A method of selecting an active factor and its robustness against correlation in the data

Shu Yamada
Department of Management Science
Tokyo University of Science
Kagurazaka, Tokyo 162-8601, Japan,
E-mail : shu@ms.kagu.tus.ac.jp

and

Jun Harashima
Department of Production, Information and Systems Engineering
Tokyo Metropolitan Institute of Technology
Asahigaoka 6-6, Hino, Tokyo 191-0065, Japan,

Abstract

A reducing variation of quality characteristics is a typical example of quality improvement. In such a case, we treat the quality characteristic, as a response variable and need to find active factors affecting the response from many candidate factors since reducing the variation of the response will be achieved by reducing variation of the active factors. In this paper, we first derive a method of selecting an active factor by linear regression. It is well known that correlation between factors deteriorates the precision of estimators. We, therefore, examine robustness of the selecting method against the correlation in the data set and derive an evaluation method of the deterioration brought by the correlation. Furthermore, some examples of selecting and evaluation methods are shown to demonstrate practical usage of the methods.

Keywords! 'Correct selection, Deterioration by correlation, Simulation study, Variable selection,

1 Introduction

In many situations, there is a need to find active factors that affect the response variables. An example is reducing variation of quality characteristic as a quality improvement. In such a case, we need to find active factors affecting to the quality characteristic from many factors because reducing variation of the quality characteristics will be achieved by reducing variation of its active factors. Another example is questionnaire surveys of customer satisfaction. For example, many hotels provide questionnaire to consult customer satisfaction for their service quality. The questions can be classified into satisfaction of service elements and overall satisfaction. In order to improve overall satisfaction, we will take actions on some elements of service. Therefore, the main goal of the questionnaire is to find service elements that affect he overall satisfaction.

In the data analysis of the above examples, quality characteristic and overall satisfaction are treated as a response variable. The goal of the data analysis can be regarded as selection of active factors. This selection problem is sometimes called "Screening problem" (Box and Draper (1987)).

As regard variable selection in linear regression, many methods have been proposed such as the forward selection based on the partial F statistic, Final Prediction Error aiming to select factors for predicting response value. However, the variable selection techniques have been aiming select factors for predicting a response value. Thus it is not ensured that these techniques are appropriate for selection of active factors. We, first, evaluate some selection methods by simulation study in terms of correctness of selection.

In the practical situation, it is impossible to avoid correlation in the collected data set, for example satisfactions of two service elements are correlated each other. In addition, some techniques of design of experiments do not ensure the orthogonality among paired factors, such as D-optimal design, composite design, supersaturated design. It is well known that correlation between factors deteriorates the precision of the analyzed result in regression analysis. Thus, the second problem discussed in this paper is examination of robustness on the selecting method against correlation. Furthermore, two examples are shown to demonstrate the application of the estimates in practical situations.

2 Selecting methods

Let y and $x_1, \ldots, x_p$ be a response variable and its factors that are supposed to affect the response, respectively. Let y_i and $x_{i1}, \ldots x_{ip}$, $i = 1, \ldots, n$, be independent n observations of the response and factors, respectively. Without loss of generality, we assume that $x_{1j}, \ldots, x_{nj}$ had been standardized with zero mean and unit variance for each variable.

A linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \tag{1}$$

is assumed to the relationship between the response and the factors. In the model, β_0 and $\beta_1, \ldots, \beta_p$ denote unknown parameters of general mean and factor effects, respectively.

Consider a factor whose effect is the largest among $x_1, \ldots, x_p$. Let τ be the suffix of the factor that provides $\max\{|\beta_j|\}$ among p factors, such that

$$\tau = \arg\max_j \{|\beta_j| \mid j = 1, \ldots, p\}. \tag{2}$$

The goal of the data analysis is to find x_τ. We evaluate the following three methods to find the factor x_τ, because of the popularity and simplicity of those methods.

(1) Estimates method : This method selects a factor whose absolute value of the estimate of regression parameter is the largest as an active factor based on the model including all factors to be considered. Let $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)^\top = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ be a least squares of $\boldsymbol{\beta}$, where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})^\top$. This method selects a factor whose absolute value is the maximum such that

$$\hat{\tau}_E = \arg_j \max\{|\hat{\beta}_j| \mid j = 1, \ldots, p\}. \tag{3}$$

(2) Correlation coefficient method : This method selects a factor whose absolute value of correlation coefficient between y is the maximum among the factors. Let r_{yj} be a correlation coefficient between y and x_j $(j = 1, \ldots, p)$. This method selects a factor whose absolute value of correlation is the maximum such that

$$\hat{\tau}_C = \arg\max_j \{|r_{yj}| \mid j = 1, \ldots, p\}. \tag{4}$$

(3) Partial F method : This method selects a factor whose partial F test statistic of factor effect based on the full model is the maximum among the factors. Let F_j be the partial F statistic for the null hypothesis $H_0 : \beta_j = 0$ and the alternative hypothesis $H_1 : \beta_j \neq 0$ based on the regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$. This method selects a factor whose F statistic is the maximum such that

$$\hat{\tau}_F = \arg\max_j \{F_j \mid j = 1, \ldots, p\}. \tag{5}$$

Let $P_E = \Pr\{\hat{\tau}_E = \tau\}$, $P_C = \Pr\{\hat{\tau}_C = \tau\}$ and $P_F = \Pr\{\hat{\tau}_F = \tau\}\}$ denote the probabilities of correct selection by Estimate, Correlation and Partial F methods, respectively. We evaluate the above three methods by the probabilities of correct selection of an active factor.

