CESIS
Electronic Working Paper Series

Paper No. 114

PERFORMANCE EVALUATION BASED ON THE ROBUST MAHALANOBIS DISTANCE AND MULTILEVEL MODELLING USING TWO NEW STRATEGIES

S. Hussain, M. A. Mohamed, R. Holder, A. Almasri and G. Shukur

February 2008

Performance Evaluation Based on the Robust Mahalanobis Distance and Multilevel Modelling Using Two New Strategies

By

S. Hussain[1]; M. A. Mohamed[2]; R. Holder[3]; A. Almasri[4]; and G. Shukur[5].

[1&3] Division of Primary Care and General Practice, School of Medicine, University of Birmingham,
[2] Department of Public Health, University of Birmingham, UK
[4] Department of Economics and Statistics, Karlstad University, Sweden
[5] Department of Economics and Statistics, Jönköping University, Sweden, and Centre for Labour Market Policy (CAFO), Department of Economics and Statistics, Växjö University, Sweden

Abstract.

In this paper we propose a general framework for performance evaluation of organisations and individuals over time using routinely collected performance variables or indicators. Such variables or indicators are often correlated over time, with missing observations, and often come from heavy tailed distributions shaped by outliers. Two double robust strategies are used for evaluation (ranking) of sampling units. Strategy 1 can handle missing data using residual maximum likelihood (RML) at stage two, while strategy two handle missing data at stage one. Strategy 2 has the advantage that overcomes the problem of multicollinearity. Strategy one requires independent indicators for the construction of the distances, where strategy two does not. Two different domain examples are used to illustrate the application of the two strategies. Example one considers performance monitoring of gynaecologists and example two considers the performance of industrial firms.

Key Words: Ranking indicators, performance, robust statistics, multilevel estimation, Mahalanobis distance.

[1] Dr. Shakir Hussin, s.hussain@bham.ac.uk; Division of Primary Care and General Practice, School of Medicine, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

1. Introduction

The development of performance or ranking indicators has a wide range of application in several disciplines. It constitutes a central topic in (e.g., economics, health, sociology, education) and modern formalized modeling has dealt with it for some considerable time. The formalized modeling of this topic has led the way towards statistical application and evaluation, and it is such application and evaluation that form the main theme of this paper. At this stage it is important to stress that each and every area of application has its own characteristics regarding the nature and definition of the problem and the observations and data collected thereby.

In the near past, numerous studies in many areas, starting with education; economics; health; and other social sciences, have been conducted in the developing of performance indicators where quantitative comparisons between institutions have been developed to introduce efficiency among the activities in those areas. Many of these studies, however, did not pay attention to possible complexities associated with real data like, for example missing values and the nature of this missingness; heavy tailed data as a result of outliers; among the *independent* variable (note that in many cases there is no dependent variable involved in the analysis); possible linear trends. However, Goldstein and Spiegelhalter (1996) addressed in some detail the need to take account of model based uncertainty in making comparisons, to establish appropriate measures of institutional outcomes and base-line measures, and to exercise care and sensitivity when interpreting apparent differences.

Among the most recent studies on performance are those of Behrenze and Althin (2005) in the area of labour market and unemployment, and Harley et al (2005) in the area of health. In the first paper the authors studied the efficiency and productivity of employment offices in Sweden. They applied the so-called Malmquist productivity index, which measures the distance from every office to the production frontier and the shifting of the frontier over time. In the second paper, however, the authors have used a two stages statistical method for evaluating health indicators using routine hospital data to identify gynecologists' performance. The robust Mahalanobis distance is first used to reduce the dimension of the indicators, after which the robust residuals from the level-two estimation were investigated to indicate the outliers. Note that, in the first

paper, the analysis has been done in the micro level and no consideration have been taken for the possible existence of any hierarchy structure in the data, while in the second paper possible hierarchy is taken into account.

However, increasing attention is nowadays paid to multilevel modelling especially in modelling dynamic relationships of hierarchical structures. The purpose of this study is to provide two strategies, for performance evaluation of individuals and organisations over time, that take into consideration the problems mentioned associated with real data which other previously studies did not. Here we combine robust multivariate methods together with multilevel estimation methods that allow for missing observations among the data. By following such strategies one can avoid inadequate methods that might lead to extremely misleading results and inferences. To demonstrate these strategies, we use two examples from two different areas. The first example is regarding evaluating health indicators to assist health authorities in decision-making, while the second example evaluates the performances of industrial firms. These examples illustrate how our approach can disentangle issues regarding producing performance indicators when the data under consideration suffer from missing values, outliers, are fat tailed, and different levels of multicollinearity in the main structure.

The rest of the paper is organised as follows: Section 2 presents a description of the data and the theoretical assumptions; Section 3 describes the methodology and estimation procedures; while the results are presented in Section 4. Section 5, finally, gives conclusions and a summary of findings.

