Prediction Models for the In-control ARL in Phase II of a T^2 Control Chart

by

Wilfredo F. Yushimito-Del Valle

A thesis submitted in fulfillment of the requirements for the degree of

Master of Science
in
Industrial Engineering

University of Puerto Rico
Mayagüez Campus
2006

Approved by:

_____	_____
Edgar Acuña, Ph.D. Member, Graduate Committee	Date
_____	_____
David González-Barreto, Ph.D. Member, Graduate Committee	Date
_____	_____
Noel Artiles-León, Ph.D. President, Graduate Committee	Date
_____	_____
Loida Rivera-Betancourt, Ph.D. Representative of Graduate Studies	Date
_____	_____
Agustín Rullán-Toro, Ph.D. Chairperson of the Department	Date

ABSTRACT

Through simulation, an analysis of the run-length in Phase II reveals that the commonly used formula for the upper control limit (UCL) of Hotelling's T^2 control chart, proposed by Alt (1976), is not exact. This study also shows that the in-control ARL depends on: (1) the total samples used to estimate parameters of Phase I, (2) the condition number of the estimated correlation matrix and (3) the desired in-control ARL.

This research provides regression models, based on the condition number of the estimated correlation matrix, number of samples, number of variables and the desired in-control ARL, for the prediction of the in-control average run length (ARLo) in Phase II of T^2 Control Charts.

When large samples are not available, the regressions can be used to correct the UCL to achieve values that are more exact. The corrections obtained are more conservative than the ones recently obtained using simulation by Champ et al. (2005).

RESUMEN

El “run-length” es analizado a partir de los resultados de una simulación. Esta simulación muestra que el límite de control tradicional propuesto por Alt (1976) no es exacto. Este estudio también muestra que el “average run-length” en control depende de: (1) el número total de muestras usadas para estimar los parámetros en la fase I, (2) el número de condición de la matriz de correlación estimada y (3) el valor del “average run-length” en control que se desea obtener.

Este trabajo presenta modelos de regresión para la predicción del “average run-length” (ARLo) en control de la fase II para las gráficas de control T^2. Los modelos están basados en el número de condición de la matriz de correlación estimada, el número total de muestras usadas en la fase I, el número de variables y el valor objetivo del “average run-length” en control.

Cuando no se dispone de un gran número de muestras, las regresiones que se proponen en este estudio, llevan una corrección en el límite de control (UCL). La corrección resultante es mas conservadora que la propuesta recientemente por Champ et al. (2005).

Es en ese espíritu nacional, pero el positivo, el que se abre al mundo, el que se cuestiona, el que tolera, el que abraza, el que integra, el que aplaude el éxito; y no el nacionalismo que se lamenta, que condena, el que divide, el que se encierra y protege la mediocridad, en donde finalmente los peruanos alcanzaremos el rostro definitivo de nuestra nación y con él, finalmente, la tan ansiada prosperidad.
(Gastón Acurio en el inicio del año académico en la Universidad del Pacífico – Lima, 30 de marzo 2006)

Different context, different narrative, different imagination.
(Thomas Friedman in "The World is Flat").

To my wife Ruth for filling my life with love, sanity, opportunity and industry

ACKNOWLEDGEMENTS

I would like to thank to my advisor, Noel Artiles-León, for his guidance in this two years. He was not only an advisor but a friend and an excellent teacher in all the sense of the word. “Office hours” at Buffallo’s has been the best time to share semi-philosophical discussions. Also thanks to my committee: Professor David González and Professor Edgar Acuña. I appreciate the time spent in the review of this work and your recommendations and comments. I also want to thank to the Industrial Engineering department for the support, in special to Mayra, Laura, Edwin, Jackie, Professor William Hernandez and Professor Fethi Calisir and his wife Ferah.

A few people who shared the past two years with me deserve a special mention. In first place, Andrés Uribe has been a great friend, always helpful and supportive. Joel Rivera has been another great friend. He generously gave me advices to make my Matlab code more efficient and let me use the Robotica Lab where most of this thesis was conceived and written. My friends and mates Juan Guillermo Gómez, Jesús Gómez, Efraín Montes, Giselle Castro, Liliana Martínez, Verónica Rodríguez, Geovanie Galán and my roommate Alfredo Cuello. All of them beautiful people with big hearts.

In the past year, I was lab instructor of four Inin 4078 sections, a big thank-you to each one of my now former students.

There are many people in Lima (or wherever they are now) that I would like to thank. Ana María Reátegui in my worst years looked after me and “no dejó que me pierda”. My

big brother Alex Huerta-Mercado Tenorio and my friends Sandro Paz and Percy Summers encouraged me to start graduate studies. My friends Juan Pablo Higuchi, Cecilia Pomar, Iván Namihas, Mario Mizushima and Eliana Chávez, have been supportive friends and, all of them, through e-mails and the messenger, make me feel as if I never left them. Gracias a Carlos Montes, Laura Salas y Marco Purizaca porque aún a la distancia seguimos brindando por la victoria, el empate y el fracaso.

Also, thanks to Pedro Reátegui-Roselló and his family, for the summer of 2001 in Playa Blanca when I started to believe in myself.

Finally, I want to thank to my parents Wilfredo and Gloria, my brother Carlos and my sister Silvia for their love and their support in all my professional, personal and academic life.

Table of Contents

Table List

Tables **Page**

Figure List

Figures **Page**

1 Introduction

1.1 Motivation and General Purpose

Multivariate quality control charts are statistical techniques used to monitor two or more quality characteristics at the same time. The most popular multivariate control chart is the T^2 control chart used when the true population values are unknown and the underlying distribution is multivariate normal. This control chart is based on the Hotelling's T^2 statistic. The chart signals an out-of-control signal when the calculated statistic T^2 for the process exceeds the upper control limit (UCL). These concepts are discussed in detail in Chapter 2.

The run-length (RL) is the number of samples before a control chart signals an out-of-control signal. The average run-length (ARL) refers to the expectation of the RL. For simple control schemes, the ARL can be computed as

$$ARL = \frac{1}{p} \tag{1.1}$$

where p is the probability of any point plots outside the control limits. When the chart signals a false alarm (out of control signal when process actually is in control), there is a Type I Error (α) and the ARL can be calculated using Equation (1.2), and it is called in-control ARL (ARL_0). The probability of a true out of control signal can be calculated as 1- Pr (Type II Error) (or 1-β) and the ARL (called out-of-control ARL or ARL_1) is calculated using Equation (1.3).

$$ARL_0 = \frac{1}{\alpha} \tag{1.2}$$

$$ARL_1 = \frac{1}{1-\beta} \tag{1.3}$$

Simulations presented in the following chapters show that, for the Hotelling's T^2 control chart, the in-control ARL, shows a large deviation from the theoretical in-control

ARL when the parameters are estimated from small samples confirming the problem stated by Champ et al. (2005). However, Champ et al. (2005) did not noted that deviations seem to depend on how well the covariance matrix is estimated. This not only affects the expected value but also the probability of getting early false alarms.

This is an important problem to consider because in the industry large samples are not usually available, also early false alarms makes the control chart mistrust able.

This thesis is focused in finding prediction models for the in-control ARL with parameters estimated of multivariate control charts. The idea is to find and provide a model that helps the industry to predict the in-control ARL especially when large samples are not available.

This work has also identified that one source of this problem is the condition number of the estimated correlation matrix. Based in this factor, the models lead to correction in the upper control limit (UCL). The corrections are compared to the ones obtained by Champ et al. (2005) using simulation, resulting that the regression's ones are more conservatives.

1.2 Literature Review

The first introduction of an extension of the $\bar{x}$ control chart was presented, as noted by Hauck, Runger and Montgomery (1999), by Shewhart in 1931. Assuming that there are p variables with a multivariate normal distribution, this chart signals if there is significant change in the mean or equivalently if

$$\chi_i^2 = (X_1 - \mu_0)\Sigma^{-1}(X_1 - \mu_0)' > UCL \tag{1.4}$$

where UCL is the control limit, μ_0 and Σ are the in-control values, which mostly are unknown. When these parameters are estimated by $\overline{X}$ and S, expression (1.4) results in the Hotelling's T^2 statistic, which, as will be explained in Section (2.2), follows an F distribution.

The traditional control limit for the T^2 control chart for subgroups was established by Alt in his master's thesis in 1973 (Ryan, 2000) and then published in 1976 (Alt, 1976). These limits are presented in Section 2.3 (Equations 2.8 and 2.23). He also proposed two steps, the retrospective phase (Phase I) and the future observations phase or Phase II (see Section 2.3).

The literature in multivariate control charts has been focused, in the Phase I, in obtaining in-control parameters after finding and eliminating out-of-control subgroups and in finding the cause of the out-of-control signal in Phase II.

For example, Sullivan and Woodall (1996) focused in the estimation of the covariance matrix for a process with *individual observations*. For this case, considering that there can be special causes (i.e.: a shift or a trend) in the historical data, those cannot be detected by the common Hotelling's T^2 control chart because of the use of the pooled covariance matrix. The authors tested many alternatives of covariance matrix estimation and finally recommended an alternative analogous to the moving range in the univariate case (difference of successive observations). This estimator was evaluated then by Vargas (2003) trying to find outliers in the data for Phase I. He proposed the use of robust estimators for the mean and covariance matrix. He compared the results from six methods: the common covariance matrix, pooled covariance matrix, the minimum volume ellipsoid (MVE), the minimum covariance determinant (MCD), and two methods proposed by Sullivan and Woodall (1996), the one described previously and one outlier detection method based in Atkinson and Mulira (1993). His simulation study showed that the T^2 control chart using MVE is more effective in the detection of outliers but he also recommends Sullivan and Woodall's method to detect shifts in the mean vector.

For Phase II, most papers are concentrated in finding the cause of the deviation from the in-control state. Montgomery's (2001) and Yai and Trewn's (2003) books provide a collection of methods to find the cause of the out-of-control signals. Between the methods mentioned are discrimination analysis, principal components and statistic decomposition.

1.2.1 The problem of parameters estimation and ARL

The ARL has been in discussion in recent years, Ryan (2000) states that the reasons are (1) the ARL is not a typical run-length (because its skewness) and (2) the standard deviation is large. This problem seems to be bigger when parameters are estimated.

Quesenberry (1993) has discussed the problem of parameter estimation on the ARL for univariate control charts. If the parameters are estimated, control limits change with the estimation of the mean or standard deviation and this leads to disturb the run-length. In general, the conclusion of parameter estimation in univariate control charts is that the ARLs are overestimated, but, even if the parameters are estimated or not, the standard deviation of the run-length is always disturbing.

Ryan (2000) noted that the problem of estimation parameters in the ARL, is also extended to all multivariate control charts as the variance-covariance matrix and the mean vector must be estimated, but the effect is not easy to describe. He cites an unpublished work of Bodden and Rigdon that explain: (1) that large samples are needed for T^2 control chart for individuals to perform as known parameters unless p (number of variables is small) and (2) the effect of overestimation reduces the in-control ARL more than underestimation increases the in-control ARL.

Other approaches to solve the problem can be found in the correction of the UCL. The next section discusses some the authors that have covered the problem from the sample size point of view.

1.2.2 Samples and UCL correction

Neduraman, G, and Pignatiello Jr., J. J. (2000) discuss the problem that the approximation of the UCL proposed by Alt (1988) in Phase I is not exact. They develop a Montecarlo experiment to prove that the number of false alarms is not exactly the same as the one that can be achieved given a Type Error I. They propose the use of

Bonferroni's adjustment in which the false alarm probability was set at α / m where α is the overall false alarm probability for all m initial subgroups which results to obtain better results. They found that between $800p/3(n-1)$ and $400p/3(n-1)$ subgroups of size n, the T^2 with estimated parameters performs like if the parameters were known using the χ^2 control limit.

Another similar work was done by Lowry and Montgomery (1995); they find the best combination of subgroups (m and n) and number of variables (p) for which the UCL calculated by Equation (2.23) is within 10% of the limits as if the parameters were known using the approximation $(1-\alpha)^{th}$ percentile of the χ^2 distribution with p degrees of freedom. This work is important because it implies that the approximation of UCL to the percentile of χ^2 distribution for the χ^2 control charts is related to the number of variables being monitored and the subgroup size.

A recently published paper by Champ et al. (2005) presents the problem that is described in Chapter 3: Alt's approximation does not lead to the exact in-control ARL of $1/\alpha$. Their proposal is that the in-control run-length does not depend on the unknown parameters. They show that large number of samples is necessary to Alt's UCL approximation to the F-distribution gets the exact in-control ARL; the total number of samples, mn, ranges from 900 – 550 for p from 2 to 10. Also by simulation, they present corrections to the UCL to achieve correctly the in-control ARL with a standard error of 2% using combinations of few sample numbers of small size.

1.3 Organization

After this introduction, basic concepts and the notation necessary to understand the problem are presented in Chapter 2. The third chapter establishes the problem. Chapter 4 presents the experiment and some additional concepts including the effect of the condition number on the T^2. The fifth chapter has the analysis of the results and the models based in multiple regression. Chapter 6 presents the procedure for determining

the in-control ARL using the models depending on the condition number with an example and a comparison with the results of Champ et al (2005). The last chapter establishes final conclusions and guidelines for future work.

Additional to the chapters, this work has attached 8 appendices. Appendix A one has information about the condition number and also presents a procedure to find matrices with a predefined condition number. Appendix B has the evaluation of the proposed design, The fractional designs used in the simulation for p = 4, 5 and 6 are shown in Appendix C. Appendix D shows the sample size necessary to a better estimation of the correlation, the fourth appendix has tables with information about the percentiles of the run-length distribution. Appendix E has the tables containing the percentiles of the run-lengths obtained by simulation. The Box-Cox transformation procedure is presented in Appendix F, and the details of the regression models are presented in Appendix G and H, and the last appendix (I) has the Matlab code used in the simulation.

2 Basic Concepts

This chapter discusses several concepts: It starts with an introduction to the concept of control chart, multivariate hypothesis testing and the T^2 statistic, and concludes with multivariate control charts as an extension of multivariate hypothesis testing. The notation, necessary for the following chapters, is also covered.

2.1 An Introduction to Control Charts

A control chart is a tool used in statistical quality control (SQC) to monitor quality characteristics in a process. Control charts can be classified as univariate control charts or multivariate control charts, depending on the number of quality characteristics of interest.

In general, control charts use fixed size samples taken at fixed intervals of time or can use individual observations. For both cases, a quality characteristic(s) is(are) inspected and, with this(these) value(s), a statistic is calculated. This statistic is used to perform a hypothesis test to determine whether the process appears to be "in control".

"In control" is defined as the expected or desired value of the variable being monitored with a determined variation. If, one or more variables appear to deviate from its desired value, or if the variation in one or more of the variables seems to have increased, the process is considered out of control. Specifically, univariate control charts controls changes in the mean or variance of the process, and multivariate control charts monitors not only the mean vector but also the correlation structure of the variables. The detection of a change in the mean vector or correlation structure of the chart variables is done through control limits. Control limits are set in the attempt to satisfy:

- A low probability that the chart signals when the statistic falls outside of the control limits when the process is in control or probability of a Type I error (α).

- A high probability that the chart signals when the process is out of control or probability of a Type II error (β).

The value of α determines the in-control average-run-length (ARL), see Equation (1.2), which is the expected number of samples until a signal occurs. A signal in the in-control case is a false alarm. Therefore, it is desirable to have a large in-control ARL.

The value of β determines the out-of-control ARL (see Equation 1.3). A signal in the out-of-control case should occur quickly. Therefore, a very low out-of-control ARL is desirable. The in-control ARL and the out-of-control ARL are the key performance measures of both univariate and multivariate control charts. These are the most common efficiency performance measurement used to evaluate control charts.

Multivariate control charts have the advantage that are able to monitor multiple quality characteristics simultaneously for both changes in the mean value and the correlation structure while maintaining a lower probability of Type I error. Instead of using a multivariate control chart, it may seem reasonable to maintain a separate univariate control chart for each quality characteristic. However this is inappropriate. Each of the separate univariate charts will have its own characteristic value. When the charts are aggregated, the overall probability of a Type I error increases significantly, decreasing the in-control ARL. A multivariate chart uses a multivariate distribution to set the control limits.

Multivariate control charts disadvantage lays in their difficulty to identify which subset of the quality characteristics are responsible for a signal since any single characteristic or combination of characteristics could have experienced a shift in mean value, variance, or correlation. This makes it hard to gain insight from a signal in a multivariate chart that would lead to the determination of the source of change in the process behavior. Commonly used multivariate control charts are the χ^2 chart, for a known covariance matrix, and Hotelling's T^2 chart, for an unknown covariance matrix.

2.2 Multivariate Hypothesis Tests and T^2 Statistic

Consider the following test of hypothesis:

$$H_0 : \mu = \mu_0 \qquad vs. \qquad H_1 : \mu \neq \mu_0$$

If $\{X_1, X_{21}, ..., X_n\}$ is a random sample from a normal population. The test statistic is

$$t = \frac{\overline{X} - \mu_0}{s / \sqrt{n}} \tag{2.1}$$

where

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j \tag{2.2}$$

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n} (X_j - \overline{X})^2 \tag{2.3}$$

The t-statistic follows a *t-student* distribution with *n-1* degrees of freedom (d.f.) and we reject H_0 if $| t |$ exceeds a critical value t for a significance level of α. Considering that rejecting H_0 for large values of $| t |$ is equivalent to reject H_0 for large values of t^2, Equation (2.1) can also be expressed as

$$t^2 = \frac{(\overline{X} - \mu_0)^2}{s^2 / n} = n(\overline{X} - \mu_0)(s^2)^{-1}(\overline{X} - \mu_0) \tag{2.4}$$

Equation (2.4) can be generalized to p variables becoming

$$T^2 = (\overline{X} - \mu_0)'(\frac{1}{n}S)^{-1}(\overline{X} - \mu_0) = n(\overline{X} - \mu_0)'S^{-1}(\overline{X} - \mu_0) \tag{2.5}$$

where

$$\overline{X} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ ... \\ \overline{x}_p \end{bmatrix}, \qquad \mu_0 = \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ ... \\ \mu_{p0} \end{bmatrix} \text{ and}$$

S is the nonnegative definite sample variance-covariance matrix and n is the sample size for each calculated $\bar{x}$.

If one assumes that $X \sim N_p(\mu, \sum)$ is a random variable with a multivariate normal distribution then the covariance matrix S is proportional to $X^T X$ which follows a Wishart distribution $W_p(n,\Sigma)$, hence $S \sim W_p(n, \sum)$. If X and S are independent, then the probability distribution of T^2 when $\mu = \mu_0$ (simultaneously tested) is the **Hotelling's T^2 distribution**

$$T^2 \sim \frac{(n-1)p}{n-p} F_{p,n-p}$$

where $F_{p,n-p}$ is a random variable with an *F-distribution* with p and n-p d.f.; and, as in the univariate case, if the observed T^2 statistics is "too" large the null hypothesis ($H_0 : \mu = \mu_0$) is rejected.

2.3 Multivariate Control Charts

Hotelling's chart uses Hotelling's T^2 statistic to monitor several variables simultaneously with a specified probability of Type I error. In the multivariate case, p variables are desired to be controlled. Hotelling's T^2 statistic is used when the covariance matrix of the correlated quality characteristics variables is assumed to be unknown. When the covariance matrix is assumed to be known the χ^2 statistic is used. As Montgomery (2001) points out, Hotelling's T^2 control chart is the analogous of the Shewhart $\bar{x}$ chart for multiple characteristics with true population values unknown. It is also assumed that the joint distribution of these p variables is the p-variate normal distribution. There are two versions of this control chart, one for individual observations and other for grouped data. For both models, Alt (1976) and then Alt and Smith (1988) defined two phases.

2.3.1 Phase I: Estimation of process parameters

Phase I, also known as ***retrospective analysis***, is used to establish control. Its objective is to obtain an in-control set of observations to estimate process parameters and to calculate the control limits for Phase II. The procedure for this Phase is shown in Figure 2.3-1.

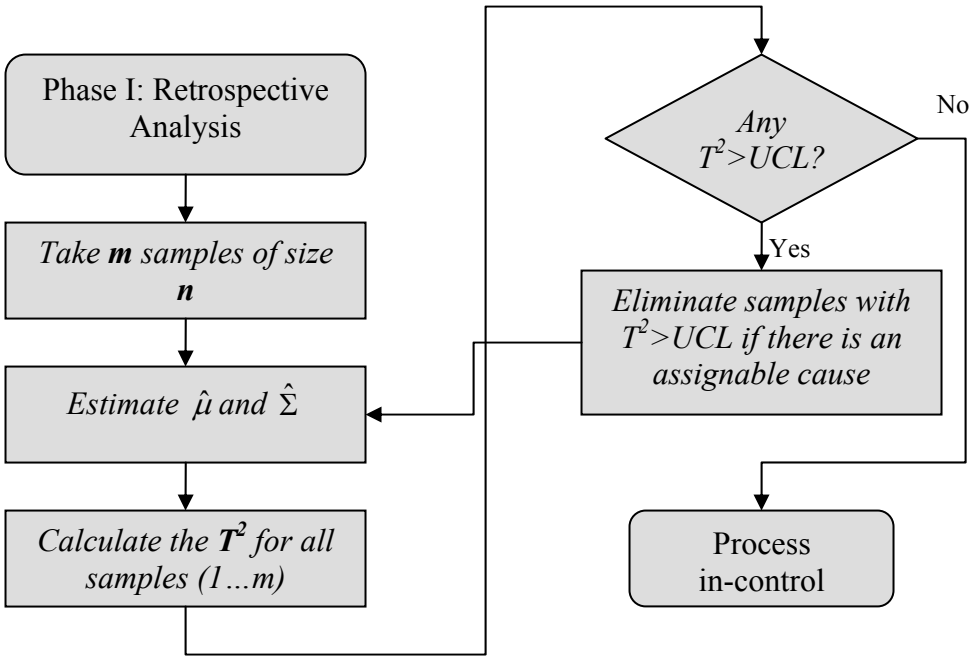


Figure 2.3-1. Phase I: Retrospective Analysis Procedure

In this phase the important part is **the estimation of the in control parameters** μ_0, Σ_0. A few recommendations for determining in-control parameters can be found in Alt and Smith (1988) and Duncan (1974). The former states that μ_0 and Σ_0 can be evaluated from a large amount of historic data. The latter proposes that those parameters can be targets selected by management.

The control limits for Phase I are different for grouped data and for individual data. Equations (2.8) and (2.9) show the control limits for the subgrouped data case. Control limits for the individual observations case are given by Equations (2.10) and (2.11).

$$UCL = \frac{p(m-1)(n-1)}{mn-m-p+1} F_{\alpha,p,mn-m-p+1} \qquad \text{(2.8) and (2.9)}$$
$$LCL = 0$$

$$UCL = \frac{(m-1)^2}{m} \beta_{\alpha,p/2,(m-p+1)/2} \qquad \text{(2.10) and (2.11)}$$
$$LCL = 0$$

where: m is the number of samples, n is the subgroup size or sample size, p is the number of variables or quality characteristics and α is the type I error.

It is clear that the control limits are the upper and lower percentile of the approximation of the T^2 to the F distribution for sub-grouped data. The approximation of is given by Alt (1988).

For the individual observations case, the *Beta* distribution is used. Tracy et al. (1992) has a complete discussion about this. In this case, both parameters are usually estimated with the pooled estimators of the average, see Equation (2.12) and the covariance matrix, Equation (2.13).

$$\hat{\mu}_0 = \bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i \qquad (2.12)$$

$$\hat{\Sigma}_0 = S = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})(x_i - \bar{x})' \qquad (2.13)$$

Using these parameters, the T^2 can be calculated as

$$T^2 = (x - \mu_0)' \Sigma_0^{-1} (x - \mu_0) \qquad (2.14)$$

The procedure calls for removal of observations with T^2 over *UCL*, calculated in Equations (2.8) or (2.10), to achieve an in-control set of observations. However, Hotelling's T^2 control chart does not detect shifts, trends or outliers efficiently and many authors have focused their work in the estimation of these process parameters, as it was

discussed in Section 1.2, Sullivan and Woodall (1996) and Vargas (2003) offers alternatives to solve this problem.