Remarks:
(a) These three methods select the same factor when the collected data of the factors are orthogonal.
(b) Correlation method selects the same factor with the forward selection based on the partial F statistics of the null hypothesis $H_0 : \beta_j = 0$, and the alternative hypothesis $H_1 : \beta_j \neq 0$ by a linear regression model including only x_j.
(c) The above three methods select the same factor when $p = 2$. Therefore, we evaluate these three methods in the case of $p \geq 3$.

3 Outline of evaluation

The probabilities P_E, P_C, P_F are evaluated under various conditions as follows:

1. The number of design variable p is $\{3, 4, 5\}$.

2. Two types of distribution of effects are utilized in this study. The first type, called "One active", implies that only one factor is active in p candidate factors such that

$$\beta_j = \begin{cases} \beta_{\max} & (j = \tau) \\ 0 & (\text{others}) \end{cases}, \tag{6}$$

where $\beta_{\max} = 0.25(0.25)1.5$, 2.0, 3.0 and $\tau = 1$. The other type, called "Step", supposes that factor effect distributed as $\beta_{\max} = \beta_{[1]} \geq \beta_{[2]} \geq \cdots \geq \beta_{[p]}$ where $\beta_{\max} = 1(1)6$ and $\beta_{[k]} = (1 - (k-1)/(p-1))\beta_{[1]}$ $(k = 1, \ldots, p)$, where $\beta_{[k]}$ is the k-th largest factor effects in terms of the absolute value $|\beta_{[k]}|$.

3. The correlation matrix of factors $\boldsymbol{R} = (n-1)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \{r_{ij}\}$ are determined by the followings:
For $p = 3$, the correlation coefficient r_{ij} is randomly determined by uniform random number [0, 1]. For $p = 4$, we put a constraint $r_{13} = r_{14} = r_{23} = r_{24}(= r)$, where r_{12}, r_{34}, r is randomly determined. Furthermore, for $p = 5$ we put a constraint that $r_{13} = r_{14} = r_{15} = r_{23} = r_{24} = r_{25}$, $r_{31} = r_{35} = r_{45}$ and other correlation coefficients are determined uniform random number. The reason of the constraint is that it may generate various $\boldsymbol{R}$ in terms of the value of $|\boldsymbol{R}|$. When the matrix constructed by the random numbers is not a positive definite matrix, the matrix is discarded

for calculation. We generate 10 correlation matrices for each intervals such that $0 \leq |\boldsymbol{R}| \leq 0.1$, $0.1 < |\boldsymbol{R}| \leq 0.2$, ..., $0.9 < |\boldsymbol{R}| \leq 1.0$, so totally 100 correlation matrices.

4. The number of observations, n, is selected from the set $\{10, 20, 50, 100\}$.

The algorithm for calculating of probability consists of two loops. The outer loop determines calculation settings such that n, p, $\boldsymbol{\beta}$ and $\boldsymbol{R}$. The inner loop calculate the probabilities P_E, P_C, P_F by the Monte Calro method where the number of repetition is 10,000 under pre-specified conditions determined by the outer loop.

Figure 1 summarizes the differences of the probabilities of correct selection $P_E - P_C$ and $P_E - P_F$ under the condition of p=3, $\beta_{\max} = 2.0$ and n=20. This figure implies an advantage of Estimates method than the other two methods in terms of correct selection. In order to examine more details, Tables 1 and 2 summarize the differences of probabilities $P_E - P_C$, $P_E - P_F$ with respect to $|\boldsymbol{R}|$ under some conditions. These tables also suggest the advantage of Estimates method in many cases. It is concluded that Estimates method is appropriate for selection of one active factor in terms of correct selection, since the probability of correct selection is generally higher than probabilities of other two methods.

4 Robustness of Estimates method against correlation

4.1 A criterion of robustness

It is well known that correlation between factors in the collected data set deteriorate the precision of estimated parameters in linear regression analysis. We, therefore, examine robustness of Estimates method against the correlation in the data set and derive an evaluating method of the deterioration. We introduce the difference of the probabilities of correct selection:

$$\Delta P_E = P_E\left(\boldsymbol{I}_{p\times p}\right) - P_E\left((n-1)^{-1}\boldsymbol{X}^\top \boldsymbol{X}\right) \tag{7}$$

as a criterion to measure the robustness against correlation. In this criterion, $P_E(\cdot)$ denotes the probability of correct selection by Estimate method under correlation structure in the brackets $(\cdot)$ and $\boldsymbol{I}_{p\times p}$ denotes the $p \times p$ identity matrix. This criterion measures the differences of the probabilities under orthogonal and correlated design matrices.