2. Data Description

In this section we describe the nature of the two data sets that we use to illustrate our modelling strategy.

Example one:

The data set here consists of routine hospital data of seven clinically relevant indicator variables from hospital episode statistics for 143 gynaecologists. These indicators are: % of finished

gynaecologist episodes with complications; mean length of spell (in days); % of finished gynaecologist episodes with more than two operations; % of finished gynaecologist episodes where spell is longer than episode; % of finished gynaecologist episodes for dilatation and curettage on women aged less than 40 years; % of finished gynaecologist episodes for sterilisation on women aged less than 25 years; and % of finished gynaecologist episodes for hysterectomy on women aged less than 30 years. The data are collected over five years period (from 1991:2 to 1995:6, but only 68 gynaecologists appear in all five years. High proportions may potentially indicate sub-optimal performance, although this requires further study.[2]

Example two:

The dataset contains all Swedish industrial firms that are registered on the stock market and with available accounts during the resent five years (1999 to 2003). Data about 2004 are unfortunately not completely recorded yet. With help of the database OSIRIS, 57 firms were identified.[3] The following ten variables were formulated: cost of goods sold (COGS) to sales; free cash flow to current liabilities; growth in gross investment; growth in net turnover; growth in shareholders' equity; growth in total debt; R&D to sales; sales to total assets; solvency ratio; and cost of employees.

COGS to sales is a profitability ratio where COGS can be seen as the cost of doing business, for example the cost of raw materials. The variable is weighted by sales and measured in percentage. *Free cash flow to current liabilities* is a liquidity ratio. Free cash flow is how much cash a firm has after paying its bills for ongoing activities and growth. Current liabilities are the bills that are due in less than one year. *Growth in gross investment* measures the growth in capital goods; the variable is measured in percentage terms. *Growth in net turnover* measures the growth in net income or revenue from the sale of goods and services, in percentage. *Growth in shareholders' equity* determine in percentage the growth of the amount by which the firm is financed through common and preferred shares. *Growth in total debt* is a variable that measure the growth in percentage of the total debt, i.e. the sum of short-term and long-term debt. *R&D to sales* is a variable in percentage terms that quantify the relative importance of R&D investments in the

[2] The data however is described in more details in Harley et al (2005).

[3] OSIRIS is a comprehensive database of listed companies, banks and insurance companies around the world.

firm. The *sales to total assets* is a ratio that indicate the firm's efficiency at using its assets in generating sales or revenue. *Solvency ratio* is a measure of the firm's ability to meet long-term obligations. The variable is calculated by taking a company's net worth to total assets.

3. Methodology and estimation procedures

This paper is designed to provide frameworks that encompass two stages statistical methods that evaluate health and firm indicators thereby assisting health- and other authorities in decision-making. Two stages of data reductions are proposed using two strategies. These strategies consist of two major components, namely the multivariate Mahalanobis distance (MD) and the level two standardised residuals estimates from the multilevel modelling. In the following we present a brief description of these two components.

Suppose that $x = (x_1, x_2,\ldots, x_n)^T$ is a set of n observations on p random variables. The classical MD is defined as:

$$MD(x_i) = \sqrt{(x_i - \mu)V^{-1}(x_i - \mu)^t} \quad . \qquad (1)$$

Where μ is the arithmetic mean vector and V is the covariance matrix. The classical squared Mahalanobis distance $(MD)^2$ is not ideally suited to multivariate outlier detection because it is not resistant to outliers. Rousseeuw and Leroy (1987) recommend using distance based on robust estimators of multivariate location and scatter (μ_R, V_R) to avoid masking effect. A cutoff point of $\sqrt{\chi^2_{p,0.975}}$, (p is the number of the variables), used to determine points above as outliers. The minimum covariance determinant (MCD) method of Rousseeuw (1985) aims to find (h) observations out of (n) whose covariance matrix (V_R) has the lowest determinant. The development of MCD is mentioned in Rousseeuw and Van Driessen (1999) under the name Fast MCD algorithm, where lower determinant of MCD can be approximated from the initial MCD.

Assuming that the fraction of outliers is at most α, (0< α≤ ½), e.g. 50%, then α can be chosen to equal $\chi^2_{p,0.50}$ where, except for the extreme values cases, we expect the majority of the data to come from a normal distribution. Let the halve contain h = (n + p +1) / 2 observations with (n) total sample size and (p) number of variables, however the determinants of the covariance matrix (V_R) will be minimized subject to the inequality

$$\{i,(x_i-\mu_R)V_R^{-1}(x_i-\mu_R)^t \le \alpha^2\} \ge h \qquad (2)$$

Finally the robust MCD distance can be written as

$$RMD(x_i)=\sqrt{(x_i-\mu_R)V_R^{-1}(x_i-\mu_R)^t} \qquad (3)$$

Where μ_R is now our first moment vector and V_R is the robust covariance matrix. Rocke and Woodruff (1996) proposed the robust M estimator that uses the fast minimum covariance determinant estimator as an initial robust estimate then the estimate refines with M iterations using the translated bi-weight function that is described in Rocke (1996). In computing the Robust Distance we use either the Fast MCD or the M estimator that is available in the robust library of S+ 8.