For *subgrouped data*, following Montgomery (2001), the sample means and variances are calculated for a simple sample, that is:

$$\bar{x}_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ijk} \qquad \begin{cases} j = 1,2,...,p \\ k = 1,2,...,m \end{cases} \tag{2.15}$$

$$S_{jk}^{\ 2} = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ijk} - \bar{x}_{jk})^2 \qquad \begin{cases} j = 1,2,...,p \\ k = 1,2,...,m \end{cases} \tag{2.16}$$

where x_{ijk} is the ith observation on jth quality characteristic in the kth sample. The covariance between quality characteristics j and h in the kth sample is:

$$S_{jhk} = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ijk} - \bar{x}_{jk})(x_{ihk} - \bar{x}_{hk}) \qquad \begin{cases} k = 1,2,...,m \\ j \neq h \end{cases} \tag{2.17}$$

Then $\bar{x}_{jk}, S_{jk,}^{2}$ and S_{jhk} are averaged over all m samples to obtain:

$$\bar{\bar{x}}_{j} = \frac{1}{m}\sum_{k=1}^{m} \bar{x}_{jk} \qquad j = 1,2,...,p \tag{2.18}$$

$$\bar{S}_{j}^{\ 2} = \frac{1}{m}\sum_{k=1}^{m} S_{jk}^{\ 2} \qquad j = 1,2,...,p \tag{2.19}$$

$$\bar{S}_{jh} = \frac{1}{m}\sum_{k=1}^{m} S_{jhk} \qquad j \neq h \tag{2.20}$$

Finally the $\{\bar{\bar{x}}_j\}$ are the elements of the vector $\bar{\bar{\bar{x}}}$, and the sample covariance matrix S is formed as

$$S = \begin{bmatrix} \bar{S}_1^2 & \bar{S}_{12} & \bar{S}_{13} & \dots & \bar{S}_{1p} \\ \bar{S}_{12} & \bar{S}_2^2 & \bar{S}_{23} & \dots & \bar{S}_{2p} \\ \bar{S}_{13} & \bar{S}_{23} & \bar{S}_3^2 & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \bar{S}_{1p} & \bar{S}_{2p} & \dots & \dots & \bar{S}_p^2 \end{bmatrix} \tag{2.21}$$

and the T^2 for subgrouped data can be calculated using the following formula

$$T^2 = n(\bar{x} - \bar{\bar{x}})' S^{-1} (\bar{x} - \bar{\bar{x}}) \tag{2.22}$$

2.3.2 Phase II: Future Observations

Once, the process parameters have been estimated, the phase of future observations monitoring starts. The procedure is the same as in Shewhart Control Charts (see Figure 2.3-2).

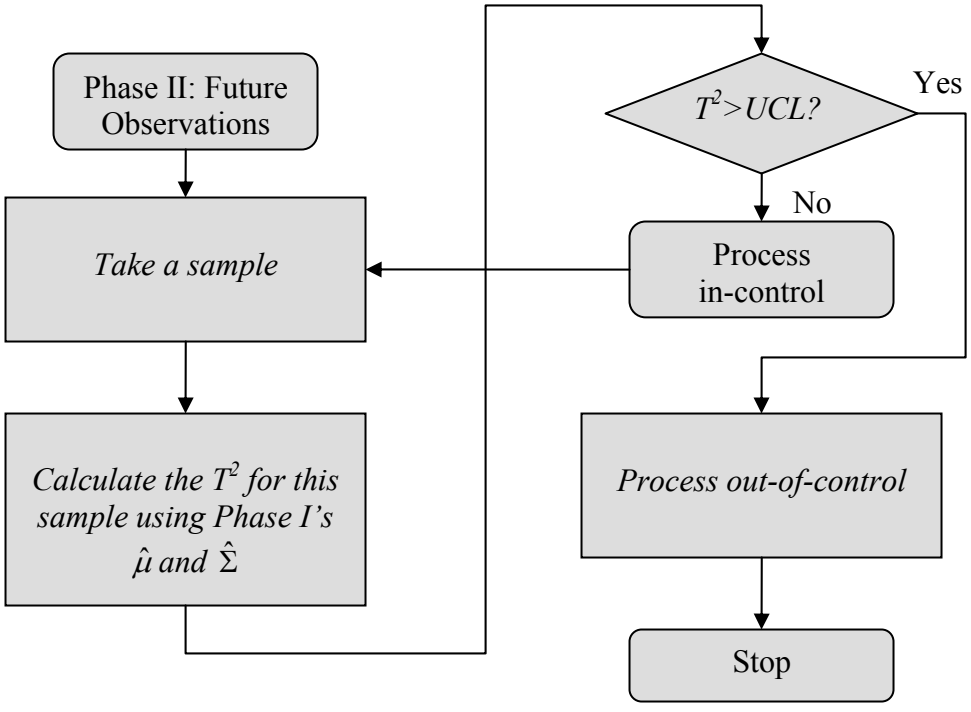


Figure 2.3-2. Phase II: Future Observations Procedure

For this phase the control limits for subgrouped data are shown in Equations (2.23) and (2.24). For individual observations the control limits are in Equations (2.25) and (2.26).

$$UCL = \frac{p(m+1)(n-1)}{mn-m-p+1} F_{\alpha,p,mn-m-p+1} \qquad \text{(2.23) and (2.24)}$$
$$LCL = 0$$

$$UCL = \frac{p(m+1)(m-1)}{m^2-mp} F_{\alpha,p,m-p} \qquad \text{(2.25) and (2.26)}$$
$$LCL = 0$$

These limits are used as an analogous to the Shewhart $\overline{X}$ control chart. Their interpretation, however, is not as easy as in univariate control charts. Many authors have studied and proposed methods for understanding out of control signals: $\bar{x}$ charts with Bonferroni-type limits (Alt, 1988), decomposition of T^2 to see the variable that contributes the most (see Runger, Alt and Montgomery, 1996 for example) and many more. But, regardless of the method, the basic idea is to find an unusual behavior in the variables being monitored. These control limits must be regularly updated, Alt (1976) suggests updates after *m+5*, *m+10*, *m+25*, *m+50* and *m+100* subgroups.

3 Problem Description

Section 1.2 has pointed out the problem of estimation in the ARL. Ryan (2000) proposes that ARL for Phase II with estimated parameters can only be obtained by simulation. Under this recommendation, a simulation was developed to show that in Hotelling's T^2 control charts, the in-control ARL has a deviation from the design value. This problem is more evident when few samples are being used to estimate the parameters. The experiment used to show this problem is described in detail.

The Chapter is organized in two sections, the first one describes the procedure to simulate samples, estimate the parameters and calculate the in-control ARL for a Hotelling's T^2 control chart. The second part shows the results for a particular example to illustrate the problem.

3.1 In-control ARL Simulation Procedure

The procedure followed was divided in two parts, based in the Phases proposed by Alt (see Chapter 2, Section 2.3).The fist part deals with the estimation of the parameters and the second is the computation of the in-control ARL.

3.1.1 Step 1: Estimation of parameters

The objective of this part is the estimation of the in-control parameters: the mean ($\hat{\mu}$) and covariance matrix ($\hat{\Sigma}$):

a. Fix in control values for the mean (μ_0) and covariance matrix (Σ_0).

b. Using the in-control values μ_0 and Σ_0, generate *m* multivariate samples, each one of size *n* following a multivariate normal distribution with the parameters in control.

c. For each multivariate sample of size *n*, estimate the mean and covariance matrix using Equations (2.15), (2.16) and (2.17).

d. Estimate the process mean and covariance matrix using Equations (2.18), (2.19) and (2.20).

Since all data are generated using the in-control parameters, we do not check whether the T^2 for each sample lower than the UCL.

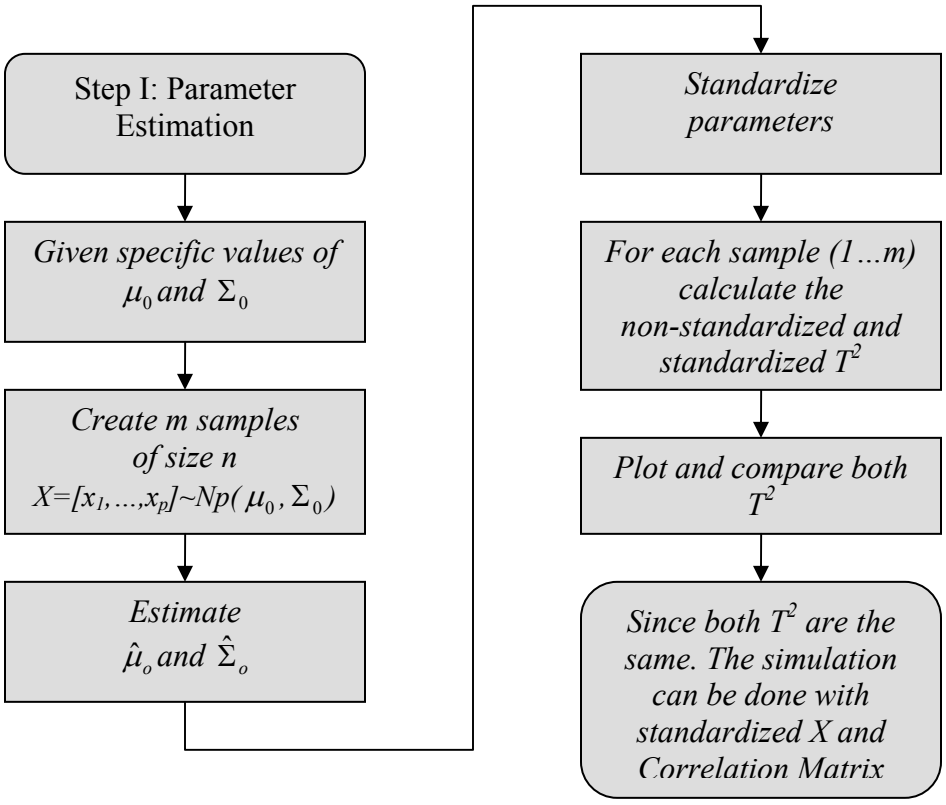


Figure 3.1-1. Procedure for the estimation of parameters

For this step a standardization of the parameters has been considered. The procedure for standardization is the same as used by Hauck, Runger and Montgomery (1999):

- The covariance matrix is "standardized" to obtain the correlation matrix (R).

$$R = \begin{bmatrix} r_{11} = \dfrac{\bar{S}_{11}}{\bar{S}_{11}} & r_{12} = \dfrac{\bar{S}_{12}}{\sqrt{\bar{S}_{11}\bar{S}_{22}}} & \cdots & r_{1p} = \dfrac{\bar{S}_{1p}}{\sqrt{\bar{S}_{11}\bar{S}_{pp}}} \\ r_{21} & r_{12} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p2} & r_{p2} & \cdots & r_{pp} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p2} & r_{p2} & \cdots & 1 \end{bmatrix} \quad (3.1)$$

  The values of $\bar{S}_{jj}$ are obtained from Equation (2.21).

- Sample means are standardized to

$$Z = [z_1, z_2, ..., z_j, ..., z_p] \quad with \quad z_j = \frac{\bar{x}_j - \bar{\bar{x}}_j}{\sqrt{\bar{S}_{jj}}}, \quad j = 1, ..., p$$

  Where $\bar{\bar{x}}_j$ is the j-component of the $\hat{\mu}_0$ vector.

Since the parameters are estimated, the standardization has three effects:

1. $E(z_j) = 0$

2. $Var(z_j) = 1$

3. $Cov(z_i, z_j) = r_{ij} = \dfrac{\bar{S}_{ij}}{\sqrt{\bar{S}_{ii}}\sqrt{\bar{S}_{jj}}}$

Equation (2.14) now becomes

$$T^2 = n\overline{Z}' R^{-1} \overline{Z} \tag{3.3}$$

The numerical example (Section 3.2) shows that there is no difference calculating the T^2 so the next steps use standardized values instead of the original ones.

3.1.2 Step 2: Run-length and Average run-length calculation

In this step, the ARL is calculated after calculating the RL. The procedure to calculate a false alarm signal is shown in Figure 2.3-1.

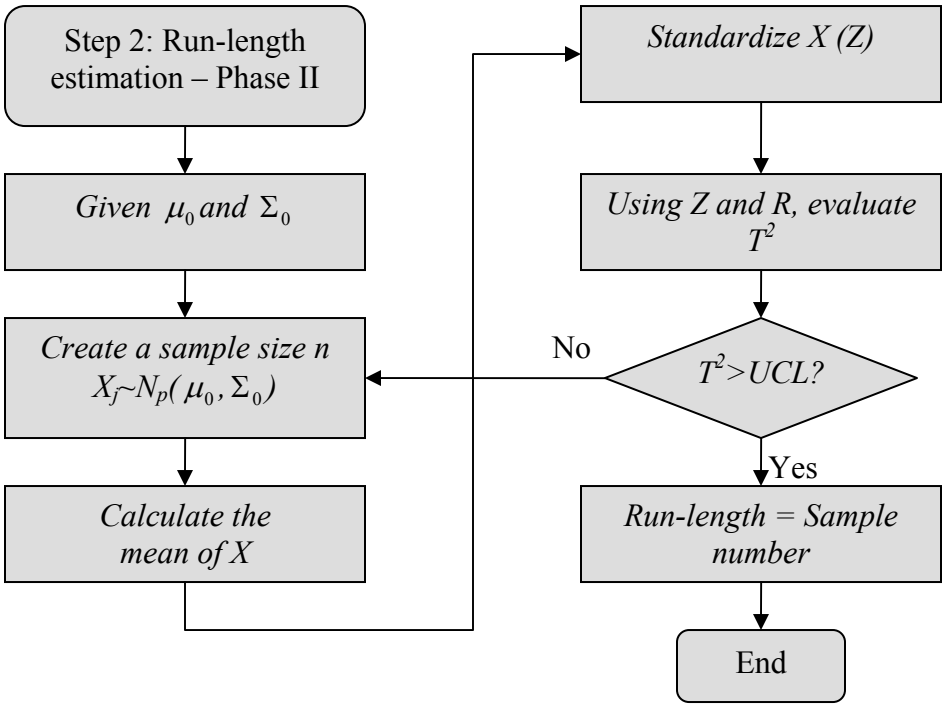


Figure 3.1-2. Procedure for Run-length calculation

A sample of size n following the $Np(\mu_0,\Sigma_0)$ distribution is generated, then standardized statistic T^2 is obtained, this value is compared with the UCL using Equation (2.23).

The RL would be the number of samples until a T^2 exceeds the UCL. The ARL is the mean of the RLs after repeating the previous procedure a considerable number of times.

The T^2 is estimated using Equation (3.3). Note that now, in the Equation (2.23), μ_0 is replaced by Z and the estimated covariance matrix is replaced by the correlation matrix.

3.2 Numerical Example

A numerical example to illustrate the procedure of Section 3.1 was coded in Matlab 7.0.1(see Appendix H). The in-control parameters considered are

$$\mu_0 = \begin{bmatrix} 10 & 15 & 20 \end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix} 1 & -0.5 & 0.5 \\ -0.5 & 1 & 0.25 \\ 0.5 & 0.25 & 1 \end{bmatrix}$$

After using $n = 5$ and $m = 100$, in-control mean and covariance matrix were estimated as

$$\hat{\mu}_0 = \begin{bmatrix} 9.9551 & 14.9903 & 19.9676 \end{bmatrix}$$

$$\hat{\Sigma}_0 = \begin{bmatrix} 1.0320 & -0.4556 & 0.5348 \\ -0.4556 & 0.9584 & 0.2399 \\ 0.5348 & 0.2399 & 1.0061 \end{bmatrix}$$

$$\hat{\Sigma}^{-1}{}_0 = \begin{bmatrix} 2.6992 & 1.7465 & -1.8510 \\ 1.7465 & 2.2398 & -1.4623 \\ -1.8510 & -1.4623 & 2.3264 \end{bmatrix}$$

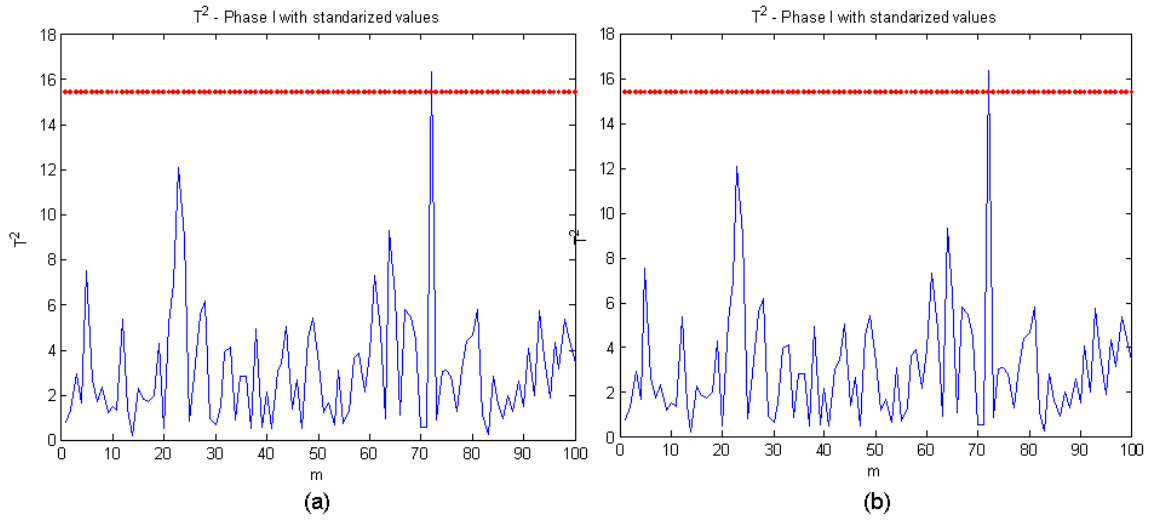


Figure 3.2-1. Comparison between Phase I's T^2 for (a) non standardized values and (b) standardized values using (m=100).

After standardization, we have

$$E(Z) = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

$$R = \begin{bmatrix} 1.0000 & -0.4581 & 0.5248 \\ -0.4581 & 1.0000 & 0.2443 \\ 0.5248 & 0.2443 & 1.0000 \end{bmatrix}$$

$$R^{-1} = \begin{bmatrix} 2.7855 & 1.7369 & -1.8862 \\ 1.7369 & 2.1465 & -1.4359 \\ -1.8862 & -1.4359 & 2.3407 \end{bmatrix}$$

Now, the estimated mean is a vector of zeros, and the T^2 is evaluated using Equation (3.3). All 100 T^2 with standardized and non-standardized were calculated and plotted in Figure 3.2-1. The out-of-control point is, as it was previously explained, a false alarm signal.

The sample size (m) used in Phase I to estimate the parameters for this experiment was **100**. The ARL was estimated as the average of **100** Run-lengths and the subgroup size, n, was fixed at **5** and the number of variables considered was **3**.

Another possible cause of the difference between the observed ARL and Alt's approximation is error in the correlation matrix. This simulation uses different correlation matrices. Let us assume that *fr* is the estimator of the true value for a correlation *r*, where *f* is a multiplicative constant that accounts for the estimation error. For example, if the true correlation is *r = 0.*5, but is overestimated as *0.55* (=*fr*), then the value of *f* is 1.1. If the correlation is underestimated, then *f<1*; however, if it is overestimated, then *f>1*. Obviously, in practice, the value of *f* is unknown, but in this simulation, the true value of the correlations are known, consequently the value of *f* can be computed as the ratio of the estimated correlation to the true correlation. Equation (3.4) summarizes these calculations for any particular element of the correlation matrix. In this equation, $r^{(+)}$ denotes the overestimated correlation and $r^{(-)}$ denotes the underestimated correlation. Moreover, to assure that the experiment runs with both underestimated and overestimated values of the correlations, for each overestimated (underestimated) correlation, the simulation computes *r/f*, the corresponding underestimated (overestimated) correlation. In the numerical example, besides running the computational experiments with the overestimated correlation of 0.55, the experiments also runs with the underestimated correlation 0.5/1.1 = 0.4545.

$$\hat{r}_{ij}^{+(-)} = \frac{r_{ij}^2}{\hat{r}_{ij}^{-(+)}} \tag{3.4}$$

For this simulation, since there are 3 correlations (r_{12}, r_{13} *and* r_{23}) the simulation of the in-control ARL was done with all combinations of their opposite correlations:

- Correlation Matrix 2, changes all the correlations.
- Correlation Matrix 3, changes r_{12} and r_{21}.
- Correlation Matrix 4, changes r_{13} and r_{31}.
- Correlation Matrix 5, changes r_{23} and r_{32}.
- Correlation Matrix 6, changes r_{12}, r_{21}, r_{13} and r_{31}.
- Correlation Matrix 7, changes r_{12}, r_{21}, r_{23} and r_{32}.
- Correlation Matrix 8, changes r_{13}, r_{31}, r_{23} and r_{32}.

The idea is to observe the effect in the in-control ARL of all possible combinations of correlations in the Correlation Matrix.

Table 1. Correlation Matrices (m = 100)

	Correlations								
Correlation Matrix	r_{11}	r_{12}	r_{13}	r_{21}	r_{22}	r_{23}	r_{31}	r_{32}	r_{33}
0: Real	1	-0.5	0.5	-0.5	1	0.25	0.5	0.25	1
1: Estimated	1	-0.4581	0.52479	-0.4581	1	0.24432	0.52479	0.24432	1
2 : Opposite	1	-0.54573	0.47638	-0.54573	1	0.25581	0.47638	0.25581	1
3	1	-0.54573	0.52479	-0.54573	1	0.24432	0.52479	0.24432	1
4	1	-0.4581	0.47638	-0.4581	1	0.24432	0.47638	0.24432	1
5	1	-0.4581	0.52479	-0.4581	1	0.25581	0.52479	0.25581	1
6	1	-0.54573	0.47638	-0.54573	1	0.24432	0.47638	0.24432	1
7	1	-0.54573	0.52479	-0.54573	1	0.25581	0.52479	0.25581	1
8	1	-0.4581	0.47638	-0.4581	1	0.25581	0.47638	0.25581	1

The results of the in-control ARL estimated using correlations matrices 1 to 8 are shown in Table 2. In-control ARLs were estimated for α = 1/100, 1/300, 1/500,…,1/1300 which under Alt's 2.23 would give in-control ARLs (Theoretical ARL) of 100, 300, 500,…, 1300.

Table 2. ARL results (m = 100, n = 5 and p = 3)

	Simulated ARL (Average of 100 run-lenghts)								
Theoretical ARL (Alt)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Average**
100	123.15	75.64	50.77	143.32	118.2	78.05	45.92	138.19	96.66
300	338.54	206.22	124.98	422.23	297.08	234.06	109.74	438.76	271.45
500	597.33	359.68	209.14	870.98	535.83	371.93	171.74	836.79	494.18
700	992.64	503.57	292.42	1432	933.93	590.83	242.07	1349.1	792.07
900	1326.5	703.07	313.46	1459.5	1125.2	781.2	264.35	1417.1	923.8

	Simulated ARL (Average of 100 run-lenghts)								
Theoretical ARL (Alt)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Average**
1100	1767.9	1029.4	465.38	2287.5	1696.2	1104.2	361.41	2154.7	1358.34
1300	1643.8	793.22	473.2	2204.8	1482.3	914.91	392.25	2163.4	1258.49

Figure 3.2-4 shows the plotted values compared with the theoretical ARL.

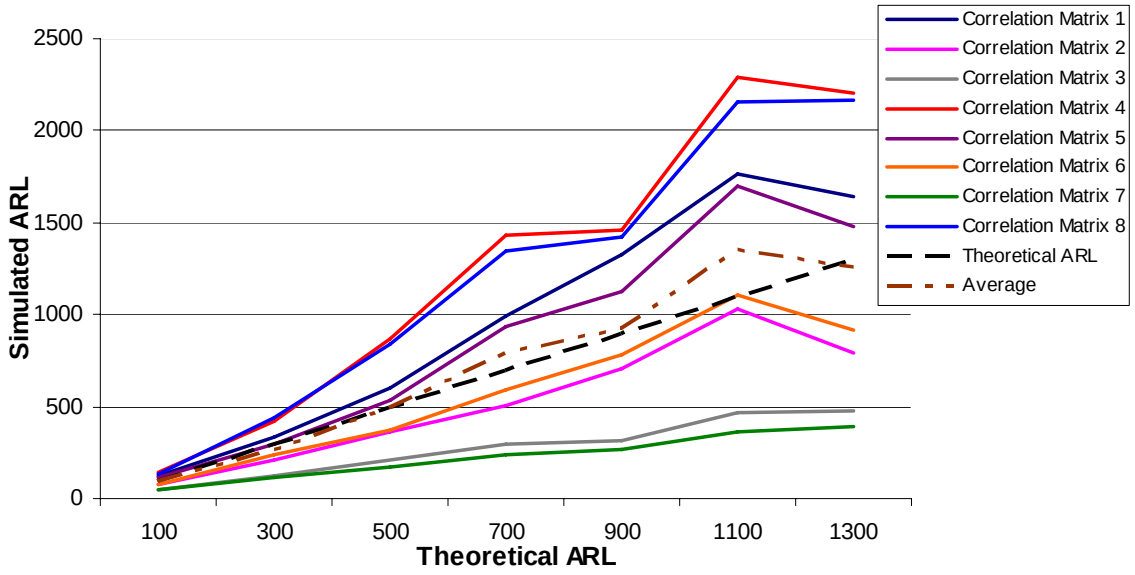


Figure 3.2-2. In-control ARL (*m* = 100, *n* = 5)

It is clear that, even when the average is close to the theoretical value, the deviation of the in-control ARL to the theoretical value increases as α decreases. Observing the average of ARLs, Alt's approximation seems to be good for obtaining in-control ARLs under 300. The following figure shows the scaled difference between the theoretical ARL and the simulated ARL.