4.2 Robustness of Estimates method against correlation

The deterioration of the probability by correlation is calculated by a similar Monte Carlo method in the previous section. The result of calculation of ΔP_E is summarized in Tables 3 and 4. These tables imply that the correlation makes the precision worse, in particular small value of $|\boldsymbol{R}|$. For example, the probability of selecting active factors decreases around 30% when $0 \leq |\boldsymbol{R}| \leq 0.1$. On the other hand, the probability of correct selection is not deteriorated so much under small correlation, such as less than 10% under $0.5 \geq |\boldsymbol{R}|$, in general. According to the result of Tables 3 and 4, we need to estimate the deterioration of the probability of Estimate method.

4.3 Estimation of deterioration of probability

In order to derive an estimate of the decrease of the probability of correct selection, we apply an approximation of

$$P_E\left((n-1)^{-1}\boldsymbol{X}^\top\boldsymbol{X}\right) = \Pr\{|\hat{\beta}_\tau| \geq |\hat{\beta}_j| \mid p \in \mathcal{P}\setminus\tau\} \tag{8}$$

to

$$P_E^*\left((n-1)^{-1}\boldsymbol{X}^\top\boldsymbol{X}\right) \approx \prod_{j\in\mathcal{P}\setminus\tau} \Pr\{|\hat{\beta}_\tau| \geq |\hat{\beta}_j|\}, \tag{9}$$

because of the difficulty to describe an explicit form due to the absolute values of the random variables, where $\mathcal{P} = \{1,\ldots,p\}$. The right-hand side of Equation (9) is calculated by using a property that the estimates of regression parameters has p dimensional normal distribution. For example, $\hat{\beta}_\tau - \hat{\beta}_j$ follows 2-dimentional normal distribution with mean $\beta_\tau - \beta_j$, $\beta_\tau + \beta_j$ and variance $V(\hat{\beta}_\tau) + V(\hat{\beta}_j) - 2Cov(\hat{\beta}_\tau, \hat{\beta}_j)$, and so on.

Let

$$\Delta P_E^* = P_E\left(\boldsymbol{I}_{p\times p}\right) - P_E^*\left((n-1)^{-1}\boldsymbol{X}^\top\boldsymbol{X}\right) \tag{10}$$

denotes the approximated difference the probabilities of correct selection. The approximation error of ΔP_E^* to ΔP_E is evaluated by Monte Calro method in the similar conditions with the previous section on $\boldsymbol{R}$, p , $\boldsymbol{\beta}$ and n. Table 5 summarizes the approximation error. This result suggests that the approximation error is at most 15% for all settings and basically less than 5%. It is concluded that the approximation has enough accuracy because the error is less than 5% in general.

Next, we consider estimation of the deterioration of probability by correlation between factors in the given data set. Let $\hat{\Delta}P_E^*$ be an estimate of ΔP_E^* by substituting $\hat{\beta}_j$ and $\hat{\sigma}$ into β_j and σ at ΔP_E^*. The bias $E(\hat{\Delta}P_E^*) - \Delta P_E^*$ and standard deviation $\sqrt{V(\hat{\Delta}P_E^*)}$ are calculated by a similar Monte Carlo method with the previous section. Tables 6 and 7 show the bias and standard deviation of the estimator. These tables show the followings:

1. Tables 6 and 7 indicate that the bias and the standard deviation of estimator are strongly depend on the correlation of the collected data set. Specifically, the bias and standard deviation of are worse when $\boldsymbol{R}$ is closed to zero. This trend is similar with the results in other settings.

2. The bias of the estimator is almost negative although there are some exceptions. In particular, the bias is less that -10% when $|\boldsymbol{R}|$ is closed to zero. This fact implies that a bias correction based on $|\boldsymbol{R}|$ is required in practical situation.

3. The standard deviation also depends on the $|\boldsymbol{R}|$. Therefore, it is also required to evaluate the estimates based on $|\boldsymbol{R}|$.

4. The bias and standard deviation of the estimates slightly depend on n, p and the type of $\boldsymbol{\beta}$.

From the above results, we consider the bias correction and interpretation of precision of estimates based on $|\boldsymbol{R}|$, n and p. For example, the results show that the bias of the estimates of the deterioration is at most 5% for almost cases of $p = 2$ and $|\boldsymbol{R}| > 0.6$.

Furthermore, the standard deviation of the estimates of the deterioration is at most 5% for almost cases of $p = 2$ and $|\boldsymbol{R}| > 0.6$. In other words, the number of factors p and the determinant of the correlation matrix $|\boldsymbol{R}|$ derives bounds for the bias and the standard deviation of the estimates of for any $\boldsymbol{\beta}$ and σ. In the same manner, Tables 6 and 7 derives the bounds of bias and standard deviation for various combinations of n, p and $|\boldsymbol{R}|$. Thus, it enables to adjust the bias and interpret the standard deviation of the estimates of deterioration based on Table 6 and Figure 7. Of course, this is not a strict quantitative way. However, it will be a guideline to use the methods. The bias correction and interpretation of standard deviation of the estimates seems to be conservative, because the bound is determined by p, n and $|\boldsymbol{R}|$, not using $\boldsymbol{\beta}$. The interpretation will be demonstrated in the next section.