The data we study here consists of repeated performance measures of random samples of item (gynaecologists /firms). We start by writing a simple empty model

$$y_{ij}=\overline{\beta}+\beta_{0j}+e_{ij} \qquad (4)$$

where y_{ij} denotes the MD of the i[th] year of the j[th] gynaecologist, $\overline{\beta}$ represent the mean MD distance, β_{0j} is a random variable representing "between-items" variability, and e_{ij} is a random variable representing "within- items" variability. The distributions of the random variables are

$$\beta_{0j} \sim \mathrm{N}(0, \sigma_b^2) \qquad e_{ij} \sim \mathrm{N}(0, \sigma_e^2), \tag{5}$$

where σ_b^2 and σ_e^2 are the residuals variances of the between items (level two) and within items (level one) effects, respectively. The so-called Huber/White or sandwich estimator is then used to obtain robust tests and confidence intervals by correcting the asymptotic standard errors.

Langford and Lewis (1998) propose downward residual checking starting from the heights level and continuing with each next lower level for the purpose of outlier inspection and ranking. Here, we consider both level one residuals, e_{ij} and level two residuals β_{0j} for this purpose.
The underlying assumptions of the model in formulas 4 and 5 suffer from lack of robustness against outlying observations since both residuals are based on the Gaussian distribution, Seltzer and Choi (2003). According to Pinheiro and Wu (2001), an interesting feature of mixed-effects models is outliear may occur either at the level of the within subject error e_{ij} called e-outliers, or at the level at random effect β_{0j} called $\beta - outliers$. In this paper our concern is the level two subjects $\beta - outliers$. Empirical Bayesian (EB) method is used to predict level two residuals. One disadvantage is its strong dependence on the model assumptions and the other disadvantage is when the number of sampling units in the level two is small, the EB approach in this case can result in underestimating of uncertainty.

The Fully Bayesian (FB) approach bases inferences on the marginal posterior distribution of all the parameters in the model and the 0.025 and 0.975 quantiles of this distribution would provide the Bayesian analogue of a 95% confidence intervals, see Seltzer and Choi (2003).

Using the scale mixture of normal representation presented by Seltzer and Choi (2003), the normal representation of the t distribution with mean 0, scale parameter equals to 1 and γ

degrees of freedom, $t(0,1,\gamma)$ can be expressed by $\frac{Z}{\omega^{1/2}}$. Z here is standard normal with mean 0 and variance 1 and (ω) is chi-squared distributed divided by its degrees of freedom and the distribution of $\chi^2_\gamma / \gamma = Gamma(a,b)$ where $a = \frac{\gamma}{2}$ and $b = \frac{\gamma}{2}$, see Gelmn et al (1995). Using the above argument and if we assume that the random effects has $t(0,\sigma^2_{\mathrm{b}},\gamma)$ then β_{0j} in (2) has the form:

$$\beta_{0j} \sim N(0, \frac{\sigma^2_b}{s_i}), \text{ where } s_i \sim Gamma(\gamma/2, \gamma/2). \qquad (6)$$

For the level one residuals, e_{ij}, if we assume $t(0,\sigma^2_{\mathrm{e}},\gamma_1)$ then the definition under normality assumption is:

$$y_{ij} \sim N(\overline{\beta}, \frac{\sigma^2_e}{r_{ij}}), \text{ where } r_{ij} \sim Gamma(\gamma_1/2, \gamma_1/2). \qquad (7)$$

Note that the mean $\overline{\beta}$ is a constant that may take any value based on the data we analyse. The direct robust assumption of t distribution for the two residuals shows different effect on ranks and outliers. Here we will conduct a simple MCMC simulation study to show the difference between normal, the scale mixture (see Peel and Mclachlan, 2004), and the direct t distribution assumption of the two robust approaches. The construction of the proper normal prior for the above simulation study will consider minimally informative prior for $\overline{\beta}$ which we choose to be:

$$\overline{\beta} \sim \mathrm{N}(0, 1*10^6). \qquad (8)$$

This type of normal prior adds no information to the data since the data have a range of {0.88, 12.72}.

Now, in the first stage of strategy one, a robust multivariate MD (RMD) is computed from several orthogonal indicators and assigned to each subject over the years. These distances are used in the second stage as the outcome in a multilevel model (level 1: design variable for distances over time, level 2: design variable for sampling units) to obtain ranks of the units. In other words, the rank from the second level robust standardized residual of a multilevel model of the repeated RMD is computed for each individual or firm. This can be done according to the following:

1- Compute the RMD for all the variables over the years.