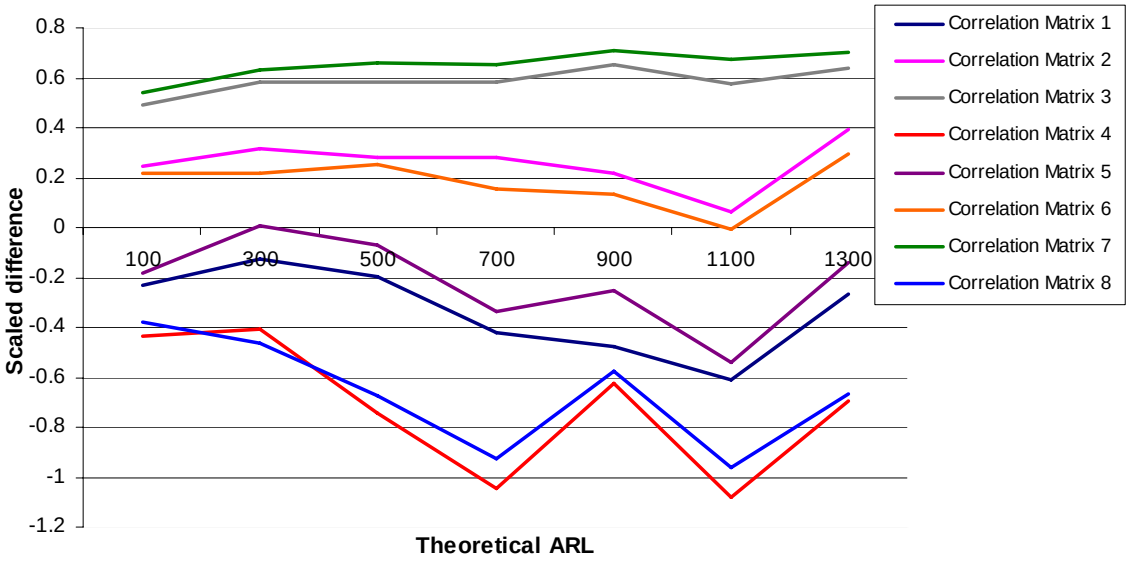


Figure 3.2-3. Scaled difference of the in-control ARL (*m* = 100, *n* = 5)

Where the scaled difference has been computed as

$$\frac{ARL_{theoretical} - ARL_{simulated}}{ARL_{theoretical}} \tag{3.5}$$

The average departure, evaluated after applying the absolute value to each scaled difference is 45%, which is considerably significant.

The same experiment was performed now with $m = 30$ and $m = 1000$, also fixing p and n in 3 and 5 respectively. The results are shown in Tables 4 and 6. The correlations are in Tables 3 and 5.

Table 3. Correlation matrices (m = 30)

	Correlations								
Correlation Matrix	r_{11}	r_{12}	r_{13}	r_{21}	r_{22}	r_{23}	r_{31}	r_{32}	r_{33}
0: Real	1	-0.5	0.5	-0.5	1	0.25	0.5	0.25	1
1: Estimated	1	-0.52237	0.49577	-0.52237	1	0.1837	0.49577	0.1837	1
2 : Opposite	1	-0.47859	0.50427	-0.47859	1	0.34023	0.50427	0.34023	1
3	1	-0.47859	0.49577	-0.47859	1	0.1837	0.49577	0.1837	1
4	1	-0.52237	0.50427	-0.52237	1	0.1837	0.50427	0.1837	1
5	1	-0.52237	0.49577	-0.52237	1	0.34023	0.49577	0.34023	1
6	1	-0.47859	0.50427	-0.47859	1	0.1837	0.50427	0.1837	1
7	1	-0.47859	0.49577	-0.47859	1	0.34023	0.49577	0.34023	1
8	1	-0.52237	0.50427	-0.52237	1	0.34023	0.50427	0.34023	1

Table 4. ARL results (m = 30, n = 5 and p = 3)

	Simulated ARL (Average of 100 run-lenghts)								
Theoretical ARL (Alt)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Average**
100	64.52	35.01	75.16	59.33	22.19	69.93	36.72	20.42	47.91
300	216.74	102.98	267.57	206.09	54.19	269.52	114.24	45.05	159.55
500	402.58	150.02	498.03	390.02	81.04	511.04	187.49	62.29	285.31
700	513.87	185.32	717.78	499.75	97.95	634.03	233.43	85.45	370.95
900	745.18	188.38	843.43	673.17	79.49	818.75	237.35	58.73	455.56
1100	782.21	252.91	954.31	730.21	135.11	934.56	299.53	96.33	523.15
1300	1002.4	300.06	1212.9	983.32	139.61	1215	362.55	123.44	667.41

Table 5. Correlation Matrices (m = 1000)

	Correlations								
Correlation Matrix	r_{11}	r_{12}	r_{13}	r_{21}	r_{22}	r_{23}	r_{31}	R_{32}	r_{33}
0: Real	1	-0.5	0.5	-0.5	1	0.25	0.5	0.25	1
1: Estimated	1	-0.49665	0.49753	-0.49665	1	0.25609	0.49753	0.25609	1
2 : Opposite	1	-0.50337	0.50249	-0.50337	1	0.24406	0.50249	0.24406	1
3	1	-0.50337	0.49753	-0.50337	1	0.25609	0.49753	0.25609	1
4	1	-0.49665	0.50249	-0.49665	1	0.25609	0.50249	0.25609	1
5	1	-0.49665	0.49753	-0.49665	1	0.24406	0.49753	0.24406	1
6	1	-0.50337	0.50249	-0.50337	1	0.25609	0.50249	0.25609	1
7	1	-0.50337	0.49753	-0.50337	1	0.24406	0.49753	0.24406	1
8	1	-0.49665	0.50249	-0.49665	1	0.24406	0.50249	0.24406	1

Table 6. ARL results (m = 1000, n = 5 and p = 3)

	Simulated ARL (Average of 100 run-lenghts)								
Theoretical ARL (Alt)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Average**
100	113.56	109.39	109.31	112.19	116.59	107.05	114.78	116.05	112.37
300	317.92	310.31	300.63	305.95	336.61	295	320.18	318.21	313.10
500	511.74	513.63	486.68	504.84	519.33	465.1	514.13	526.27	505.22
700	694.74	679.92	618.63	641.51	750.76	600.94	697.84	708.71	674.13
900	918.14	864.36	819.73	843.12	958.46	779.41	939.99	910.52	879.22
1100	1220.6	1240.6	1144.7	1156.8	1292.8	1108.6	1285	1293.2	1217.79
1300	1369.2	1377.3	1297.6	1310.5	1500.3	1256	1451.6	1412.2	1371.84

Figure 3.2-4 and Figure 3.2-5 display the plot of those values versus the theoretical Alt's in-control ARL.

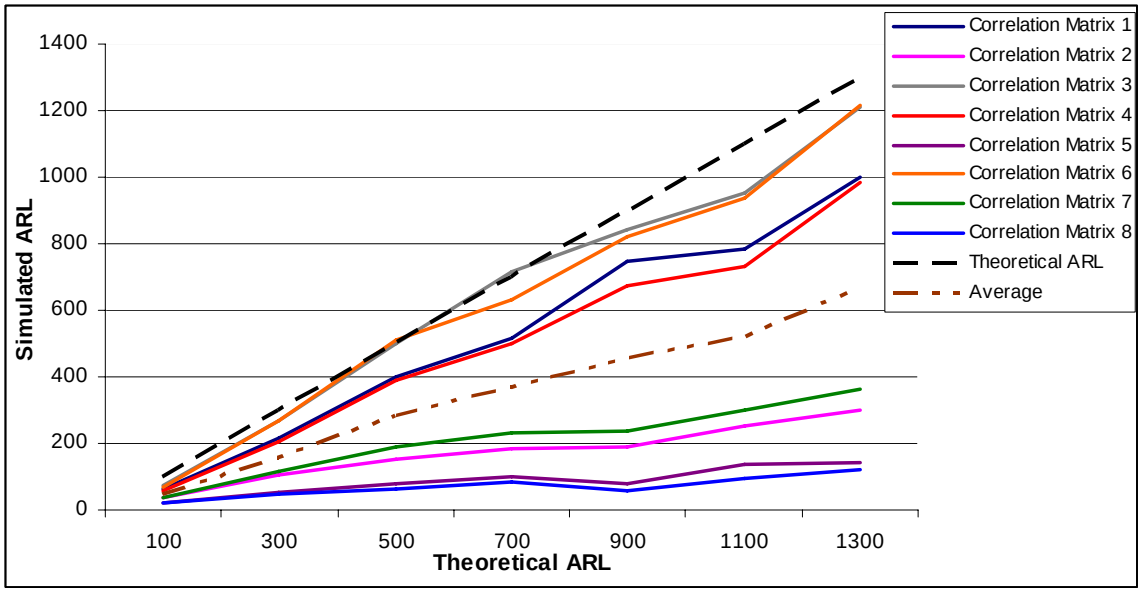


Figure 3.2-4. In-control ARL (*m*=30, *n* = 5)

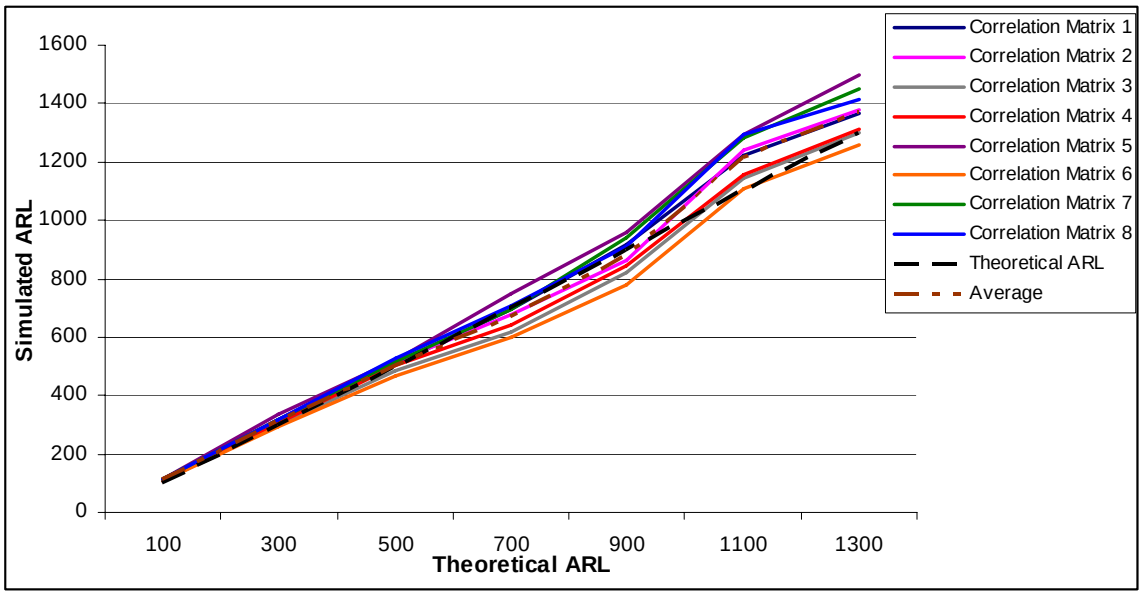


Figure 3.2-5. In-control ARL (*m*=1000, *n* = 5)

When correlations matrices are estimated from small samples the simulated in-control ARL is underestimated, that means that the chart will signal a false alarm sooner than expected. On the other hand, if correlation matrices are estimated from a large sample, the in-control ARL seems to produce very close results to the expected.

Checking the scaled differences, the average scaled difference (after applying absolute value to the scaled differences), give us an average departure of 48.64% for the matrices estimated with small samples (m = 30 and n = 5) and only an average departure of 7% for matrices estimated with large samples (m = 1000 and n = 5).

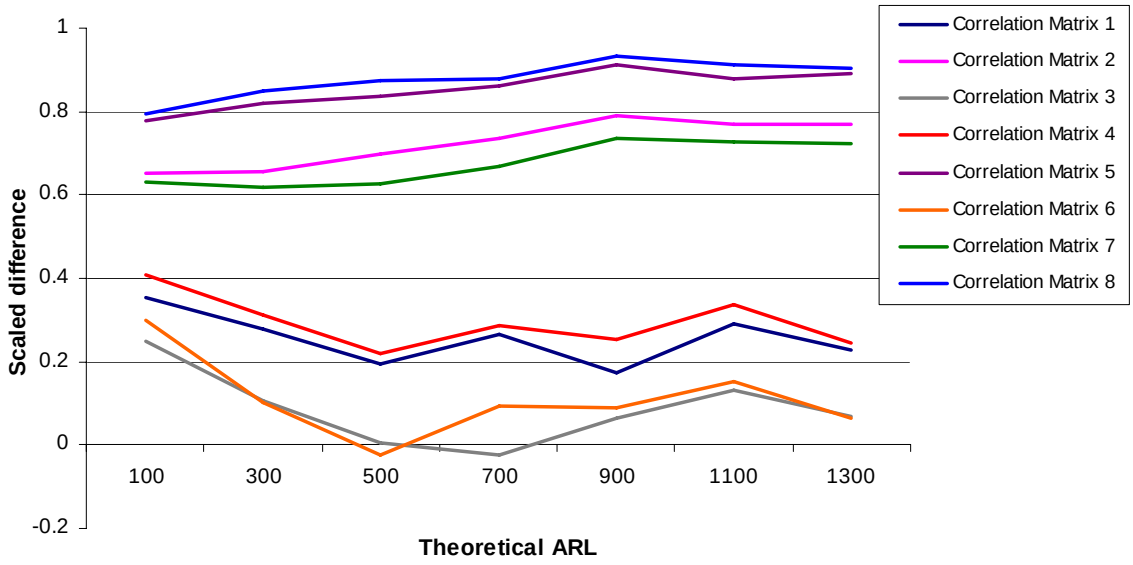


Figure 3.2-6. Scaled difference of the in-control ARL (m = 30, n = 5)

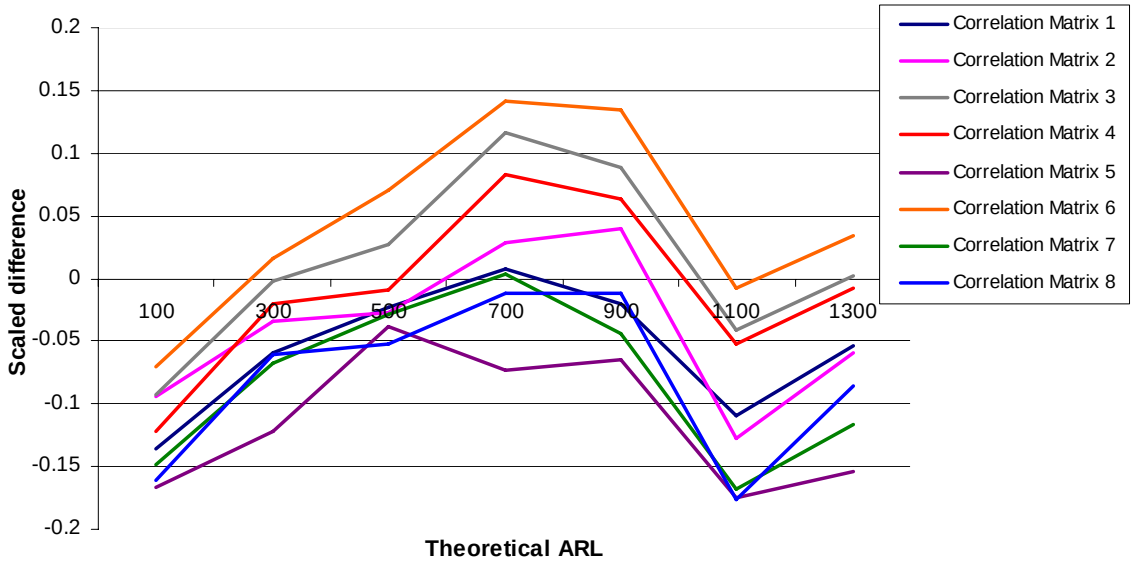


Figure 3.2-7. Scaled difference of the in-control ARL (m = 1000, n = 5)

3.3 Observations

From this first simulation, we observe that:

1. Alt's recommendation (Alt, 1988) of using a large amount of historical data to obtain the in control parameters seems to be good (see results when m = 1000 and n = 5). This results confirms the work by Champ et al. (2005) who, as it was explained in Chapter I, has recently find the combination of mn that makes Alt's approximation work the better.

2. Even when it is necessary to perform another experiment to determine exactly the parameters that influence in this problem, it can be inferred from the previous simulation, that the estimation of the Correlation Matrix (how far is from the real Correlation Matrix) has serious influence in the in-control ARL estimation. The next chapter deals with this, under the assumption that the condition number is a good characterization of the correlation matrix.

4 Experimental Methodology

To determine what variables affect the previously described problem, it is necessary to perform an experiment that not only determines the factors but also help us to fit a prediction model. This Chapter discusses the experiment methodology and the additional theoretical background to understand the procedure employed. The first part contains an explanation of response surface modeling, a brief explanation of the most common designs in this family, and the justification of the chosen design: the Central Composite Design (CCD). The second part explains how to analyze a CCD. The third part shows in detail the design of the experiment employed in this work. Finally, the last part explains the simulation developed to run the experiment.

4.1 Response Surface Methodology (RSM)

4.1.1 Definition

RSM is a methodology used when, in addition to the identification of main factors, we want to find a model that helps us to predict. The RSM model has the following form

$$\begin{aligned} y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \beta_{12} x_1 x_2 + \ldots + \beta_{1n} x_1 x_n \\ + \ldots + \beta_{n-1,n} x_{n-1} x_n + \beta_{11} x_1^2 + \ldots + \beta_{nn} x_n^2 + \varepsilon \end{aligned} \tag{4.1}$$

It is clear that this experiment is designed to fit a second order model that includes interaction and quadratic effects of the factors under study, which is the reason why this method has been selected for this work.

4.1.2 Common designs

The two most used designs used in RSM are the central composite design and the Box-Behnken design. In these designs, the inputs take on three or five distinct levels, but not all combinations of these values appear in the design.

4.1.2.1 Central Composite Design (CCD)

The CCD uses a factorial or fractional factorial design with center points and adds points to estimate curvature. The idea is to choose points that their values maintain rotatability in the design. Usually this value is called the "axial distance" or "axial point", denoted by ρ. The recommendation (see Montgomery & Myers, 1995) says that if the distance from the center point to the design space is ± 1 then $|\rho| > 1$ and its value depends on the design and the number of factors:

$$\rho = (number\ of\ \ factorial\ runs)^{1/4} \tag{4.2}$$

The following figure shows the design for 2 factors

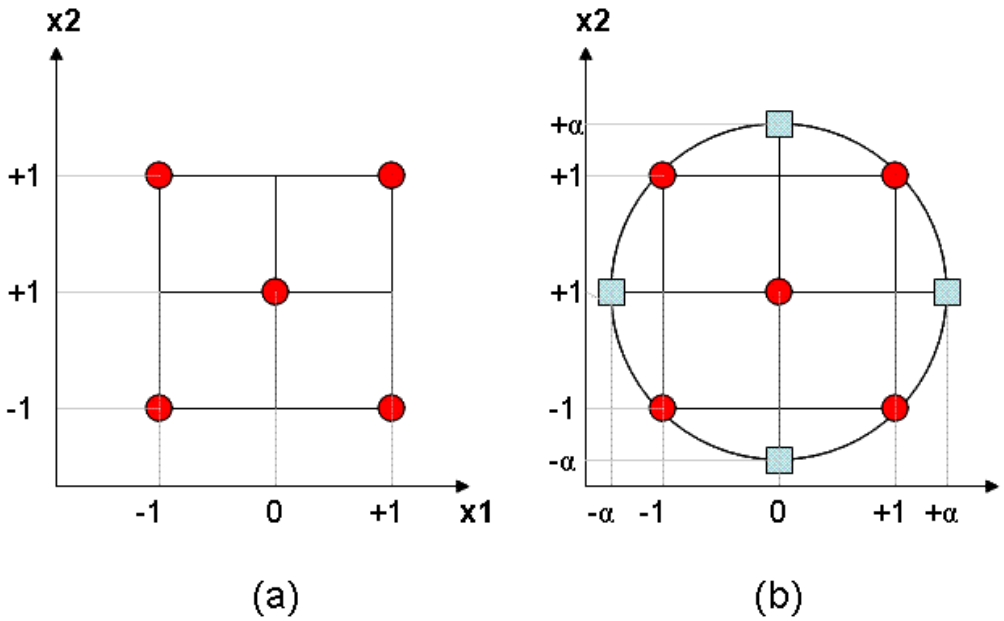


Figure 4.1-1. (a) Factorial design (2 factors) with center point (b) Central Composite Design (2 factors)

4.1.2.2 Box-Behnken Designs

These are designs that take only three levels instead of five. Each combination of the extreme values of two of the variables is tested, the remaining variables taking a coded level of zero. Figure 4.1-2 shows the design for three factors. These designs require fewer treatment combinations than a central composite design in cases involving 3 or 4 factors.

Montgomery and Myers (1995) point out that the Box-Behnken design is nearly rotatable but it contains regions of poor prediction quality. It is usually recommendable only when the experimenter wants to avoid combination of factor extremes.

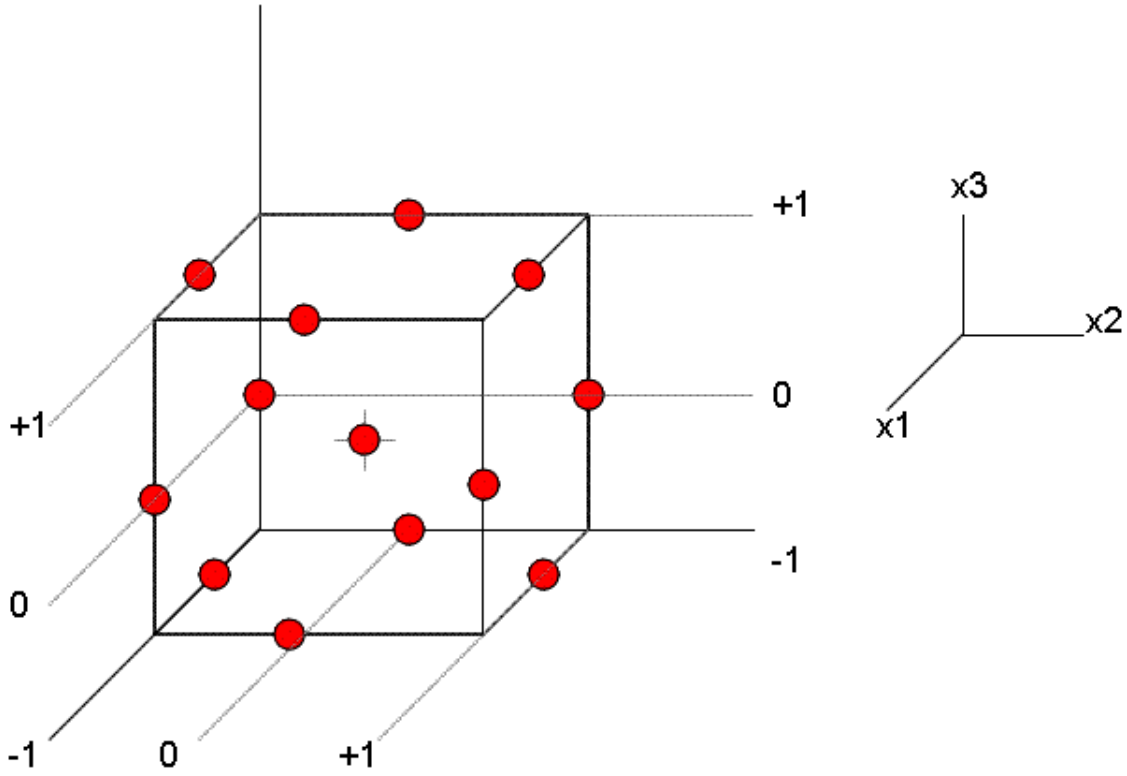


Figure 4.1-2. Box-Behnken Design (3 factors)

4.2 Analysis of a Surface Design

Linear regression and analysis of variance (ANOVA) are two of the most widely used statistical techniques. Regression describes the relationship between a response variable and one or more continuous independent variables by using least squares to determine the quantitative relationship. ANOVA determines whether a response variable differs among discrete values of the independent variable(s) which implies that **ANOVA is not designed to be used as a prediction model only to find relevant factors**.

4.2.1 Linear Regression

The aim of this technique is to establish a relationship, between a response variable (y) and one or more independent variables, in the form

$$y = \beta X + \varepsilon \tag{4.3}$$

where X is a matrix of mxp, β is a vector of $px1$ of the regression coefficients and ε is the random error, assumed to be normal, independently distributed with constant variance.