5 Examples

(1) Metal bonding process data (Myers (1990))

Myers (1990) listed a data set in an elastomery metal bonding process. The response variable y : is a length of debonding (cm). The factors are x_1: time (min), x_2 voltage (volts), x_3: pH at time of bonding (pH) and x_4: temperature (°F). The correlation matrix of the data $\boldsymbol{R}$ is shown in the upper part of Table 8. Some strong correlations exist, such as $r_{12} = -0.54$ in Table 8. In other words, this table suggests that correlation may deteriorate the estimates. Let us consider a situation to find an active factor out of $x_1, \ldots, x_4$ for reducing variation of y. The estimated regression model based on the data is

$$\hat{y} = 5.670 + 2.237x_1 + 0.406x_2 - 0.683x_3 + 1.396x_4, \tag{11}$$

where $\hat{\sigma} = 2.278$ and the data of factors are standardized to zero mean and unit variance. Since the absolute value of estimated coefficient of x_1 is the maximum, x_1 is selected as an active factor that affect the response variable seriously.

Next, we consider deterioration by correlation of paired factors. The estimates of deterioration is $\hat{\Delta}P_E^* \approx 0\%$ by substituting $\hat{\beta}_j$ and $\hat{\sigma}$ into ΔP_E^*. It implies that we may be able to ignore the deterioration in a sense of point estimation.

Consider the precision of the estimates of deterioration. Since $p = 4$ and $|\boldsymbol{R}| = 0.51$ in this data set, Table 6 and Figure 7 show that the bias and the standard deviation of the estimate are 0% and 3%, respectively. Thus, it is able to ignore the deterioration because the 2 × standard deviation is around 6% that can be regarded as small error.

(2) Weight of pine tree data (Draper and Smith (1981))

Draper and Smith (1981) shows a data set of pine weight and its factors tree such that y and $x_1, \ldots, x_5$. The correlation matrix of the data set is shown in Table 8. The estimated regression model is

$$y = 0.525 + 0.008x_1 + 0.001x_2 - 0.029x_3 - 0.009x_4 + 0.016x_5, \tag{12}$$

where $\hat{\sigma} = 0.018$ and the data set of factor are standardized with zero mean and unit variance as well as the previous example. This equation implies the possibility that x_3 is

the most important factor. The estimates of regression parameters derive $\hat{\Delta}P_E^* = 15\%$ as a point estimation of the deterioration. This deterioration is larger the previous example. Since $p = 5$ and $| \boldsymbol{R} | = 0.15$, Tables 6 and 7 indicates that the bounds for the bias and standard deviation on the deterioration are -3% and 11%, respectively. This result implies that we may not be able to ignore the deterioration by the correlation. Thus the additional data collection, in particular orthogonal data, would be required to derive more precise conclusion.

6 Concluding remarks

In this paper, we consider a method of selecting active factors and its robustness against correlation. The simulation study shows that Estimates method is the best among three methods in many cases. The robustness of Estimates methods is evaluated in terms of deterioration of the probability of correct selection by a simulation study. The result implies that the deterioration of the probability strongly depend on the number of factors and the determinant of correlation matrix of factors. Furthermore, we show a method to evaluate the deterioration of the probability of correct selection. This evaluation method consists of a point estimation of the deterioration of the probability and a bound for the bias and the standard deviation of its estimator. Finally, two examples are discussed to indicate the practical usage of our methods.

References

[1] Box, G. E. P. and Draper, N. R., (1987), *Empirical Model-Building and Response Surfaces*, New York: John Wiley, pp.10–14.

[2] Draper, N. R. and Smith, H., (1982), *Applied Regression Analysis*, 2nd ed., New York: John Wiley, p.404.

[3] Lin, D. K. J., (1993a), A New Class of Supersaturated Designs, *Technometrics*, **35**, pp.28–31.

[4] Lin, D. K. J., (1993b), Another Look at First-Order Saturated Designs: The p-efficient Designs, *Technometrics*, **35**, pp.284–292.

[5] Montgomery, D. C. and Peck, E. A., (1992), *Introduction to Linear Regression Analysis*, 2nd ed., New York: John Wiley, pp.305–365.

[6] Myers, R. H., (1990), *Classical and Modern Regression with Applications*, 2nd ed., Boston: PWS-Kent, p.363.

[7] Plackett, R. L. and Burman, J. P., (1946), The Design of Optimum Multifactorial Experiments, *Biometrika*, **33**, pp.303–325.