2- Use these RMD as the outcome in the robust multilevel model (level one is the repeated RMD while level two is the gynaecologist).

3- Obtain the rank from the robust level two standardised residuals.

4- Compute the uncertainty of each rank using the posterior distribution of the MCMC.

For strategy two, in the first stage and to avoid possible multicollinearity, we consider one variable at a time. For each variable, we apply the multilevel modelling to achieve a reduction over time to one single point. In order to obtain robust standardised residuals for a specific gynaecologist or firm, the level two subjects $\beta - outliers$ will be the outcome of this modelling. We then repeat this procedure for the rest of the variables and obtain as many points as the number of variables. In the second stage, we compute a RMD to obtain the rank of those units. Processing in this manner we avoid the problem of dealing with dependency among the variables. Strategy two can be summarised according to the following:

1- Compute the robust level two residuals for each indicator or variable separately over the years.

2- Use these level two residuals in the second stage to compute the RMD.

3- Obtain the rank from the RMD.

4- Compute the uncertainty of each rank using the posterior distribution of the MCMC.

Note that, one can also apply the idea of strategy one to dependent data but MD requires independent indicators for the construction of the distances. However, robust principle components can help us to decide which strategy to use by detecting independence in the data structure. In what follows, we confine ourselves to brief descriptions of the methodological issues used in this study, and further details are found in cited references.

3.1. Robust Multivariate Outliers Detection

The classical squared Mahalanobis Distance $(MD)^2$ is one approach to multivariate outlier detection based on arithmetic means and covariance matrix, MD measures how far a random vector is from the middle of its distribution. This will provide a reasonable summary measure of the distance of each item (individual or firm) from the mean. Points in multivariate data with large MD and greater than $\sqrt{(\chi^2_{p,0.95})}$ are approximately considered outliers, where p here denotes the number of variables with 0.95 χ^2 quantile (see Rousseeuw and Leroy, 1987).

Note that occasionally one might be faced with cases where one outlier is too extreme that it may mask other outliers as a result. To overcome this problem in MD, Rousseeuw and Leroy (1987) proposed a robust estimation of covariance (M-type robust of covariance). If the covariance matrix for example is not estimated robustly then the underlying structured parameters are not robust. Using a method called “C-step”, Rousseeuw and Van Driessen (1999) developed a fast algorithm for Minimum Covariance Determinant (MCD) to approximate the minimum covariance determinant estimator of the MD (see the S+ 8 Robust Library (2007) for use of MCD algorithm and other algorithms).

3.2. MANOVA and Multilevel Model

The data collected in this study are unbalanced repeated measures over time. For strategy one, for example, when modelling the outcome measure MD, this repeated measurement data could be viewed as multilevel data with repeated MD nested within subjects. This leads to a two level

model with the series of repeated MD as the lowest level and the individual subjects as the highest level. Multivariate ANalysis Of VAriance (MANOVA) may be used to model the MD data. The advantages of MANOVA are that no assumptions about the covariance matrix need to be made and that under normality assumption it yields exact tests. The assumptions are related to independent and identical distributions within treatment groups, homoscedasticity between groups, multivariate normality and complete data. When group sizes are different MANOVA will suffer from lack of uniqueness, Searle Casella and Mcculloch (1992). The assumption of homoscedasticity is unlikely to occur in the structure of the data that we are dealing with, however, since a group of subjects may show more variation over time than other subjects and this variation may lead the observation to be an outlier. This example shows the violation of the homoscedasticity assumption.

The multilevel model in this study is composed of three components; first, the fixed part that represents the fixed effect of the intercept and the trend, second, the random effect at level one, and finally we have the random effect at level two. The measurement occasions are nested within subjects, level one units are occasions and level two units are subjects. The random coefficient approach to repeated measures is usually based on polynomial trend to a model that is a polynomial function of time. Depending on the research topic one may use any other linearly independent set of functions. At this stage it is important to mention that there are two estimation procedures for these purposes, namely, the Maximum Likelihood (ML) and the Residual Maximum RML. RML estimation takes into account the loss of degree of freedom resulting from the estimation of the parameters of the fixed part and also allows for missing values. Hence, in this paper, we are considering the RML as our estimation method.

4. Results

In this section we present our main results for our methodology for performance evaluation and giving our two empirical examples. For strategy one, when we use the assumption of normality and in the case of almost independent or weak dependent data, we start the analysis by obtaining the MD that we then use as the outcome in a multilevel model to obtain ranks of sampling units,

(the outcomes from different time point are unbalance). On the other hand, and when some dependency is existing among the variables, we start by calculating the principle components (to detect independence) and then we obtain the RMD that we then use as the outcome in a multilevel model using RML to obtain ranks of sampling units.