In linear regression, the main idea is to find a model that minimizes the sum of square error (SS_{ERROR}) defined by the square of the difference between the predicted value of y and its real value

$$SS_{ERROR} = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{4.4}$$

The coefficients that minimizes this value are obtained from

$$\hat{\beta} = (X'X)^{-1} X'y \tag{4.5}$$

The significance of the regression is evaluated with the Analysis of Variance that is explained in the next section. But the most important measure is the coefficient of multiple determination R^2 defined as

$$R^2 = 1 - \frac{SS_{ERROR}}{SS_{TOTAL}} \tag{4.5}$$

This coefficient measures the amount of variation in y explained by using the regressor variables included. However, the R^2 always increases as we add terms in the

model, on the contrary many recommend the use of the adjusted R^2_{adj} which decreases as we add unnecessary variables.

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-p} \right)(1 - R^2) \tag{4.5}$$

4.2.2 Analysis of Variance (ANOVA)

ANOVA is but a special case of regression, for example in one-way ANOVA we have the following model: every population mean contains μ and one effect treatment τ_i. The observations y_{ij} follows the linear model

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \tag{4.6}$$

ANOVA compares means by dividing the total variance in different parts. It divides the total sum of squares (SS_{TOTAL}) in several components each one related to the effect or factor used in the model.

$$SS_{TOTAL} = SS_{ERROR} + SS_{effect1} + ... + SS_{effectn} \tag{4.7}$$

The same is applied to the degrees of freedom to construct the mean square (*MS*), which is the ratio between the sum of squares and its related degrees of freedom.

The MS_{ERROR} is considered an estimation of the variance and its square root is an estimate of the standard deviation of the model.

The significance of the factors is evaluated by using the *F* distribution, the *MS* of each factor is divided by the MS_{ERROR} to construct an statistic that follows an *F* distribution. If the statistic exceeds the critical value, then the factor is significant.

In the case of the linear regression, the factors are grouped in one source of variation and the significance of the regression is obtained from the following statistic

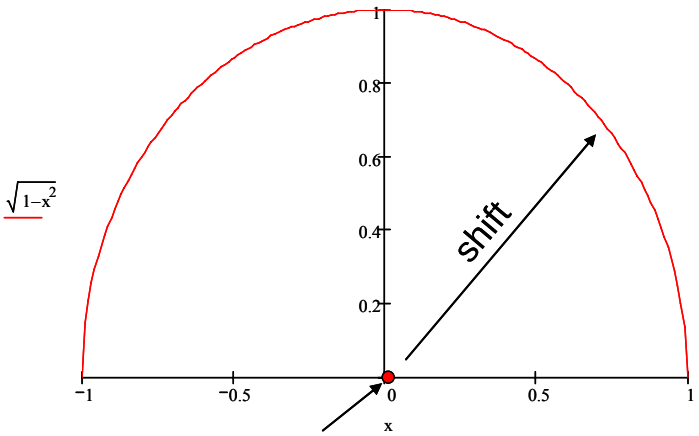
$$F_0 = MS_{REGRESSION} / MS_{ERROR} \tag{4.8}$$

4.3 The effect of the condition number (k) in the computation of the T^2

One of the conclusions of the illustrative example in Section 3.3 is that correlation matrix seems to have an effect in the in-control ARL. One way to characterize the correlation matrix is using the condition number (see Appendix A for a more detailed discussion). The effect of the condition number on T^2 can be illustrated using the following example. Consider an in-control correlation matrix S_{01} with a condition number (k) = 5.00.

$$S_{01} := \begin{pmatrix} 1 & -0.571 & 0.377 \\ -0.571 & 1 & -0.059 \\ 0.377 & -0.059 & 1 \end{pmatrix}$$

Assume that we want to calculate the Hotelling's T^2 statistic for a shift in the mean vector from $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ to $\begin{bmatrix} x & 0 & \sqrt{1-x^2} \end{bmatrix}$ where $1 \leq x \leq 1$. This is a shift of one standard deviation from the in-control value in many directions on the plane x_1x_3. The following picture illustrates the change in the vector mean.



In control vector:

$[x_1=0\ x_2=0\ x_3=0\ ]$

Figure 4.3-1. Shift in the x_1x_3 plane

In this case, applying Equation (2.22), the statistic T^2, with n = 1, ranges approximately from 0.82 and 2.2 for $-1 \leq x \leq 1$. Now suppose that the matrix S_{01} has been estimated by

$$S_{11} := \begin{pmatrix} 1 & -0.542 & 0.358 \\ -0.542 & 1 & -0.056 \\ 0.358 & -0.056 & 1 \end{pmatrix}$$

This matrix has a reasonable error in the estimation (5%) of the correlations. Computing the T^2 with this estimate correlation matrix, its values, when $-1 < x < 1$, are approximately between 0.82 and 2.00 which are practically the same results obtained with the original matrix. Figure 4.3-2 shows the robustness of the process.

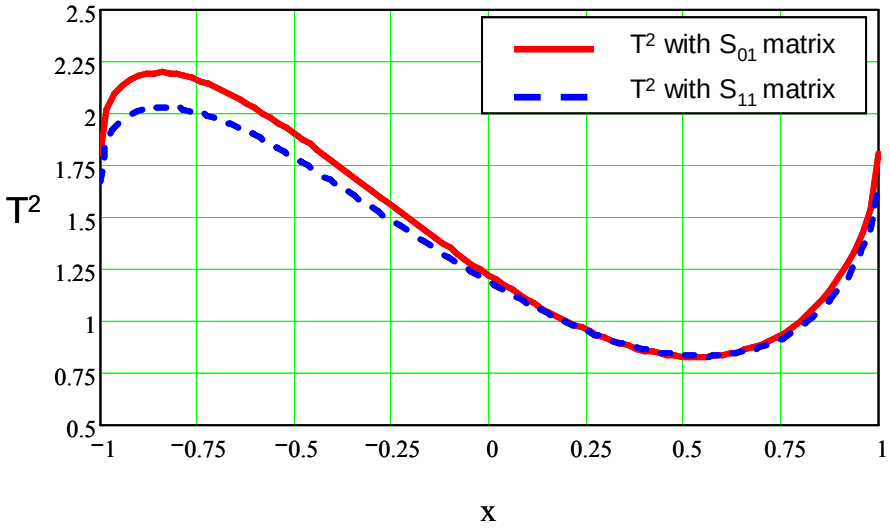


Figure 4.3-2. T^2 by x under shifts in the x_1x_3 plane (p =3, n = 1)

Now let's consider an in-control matrix, S_{02}, with a large condition number , k = 100

$$S_{02} := \begin{pmatrix} 1 & 0.893 & 0.893 \\ 0.893 & 1 & 0.664 \\ 0.893 & 0.664 & 1 \end{pmatrix}$$

Now, using the same vector of shifts, from $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ to $\begin{bmatrix} x & 0 & \sqrt{1-x^2} \end{bmatrix}$, and n = 1, the corresponding T^2 values range from 1.4 and 31.4. Now let's see the effect of estimation in S_{02}. Consider the matrix S_{12} which is an estimate of S_{02} with an estimation error of 5% in the correlations (as before):

$$S_{12} := \begin{pmatrix} 1 & 0.848 & 0.848 \\ 0.848 & 1 & 0.631 \\ 0.848 & 0.631 & 1 \end{pmatrix}$$

One would expect a similar situation as in the case with a low condition number. However, the T^2 computed using S_{12} now ranges from 1.4 to 11.5 (see Figure 4.3-3), considerably different from the values of T^2 with S_{20}.

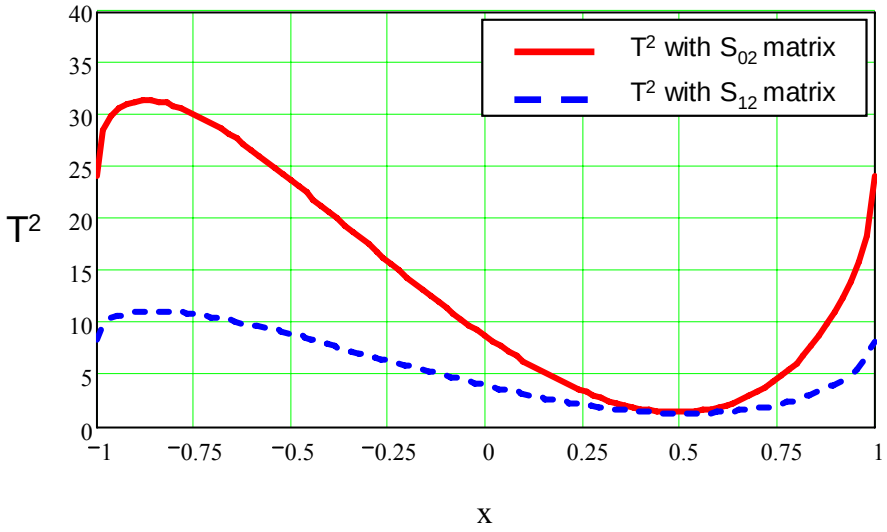


Figure 4.3-3. T^2 by x under shifts in the XZ plane (p =3, n = 1)

From this example, two observations about the condition number can be noted:

a. The effect of estimation is not a problem for well-conditioned matrices.
b. The effect of estimation is a serious problem for ill- conditioned matrices.

Hence, the condition number can be considered as an important factor for the design of the experiment presented in the following section.

4.4 Design of the Experiment

4.4.1 Objectives

As it was mentioned at the beginning of this Chapter, the purposes of the experiment are

1. To find the relevant factors that affect the expected value of the ARL

2. To adjust a model that can be used to predict the expected value of the ARL

4.4.2 Factors

The factors considered are

1. The number of quality characteristics (variables), p.
2. The number of samples in Phase I, m, used to estimate the mean vector and the correlation matrix.
3. The sample size, n.
4. And, finally the second condition number, k, which is a measurement of how ill-conditioned is a matrix. The previous section has shown the effect of the condition number on the T^2. The condition number can be used as a characterization of the correlation matrix. Ill-conditioned matrices are difficult to estimate (see Appendix A) and this affects the ARL. Appendix A has a more detailed definition of the condition number but, simplifying it, it can be estimated by

$$\left|\frac{\lambda_{\max}}{\lambda_{\min}}\right| \tag{4.9}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximal and minimal eigenvalue respectively of the matrix.

The experiment also has some important remarks: the first three parameters are discrete, which means that these parameters can only take integer values; by the other hand, the condition number is a continuous variable.

The condition number k was fixed at 5 values (5, 10, 20, 50 and 100). The values were chosen considering that under 100, there is no serious problem of multicollinearity (see Montgomery, Peck and Vinning, 2001). Appendix A explains the procedure to obtain matrices with a determined k that also satisfies some conditions explained in detail

also in this Appendix. Those values are condition number of the true matrices, the simulation also contains the condition number of the estimated matrices, $\hat{k}$, as another factor of the experiment.

For each condition number a central composite design (CCD) of the three discrete factors was chosen.

4.4.3 Adaptation of the CCD to discrete factors

The CCD uses a factorial or fractional factorial design with center points and adds points to estimate curvature. In this experiment, there are 3 factors each one at two levels so the axial point, after applying Equation (4.2), is

$$\rho = \left(2^3\right)^{1/4} = 1.682$$

The problem is that given the fact that the three factors are integer numbers, it is not possible to obtain the exact axial point. Hence, the axial point was rounded to $\rho = 2$. Appendix B shows that this axial point leads to a similar orthogonality to the optimum axial point. This axial point is also rotatable even when its efficency is about 34% less than the optimum axial point.

Another option could be the use of the Box-Behnken's Faced Centered design (CCF), but this design also has two problems: The first one is that it lacks of rotatability, see Montogomery and Myers (1995) for a more detailed explanation. The second is that it uses only 3 levels. So, choosing $\alpha = 2$, the loss in efficiency of rotatability of the modified CCD was compensated by increasing the experimental region to 5 levels which is better for the sake of the experiment's objective (prediction).

The experimental conditions with the levels that satisfies the proposed CCD are shown in the following table

Table 7. CCD design for k = 5, 10, 20, 50 and 100

	Coded Factors			Uncoded Factors		
Run	*p*	*m*	*n*	*p*	*m*	*n*
1	-1	1	-1	3	765	4
2	0	2	0	4	1030	6
3	1	1	-1	5	765	4
4	0	-2	0	4	30	6
5	2	0	0	6	500	6
6	0	0	0	4	500	6
7	0	0	-2	4	500	2
8	-1	-1	-1	3	265	4
9	0	0	0	4	500	6
10	0	0	0	4	500	6
11	1	-1	1	5	265	8
12	1	-1	-1	5	265	4
13	0	0	0	4	500	6
14	1	1	1	5	765	8
15	0	0	2	4	500	10
16	0	0	0	4	500	6
17	-2	0	0	2	500	6
18	0	0	0	4	500	6
19	-1	1	1	3	765	8
20	-1	-1	1	3	265	8

4.4.4 Considerations for the experiment

The procedure for this experiment is the same of Chapter 3's. After selecting correlations matrices (see Appendix A for a detailed explanation of the procedure) with a defined condition number, then the correlation matrix and the mean vector are estimated.

Phase II's simulation creates in-control values until one value exceeds the UCL and this sample is stored in the run-length array. As in Chapter 3, the intention is to use not only the estimated matrix but also the true matrix and matrices with variations of the correlations. **The idea is to balance the experiment with both underestimated and overestimated matrices**. The procedure is similar to the one used in Chapter 3, but in

this case, Equation (3.4), was modified in order to prevent poor deviations when the correlation is around 0. For example, suppose that the correlation is 0.5 and that this correlation has been underestimated as 0.4. Applying Equation (3.4) results in an overestimated correlation of 0.625. Now suppose that the correlation 0.5 has poorly been underestimated as 0.2, now the overestimated correlation would be 1.25 which, given that the values are correlation, is not acceptable.

A better approach can be found using

$$\hat{r}^{+/-} = \tanh(2 * a \tanh(r) - a \tanh(\hat{r}^{-/+})) \tag{4.10}$$

which is based in the transformation used to find the confidence interval of the correlation (see Montogmery, Peck and Vining, 2001). Using the previous example, if the correlation is 0.5 and the estimation leads to a correlation of 0.4, applying Equation (4.10) the resulting correlation will be 0.588, which is the upper side of the confidence interval. Even if 0.5 has been poorly underestimated as 0.2, the overestimation would be now 0.71 which is a more realistic value than the obtained with Equation (3.4).

Another problem noted using additional matrices is that the number of these matrices is function of the number of correlations, which is also related to the number of variables monitored. For example, if we have three variables being monitored, the number of correlations is three. The variations introduced by Equation (4.10) implies that now there are two levels in each correlation (one when is overestimated and the other when is underestimated) and there will be 2^3 matrices. For more complex matrices such as a *6x6* matrix, the number of correlations is 15 and 2^{15} matrices must be evaluated which, computationally, could be time consuming. Consequently, a fractional factorial design was used for $p = 4$, $p = 5$ and $p = 6$, the size of the fraction was chosen so no more than 32 estimated matrices were evaluated not including the real correlation matrix. The complete experiment design is shown in the following picture.

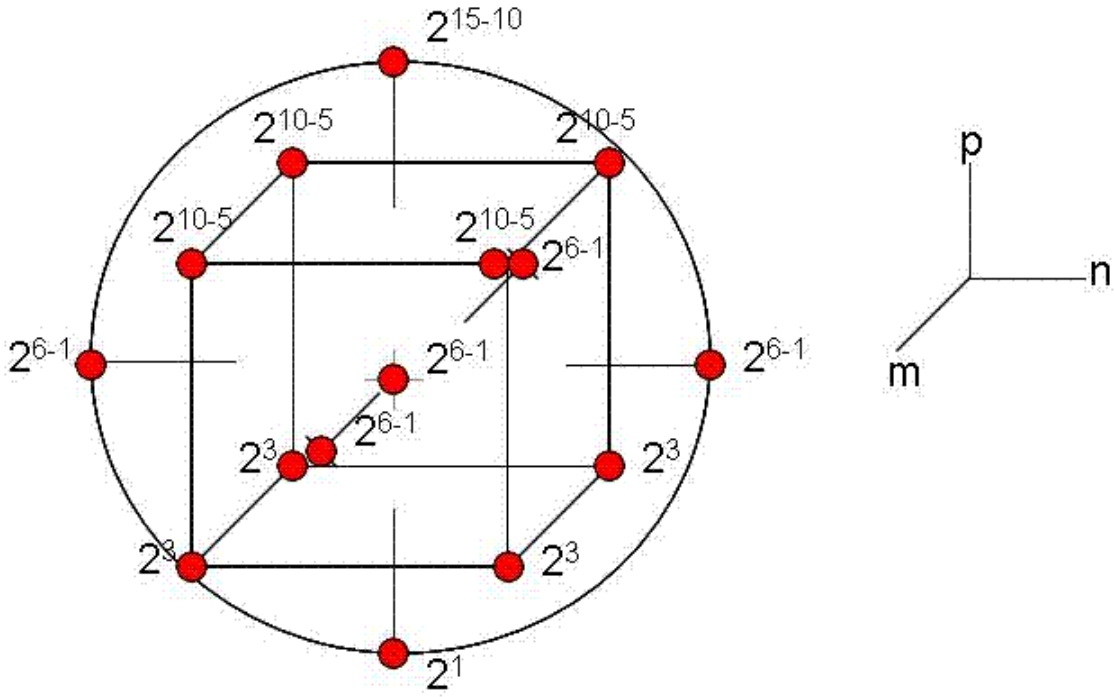


Figure 4.4-1. Complete view of the design of the experiment

4.5 Simulation procedure

4.5.1 Pseudo-Code

This part presents the pseudo-code for the simulation used to run the experiment. Practically, is the same program used to in Chapter 3 but with using Equation (4.10) instead of Equation (3.4) when determining the variation in the estimated matrix.

The computer program was developed in Matlab 7.1 (see Appendix I) and can be summarized as follows

1. Provide in control values μ_0, Σ_0 and n and m.

2. Using the Matlab function ***mvrnd*** create m samples of size n following the multivariate normal distribution with μ_0 and $\sum_0$.
3. Estimate $\hat{\mu}$ and $\hat{\sum}$.
4. Standardize to work with correlation matrices.
5. Create variations in the estimated correlations.
6. Create an array of estimated matrices using the variations.
    a. If $p \geq 4$ use the fractional factorial design provided to select matrices.
    b. Else, use all possible combinations.
7. Create an array of the inverse of the chosen matrices.
8. For each theoretical ARL
    a. Calculate UCL.
    b. Create in control samples of size n following the multivariate normal distribution with μ_0 and $\sum_0$.
        i. For each matrix
        ii. Compute the T^2 until one falls over the UCL.
    c. Save this value in an array of Run-lengths.
    d. Repeat the procedure 50 times and evaluate ARL.

The simulation run for 7 different values of theoretical ARLs: 200, 400, 600, 800, 1000, 1200 and 1400. The ARL for each condition and matrix was estimated as the average of 50 run-lengths. Also, every condition has 10 replicates.

Each replicate has different seeds. However, the random number generation was blocked using the same seed for each ARL, for instance, the same seed was used for a theoretical ARL = 200 in every experimental condition, this seed changes for different

ARLs that was used in the same experimental condition. This does not apply for the center point, which uses different seeds in each replicate.

5 Analysis and results

The analysis includes an ANOVA to find relevant factors affecting the in-control ARL. The problem of estimation is discussed in detail and the in-control ARL with both the true matrix (which can be understood as a perfectly estimated or, equivalently, a matrix where the total observations is larger or infinite) and the estimated matrices are analyzed. Following these, multiple regression has been used to fit the data to a model. The results of the effects using regression models are also presented. The whole analysis was performed using STATA 7.0.

5.1 Finding relevant factors

An ANOVA was performed to find the factors that affect the ARL. The model considers the following variables:

- Response: Logarithm of the Simulated ARL (larl_sim)
- Factors:
    - Logarithm of the Theoretical ARL (larl_theo)
    - Matrix: This is a categorical variable that assumes the value of 1 if the matrix is estimated, if the matrix is the true matrix, this variable gets the value of 0. It is the change in the response by the effect of estimation.
    - Logarithm of the condition number of the matrix used in the simulation (logk)
    - The number of samples (m_u)

- Sample size (n_u)
- The number of variables (p_u)
- The interactions:
    - Number of samples with sample size (mn1)
    - Number of samples with number of variables (mp1)
    - Sample size with the number of variables (np1)
- Quadratic effects:
    - The square of number of samples (m2)
    - The square of sample size (n2)
    - The square of number of variables (p2)

Table 8 presents the ANOVA for the logarithm of the simulated in-control ARL. Observe that the Sum of Squares primarily depends on two factors: the desired in-control ARL and the condition number. Other factors that have some influence in the in-control ARL are the effect of the estimation (if the matrix has been estimated or not), the number of samples (m2) and the sample size (n). The other factor and interactions are practically not relevant (observe that their *F* statistic is too small compared with the ones from the relevant factors).

Table 8. ANOVA of the log(simulated ARL)

```
            Number of obs =  197400     R-squared     =  0.3751
            Root MSE      = .877524     Adj R-squared =  0.3751

   Source |  Partial SS     df       MS           F     Prob > F
----------+----------------------------------------------------
    Model |  91234.0823     12  7602.84019    9873.19     0.0000
          |
     logk |  18593.3989      1  18593.3989   24145.75     0.0000
larl_theo |  71053.7332      1  71053.7332   92271.74     0.0000
      m_u |  48.1028176      1  48.1028176      62.47     0.0000
      n_u |  115.931398      1  115.931398     150.55     0.0000
      p_u |  .002376733      1  .002376733       0.00     0.9557
      mn1 |  43.7337661      1  43.7337661      56.79     0.0000
      mp1 |  35.6778549      1  35.6778549      46.33     0.0000
      np1 |  31.1762264      1  31.1762264      40.49     0.0000
       m2 |  258.120453      1  258.120453     335.20     0.0000
       n2 |  33.9514503      1  33.9514503      44.09     0.0000
       p2 |   2.2919788      1   2.2919788       2.98     0.0845
   matrix |   328.50253      1   328.50253     426.60     0.0000
          |
 Residual |  151997.596 197387  .770048665
----------+----------------------------------------------------
    Total |  243231.678 197399  1.23218293
```

5.2 Effect of estimation and the condition number

5.2.1 General effect in the average run-length

Section 4.3 presents a problem related with the estimation on the T^2 statistic. This problem is quite serious when the condition number is large. This section shows that the problem of estimation also extends to the in control-ARL. Figure 5.2-1 shows the logarithm of the simulated in-control ARL, computed using estimated and true matrices, by theoretical ARL and grouped by the condition number of the true matrix. From this figure, it is clear that results when the true matrix has a condition number of 50 o more the range of values increases approximately 2 times the range with small condition number (20 or less).

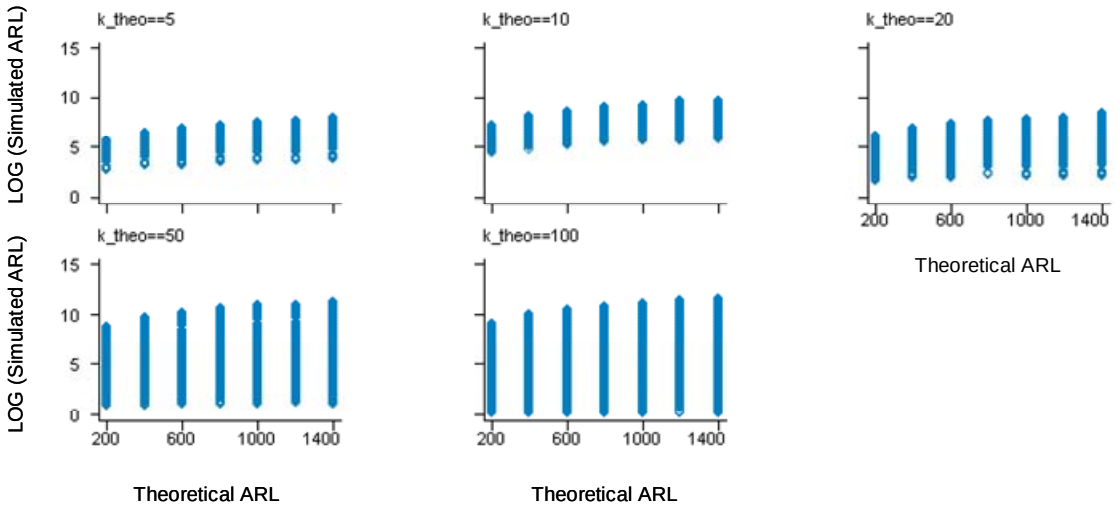


Figure 5.2-1. Log(Simulated In-control ARL) by condition number of the true matrix.

True matrices

If only the simulated in-control ARL, that comes from using the true matrices only, is considered, there is practically a minimal influence from the condition number (see Figure 5.2-2). The results are similar, even when the theoretical $k = 10$ has many outliers, the ANOVA (Table 9) using only the true matrices tell us that the influence of factors different from the theoretical ARL are irrelevant. Observe the value of F statistic and the Partial Sum of Squares in Table 9. The theoretical ARL, alone, explains about the 82% of the total variance of the logarithm of the in-control ARL.