[8] Yamada, S., Effects of Correlation Between Explanatory Variables in the Linear Calibration Problem, *Quality*, **27**, pp.117-124. (in Japanese)

Table 1: The differences of the probabilities of correct selection: $P_E - P_C$

		n=10				n=20			
		One factor		Step		One factor		Step	
p	$\|\boldsymbol{R}\|$	$\beta_{\max} = 0.5$	1	1	2	0.5	1	1	2
3	0.1	-0.06	-0.17	-0.03	0.10	-0.10	-0.08	-0.07	0.07
3	0.2	0.02	-0.04	0.27	0.52	-0.09	-0.07	0.31	0.56
3	0.3	-0.04	-0.05	0.06	0.23	-0.02	-0.04	0.38	0.61
3	0.4	-0.03	-0.05	0.04	0.16	-0.03	-0.02	0.10	0.29
3	0.5	-0.06	-0.05	0.19	0.47	-0.04	-0.02	0.13	0.38
4	0.1	-0.18	-0.32	0.00	0.13	-0.12	-0.20	0.04	0.16
4	0.2	-0.12	-0.16	-0.10	0.01	-0.13	-0.09	0.08	0.24
4	0.3	-0.04	-0.08	0.01	0.14	-0.09	-0.06	0.11	0.30
4	0.4	-0.06	-0.06	0.01	0.16	-0.10	-0.04	0.16	0.34
4	0.5	-0.06	-0.05	-0.02	0.05	-0.05	-0.02	0.14	0.22
5	0.1	-0.06	-0.17	-0.08	-0.06	-0.22	-0.25	-0.20	-0.05
5	0.2	-0.11	-0.14	0.00	0.13	-0.15	-0.08	-0.06	0.07
5	0.3	-0.08	-0.10	-0.04	0.06	-0.07	-0.04	0.05	0.18
5	0.4	-0.05	-0.07	0.01	0.10	-0.08	-0.03	0.11	0.25
5	0.5	-0.02	-0.04	-0.04	-0.01	-0.05	-0.02	0.00	0.04

		n=50				n=100			
		One factor		Step		One factor		Step	
p	$\|\boldsymbol{R}\|$	$\beta_{\max} = 0.5$	1	1	2	0.5	1	1	2
3	0.1	-0.28	-0.21	0.29	0.47	-0.23	-0.11	0.26	0.40
3	0.2	-0.10	-0.04	0.47	0.71	-0.06	0.00	0.44	0.54
3	0.3	-0.07	-0.02	0.35	0.54	-0.04	0.00	0.14	0.25
3	0.4	-0.05	0.00	0.25	0.42	-0.02	0.00	0.48	0.59
3	0.5	-0.03	0.00	0.35	0.49	-0.01	0.00	0.26	0.30
4	0.1	-0.27	-0.20	0.19	0.35	-0.24	-0.14	0.22	0.32
4	0.2	-0.12	-0.03	0.37	0.60	-0.09	-0.01	0.35	0.43
4	0.3	-0.07	-0.01	0.03	0.19	-0.03	0.00	0.42	0.54
4	0.4	-0.06	0.00	0.05	0.13	-0.02	0.00	0.39	0.46
4	0.5	-0.03	0.00	0.17	0.26	-0.01	0.00	0.48	0.59
5	0.1	-0.24	-0.12	0.25	0.49	-0.21	-0.07	0.19	0.36
5	0.2	-0.12	-0.02	0.13	0.31	-0.07	0.00	0.22	0.35
5	0.3	-0.09	-0.01	0.15	0.30	-0.03	0.00	0.25	0.25
5	0.4	-0.05	0.00	0.12	0.20	-0.02	0.00	0.28	0.37
5	0.5	-0.04	0.00	0.21	0.31	-0.01	0.00	-0.06	0.00

Table 2: The differences of the probabilities of correct selection: $P_E - P_F$

		n=10				n=20			
		One factor		Step		One factor		Step	
p	$\|\boldsymbol{R}\|$	$\beta_{\max} = 0.5$	1	1	2	0.5	1	1	2
3	0.1	0.00	-0.03	-0.01	0.07	0.01	0.03	0.07	0.26
3	0.2	0.07	0.05	0.16	0.31	-0.03	-0.01	0.05	0.22
3	0.3	0.01	0.01	0.01	0.09	0.05	0.02	0.17	0.26
3	0.4	0.01	0.01	0.03	0.06	0.02	0.01	0.10	0.18
3	0.5	-0.04	-0.02	-0.06	-0.04	0.00	0.00	-0.01	0.01
4	0.1	-0.05	-0.07	0.03	0.15	0.06	0.04	0.14	0.30
4	0.2	-0.07	-0.07	-0.11	-0.09	-0.04	-0.02	-0.01	0.08
4	0.3	0.00	0.00	-0.01	-0.01	-0.01	0.00	0.03	0.09
4	0.4	-0.02	-0.01	-0.03	0.00	-0.06	-0.02	-0.10	-0.03
4	0.5	-0.03	-0.02	-0.05	-0.05	-0.02	0.00	-0.04	0.00
5	0.1	0.06	0.05	0.05	0.08	-0.04	-0.03	0.01	0.10
5	0.2	-0.06	-0.05	-0.08	-0.04	-0.07	-0.03	-0.13	-0.09
5	0.3	-0.03	-0.03	-0.07	-0.05	0.00	0.00	0.03	0.09
5	0.4	0.00	0.00	0.03	0.06	-0.03	-0.01	-0.01	0.04
5	0.5	0.02	0.02	0.03	0.05	-0.01	0.00	0.00	0.02

