linear mixed model to the i^{th} response with a random intercept at the cluster level, and using all the covariates. Then $\hat{\rho} = (1/p)\sum_{i=1}^{p} \hat{\rho}_i$.

To estimate Σ, we use the sample covariance matrix of the residuals obtained from the same response–wise linear mixed models.

5.2 Design of the simulation

Data are generated according to the following multivariate linear mixed model with $p = 3$ or 7 responses and d clusters:

$$\mathbf{Y}_{ij} = X_{1ij}\mathbf{1}_p + X_{2ij}\mathbf{1}_p + X_{3ij}\mathbf{1}_p + \alpha_i + \varepsilon_{ij}, \quad i = 1,\ldots,d, \quad j = 1,\ldots,m_i.$$

In this model, $\mathbf{Y}_{ij}$ is the p-variate response vector for the j^{th} observation in cluster i, the α_i's are i.i.d. p-variate cluster effects (random intercept), and ε_{ij}'s are i.i.d. p-variate individual error terms. The three covariates (the X's) are i.i.d. from $N(0,1)$. The covariates, random intercepts and error terms are all independent. Hence the true matrices of regression coefficients are

$$\beta = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

for the $p = 3$ and 7 cases, respectively.

The cluster design consists in 5 clusters of size 2, 5 clusters of size 3, 5 clusters of size 4, 5 clusters of size 5 and 5 clusters of size 6, for a total of 25 clusters and 100 observations.

The α_i's and ε_{ij}'s are generated from either the normal or the t_3 distribution with mean vector $\mathbf{0}$. In all cases, the scale matrix of these distribution has the form $\rho\mathbf{I}_p$ for α_i and $(1-\rho)\mathbf{I}_p$ for ε_{ij}. We then let ρ vary between 0 to 0.95 by steps of 0.05.

Four estimation methods are compared.

1. Transformation–retransformation L_1 method assuming the compound symmetry structure with equal ρs. The estimation of the parameters are described in Section 5.1. This is the proposed method.
2. Random intercept linear mixed models fitted separately to each response.
3. Basic L_1 regression applied directly to the data, neglecting the intracluster correlation.
4. Transformation–retransformation L_1 method assuming the compound symmetry structure with equal ρs. But this time we use the true values of ρ. However, we still estimate Σ as explained in Section 5.1. This method is not feasible in practice because ρ will likely never be known. But we use it as a benchmark to investigate the effect of having to estimate ρ.

The number of simulation runs is 500 for each configuration. All computations are performed in `R` (R Core team, 2013). The linear mixed models are fitted with the `lme` function in the `nlme` package (Pinheiro et al., 2014). The L_1 regressions are performed with the `mv.l1lm` function in the `MNM` package (Nordhausen and Oja, 2011).

5.3 Results

For each estimation method r (=1,2,3,4), the performance criterion is

$$P_r = (1/500)\sum_{i=1}^{500} \mathrm{vec}(\hat{\beta}_{ri} - \beta)'\mathrm{vec}(\hat{\beta}_{ri} - \beta),$$

where $\hat{\beta}_{ri}$ is the estimation from this method for the i^{th} simulation run. The results are summarized in figures 1 to 4. In each plot, the $Y-$axis gives the efficiency of each method relative to the proposed method (method 1 in Section 5.2), as a function of ρ. More precisely, it is P_1/P_r for $r = 2,3,4$. Hence, the proposed method is more efficient than the other one when the relative efficiency is below 1 and less efficient when it is above 1.

The results are reported in Figure 1. The upper-left plot corresponds to the three-variate case where both the random intercepts and the errors are normally distributed. As expected, the linear mixed model (method 2) is a bit more efficient than the proposed method in this case, and the efficiency remains constant over the

range of ρ. We can also see clearly the effect of neglecting the intracluster correlation. The performance of the basic L_1 regression (method 3) dramatically worsens as ρ increases. We also see that using the true value of ρ (method 4) has the same performance as the proposed method. Hence, nothing is lost by having to estimate this parameter.