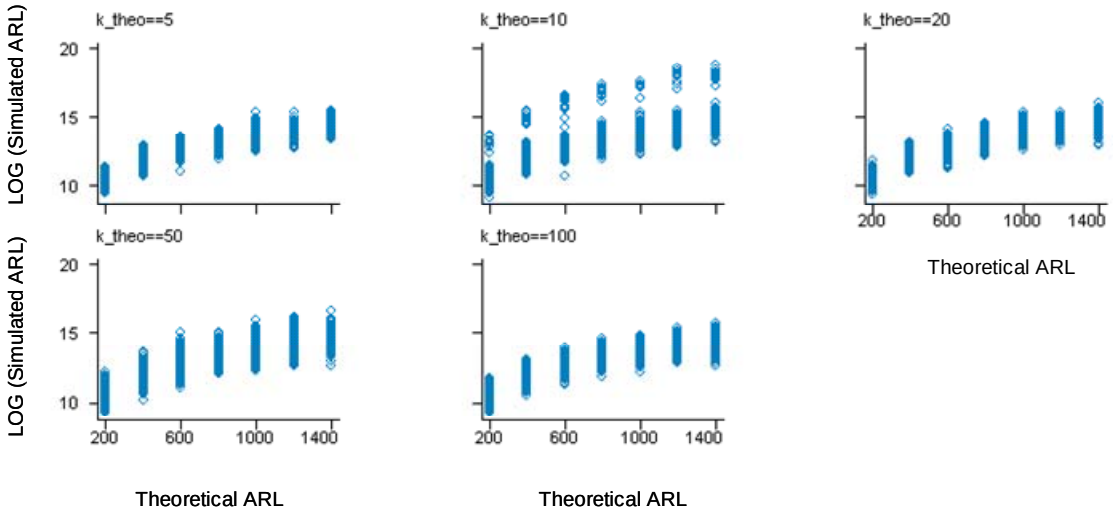
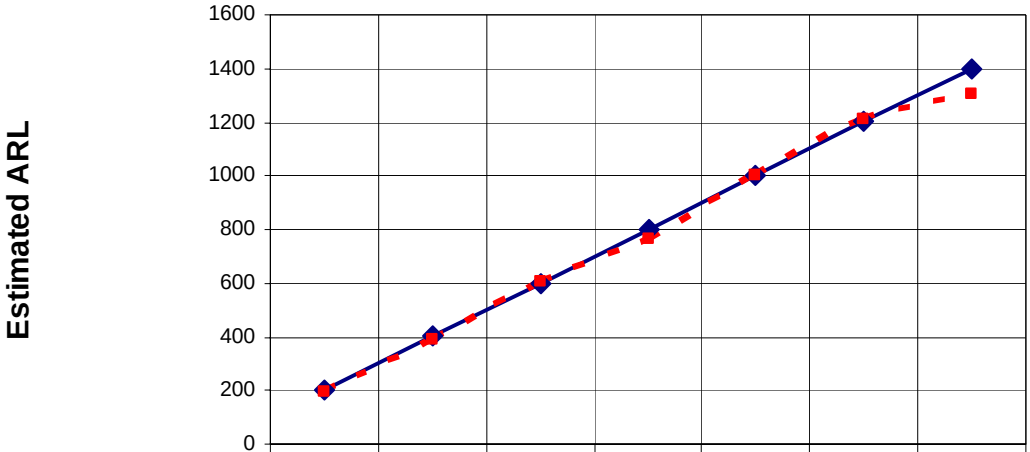


Figure 5.2-2. Log(Simulated In-control ARL) from true matrices only, by condition number of the true matrix.

Table 9. ANOVA of the log(simulated ARL) considering true matrices only

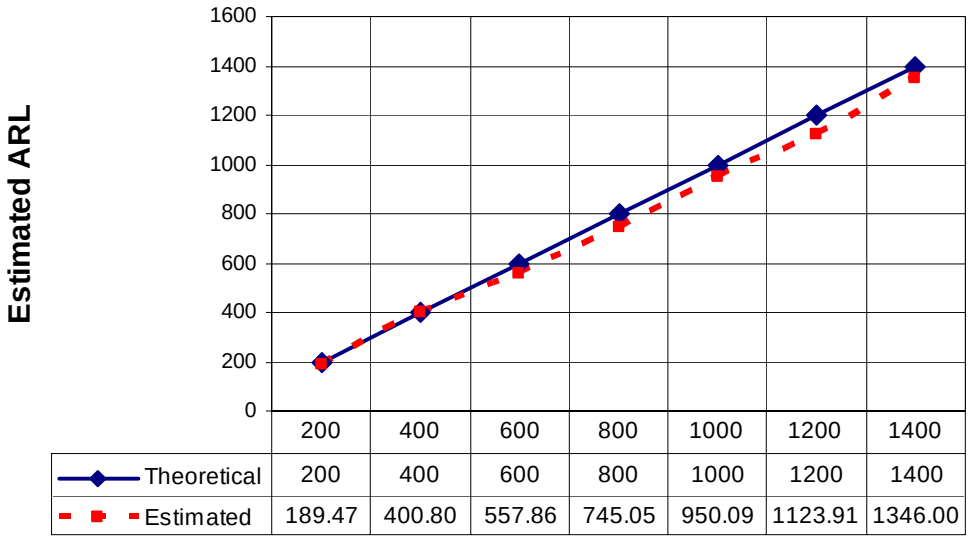
```
           Number of obs =    7000     R-squared     =  0.8500
           Root MSE      = .270606     Adj R-squared =  0.8498

   Source |  Partial SS    df       MS           F     Prob > F
----------+----------------------------------------------------
    Model |  2899.99527    11  263.635934    3600.22     0.0000
          |
     logk |  .655641954     1  .655641954       8.95     0.0028
larl_theo |  2865.32902     1  2865.32902   39128.98     0.0000
      m_u |   .14825393     1   .14825393       2.02     0.1548
      n_u |  .003044704     1  .003044704       0.04     0.8384
      p_u |  4.40292769     1  4.40292769      60.13     0.0000
      mn1 |  3.36195422     1  3.36195422      45.91     0.0000
      mp1 |  .776550964     1  .776550964      10.60     0.0011
      np1 |  2.96879617     1  2.96879617      40.54     0.0000
       m2 |  4.23127689     1  4.23127689      57.78     0.0000
       n2 |  8.65726155     1  8.65726155     118.22     0.0000
       p2 |   .78219006     1   .78219006      10.68     0.0011
          |
 Residual |  511.715874  6988  .073227801
----------+----------------------------------------------------
    Total |  3411.71115  6999  .487456943
```

Now, let's get a closer look in the simulated in-control ARL. From the 175 total experimental conditions considered, the overestimation and underestimation are distributed 54% - 46 %, as expected by the design, recall that overestimation and underestimation has been balanced through factorial designs and Equation (4.10). The results of the in-control ARLs obtained from the true matrices do not show a large deviation from the theoretical value. The average of the absolute value of scaled deviation (using Equation 3.5) by all the condition numbers is around 10%, for $p = 2$, the scaled deviation is 9.2% from the target value, for $p = 3$, is 6%, for $p = 4$ is approximately 12%, and for $p = 5$ and 6, its 8% and 13% respectively.



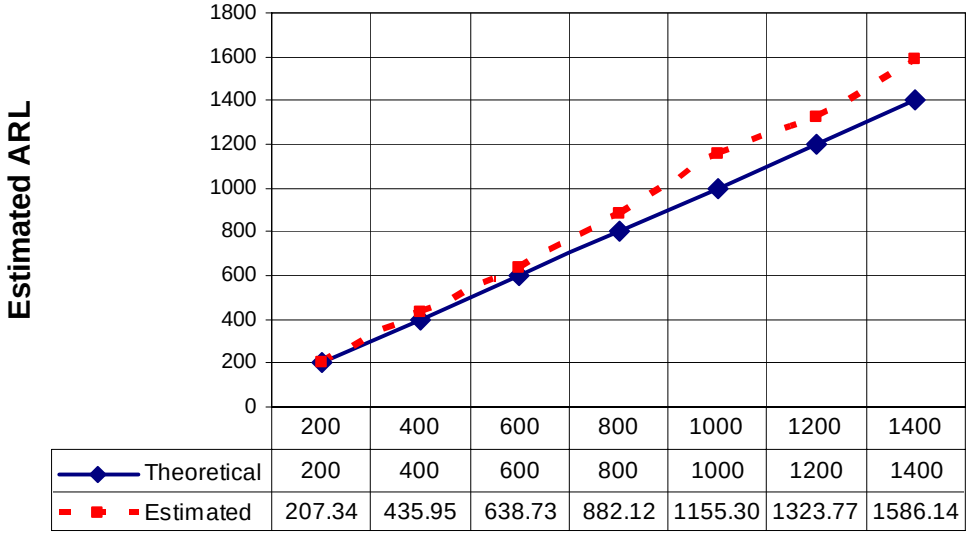
	200	400	600	800	1000	1200	1400
Theoretical	200	400	600	800	1000	1200	1400
Estimated	195.87	390.18	602.98	766.49	998.74	1210.71	1303.26

Figure 5.2-3. Theoretical in-control ARL vs. estimated in-control ARL with the true matrix ($p = 2$)



	200	400	600	800	1000	1200	1400
Theoretical	200	400	600	800	1000	1200	1400
Estimated	189.47	400.80	557.86	745.05	950.09	1123.91	1346.00

Figure 5.2-4. Theoretical in-control ARL vs. estimated in-control ARL with the true matrix ($p = 3$)



	200	400	600	800	1000	1200	1400
Theoretical	200	400	600	800	1000	1200	1400
Estimated	207.34	435.95	638.73	882.12	1155.30	1323.77	1586.14

Figure 5.2-5. Theoretical in-control ARL vs. estimated in-control ARL with the true matrix ($p = 4$)

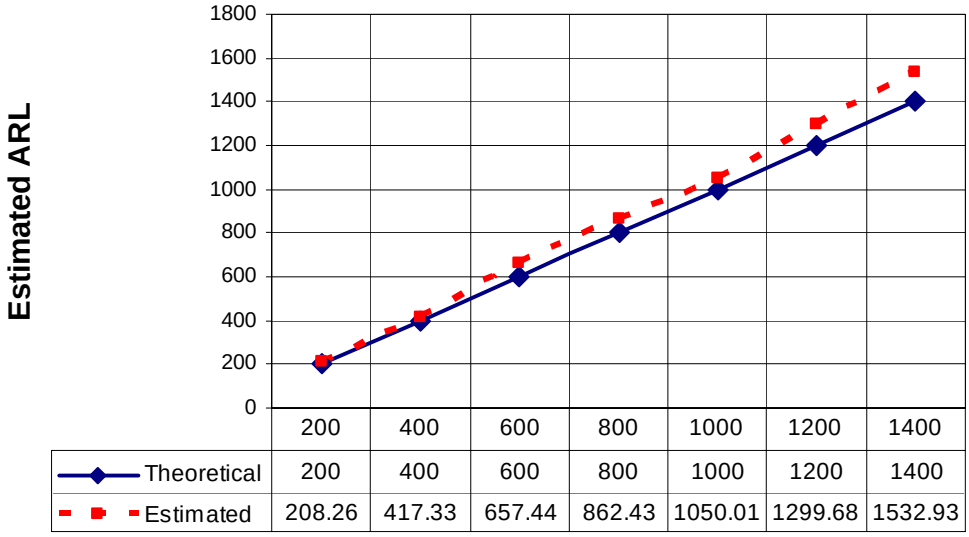


Figure 5.2-6. Theoretical in-control ARL vs. estimated in-control ARL with the true matrix ($p = 5$)

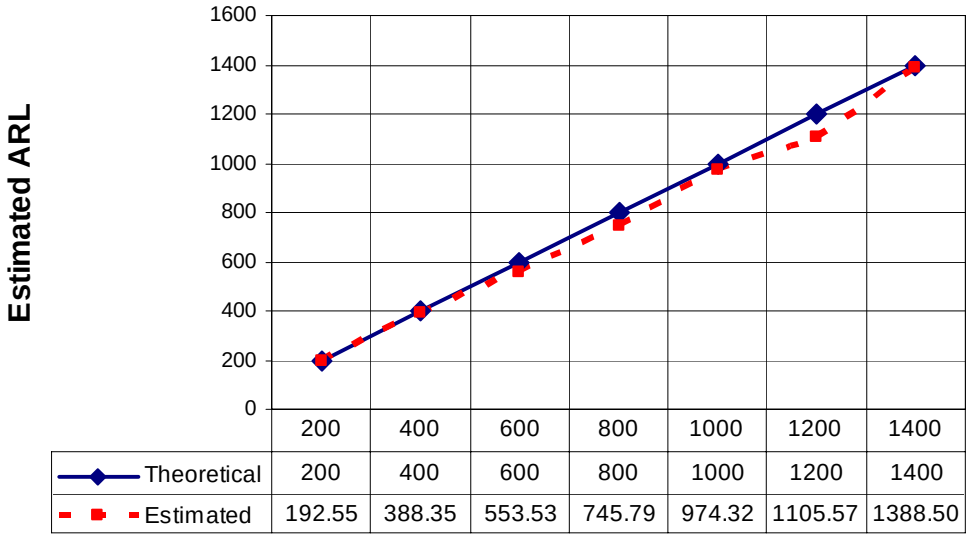


Figure 5.2-7. Theoretical ARL vs. Average of the estimated ARL with the true matrix ($p = 6$)

Estimated matrices

Now, consider the estimated matrices only. The result is very similar to the ones obtained using all matrices: there is an increase in the range of values for the estimation of the true matrices starting when k = 20. With k = 50 and 100, the range of values doubles the values of the simulated ARLs obtained from the estimation of matrices with k = 5, 10 and 20 (see Figure 5.2-8). These results are consistent with the observations about the influence of the condition number in the T^2 when the correlation matrix is estimated. As the T^2 varies it is obvious that the in-control ARL will change. The ANOVA (see Table 10) shows that when estimated matrices are used, the in-control ARL is now affected not only by the desired ARL but also by the condition number. The influence of the sample size and sample number have also increased (observe their F values).

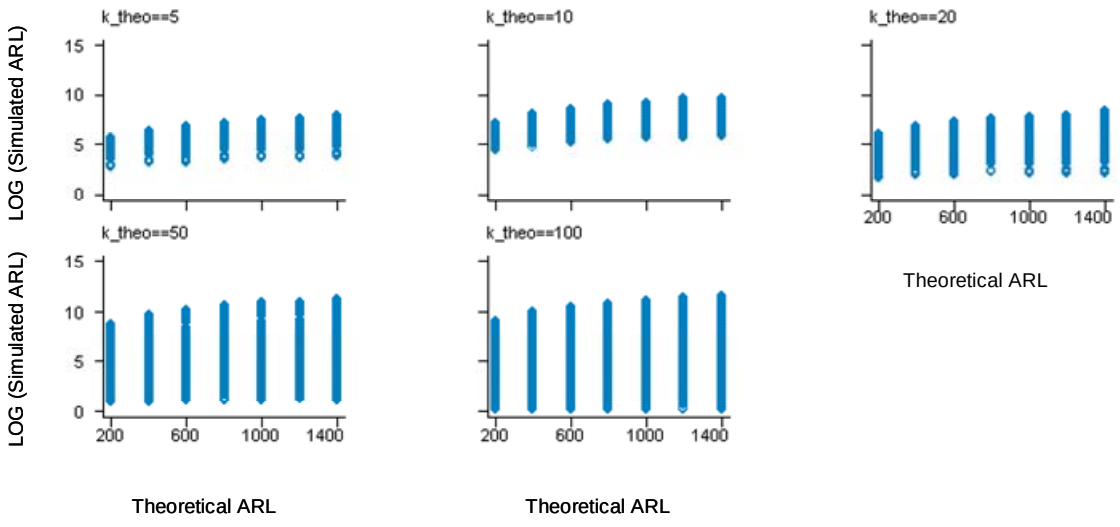


Figure 5.2-8. Log(Simulated In-control ARL) from estimated matrices only, by condition number of the true matrix.

Table 10. ANOVA of the log(simulated ARL) considering true matrices only

```
           Number of obs =  190400     R-squared     =  0.3696
           Root MSE      = .890265     Adj R-squared =  0.3695

   Source |  Partial SS     df       MS           F     Prob > F
----------+----------------------------------------------------
    Model |  88453.6678     11  8041.24252   10145.77     0.0000
          |
     logk |  19059.6698      1  19059.6698   24047.90     0.0000
larl_theo |  68199.9201      1  68199.9201   86048.98     0.0000
      m_u |  52.2228564      1  52.2228564      65.89     0.0000
      n_u |  119.270157      1  119.270157     150.49     0.0000
      p_u |  2.64541727      1  2.64541727       3.34     0.0677
      mn1 |  40.2155377      1  40.2155377      50.74     0.0000
      mp1 |  39.1316032      1  39.1316032      49.37     0.0000
      np1 |  31.2732103      1  31.2732103      39.46     0.0000
       m2 |  258.869281      1  258.869281     326.62     0.0000
       n2 |  39.7294572      1  39.7294572      50.13     0.0000
       p2 |  .090571407      1  .090571407       0.11     0.7353
          |
 Residual |  150896.002 190388  .792570972
----------+----------------------------------------------------
    Total |   239349.67 190399   1.2570952
```

A closer look to the estimated matrices, present us an averaged scaled deviation (after applying absolute values to eliminate the effect of negatives) of 23% (the details of the values are presented in Table 28 to Table 32).

In this case, the number of variables is not a relevant factor for the in-control ARL by itself. Its influence is related with the number of variables. Looking at the behavior of the in-control ARL by number of variables, see Figure 5.2-9 it seems to be that the number of samples influences in the in-control ARL. When $p = 2$, the deviation from the target value is around 10%, this deviation increases as p increases until $p = 5$ but when $p = 6$, the deviation reduces up to 20% which seems to be a contradictory result.

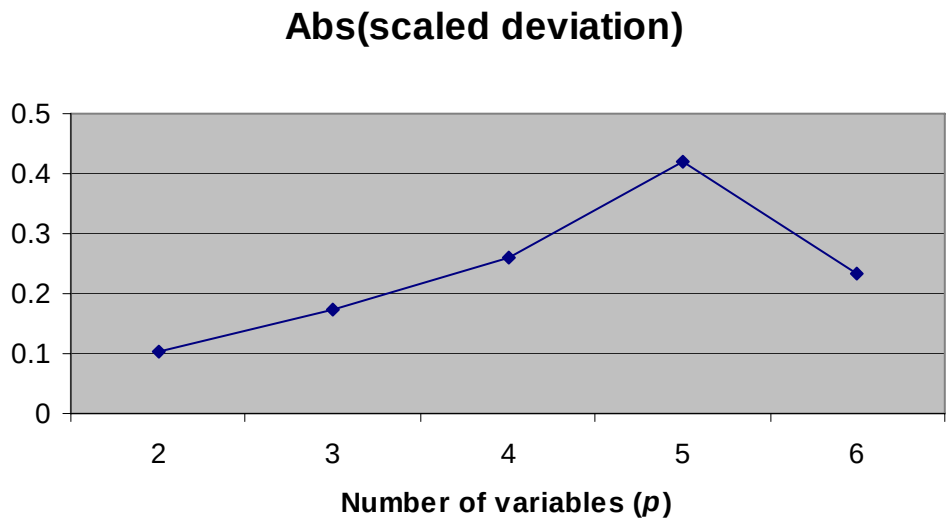


Figure 5.2-9. Average of absolute scaled deviation from the target in-control ARL by number of variables

However, this can be explained by the "error" of estimation. It is not easy to know whether the matrix has been more overestimated and underestimated because there are many correlations involved. A measurement of the "error" of estimation can be approximated by how deviated are the estimated matrices from the target value. This can be obtained by the average of the absolute values of the difference between the estimated matrices and the true matrix. Table 11 shows that the error on estimation, when $p = 6$, is slightly over the value of $p = 3$, causing the reduction of the deviation from the target ARL (see Figure 5.2-10).

Table 11. Average of the absolute values of the deviation from the target matrix.

Number of variables	Number of samples	Sample size 2	4	6	8	10
	30					
	265					
2	500			0.0023294		
	765					
	1030					
	30					
	265		0.0066598		0.005275	
3	500					
	765		0.0039379		0.0019341	
	1030					
	30			0.144781		
	265					
4	500	0.0073464		0.0040103		0.0029838
	765					
	1030			0.0024843		
	30					
	265		0.0052181		0.0037974	
5	500					
	765		0.0031583		0.002148	
	1030					
	30					
	265					
6	500			0.0023544		
	765					
	1030					

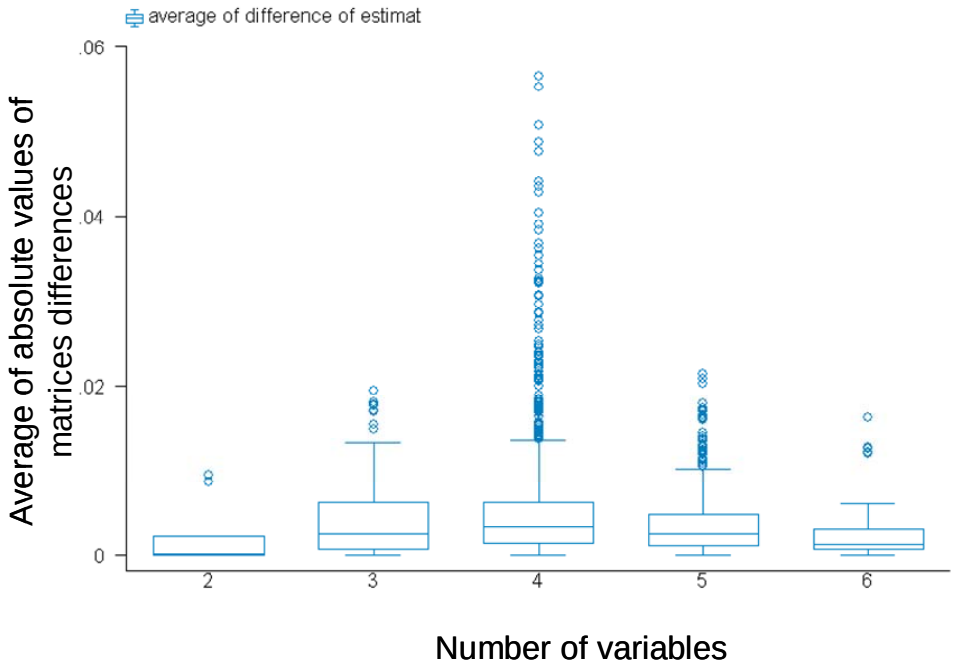


Figure 5.2-10. Boxplot of the average of the absolute values of the deviation from the target matrix by number of variables

The problem with the estimation can also be noted by plotting the results by condition number. Note that when $p = 6$, the departure is not so large like when $p = 5$ (Figure 5.2-11). It is caused by the deviation of the estimated correlation matrix from the true matrix. For example, when the condition number of the true matrix is 100, the estimated matrices when $p = 5$ has a deviation of 0.0034657 versus 0.0008409 for $p = 6$ (see Table 12).