Strategy two starts with the multilevel modelling using RML first and then the RMD in the case when normality is assumed, while we use the robust counterpart in the situation of fat tailed distribution. The only difference between this strategy and strategy one is that here we ignore the dependency between the variables since we consider only one variable at a time (missing time points is possible) by applying multilevel modelling to reduce the points over the time for the respective variable to one single point. We then repeat this procedure for the rest of variables and obtain as many standardized points as the number of variables. Finally we obtain the RMD, or the robust version of it in the presence of fat tailed, to obtain the rank of those units.

Example one

For this example, and since the data is not highly collinear, we apply both strategies to maintain the ranking performance. We first follow strategy two to achieve the results of ranking indicators of the gynaecologists and compare the last 10% of the ranking distribution with those obtained by means of strategy one. If the two strategies lead to similar results, we then concludes that they converge to each other with respect to ranking the best/worst individual in the sample. The results of the two strategies are presented in Figures 1 and 2. These figures are constructed using two different functions the mater that makes them looking different in construction.

In these figures we show all the results from our analyses but we only rank a subset of them that are quite different in performance than the others. For example using strategy one we in Figure 1 show the results from about the upper 10^{th} percentile ranking of the indicators (16 gynaecologists) that might give us a picture of bad performance for the ranked individuals. Figure 2 also shows results about the upper 10^{th} percentile ranking of the indicators (16 gynaecologists), but we use strategy two to produce these ranking indicators. The results from the two figures reveal that both strategies almost agree in this case regarding the number of individual that belong to the upper 10^{th} percentile of the ranking distribution of the gynaecologists (an agreement of about 80% in

ranking and a correlation of about 0.75%). However, both strategies totally agree with respect to the last three gynaecologists with the worst performances in the sample, namely gynaecologists with ranks, 111, 117, and 118.

Figure 1. Upper 10th percentile ranking of the indicators using strategy one

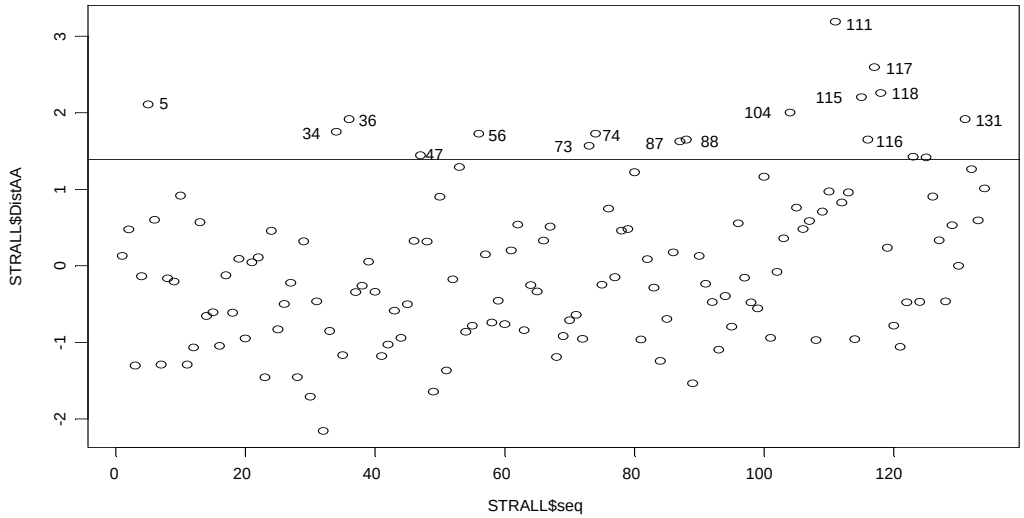
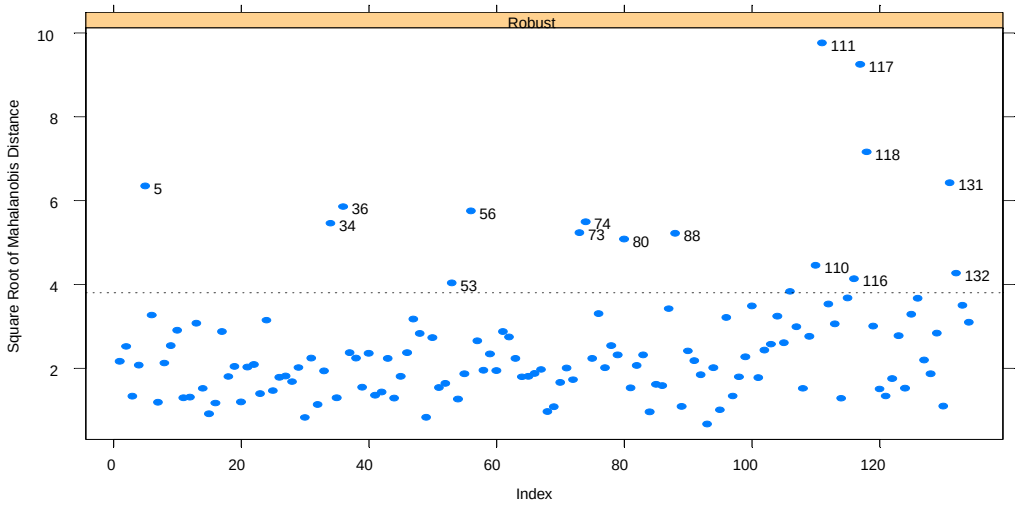