		n=50				n=100			
		One factor		Step		One factor		Step	
p	$\|\boldsymbol{R}\|$	$\beta_{\max} = 0.5$	1	1	2	0.5	1	1	2
3	0.1	0.04	0.16	-0.11	-0.07	-0.05	-0.02	0.08	0.17
3	0.2	0.11	0.27	-0.01	-0.01	-0.02	0.00	0.06	0.15
3	0.3	0.15	0.24	0.01	0.00	-0.01	0.00	0.10	0.18
3	0.4	0.03	0.15	-0.01	0.00	0.00	0.00	0.13	0.20
3	0.5	0.04	0.10	-0.01	0.00	0.00	0.00	0.09	0.09
4	0.1	0.00	0.00	0.20	0.29	-0.01	0.00	0.13	0.22
4	0.2	-0.01	0.00	0.09	0.19	0.00	0.00	0.16	0.22
4	0.3	0.00	0.00	0.07	0.14	0.00	0.00	0.16	0.23
4	0.4	-0.01	0.00	0.04	0.08	0.01	0.00	0.13	0.09
4	0.5	0.01	0.00	0.04	0.07	0.00	0.00	0.05	0.04
5	0.1	-0.02	0.00	0.12	0.30	0.01	0.01	0.22	0.38
5	0.2	-0.02	0.00	0.01	0.07	-0.01	0.00	0.02	0.05
5	0.3	-0.01	0.00	-0.04	-0.03	0.01	0.00	0.12	0.16
5	0.4	0.01	0.00	0.07	0.09	0.00	0.00	0.02	0.04
5	0.5	0.01	0.00	0.05	0.07	0.00	0.00	0.01	0.01

Table 3: Deterioration of the probability, ΔP_E, to select an active factors brought by correlation among factors ($n = 10, 20$)

		n=10				n=20			
		One factor		Step		One factor		Step	
p	$\|\boldsymbol{R}\|$	$\beta_{\max} = 0.5$	1	1	2	0.5	1	1	2
2	0.1	0.19	0.34	0.14	0.21	0.29	0.33	0.18	0.25
2	0.2	0.14	0.24	0.06	0.05	0.22	0.19	0.05	0.08
2	0.3	0.12	0.18	0.05	0.06	0.18	0.13	0.05	0.08
2	0.4	0.09	0.13	0.09	0.14	0.14	0.08	0.12	0.16
2	0.5	0.07	0.10	-0.01	-0.01	0.11	0.05	-0.01	0.01
3	0.1	0.16	0.34	0.14	0.24	0.29	0.31	0.20	0.26
3	0.2	0.15	0.24	0.12	0.16	0.23	0.17	0.14	0.15
3	0.3	0.14	0.17	0.12	0.12	0.18	0.09	0.12	0.10
3	0.4	0.10	0.14	0.08	0.11	0.14	0.08	0.10	0.11
3	0.5	0.08	0.10	0.06	0.10	0.10	0.04	0.09	0.09
4	0.1	0.20	0.39	0.19	0.28	0.32	0.34	0.25	0.28
4	0.2	0.17	0.22	0.14	0.16	0.22	0.13	0.15	0.14
4	0.3	0.13	0.15	0.09	0.09	0.17	0.08	0.10	0.09
4	0.4	0.09	0.11	0.07	0.07	0.12	0.05	0.08	0.05
4	0.5	0.06	0.08	0.07	0.06	0.09	0.04	0.07	0.05
5	0.1	0.18	0.36	0.22	0.24	0.31	0.28	0.24	0.22
5	0.2	0.11	0.22	0.13	0.13	0.19	0.14	0.14	0.11
5	0.3	0.09	0.14	0.10	0.07	0.14	0.06	0.09	0.06
5	0.4	0.06	0.10	0.08	0.04	0.10	0.04	0.06	0.02
5	0.5	0.05	0.08	0.05	0.03	0.09	0.03	0.05	0.01

Table 4: Deterioration of the probability, ΔP_E, to select an active factors brought by correlation among factors ($n = 50, 100$)