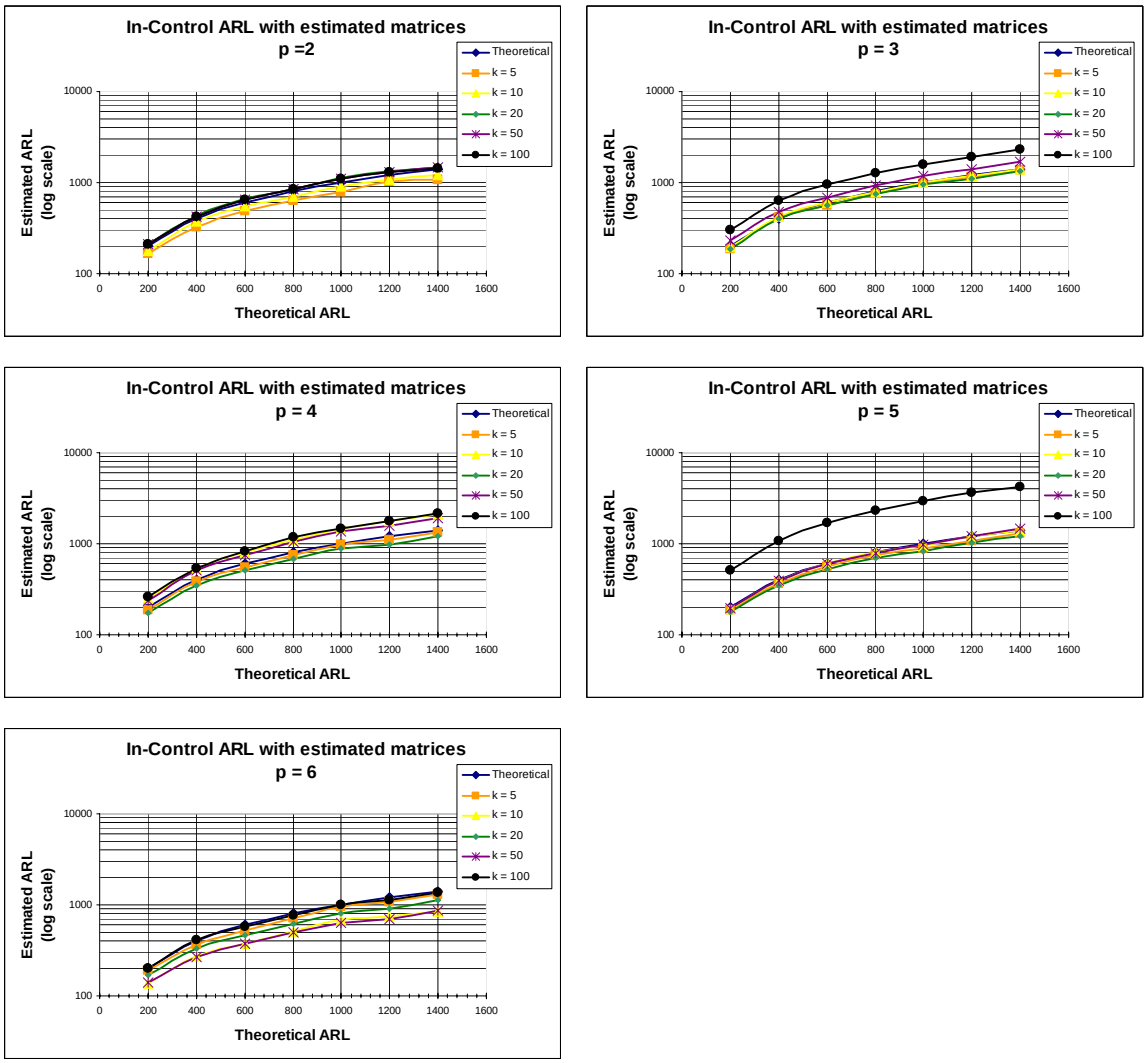


Figure 5.2-11. Estimated ARL by theoretical condition number by number of variables.

Table 12. Average of the absolute value of the difference between estimated and true matrices by theoretical in-control ARL, condition number (*k*) of the true matrix and number of variables (*p*)

		theoretical arl						
p	k	200	400	600	800	1000	1200	1400
2								
	5	.0021365	.0021365	.0021365	.0021365	.0021365	.0021365	.0021365
	10	.0089982	.0089982	.0089982	.0089982	.0089982	.0089982	.0089982
	20	.0002788	.0002788	.0002788	.0002788	.0002788	.0002788	.0002788
	50	.000149	.000149	.000149	.000149	.000149	.000149	.000149
	100	.0000846	.0000846	.0000846	.0000846	.0000846	.0000846	.0000846
3								
	5	.0034641	.0034641	.0034641	.0034641	.0034641	.0034641	.0034641
	10	.0058838	.0058838	.0058838	.0058838	.0058838	.0058838	.0058838
	20	.0044095	.0044095	.0044095	.0044095	.0044095	.0044095	.0044095
	50	.0055497	.0055497	.0055497	.0055497	.0055497	.0055497	.0055497
	100	.0029514	.0029514	.0029514	.0029514	.0029514	.0029514	.0029514
4								
	5	.0060809	.0060809	.0060809	.0060809	.0060809	.0060809	.0060809
	10	.0043604	.0043604	.0043604	.0043604	.0043604	.0043604	.0043604
	20	.0042436	.0042436	.0042436	.0042436	.0042436	.0042436	.0042436
	50	.0054545	.0054545	.0054545	.0054545	.0054545	.0054545	.0054545
	100	.0055678	.0055678	.0055678	.0055678	.0055678	.0055678	.0055678
5								
	5	.0039249	.0039249	.0039249	.0039249	.0039249	.0039249	.0039249
	10	.0028342	.0028342	.0028342	.0028342	.0028342	.0028342	.0028342
	20	.0039873	.0039873	.0039873	.0039873	.0039873	.0039873	.0039873
	50	.0036902	.0036902	.0036902	.0036902	.0036902	.0036902	.0036902
	100	.0034657	.0034657	.0034657	.0034657	.0034657	.0034657	.0034657
6								
	5	.0030623	.0030623	.0030623	.0030623	.0030623	.0030623	.0030623
	10	.0027261	.0027261	.0027261	.0027261	.0027261	.0027261	.0027261
	20	.0029748	.0029748	.0029748	.0029748	.0029748	.0029748	.0029748
	50	.0021679	.0021679	.0021679	.0021679	.0021679	.0021679	.0021679
	100	.0008409	.0008409	.0008409	.0008409	.0008409	.0008409	.0008409

On the other hand, an increment in the number of samples of sample size reduces the error (see Figure 5.2-12 and Figure 5.2-13).

.

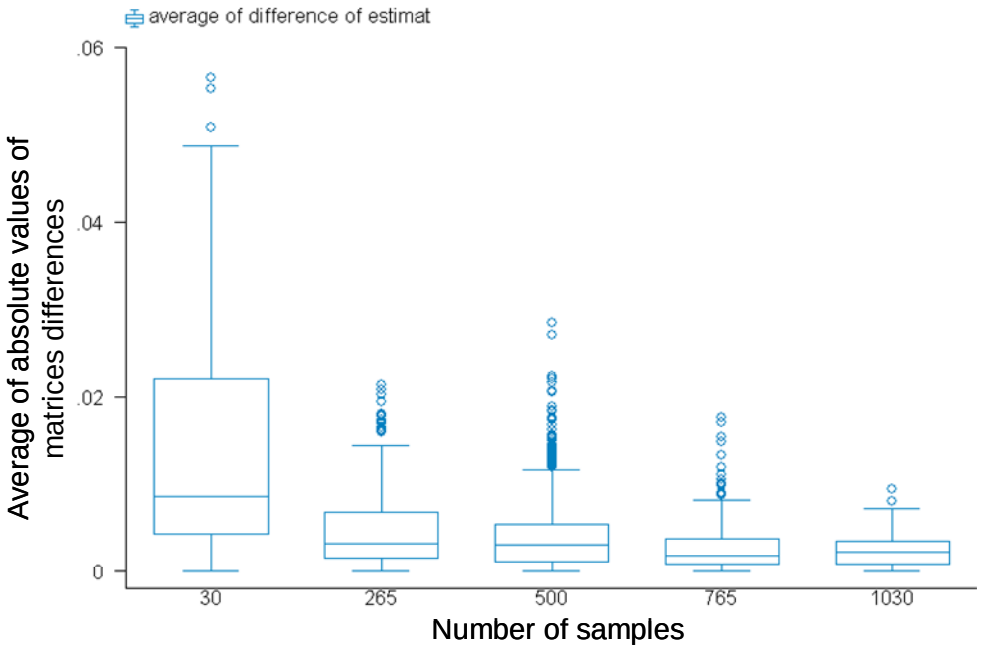


Figure 5.2-12. Average of the absolute values of the deviation from the target matrix by number of samples

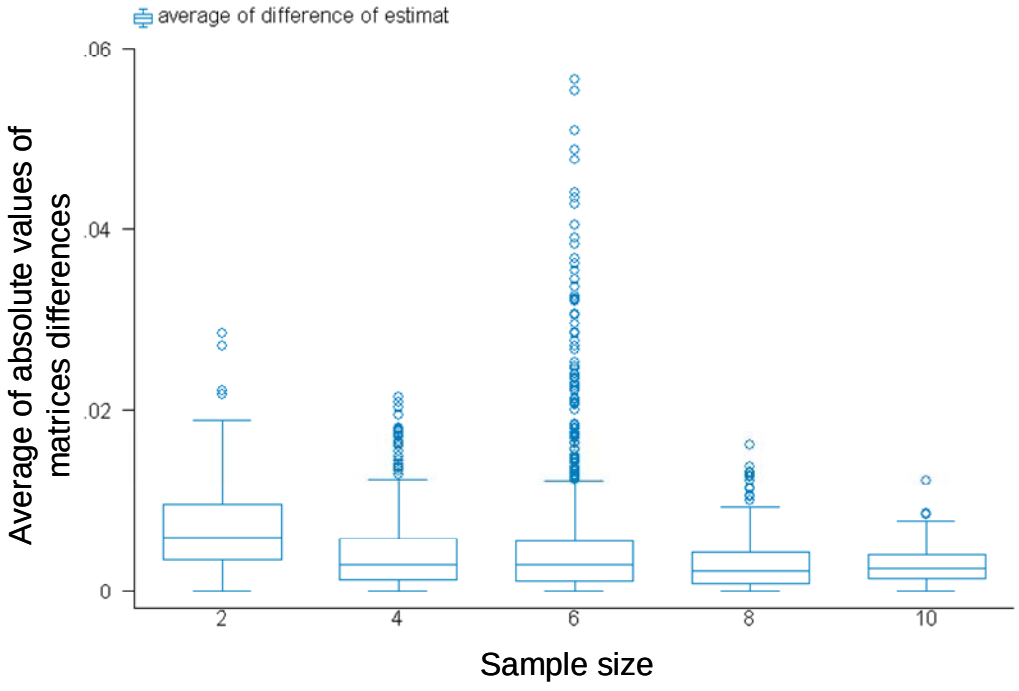


Figure 5.2-13. Average of the absolute values of the deviation from the target matrix by number of samples

5.2.2 General effect in the run-length's variance

Estimation also produces a similar effect in the variance of the run-length. Figure 5.2-14 shows that the variance increment starts when the value of the true matrix being estimated is 20. The ANOVA in Table 13 also shows that the relevant factors are the condition number, the desired value of in-control ARL (theoretical) and the effect of the matrix used.

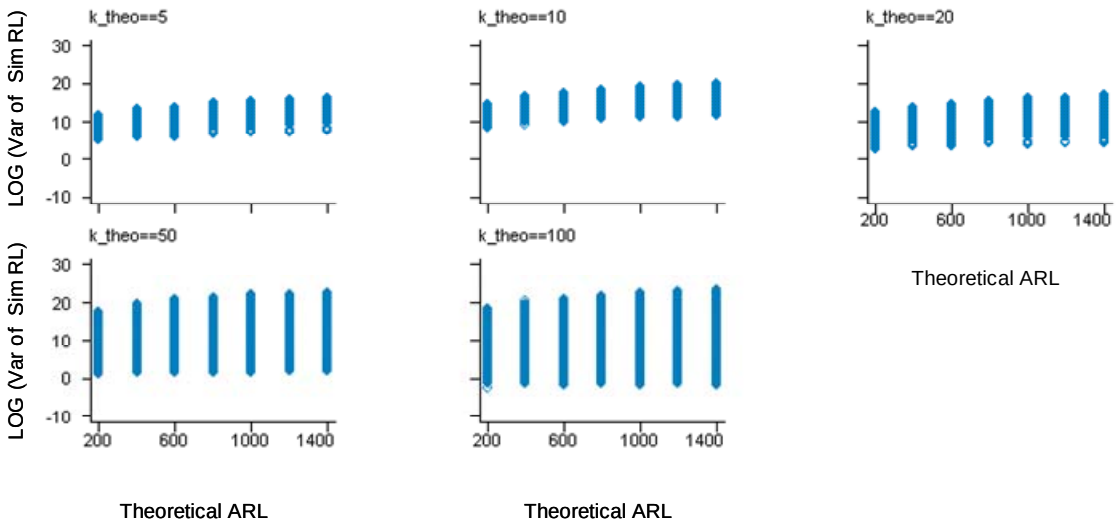


Figure 5.2-14. Log(Variance of simulated in-control run-lengths), by condition number of the true matrix.

Table 13. ANOVA of the log(Variance of RL)

```
           Number of obs =  197400     R-squared     =  0.3674
           Root MSE      = 1.79077     Adj R-squared =  0.3674

   Source |  Partial SS     df       MS           F     Prob > F
----------+----------------------------------------------------
    Model |  367705.334     12  30642.1111    9555.15     0.0000
          |
     logk |  77098.7484      1  77098.7484   24041.74     0.0000
larl_theo |  283808.573      1  283808.573   88500.18     0.0000
      m_u |  175.740154      1  175.740154      54.80     0.0000
      n_u |   504.82678      1   504.82678     157.42     0.0000
      p_u |  .017330561      1  .017330561       0.01     0.9414
      mn1 |  182.058237      1  182.058237      56.77     0.0000
      mp1 |  150.061164      1  150.061164      46.79     0.0000
      np1 |  90.2357277      1  90.2357277      28.14     0.0000
       m2 |  911.734361      1  911.734361     284.31     0.0000
       n2 |  242.297616      1  242.297616      75.56     0.0000
       p2 |  18.6289143      1  18.6289143       5.81     0.0159
   matrix |   1310.3345      1   1310.3345     408.60     0.0000
          |
 Residual |  632994.466 197387  3.20687009
 ---------+----------------------------------------------------
    Total |  1000699.80 197399   5.0694269
```

Now, analyzing separately the true matrices (Figure 5.2-15) and their estimates (Figure 5.2-16), the situation is similar as in the average run-length. When true matrices are considered alone, the variance practically depends on the desired in-control ARL (see Table 14). However, when their estimated are considered, the in-control ARL not only depends on the desired in-control ARL but also depends on the condition number (see Table 15).

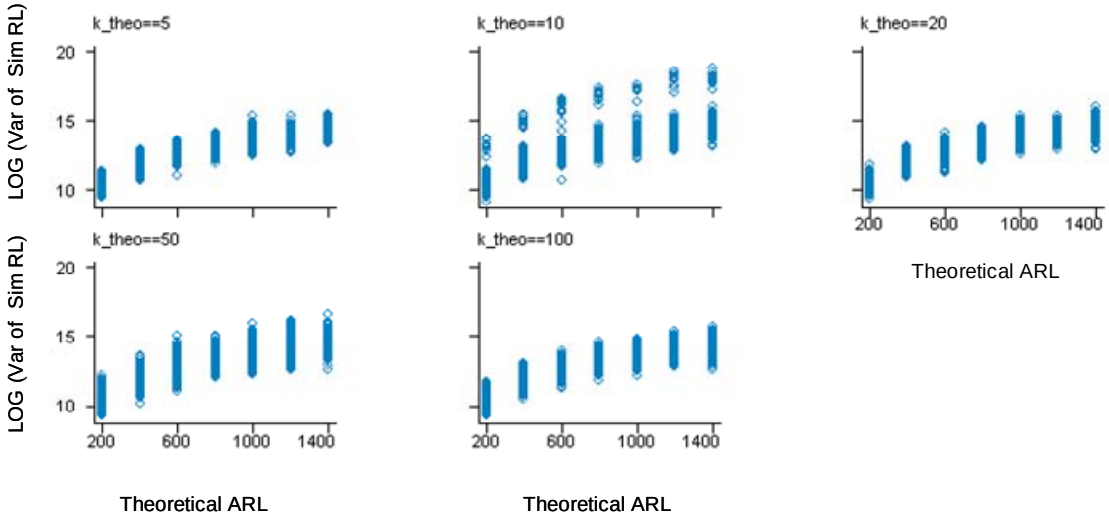


Figure 5.2-15. Log(Variance of simulated in-control run-lengths) from true matrices only, by condition number of the true matrix.

Table 14. ANOVA of the log(Variance of RL) considering true matrices only

```
           Number of obs =    7000     R-squared     =  0.8217
           Root MSE      =  .59918     Adj R-squared =  0.8214

   Source |  Partial SS    df       MS           F     Prob > F
----------+----------------------------------------------------
    Model |  11560.2844    11  1050.93495    2927.26     0.0000
          |
     logk |  2.54185351     1  2.54185351       7.08     0.0078
larl_theo |  11433.9221     1  11433.9221   31847.86     0.0000
      m_u |  .404744861     1  .404744861       1.13     0.2884
      n_u |  .008588845     1  .008588845       0.02     0.8771
      p_u |  16.2477169     1  16.2477169      45.26     0.0000
      mn1 |  13.2776334     1  13.2776334      36.98     0.0000
      mp1 |  1.91402043     1  1.91402043       5.33     0.0210
      np1 |  9.38922279     1  9.38922279      26.15     0.0000
       m2 |  13.3923802     1  13.3923802      37.30     0.0000
       n2 |  27.0656714     1  27.0656714      75.39     0.0000
       p2 |  3.94546898     1  3.94546898      10.99     0.0009
          |
 Residual |  2508.81049  6988  .359016956
----------+----------------------------------------------------
    Total |  14069.0949  6999  2.01015787
```

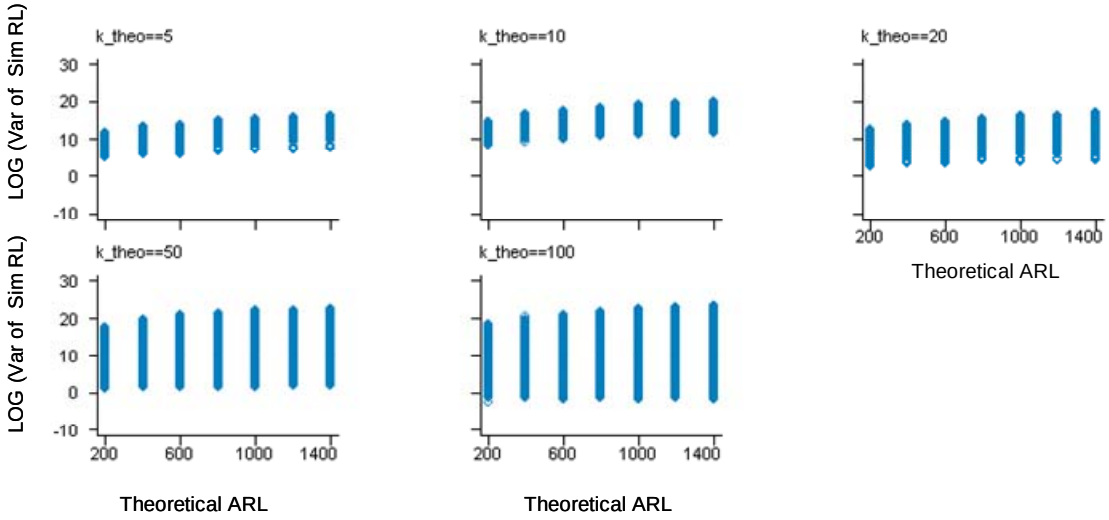


Figure 5.2-16. Log(Variance of simulated in-control run-lengths) from estimated matrices only, by condition number of the true matrix.

Table 15. ANOVA of the log(Variance of RL) considering estimated matrices only

```
            Number of obs =  190400     R-squared     =  0.3622
            Root MSE      = 1.81625     Adj R-squared =  0.3622

   Source |  Partial SS     df       MS           F     Prob > F
----------+----------------------------------------------------
    Model |  356698.352     11  32427.1229    9830.10     0.0000
          |
     logk |  79036.6619      1  79036.6619   23959.53     0.0000
larl_theo |   272419.94      1   272419.94   82582.61     0.0000
      m_u |  189.323124      1  189.323124      57.39     0.0000
      n_u |  517.003566      1  517.003566     156.73     0.0000
      p_u |  10.4090768      1  10.4090768       3.16     0.0757
      mn1 |  167.995406      1  167.995406      50.93     0.0000
      mp1 |   161.19849      1   161.19849      48.87     0.0000
      np1 |  91.2072095      1  91.2072095      27.65     0.0000
       m2 |  917.621291      1  917.621291     278.17     0.0000
       n2 |  270.709548      1  270.709548      82.06     0.0000
       p2 |  .334946096      1  .334946096       0.10     0.7500
          |
 Residual |  628043.689 190388  3.29875669
---------+----------------------------------------------------
    Total |  984742.041 190399  5.17199167
```

5.2.3 Probability of getting false alarms

An important analysis is whether the control chart has a higher probability of early false alarms. The median and the 10^{th} percentile give us an idea about it. For the true matrices only, given a combination of p and theoretical ARL, if both percentiles are smaller for different for condition numbers, it can be argued that there is an effect of the condition number in the probability of getting early false alarms.

Table 16 shows the in-control Average Run-Length (ARL), the standard deviation of the Run-Length (SDRL) and the 10^{th} (Q10), median or 50^{th} (Q50) and 90^{th} (Q90) percentiles of the run-length by the target in-control ARL and the condition number when the true matrix is used in the control chart. In any case, the lower percentiles are similar for each p and each desired in-control ARL. Hence, the probability of getting a false alarm is practically the same.

Tables 28 to 32 in Appendix E show the complete percentiles of the run-length by theoretical ARL, condition number and number of variables. Those tables show similar results as those in Table 16 for many values of theoretical ARLs and number of variables confirming that when true matrices the condition number does not affect the probailbity of getting early false alarms.

Table 16. Summary of the run-length for a theoretical ARL = 200

Number of variables	*k*	True Matrix (*mn* --> ?) Mean	SDRL	Q10	Q50	Q90
	3	163.98	172.00	12	110	386.5
	10	180.41	183.11	15	113.5	414.5
***p* = 2**	20	213.10	232.79	19	141	490
(n = 2500)	50	211.23	231.93	19	139.5	486
	100	210.63	232.07	18.5	137.5	486
	Avg	195.87	212.88	16	126.5	454
	5	187.28	183.80	22	133	431
	10	200.45	194.22	26	144	450
***p* = 3**	20	191.57	187.37	22	135	445
(n = 10000)	50	173.73	170.84	21	120	392
	100	194.31	191.73	22	135	455
	Avg	189.47	185.95	22	133	436
	5	189.05	190.79	20	130	439
	10	261.04	347.70	23	150	589.5
***p* = 4**	20	199.24	205.29	20.5	132	475
(n = 25000)	50	200.12	216.80	20	131	464
	100	187.26	192.30	20	127	433
	Avg	207.34	239.63	21	132	481
	5	192.23	187.83	22	131	435
	10	194.46	192.22	22	131	450
***p* = 5**	20	211.47	212.23	23	143.5	475
(n = 10000)	50	227.84	230.24	26	158	509.5
	100	215.31	217.38	25	147	498
	Avg	208.26	208.97	24	141	472
	5	200.44	217.20	17	128	480
	10	140.44	158.22	12	90	311.5
***p* = 6**	20	211.00	211.94	21.5	147.5	475
(n = 2500)	50	168.22	172.49	15	109	398
	100	242.67	242.83	24.5	158	565
	Avg	192.55	205.79	17	127	444.5

However, when percentiles of estimated matrices are compared with the percentiles of the true matrices, the situation is different. Table 17 contains the percentiles of the run-length for the true matrix and the percentiles of the run-length from estimations of the true matrix when the number of variables, p, is 2.

Table 17. Summary of the run-length when $p = 2$

		Estimated Matrix					True Matrix ($mn \to \infty$)				
	k	Mean	SDRL	Q10	Q50	Q90	Mean	SDRL	Q10	Q50	Q90
ARL = 200	3	163.64	171.29	12	110	388	163.98	172.00	12	110	386.5
	10	178.36	199.54	15	111	410.5	180.41	183.11	15	113.5	414.5
	20	210.44	222.91	19	138	490	213.10	232.79	19	141	490
	50	210.44	222.91	19	138	490	211.23	231.93	19	139.5	486
	100	210.44	222.91	19	138	490	210.63	232.07	18.5	137.5	486
	Avg	194.66	209.77	17	126	454	195.87	212.88	16	126.5	454

The condition number does not affect both the percentiles nor the mean. However if the number of variables increases, see Table 18, the difference is significant both in the ARL, the percentile 10^{th} and median when more variables are been monitored.

This implies that there is a higher probability of getting early false alarm when one uses estimated matrices instead of the true matrix when the chart monitors more than 2 variables. Tables 28 to 32 provide similar results for $p = 3$, 5 and 6.

Table 18. Summary of the run-length when $p = 4$

		Estimated Matrix					True Matrix ($mn \to \infty$)				
	k	Mean	SDRL	Q10	Q50	Q90	Mean	SDRL	Q10	Q50	Q90
ARL = 200	5	184.44	191.82	17	124	433	189.05	190.79	20	130	439
	10	249.87	346.51	21	146	565	261.04	347.70	23	150	589.5
	20	175.27	197.15	16	113	417	199.24	205.29	20.5	132	475
	50	238.34	746.58	13	113	453	200.12	216.80	20	131	464
	100	258.11	995.26	9	89	423	187.26	192.30	20	127	433
	Avg	221.21	591.54	15	117	460	207.34	239.63	21	132	481

5.3 Model fitting and analysis of the effects in the models

This work not only focuses in finding factors but also in obtaining a prediction model by the use of multiple regression. The main assumption is that practitioners do not know if the parameters have the true values or not. Hence, even when regressions using the true matrices only have a large R^2 (those models can be found in Appendix H), both cases are considered as a categorical variable (as explained in Section 5.1).

5.3.1 A general regression model

The general model considers all the variables presented in Section 5.1. As it was previously explained, some variables and the response were transformed to reduce the size of the sum of squares.