The upper-right plot corresponds to the same situation but with random intercepts and errors distributed as t_3. This time, the proposed method is more efficient than the linear mixed model. Again, this was expected since L_1 based methods are more efficient for heavier–tailed distributions. However, the difference between the two is more pronounced for smaller values of ρ. The performance of the basic L_1 regression worsens as ρ increases and having to estimate ρ does not hurt the performance.

The lower plots present the corresponding results for the seven-variate cases. For the normal case (lower-left), the same patterns as for the three-variate case occur, except that the difference between the proposed method and the linear mixed model is even smaller. For the t_3 case, (lower-right), the situation is similar and the same patterns as for the three-variate case also occur. Hence, from these limited results, the proposed transformation-retransformation L_1 method is clearly preferable to the direct L_1 method.

5.4 Additional simulations under unequal ρs scenarios

The results presented previously showed that the transformation-retransformation L_1 method is more efficient than the direct L_1 method. However, the particular implementation of the method used in the simulation assumes a compound symmetry structure with the same correlation for each response. Hence, it is natural to ask if it still performs well when this is not true. To investigate this, we used the same scenarios as before, except that we allowed the intracluster correlations to vary for each response. More precisely, The α_i's and ε_{ij}'s are still generated from either the normal or the t_3 distribution with mean vector $\mathbf{0}$. But this time, the scale matrix of these distribution has the form $\text{diag}(\rho_1,\ldots,\rho_p)$ for α_i and $\text{diag}((1-\rho_1),\ldots,(1-\rho_p))$ for ε_{ij}. When $p=3$, we fix $\rho_2=0.5$, $\rho_1=\rho$, $\rho_3=1-\rho$, and we let ρ vary between 0.05 to 0.5 by steps of 0.05. When $p=7$, we fix $\rho_4=0.5$, $\rho_1=\rho_2=\rho_3=\rho$, $\rho_5=\rho_6=\rho_7=1-\rho$, and we let ρ vary between 0.05 to 0.5 by steps of 0.05. Hence, the greatest variance among the ρs occur when $\rho=0$. When $\rho=0.5$, we fall back to the equal ρs case. Note that method 4 is not applicable with these scenarios. Therefore, only the first three are compared.

Figure 2 presents the results. The upper-left plot corresponds to the three-variate case where both the random intercepts and the errors are normally distributed. We see that the proposed method continues to be more efficient than the direct L_1 method even though its working correlation structure (compound symmetry with equal ρs) is not true. The comparison with the linear mixed model is similar to what we have in Figure 1. We can only see a little increase in the curve when $\rho=0.05$,

which corresponds to the case where the variance among the ρs is the greatest, and thus where we are the farthest away from the working correlation. Hence it seems that the proposed method is quite robust to this type of departure from the assumptions.

The upper-right plot corresponds to the same situation but with random intercepts and errors distributed distributed as t_3. This time, the proposed method is always more efficient than both the direct L_1 method and the linear mixed model. Similar findings can be obtained from the lower plots, which present the corresponding results for the seven-variate cases.

The conclusion from this limited simulation study is that the transformation-retransformation L_1 method is very promising as it seems more efficient than the direct L_1 method. Moreover, the advantage over the direct L_1 method seems to hold even when the working correlation structure is not well-specified. Hence, the partial information recovered by the transformation-retransformation L_1, even if coming from a wrongly specified working correlation structure, is still useful enough to beat the direct L_1 method.

6 Concluding remarks

The goal of this paper was to show the potential of a novel L_1 norm estimation method of regression coefficients for clustered data. Unlike using the L_1 norm directly on the data, the proposed transformation-retransformation method uses the clustering structure to produce more efficient estimates, as shown in a simulation study. Hence this method deserves to be investigated further. The next logical step is to study the theoretical properties of the proposed method. More precisely, we are planning to derive its asymptotic properties, including calculations of asymptotic efficiencies. Moreover, we used a simple estimation of ρ, based on response-wise linear mixed models in the particular implementation used in the simulation study. However, using an L_1-type method would be more natural and we are also planning to develop such a method.