Figure 2. Upper 10th percentile ranking of the indicators using strategy two



A Markov Chain Monte Carlo (MCMC) simulation study has been conducted using WinBugs, version 1.4 to compute the uncertainty of the ranks. We compare between 3 types of distributions for the level 1 and level 2 residuals, namely the Normal, t(4,4) and the Scale Mixture of the normal representation of the t-distributions. The results of this simulation study are presented in the following Table 1. The results presented in Table 1 are for the cases when there exist some kind of disagreement, when using these three distributions, regarding indicating the outlierness of the individuals. In this table we only include the cases of individuals that considered as outliers by at least one of used distributions (i.e., when the values of the level two residuals are greater than 2). The results, however, indicate the following: The 3 methods show negative slope over the studied period, i.e. The fixed linear trend = { -0.1135, -0.0754, -0.0717} indicating progressive shrink of the MD distance over time. The Inter Class Correlation (ICC) shows stronger clustering with the case of the Normal (0.614) and weaker in the case of t(4,4) (0.418) and scale mixture(0.428). The Tau in the table stands for the within level 1 residuals variance over the number of years, while the Tau.u2 is the between level 2 residuals variance. These variances are significant in all cases.

Using the t(4,4) and Scale Mixture**, t**he level 2 residual value (of sample size 5 years) for the individual number [75] strongly classify him/her as an outlier, indicating consistent low performance for this individual, while using the Normal distribution this individual is not considered as an outlier. This indicates that t(4,4) and scale mixture show robust outlier detection.

Table 1. Results that disagree regarding indicating the outlierness of the individuals using the 3 distributions (values, standard divinations and confidence intervals are reported)

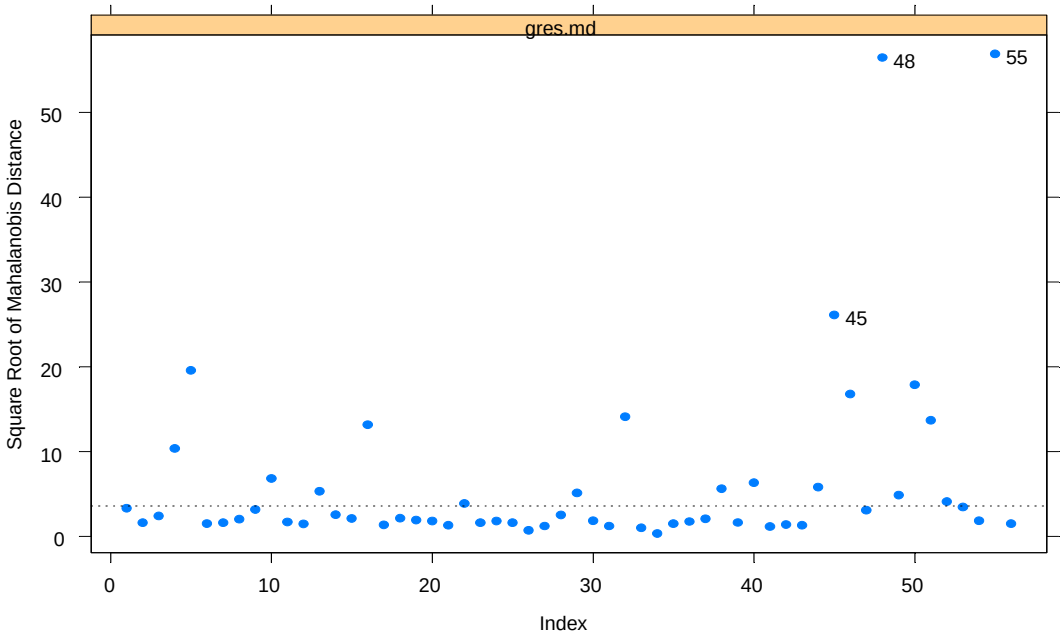
	Normal	$T_{4,4}$	Scale Mixture	Sample size
Grand mean	3.584 0.1762 (3.24, 3.934)	3.076 0.1395 (2.881, 3.353)	3.039 0.152 (2.755, 3.359)	-
Fixed linear trend	-0.1136 0.038 (-0.187, -0.038)	-0.074 0.0313 (-0.137, -0.013)	-0.0735 0.032 (-0.137, -0.010)	-
Tau	1.221 0.095 (1.049, 1.42)	1.716 0.173 (1.399, 2.075)	1.729 0.171 (1.435, 2.091)	-
Tau.u2	1.967 0.319 (1.411, 2.665)	1.246 0.259 (0.806, 1.853)	1.309 0.271 (0.833, 1.899)	-
Individual [75]	1.871 0.483 (0.914, 2.835)	2.410 0.6822 (0.951, 3.598)	2.382 0.715 (0.851, 3.581)	5
Individual [90]	2.157 0.593 (0.995, 3.35)	1.721 0.803 (0.245, 3.339)	1.755 0.888 (0.229, 3.779)	3
Individual [106]	2.578 0.880 (0.860, 4.301)	2.570 1.724 (-0.643, 5.398)	2.586 1.792 (-0.724, 5.452)	1
Individual [111]	5.966 0.953 (4.047, 7.820)	5.144 4.532 (-1.11, 11.20)	5.883 4.501 (-0.950, 11.18)	1
Individual [115]	2.765 0.887 (1.050, 4.531)	2.846 1.90 (-0.583, 5.78)	3.118 1.864 (-0.560, 5.87)	1
Individual [116]	2.203 0.599 (1.05, 3.346)	1.205 0.89 (-0.290, 3.065)	1.087 0.847 (-0.331, 2.879)	3
Individual [117]	3.765 0.902 (0.965, 5.599)	4.694 2.393 (-0.50,7.715)	3.976 2.676 (-0.835, 7.68)	1
Individual [118]	2.876 0.895 (1.108, 4.689)	2.814 2.023 (-0.760, 5.959)	3.178 1.955 (-0.425, 6.115)	1
Individual [131]	2.194 0.884 (0.480, 3.964)	2.401 1.435 (-0.434, 4.716)	2.392 1.552 (-0.599, 4.788)	1
DIC	1593.74	1505.54	1400.50	
ICC	0.614	0.418	0.428	