After eliminating not significant factors, the final model has been practically reduced to the desired ARL (larl_theo), the effect of the estimation, the condition number and the total samples used (m_u and n_u through their interactions), see Table 9. The first three factors are the most relevant factors. The number of variables is not relevant by itself, but influences through interactions.

The multiple determination coefficient (R^2) is 0.37 and the root of the Mean Square Error (MSE) for the logarithm of the in-control ARL is 0.87. This, as it was pointed out in Chapter 4, is also a measurement of the standard deviation of the model. Considering that the response variable was transformed by logarithm, the real root MSE for the in-control ARL was 2458.56.

However, the residual analysis in this case is not as good as expected. The residuals are not normally distributed. Figure 5.3-1 shows the normal q-q plot. Looking at the histogram and kernel density estimate of the distribution of the residuals (see Figure 5.3-2)

it is clear the there is a larger peek in the residuals (produced by the influence of extreme values), but note that the residuals distribution is symmetric.

A Box-Cox transformation was applied in order to improve the model (see Appendix E), but the results only led to a minimal improvement, so the original model (without transformation) was kept.

Table 19. General regression model

```
      Source |       SS       df       MS              Number of obs =  197400
-------------+------------------------------           F(  7,197392) =16898.87
       Model |  91142.9723     7  13020.4246           Prob > F      =  0.0000
    Residual |  152088.706197392  .770490728           R-squared     =  0.3747
-------------+------------------------------           Adj R-squared =  0.3747
       Total |  243231.678197399  1.23218293           Root MSE      =  .87778

------------------------------------------------------------------------------
    larl_sim        Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------------------------------------------------------------
_cons            .6643619   .0229689    28.92   0.000     .6193434    .7093805
larl_theo        .9462778   .0031161   303.68   0.000     .9401703    .9523852
logk            -.2486753   .0015999  -155.43   0.000     -.251811   -.2455396
n_u               .095881   .0028963    33.10   0.000     .0902043    .1015576
mn1             -.0000829   3.87e-06   -21.40   0.000    -.0000905   -.0000753
np1             -.0096965   .0004578   -21.18   0.000    -.0105938   -.0087992
m2               4.39e-07   2.12e-08    20.74   0.000     3.97e-07    4.80e-07
matrix
          1      .2207646    .010717    20.60   0.000     .1997595    .2417697
          2     (dropped)
------------------------------------------------------------------------------
```

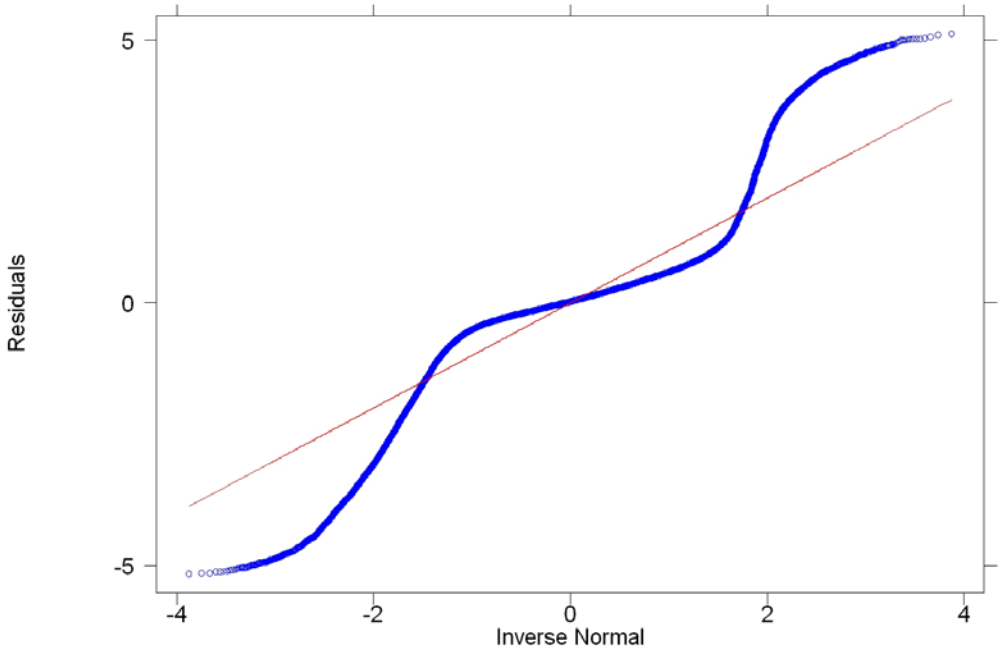
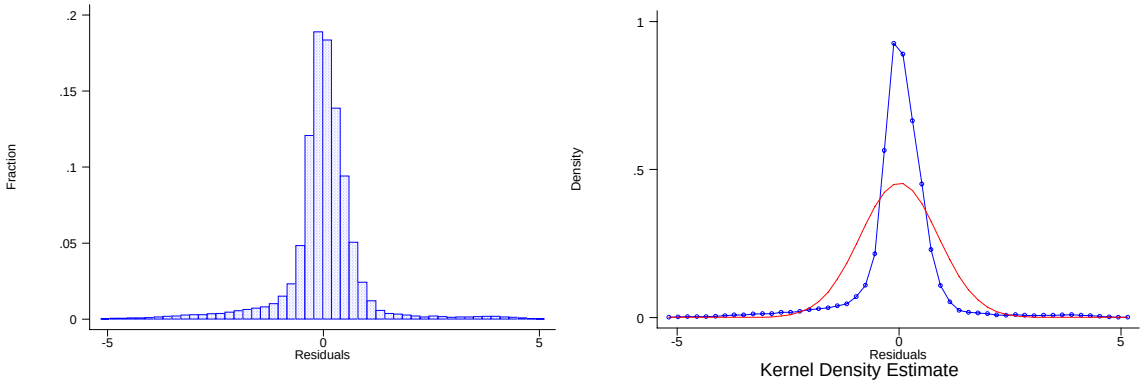


Figure 5.3-1. Normal Q-Q plot of the residuals



(a) (b)

Figure 5.3-2. (a) Histogram of the residuals (b) Kernel density of the residuals

Effects

To evaluate the effects of the considered factors, the factor under analysis has been varied fixing the other ones. For example, to analyze the effect of the sample size in the ARL, the factors: Theoretical ARL, k, m and p has been fixed at 600, 5, 500 and 4 respectively (see Figure 5.3-3), varying only the factor that is being analyzed, n. In this case, the relationship between n and the ARL is linear, and as n increases, the ARL increases, it tends to overestimate given the other factors fixed. This does not occur with the number of samples, m, which has a quadratic relationship (see Figure 5.3-4). In this case, the closest value to the theoretical 600 is about 650, reached when m is approximately 440. In both cases, the effect of the matrix makes the results get overestimated if estimated matrices are being used.

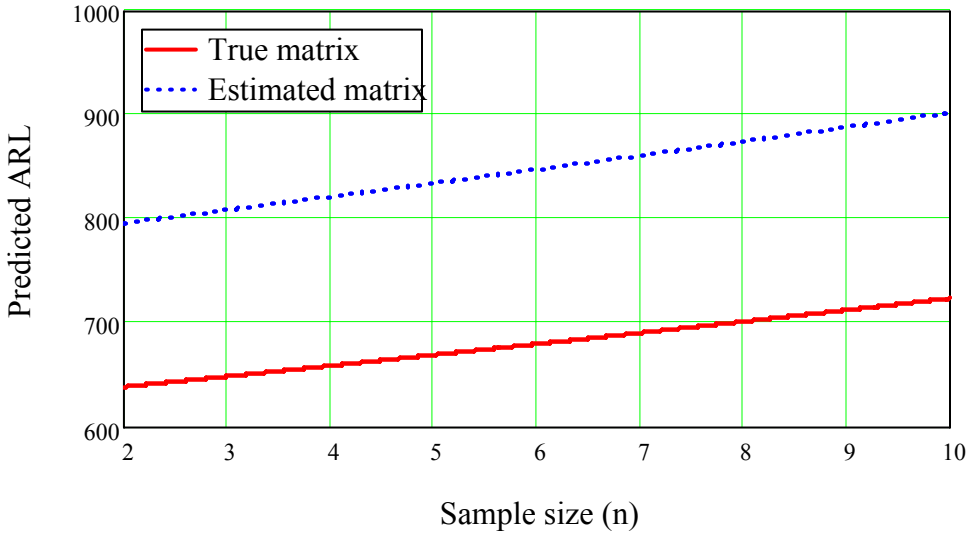


Figure 5.3-3. Effect of the sample size (Theoretical ARL = 600, k = 5, m = 500 and p = 4)

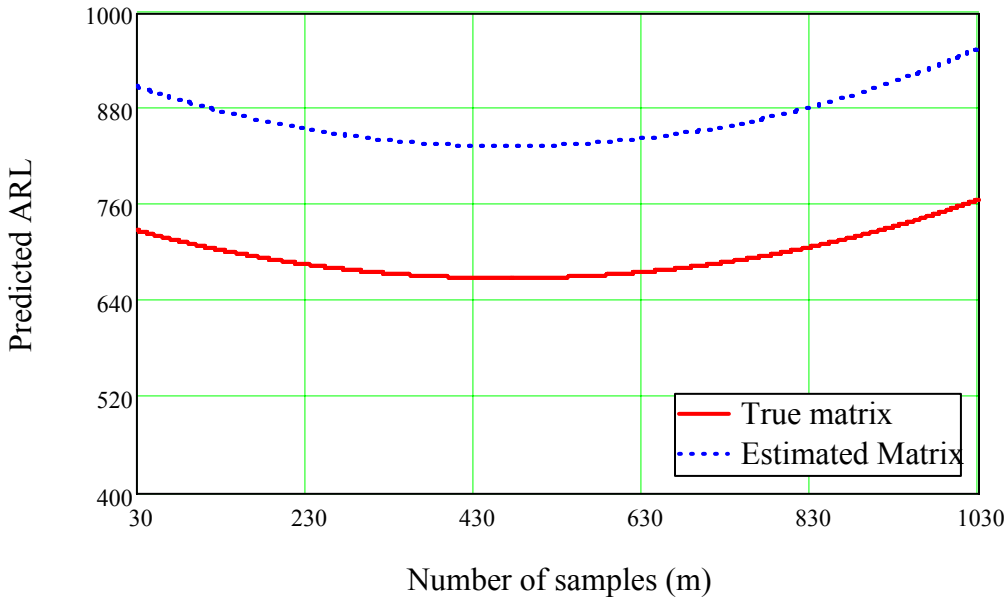


Figure 5.3-4. Effect of the number of samples (Theoretical ARL = 600, k = 5, n = 5 and p = 4)

The effect of p is similar to the one described by Champ et al. (2005), he states that as p increases, the ARL overestimation decreases. This is exactly what is happening in Figure 5.3-5. Figure 5.3-6 shows that, when the true matrix is being used, the effect of the target ARL is an overestimation that decreases as the theoretical ARL increases. Observe that the values for theoretical ARLs of 200, 400, 600, 800, 1000, 1200 and 1400 are 236, 455, 668, 877, 1084, 1289 and 1491 and the overestimation is about 18% for theoretical ARL = 200 and decreases until 6.9% when ARL = 1400. But, when the matrix is replaced by the estimated matrix, the overestimation increases as larger theoretical ARLs are desired. The effect of the condition number shows an overestimation if $k < 10$ and a underestimation for higher values, as k increases the ARL decreases (see Figure 5.3-7). When k = 7.7, m = 500, n = 5 and p = 4 the ARL reaches the exact value of the theoretical ARL (600).

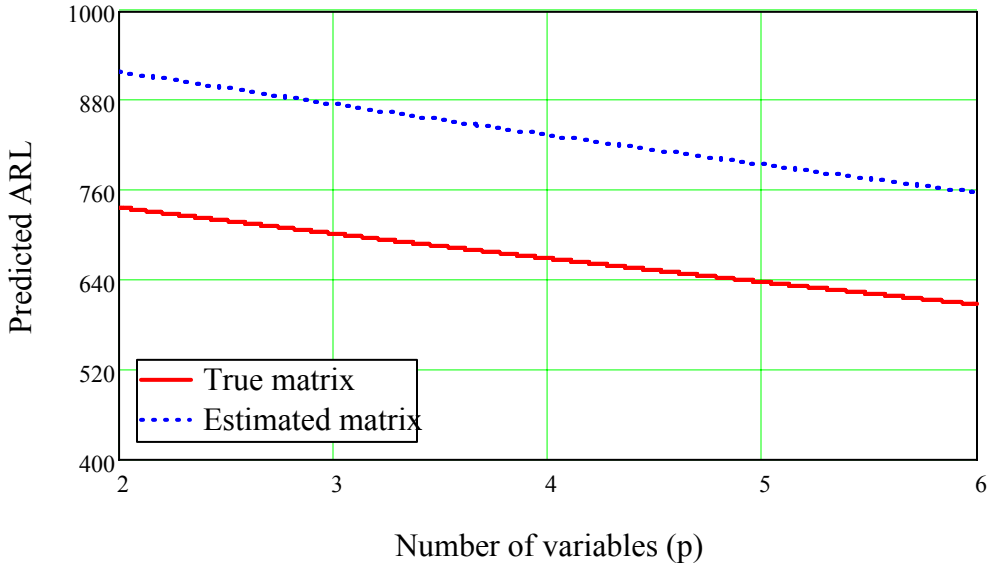


Figure 5.3-5. Effect of the number of variables (Theoretical ARL = 600, *k* = 5, *m* = 500 and *n* = 5)

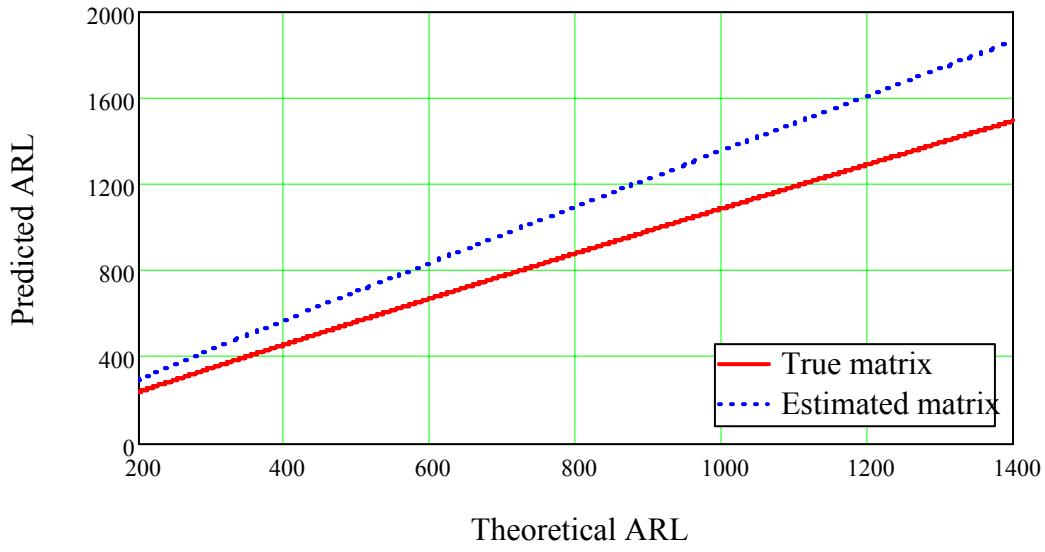


Figure 5.3-6. Effect of the Theoretical ARL (k = 5, m = 500, n = 5 and p = 4)

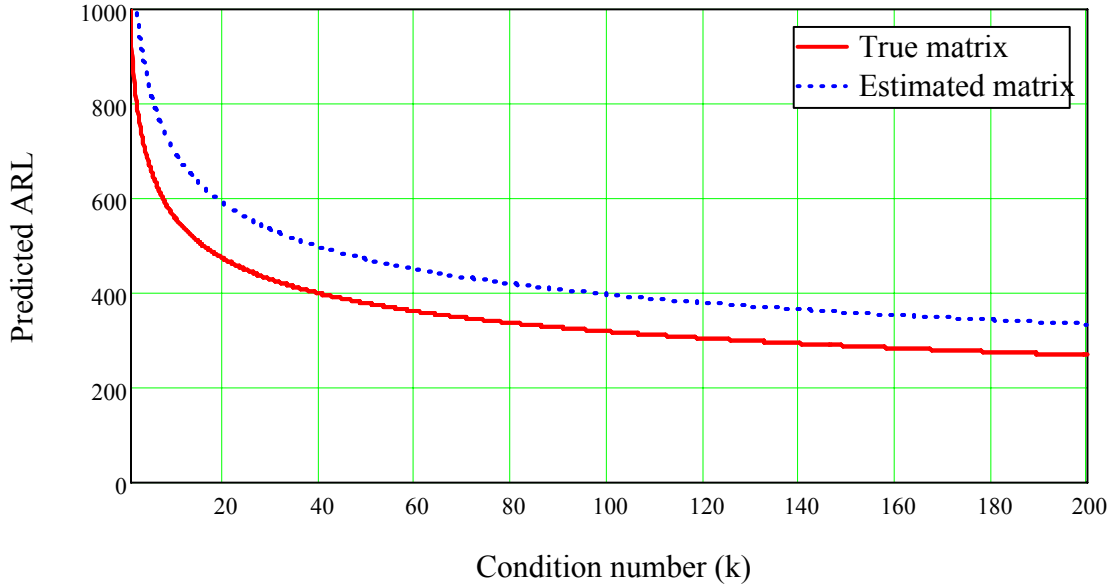


Figure 5.3-7. Effect of the condition number (Theoretical ARL = 600, m = 500, n = 5 and p = 4)

The effect of the interaction mn can be understood as the total number of observations. Figure 5.3-8 shows the contour plot when the true matrix is used to predict the in-control ARL of 600 with $p = 4$ and $k = 5$. The closest value to the theoretical ARL is around 650, and it is obtained for combinations of m and n such as $m = 200$ and $n = 4$ or $m = 600$ and $n = 6$. But, for a fixed value of m or n, if the other factor increases also the ARL increases. The effect of the matrix, as in the main effects, increases the in-control ARL.

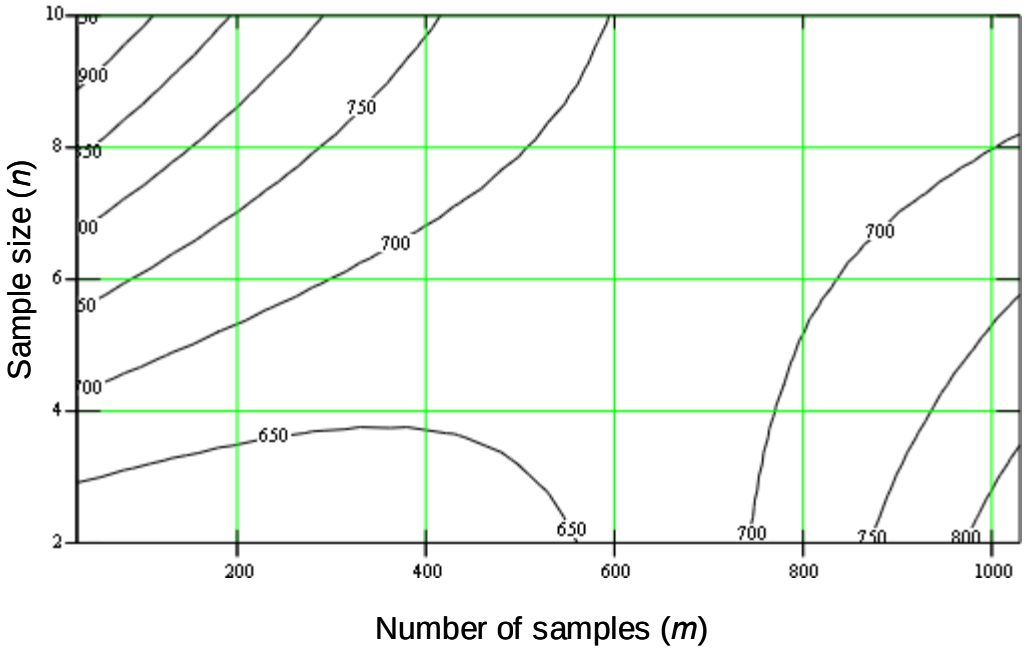


Figure 5.3-8. Effect of the interaction *mn*. Contour plot with true matrix(p = 4, k = 5 and Theoretical ARL = 600)

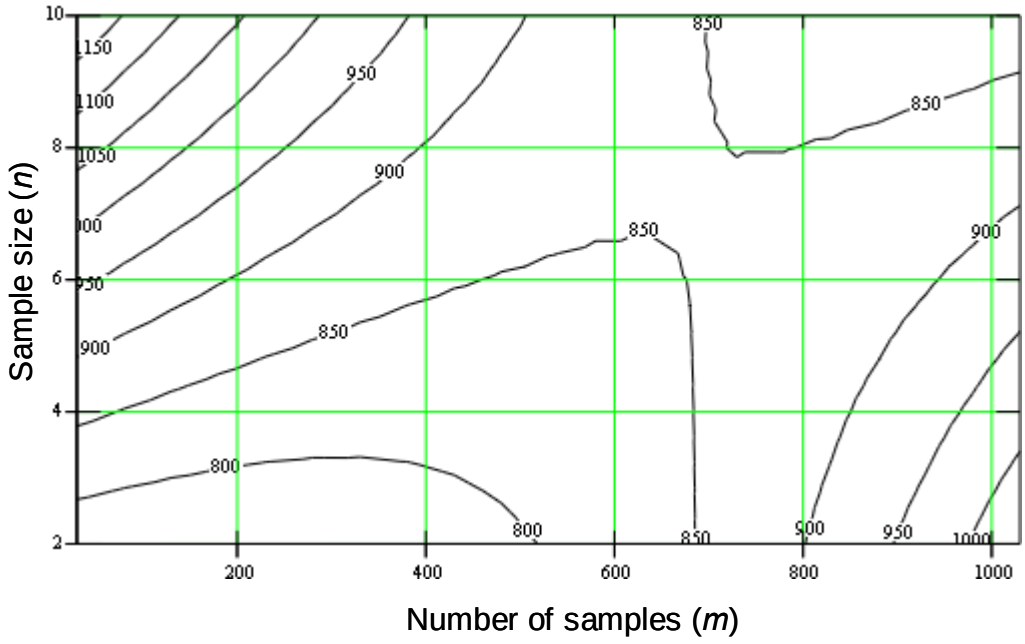


Figure 5.3-9. Effect of the interaction *mn*. Contour plot with estimated matrix (p = 4, k = 5 and Theoretical ARL = 600)

Fixing the values of *m* and *k* in 500 and 5, and for a theoretical ARL of 600, the effect of *np* tells us the relationship between the sample size and the number of variables. In this case, for a large number of variables (5 or 6) large samples sizes are better. For small number of variables large sample sizes only increases the overestimation (see Figure 5.3-10 and Figure 5.3-11).

Finally as in the other cases, estimated matrices overestimates the ARL in about 22% approximately.