Acknowledgements

This research was supported by NSERC, FRQNT and the Academy of Finland.

References

Datta, S., Satten, G.A.: Rank-sum tests for clustered data. J. Amer. Statist. Assoc. **471**, 908–915 (2005)

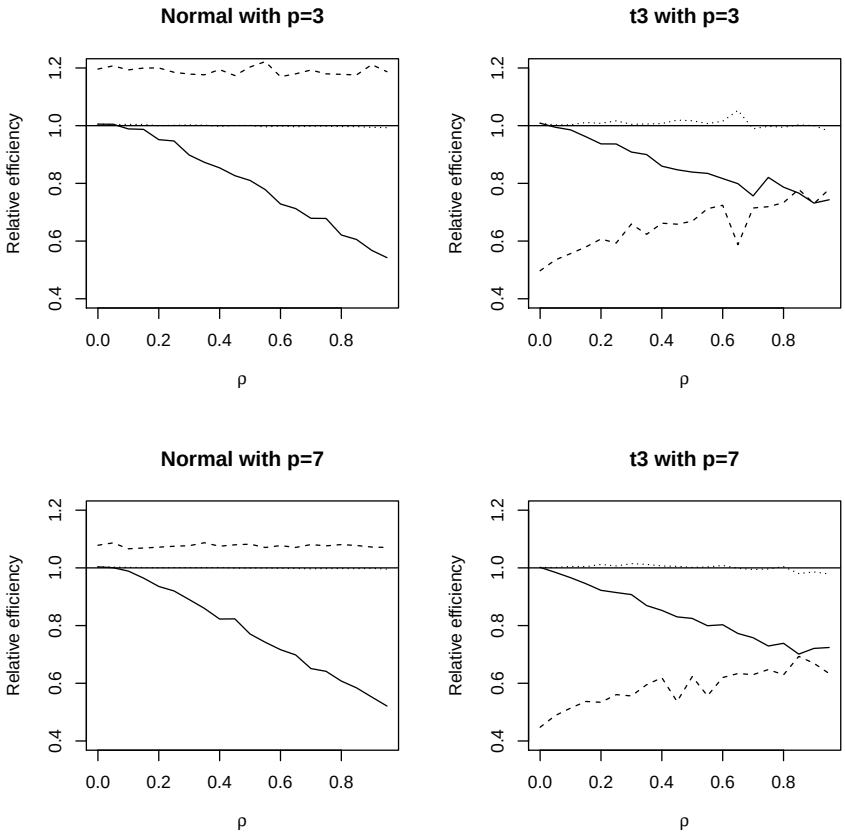

Fig. 1 Efficiency relative to the transformation-retransformation L_1 method with estimated ρ. Equal ρ case. Basic L_1 is the full line (–), mixed model is the dashed line (- - -), transformed L_1 with true ρ is the dotted line ($\cdots$).

Diggle, P., Heagerty, P., Liang, K-Y., Zeger, S.: Analysis of Longitudinal Data (second edition). Oxford University Press (2002)

Fu, L., Wang, Y.-G., Bai, Z.: Rank regression for analysis of clustered data: A natural induced smothing approach. Comp. Statist. & Data Anal. **54**, 1036–1050 (2010)

Fu, L., Wang, Y.-G.: Efficient estimation for rank-based regression with clustered data. Biometrics **68**, 1074–1082 (2012)

Jung, S.-H., Ying, Z.: Rank-based regression with repeated measurements data. Biometrika **90**, 732–740 (2003)

Jurěcková, J.: Asymptotic linearity of a rank statistic in regression parameter. Ann. Math. Statist. **40**, 1889–1900 (1969)