In the table, we have 6 individuals with only one year sample size available, namely [106], [111], [115], [117], [118] and [131]. Using the Normal distribution, these individuals have been significantly considered as outliers. On the other hand, when applying the **t**(4,4) and the Scale Mixture, these individuals are still considered as outliers, but with higher uncertainty making the results statistically insignificant. The confidence intervals associated with the estimated residuals for these individuals cover the value of zero. This means that making analysis on these individuals by means of only one observation of time might lead to misleading conclusions. This has been discovered when using the more robust t(4,4) and Scale Mixture but not the Normal distribution. The same is true for the individuals [90] and [116] for whom the level 2 residual values (of sample size 3 years) significantly indicate them as outliers using the Normal distribution, while the t(4,4) and the Scale Mixture significantly do not indicate the first as an outlier, the second however is not statistically significant.

Example two

In this example, and since the variables are highly correlated (the correlations between the variables are between 0.60 - 0.90), we only apply strategy two. Applying the first strategy, by first computing the PC to detect independence, will significantly reduce the information used in the analysis which might lead to the problem of omitting relevant information. The results shown are obtained using the variables *COGS to sales, Free cash flow to current liabilities, Growth in gross investment, Growth in net turnover, Growth in total debt, R&D to sales* and *cost of employees.* The other variables did not show any significance and were excluded from the analysis. The results of the ranking indicators are shown in Figure 3 bellow.

Figure 3. Ranking of the companies indicators regarding best performances using strategy two



Our analysis indicated the firms with the best performances among the others, these are the observations 45, 48 and 55 that stand for the firms Sandvik, Securitas and Volvo, respectively (but that Sandvik and Securitas clearly have the best performance). A closer examination of the figure, reveal that some other firms are also indicated to have fairly good performances, namely those that are associated with observations 5, 16, 32, 46, 50 and 51 (not specified in the figure) that stand for the firms Atlas Copo, Fingerprint Card, NCC, Scania, Skanska and SKF. Note that all these indicated firms are heavy industry, building and computer firms in Sweden.

5. Summary and Conclusions

The main purpose of this paper is to propose a general framework for performance evaluation of organisations and individuals over time using routinely collected performance variables or indicators. Two approaches or strategies are recommended depending on the nature of the data under consideration. It is not unusual that the data are often interdependent, correlated over time, with missing observations, or used to come from heavy tailed distribution shaped by outliers. Based on this fact, and the strength of possible interdependence between variables, we introduce

two strategies for evaluation (ranking) of sampling units. In cases when the dependency structure between the variables is weak or the variables are almost orthogonal to each other, the first strategy starts by computing the Mahalanobis distance (MD) for each sampling unit (if these units are normally distributed), or the RMD (in the case of heavy tails distributions) for all indicators over time. These distances are then used in the second stage as the outcome in a multilevel model using RML to obtain ranks of sampling units. The second strategy should be used when the dependency structure between the variables is high (but it also works in cases with weak dependency or if the variables are almost orthogonal). It starts by first applying the multilevel model with indicators as the outcome to derive robust standardised residual for each indicator variable in the first stage, and then compute an MD or RMD of all indicators for each sampling unit in the second stage.