		n=50				n=100			
		One factor		Step		One factor		Step	
p	$\lvert\boldsymbol{R}\rvert$	$\beta_{\max} = 0.5$	1	1	2	0.5	1	1	2
2	0.1	0.35	0.24	0.23	0.25	0.31	0.17	0.25	0.21
2	0.2	0.23	0.08	0.06	0.09	0.16	0.02	0.09	0.06
2	0.3	0.17	0.04	0.07	0.08	0.10	0.01	0.09	0.05
2	0.4	0.12	0.01	0.15	0.11	0.06	0.00	0.15	0.05
2	0.5	0.09	0.01	0.00	0.02	0.04	0.00	0.02	0.01
3	0.1	0.34	0.19	0.25	0.23	0.27	0.11	0.26	0.16
3	0.2	0.22	0.05	0.16	0.11	0.13	0.01	0.14	0.06
3	0.3	0.14	0.02	0.11	0.06	0.06	0.00	0.09	0.02
3	0.4	0.12	0.01	0.11	0.07	0.05	0.00	0.10	0.02
3	0.5	0.08	0.00	0.10	0.04	0.03	0.00	0.09	0.01
4	0.1	0.39	0.20	0.29	0.22	0.31	0.11	0.27	0.15
4	0.2	0.19	0.03	0.15	0.08	0.09	0.00	0.13	0.03
4	0.3	0.13	0.01	0.10	0.05	0.05	0.00	0.08	0.02
4	0.4	0.09	0.00	0.07	0.02	0.03	0.00	0.05	0.00
4	0.5	0.07	0.00	0.06	0.02	0.02	0.00	0.04	0.00
5	0.1	0.34	0.12	0.23	0.15	0.23	0.04	0.21	0.09
5	0.2	0.20	0.03	0.13	0.06	0.10	0.01	0.10	0.02
5	0.3	0.11	0.00	0.07	0.02	0.04	0.00	0.05	0.00
5	0.4	0.08	0.00	0.03	0.00	0.02	0.00	0.02	0.00
5	0.5	0.07	0.00	0.02	0.00	0.02	0.00	0.01	0.00

Table 5: The approximation error of the probability of correct selection.

p	$\|\boldsymbol{R}\|$		$n = 10$	20	50	100
3	0.1	ave	0.08	0.08	0.06	0.05
3	0.1	worst	0.13	0.13	0.13	0.12
3	0.2	ave	0.05	0.04	0.02	0.02
3	0.2	worst	0.10	0.10	0.09	0.09
3	0.3	ave	0.04	0.03	0.02	0.01
3	0.3	worst	0.08	0.08	0.07	0.07
3	0.4	ave	0.03	0.02	0.01	0.01
3	0.4	worst	0.08	0.07	0.05	0.06
3	0.5	ave	0.03	0.02	0.01	0.01
3	0.5	worst	0.06	0.07	0.05	0.05
4	0.1	ave	0.13	0.13	0.10	0.08
4	0.1	worst	0.23	0.23	0.23	0.21
4	0.2	ave	0.07	0.06	0.03	0.02
4	0.2	worst	0.16	0.15	0.13	0.14
4	0.3	ave	0.05	0.04	0.02	0.02
4	0.3	worst	0.12	0.11	0.09	0.12
4	0.4	ave	0.04	0.02	0.01	0.01
4	0.4	worst	0.09	0.08	0.06	0.08
4	0.5	ave	0.03	0.02	0.01	0.01
4	0.5	worst	0.07	0.06	0.05	0.07
5	0.1	ave	0.12	0.11	0.08	0.05
5	0.1	worst	0.23	0.23	0.23	0.19
5	0.2	ave	0.09	0.07	0.04	0.03
5	0.2	worst	0.19	0.19	0.18	0.16
5	0.3	ave	0.04	0.03	0.02	0.01
5	0.3	worst	0.10	0.09	0.08	0.10
5	0.4	ave	0.03	0.02	0.01	0.01
5	0.4	worst	0.08	0.07	0.06	0.08
5	0.5	ave	0.03	0.02	0.01	0.01
5	0.5	worst	0.07	0.06	0.05	0.07

Table 6: Bias on the estimator of deterioration of the probability of correct selection brought by correlation between factors.

p	$\|\boldsymbol{R}\|$	n=10	20	50	100
2	0.1	-0.22	-0.22	-0.19	-0.14
2	0.2	-0.11	-0.09	-0.05	-0.03
2	0.3	-0.08	-0.06	-0.03	-0.02
2	0.4	-0.08	-0.05	-0.03	-0.01
2	0.5	-0.03	-0.02	-0.01	0.00
2	1.0	-0.01	-0.01	0.00	0.00
3	0.1	-0.15	-0.14	-0.10	-0.07
3	0.2	-0.09	-0.06	-0.03	-0.01
3	0.3	-0.05	-0.03	-0.01	0.00
3	0.4	-0.04	-0.02	-0.01	-0.01
3	0.5	-0.02	-0.01	0.00	0.00
3	1.0	-0.01	0.00	0.00	0.00
4	0.1	-0.14	-0.11	-0.07	-0.04
4	0.2	-0.05	-0.02	0.00	0.00
4	0.3	-0.03	-0.01	0.00	0.00
4	0.4	-0.01	0.00	0.00	0.00
4	0.5	-0.01	0.00	0.00	0.00
4	1.0	0.00	0.00	0.00	0.00
5	0.1	-0.10	-0.07	-0.02	0.00
5	0.2	-0.03	-0.01	0.01	0.01
5	0.3	-0.02	0.00	0.00	0.00
5	0.4	-0.01	0.00	0.00	0.00
5	0.5	0.00	0.00	0.00	0.00
5	1.0	0.00	0.00	0.00	0.00

Table 7: Standard deviation of the estimator of deterioration of the probability of correct selection brought by correlation between factors.