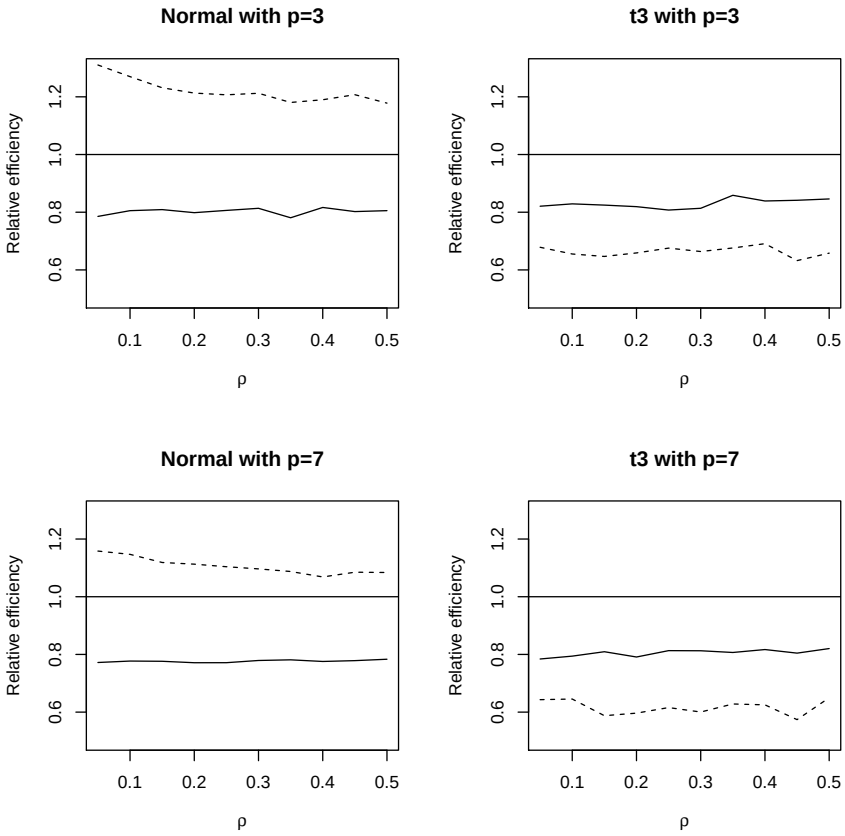

Fig. 2 Efficiency relative to the transformation-retransformation L_1 method with estimated ρ. Unequal ρ case. Basic L_1 is the full line (–), mixed model is the dashed line (- - -).

Jurěcková, J.: Non-parametric estimate of regression coefficients. Ann. Math. Statist. **42**, 1328–1338 (1971)

Kloke, J.D., McKean, J.W., Mushfiqur Rashid, M.: Rank-based estimation and associated inferences for linear models with cluster correlated errors. J. Amer. Statist. Assoc. **485**, 384–390 (2009)

Nevalainen, J., Larocque, D., Oja, H. Pörsti, I.: Nonparametric analysis of clustered multivariate data. J. Amer. Statist. Assoc. **490**, 864–872 (2010)

Nordhausen, K. and Oja, H.: Multivariate L_1 Methods: The Package MNM. J. Statist. Soft. **43**, 1–28 (2011)

Oja, H.: Multivariate nonparametric methods with R: An approach based on spatial signs and ranks. Springer (2010)

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team: nlme: Linear and nonlinear mixed effects models. R package version 3.1-117 (2014)

http://CRAN.R-project.org/package=nlme
R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing (2013)
http://www.R-project.org
Wang, Y.-G. and Zhao, Y.: Weighted rank regression for clustered data analysis. Biometrics **64**, 39–45 (2008)
Wang, Y.-G. and Zhu, M.: Rank-based regression for analysis of repeated measures. Biometrika **93**, 459–464 (2006)