To summarise, strategy one can handle missing data using robust residual maximum likelihood (RML) at stage two, while strategy two handle missing data at stage one. Running one indicator at a time in strategy two solve the multicolinearity problem. Strategy one requires independent indicators for the construction of the distances (this imply a great loss of information when the variables are not independent, however when calculating principle components and excluding dependent data to achieve independency), where as strategy two does not. Two different domain examples are used to illustrate the application of the two strategies. Example one considers performance monitoring of surgeons and example two considers the performance of industrial firms. The code for the program is included at the end of this paper and the data we used together with the initial values are available upon request from the authors.

The results from the first example reveal that both strategies almost agree regarding the number of individual that belong to the upper 10^{th} percentile of the ranking distribution of the gynaecologists (an agreement of about 80% in ranking and a correlation of about 0.75). However, both strategies totally agree with respect to the last three gynaecologists with the worst performances in the sample. In the second example, however, we are intending to rank the firms with the best performances. Since the analysed variables are very collinear, we applied strategy two in our analysis. The results show that three firms are ranked as having the best performance with respect to the sample period. These are Sandvik, Securitas and Volvo. Other firms have also

shown to have good performances, namely; Atlas Copo, Fingerprint Card, NCC, Scania, Skanska and SKF. These firms are known to be heavy industry, building and computer firms in Sweden.

References

Behrenz, L. and Althin, R. (2005), "Efficiency and productivity of employment officies: Evidence from Sweden", *International Journal of Manpower*, Vol. 26, No. 2.

Gelman, A., Carlin, J., Stern, H., & Rubin, D (1995). "Bayesian data analysis". London: Chapman & Hall

Goldstein, H., and Spiegelhalter, D. J (1996), "League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance", *J. R. Statist. Soc*. A, 159 (1996), pp. 385-443.

Harley, M., M. A. Mohammed, S. Hussain, J. Yates, and A. Almasri (2005), "Was Rodney Ledward a Statistical outlier? Retrospective Analysis Using Routine Hospital Data to Identify Gynaecologist's Performance", *BMJ*, doi: 10.1136 / bmj. 38377.675440.8F (Published 15 April 2005).

Hampel, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation", *JASA*, 62, pp 1179-86.

Huber, P. (1981), "*Robust Statistics*", New York, John Wiley.

Langford, I. H; and Lewis, T. (1998), "Outliers in multilevel data", ", *J. R. Statist. Soc*. A, 161: pp 121-160, 1998.

Marazzi, A. (1993), "*Algorithms, routines, and S functions for robust statistics*". Wadsworth & Brooks/Cole, Pacific Grove, CA.

Peel, D. and McLachlan, G. J. (2000), "Robust mixture modelling using the t distribution", *Statistics and Computing*, 10, No.4/October, 2000 pp 339-348.

Pinheiro, Jose C; Chuanhai Liu; Ying Nian Wu (2001), "Efficient Algorithms for Robust Estimation in Linear Mixed-Effects Models Using the Multivariate t-Distribution", *Journal of Computational and Graphical Statistics*, 10, No. 2, pp. 249-276.

Rocke, D. M. (1996), "Robustness Properties of S-Estimators of Multivariate Location and Shape in high Dimension, *Annals of Statistics*, Vol. 24, No. 3, pp. 1327-45.

Rocke, D. M.; Woodruff, D. L. (1996), "Identification of Outliers in Multivariate Data", *JASA*, 91, pp. 1047-61.

Rousseeuw, P. J. (1985), "Multivariate Estimation with High Breakdown Point", Mathematical Statistics and Applications B (W. Grossmann, G. Pflug, I. Vineze and W. Werz, eds.) pp. 283-297

Rousseeuw, P. J. and Van Driessen (1999), "A Fast Algorithm for the Minimum Covariance Determnint Estimator", *Technometrics*, 41, No. pp. 212-223.

Rousseeuw, P J. and A. M. Leroy (1987), "*Robust Regression and Outlier Detection*", New York, John Wiley.

Searle, R. S., Casella G., and Mcculloch. C. E. (1992), *"Variance Components"*, John Wiley & Sons, INC

Seltzer, M and K.Choi (2003) "Sensitivity Analysis for Hierarchical Models: Downweightingand Identifying Extreme Cases Using the t distribution", edited by Steven P.

Reise and Naihua Duan, "Multilevel Modeling Methodological Advances, Issues, and Application" LONDON, LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS 2003.