p	$\|\boldsymbol{R}\|$	n=10	20	50	100
2	0.1	0.19	0.19	0.18	0.17
2	0.2	0.12	0.11	0.10	0.08
2	0.3	0.10	0.09	0.08	0.06
2	0.4	0.09	0.08	0.07	0.05
2	0.5	0.06	0.05	0.04	0.03
2	1.0	0.03	0.03	0.02	0.01
3	0.1	0.19	0.19	0.17	0.15
3	0.2	0.12	0.11	0.10	0.08
3	0.3	0.09	0.08	0.06	0.05
3	0.4	0.08	0.07	0.06	0.04
3	0.5	0.06	0.06	0.04	0.03
3	1.0	0.03	0.02	0.02	0.01
4	0.1	0.20	0.20	0.18	0.16
4	0.2	0.11	0.10	0.08	0.06
4	0.3	0.08	0.07	0.06	0.04
4	0.4	0.06	0.05	0.04	0.03
4	0.5	0.05	0.04	0.03	0.02
4	1.0	0.02	0.02	0.01	0.01
5	0.1	0.17	0.16	0.14	0.12
5	0.2	0.11	0.11	0.09	0.07
5	0.3	0.06	0.06	0.04	0.03
5	0.4	0.04	0.04	0.03	0.02
5	0.5	0.04	0.04	0.02	0.02
5	1.0	0.02	0.02	0.01	0.01

Table 8: Correlation matrices on the data sets

(1)Metal bonding process data (Myers (1990)

	x_1	x_2	x_3	x_4
x_1	1.000	-0.536	0.229	-0.274
x_2	-0.536	1.000	-0.393	-0.047
x_3	0.229	-0.393	1.000	0.137
x_4	-0.274	-0.047	0.137	1.000

(2)Weight of pine tree data (Draper and Smith (1982))

	x_1	x_2	x_3	x_4
x_1	1.000	0.763	0.364	-0.240
x_2	0.763	1.000	0.569	-0.263
x_3	0.364	0.569	1.000	-0.216
x_4	-0.240	-0.263	-0.216	1.000

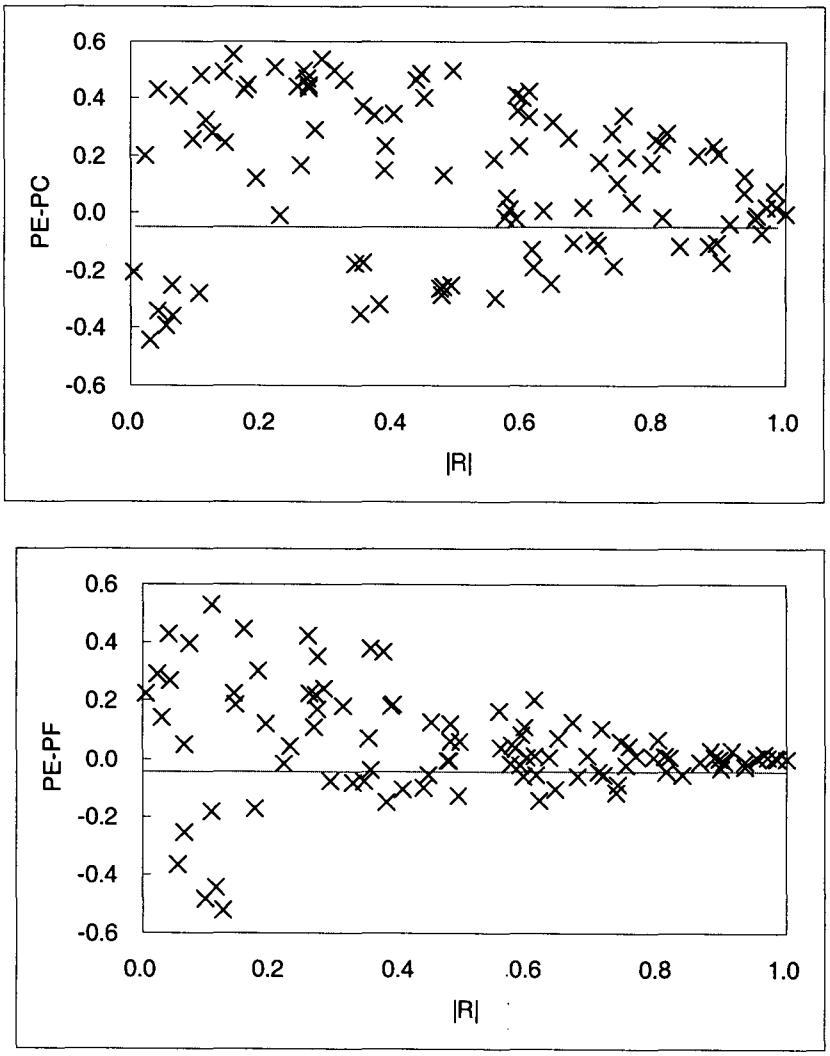

Figure 1: Differences of the probability of correct selection $(p = 3, \beta_{\max} = 2.0, n = 20)$