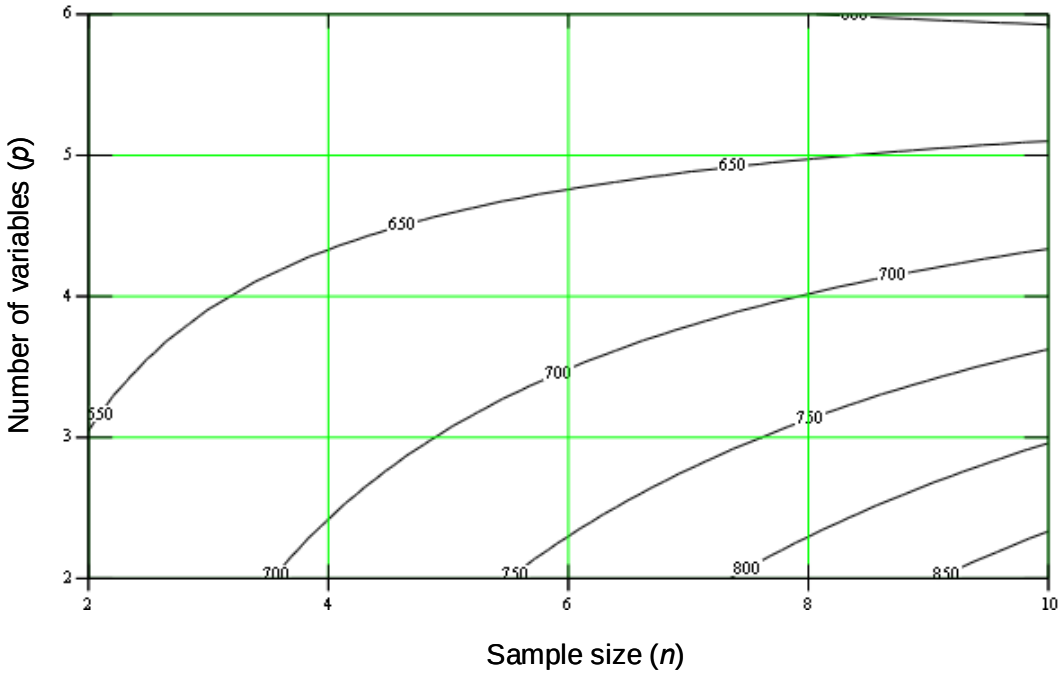


Figure 5.3-10. Effect of the interaction *np*. Contour plot with true matrix (*m* = 500, *k* = 5 and Theoretical ARL = 600)

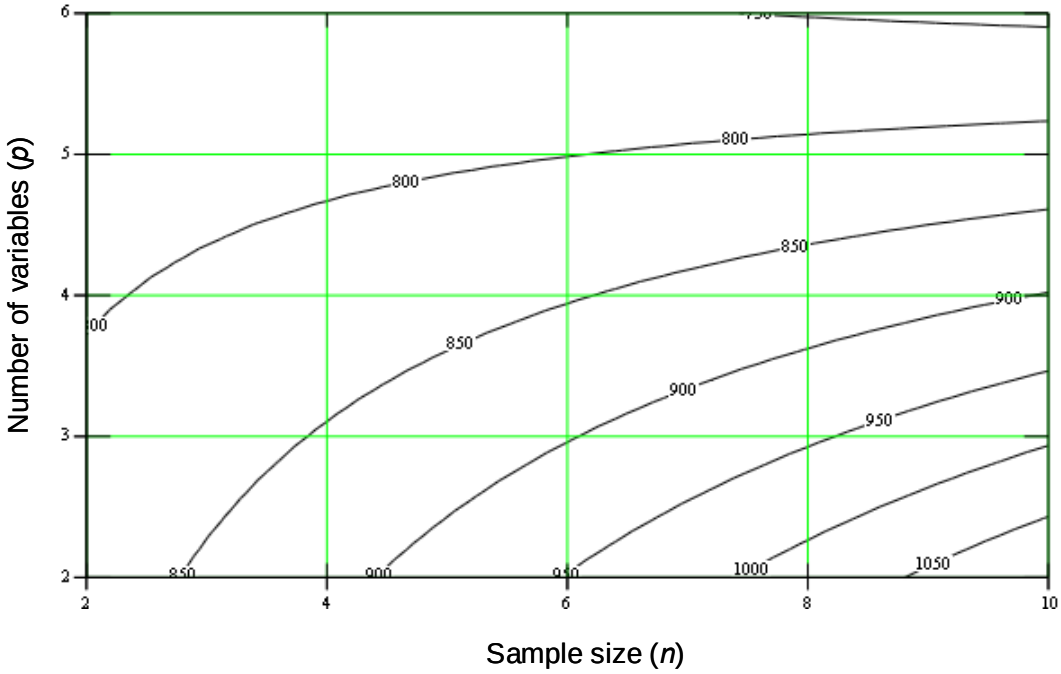


Figure 5.3-11. Effect of the interaction *np* Contour plot with estimated matrix (*p* = 4, *k* = 5 and Theoretical ARL = 600)

5.3.2 Improving the model with individual regression models (by condition number, k)

To improve the multiple determination coefficient (R^2), regression models has been evaluated after dividing the observations by the observed condition number ($\hat{k}$). The criterion used was the use of percentiles of the observed *k*. Table 10 shows its summary statistics. The 25th, 50th and 75th percentiles are used as cutting points, hence there are now 4 models: when $\hat{k}$ <10, 10≤ $\hat{k}$ <20, 20≤ $\hat{k}$ <60 and 60≤ $\hat{k}$. The rotatability and orthogonality are not affected because these cutting points are close to the initial target condition numbers chosen in the experimental design.

Table 20. Summary statistics of the observed condition number ($\hat{k}$)

condition number of estimated matrices

	Percentiles	Smallest		
1%	4.217589	2.874916		
5%	4.57041	2.874916		
10%	4.843006	2.874916	Obs	197400
25%	9.559045	2.874916	Sum of Wgt.	197400
50%	20.28739		Mean	62.02734
		Largest	Std. Dev.	313.9931
75%	61.81712	14561.97		
90%	106.1836	14561.97	Variance	98591.65
95%	152.1191	14561.97	Skewness	36.50604
99%	604.6937	14561.97	Kurtosis	1628.156

The models and their R^2 are summarized in Table 21. For three of the models, the R^2 has improved the result of the general model, but for $60 \leq \hat{k}$, the R^2 poorly reaches about the 50% of the initial model. This shows the difficulty of finding a model for ill-conditioned matrices. Practically, the prediction under this condition is impossible. The root MSE presented in this table is the root MSE for the ARL. It has been obtained by applying

$$root(MSE) = \sqrt{\frac{\sum_{i=1}^{n} \left(ARL_{simulated} - \exp(\hat{y}) \right)^2}{\#observations - \#parameters}} \tag{5.1}$$

Table 21. Summary of the regressions by model. It includes the regression coefficients and standard errors in parenthesis.

Coefficients	General	*k<10*	*10<=k<20*	*20<=k<60*	*k>=60*
constant	0.6643619	-0.4592629	-0.4328219	-0.3269787	2.913661
	(0.229689)	*(0.0192468)*	*(0.0271763)*	*(0.0343999)*	*(0.0899396)*
larl_theo	0.9462778	1.007269	0.9983577	0.9558468	0.8278754
	(0.0031161)	*(0.0024992)*	*(0.002944)*	*(0.0035772)*	*(0.0105805)*
logk	-0.2486753	---	0.119801	0.0793531	-0.6466551
	(0.0015999)	---	*(0.006834)*	*(0.0062001)*	*(0.0104033)*
n_u	0.0958810	0.0096051	0.1066148	0.0673399	0.1547571
	(0.0028963)	*(0.0008759)*	*(0.0027878)*	*(0.0033923)*	*(0.0073366)*
mn1	-0.0000829	---	-0.0001366	-0.0000337	-0.0002547
	(3.87E-06)	---	*(3.68E-06)*	*(4.65E-06)*	*(1.29E-05)*
np1	-0.0096965	---	-0.0107158	-0.0107668	---
	(0.0004578)	---	*(0.0004426)*	*(0.0005265)*	---
m2	4.39E-07	2.99E-08	5.17E-07	2.54E-07	1.25E-06
	(2.12E-08)	*(6.24E-09)*	*(2.00E-08)*	*(2.43E-08)*	*(7.30E-08)*
effect 1	0.2207646	---	0.1140992	0.1996949	0.6121385
	(-0.10717)	---	*(0.0103353)*	*(0.0110428)*	*(0.0410418)*
2	(dropped)	(dropped)	(dropped)	(dropped)	(dropped)
n	197400	59010	37170	50400	50820.00
SSE	1.19314E+12	5.02E+10	8089673145	19949410212	1.11909E+12
MSE	6044508	850580	217686.70	395884	22023710
root(MSE)	**2458.56**	**922.27**	**466.57**	**629.19**	**4692.94**
R^2	**0.37**	**0.74**	**0.76**	**0.59**	**0.17**

Effects

Model for $\hat{k} < 10$ does not include the effect of the matrix. This confirms the statement of Section 4.3 about the effect of estimation in T^2 statistic for matrices with small condition number. This model also does not include any interaction but the ARL depends mostly of the theoretical ARL desired and the condition number, the effects of m and n. The effect of the latter factors is shown in Figure 5.3-12 and Figure 5.3-13. Those figures show that large number of samples are better to reach the desired in-control ARL (600). The sample size, on the contrary, not necessary has the same effect. Only for models $\hat{k} < 10$ and $20 \leq \hat{k} < 60$ this is true but for the model $10 \leq \hat{k} < 20$ only for samples size equal to 2 the value is close to 600, if the sample size increases, the ARL gets overestimated. When $60 \leq \hat{k}$ starts underestimated and reaches optimal value approximately when $n = 8$, then it gets overestimated.

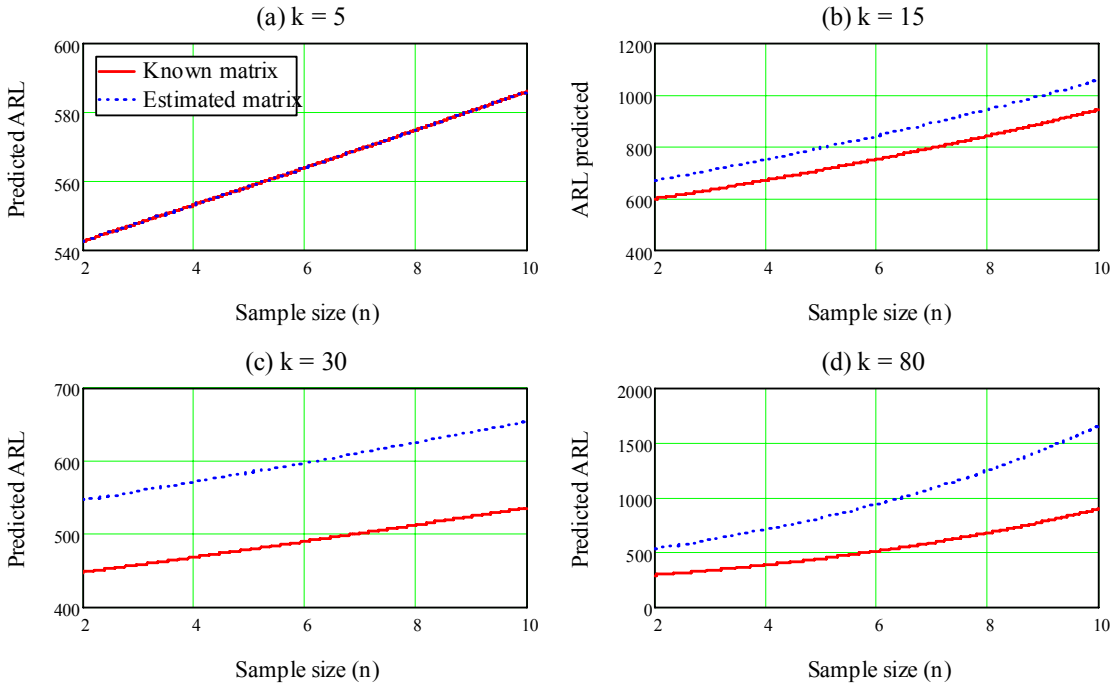


Figure 5.3-12. Effect of the sample size when (a) $\hat{k} < 10$, (b) $10 \leq \hat{k} < 20$, (c) $20 \leq \hat{k} < 60$ and (d) $60 \leq \hat{k}$ (Theoretical ARL = 600, $m = 50$ and $p = 4$

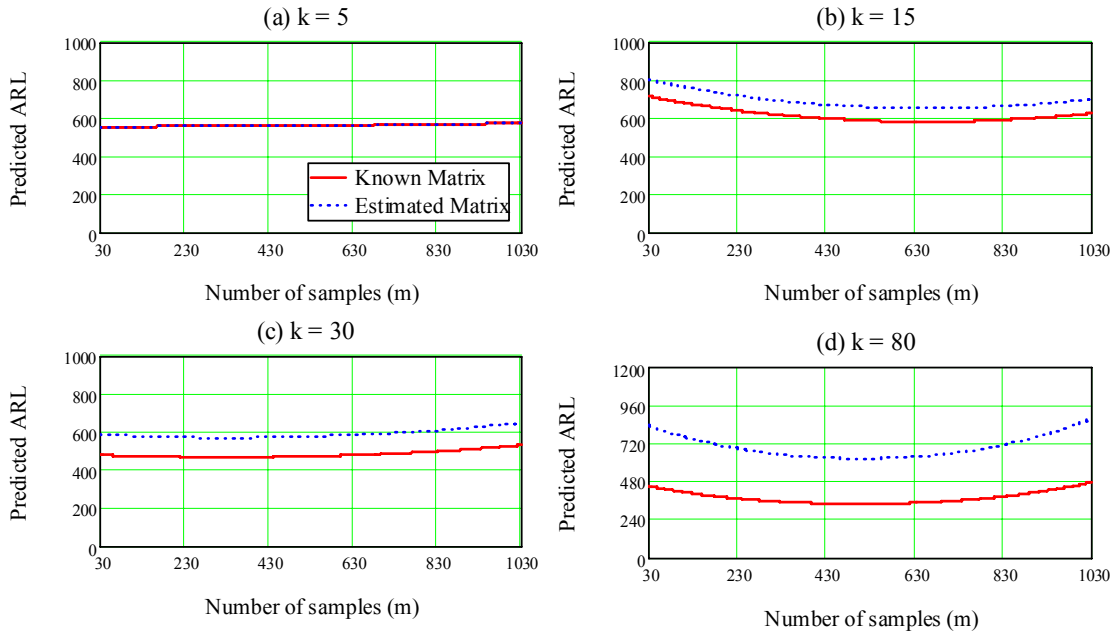


Figure 5.3-13. Effect of the number of samples when (a) $\hat{k}$ < 10, (b) 10≤$\hat{k}$ < 20, (c) 20≤$\hat{k}$ < 60 and (d) 60≤$\hat{k}$ (Theoretical ARL = 600, *n* = 5 and *p* = 4)

As it was explained, only in the model $\hat{k}$ < 10 there is not effect due to the matrix used. In the remaining models, also large *m* and *n* are better to achieve the target ARL. In those models, also, the effect of the matrix is to overestimate the results of the true matrix.

The effect of the number of variables is shown in Figure 5.3-14, in the $\hat{k}$ < 10 model there is no effect of the number of variables. In the 10≤$\hat{k}$ < 20 large number of variables is better but when 20≤$\hat{k}$ < 60, large number of variables has a decreasing effect in the ARL. In the 60≤$\hat{k}$ model, the ARL with the true matrix is underestimated by 20%.

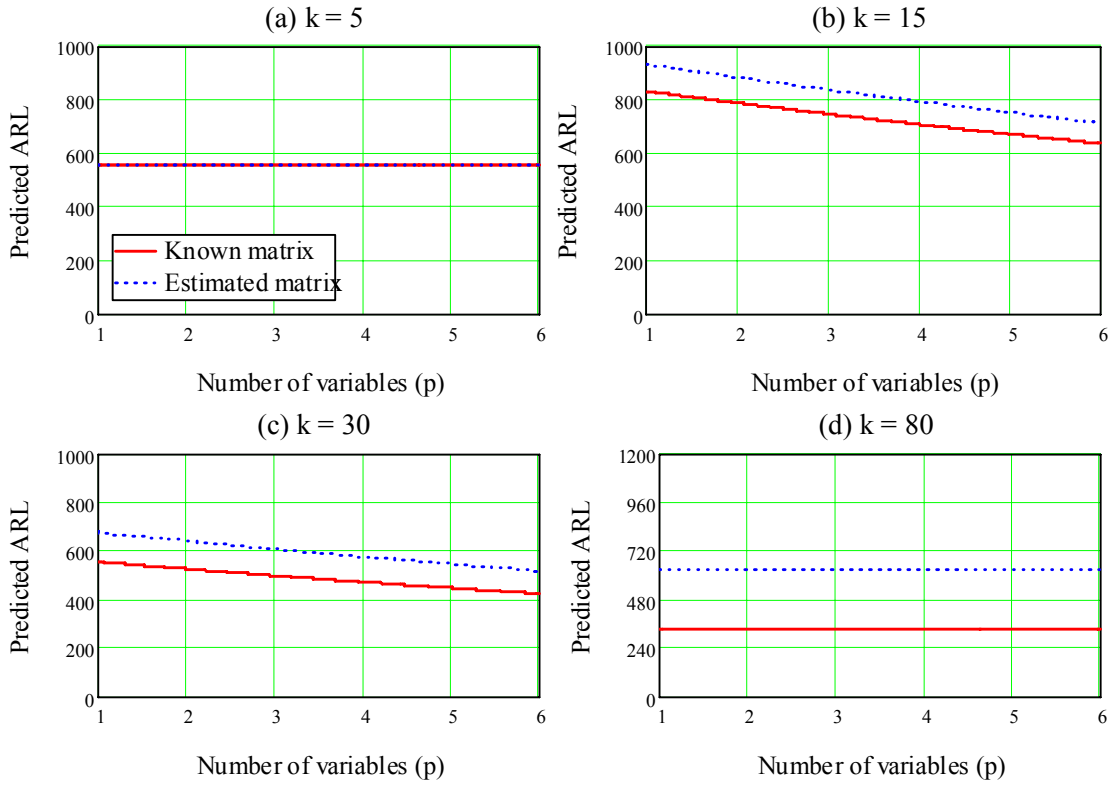


Figure 5.3-14. Effect of the number of variables when (a) $\hat{k} < 10$, (b) $10 \leq \hat{k} < 20$, (c) $20 \leq \hat{k} < 60$ and (d) $60 \leq \hat{k}$ (Theoretical ARL = 600, *m* = 50 and *n* = 5)

Figure 5.3-15 shows the effect of ARL, the values obtained shows a not serious underestimation (around 7 %) for the models where $\hat{k} < 10$ using the true matrix. The $10 \leq \hat{k} < 20$ model overestimates the ARL approximately by 18%. This overestimation increases for the $20 \leq \hat{k} < 60$ model (28%). The $60 \leq \hat{k}$ model shows underestimation that starts 12% under the theoretical value for a theoretical ARL of 200, this increases until an underestimation of 37% when the target ARL is 1400. The estimated matrices have the effect of increasing the overestimation, the overestimation in the $10 \leq \hat{k} < 20$ model increases up to 33% , for the $20 \leq \hat{k} < 60$ model, the overestimation is now about 44% and

finally for the $60 \leq \hat{k}$ model, there is not an underestimation, instead with estimated matrices, there is an overestimation that decreases as the desired ARL decreases.

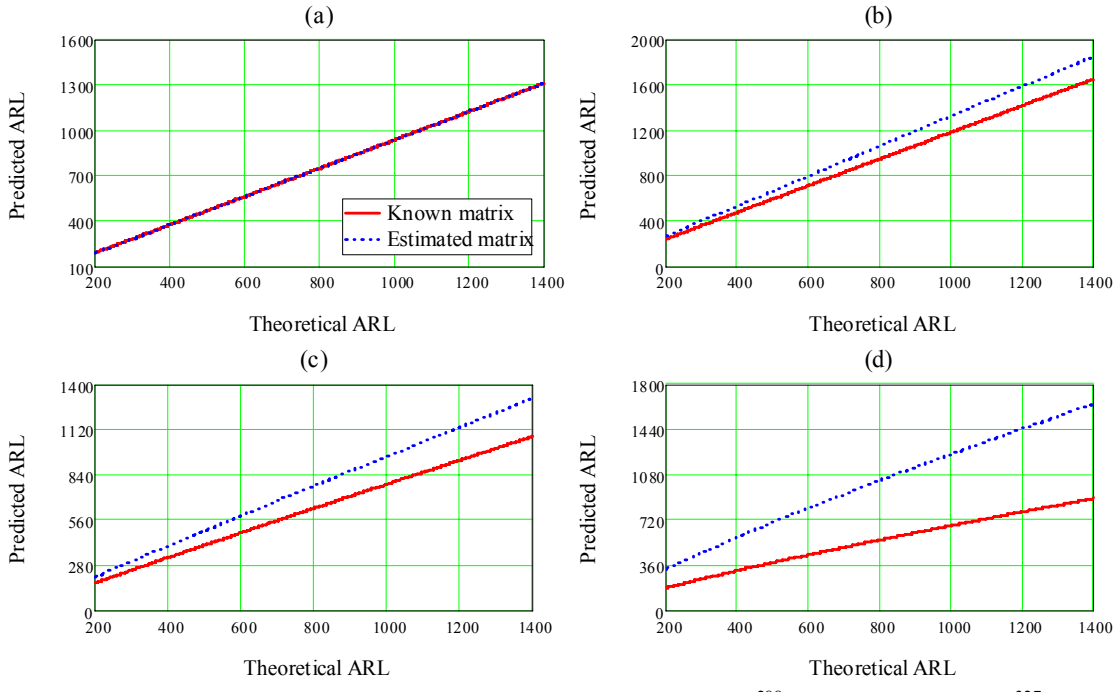


Figure 5.3-15. Effect of the Theoretical ARL (a) $\hat{k} < 10$, (b) $10 \leq \hat{k} < 20$, (c) $20 \leq \hat{k} < 60$ and (d) $60 \leq \hat{k}$ (m = 500, n = 5 and p = 4)

Table 22. In-control ARLs with true matrix (m = 500, n = 5 and p = 4)

Theoretical ARL	Models			
	$\hat{k} < 10$ ($\hat{k}$ =5)	$10 \leq \hat{k} < 20$ ($\hat{k}$ =15)	$20 \leq \hat{k} < 60$ ($\hat{k}$ =30)	$60 \leq \hat{k}$ ($\hat{k}$ =80)
200	184.65	236.78	257.28	177.61
400	371.17	473.02	513.97	315.27
600	558.4	709.05	770.45	441.02
800	746.09	944.96	1026.78	559.62
1000	934.13	1180.76	1283	673.17
1200	1122.44	1416.49	1539.14	782.85
1400	1310.98	1652.16	1795.21	889.41

The effect of the interaction *np* it is only present in models $10 \leq \hat{k} < 20$ and $20 \leq \hat{k} < 60$. In the former, it shows that small sample sizes are better and in the latter large sample sizes are preferred.

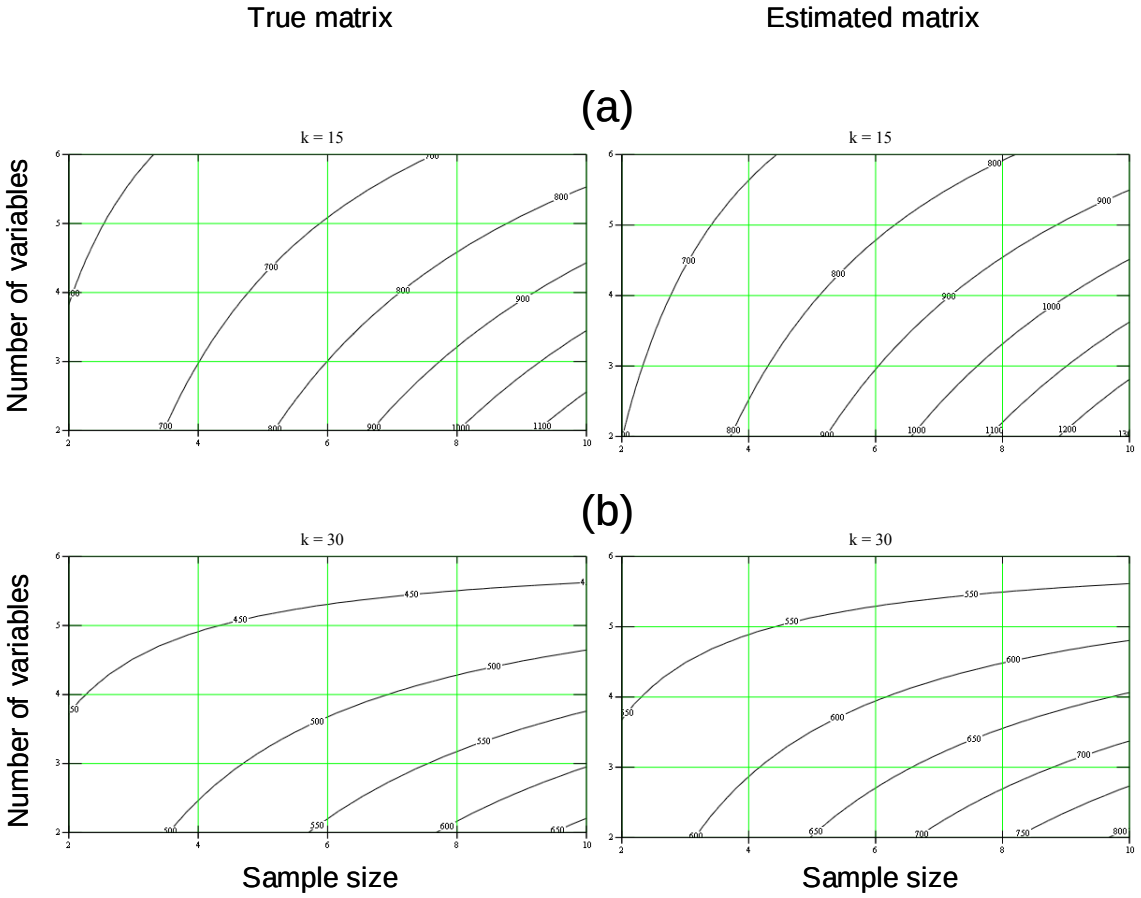


Figure 5.3-16. Effect of interaction between sample size (*n*) and number of variables (*p*) in model (a) $10 \leq \hat{k} < 20$ and (b) $20 \leq \hat{k} < 60$ for a desired ARL = 600, and *m* = 50

The effect of the interaction *mn* confirms that for $\hat{k} < 10$, large samples are better. This is also true for the $10 \leq \hat{k} < 20$ and $20 \leq \hat{k} < 60$ models. In the remaining model, large samples do not improve (underestimates) the in-control ARL with the true matrix, the estimated matrix overestimates the ARL.

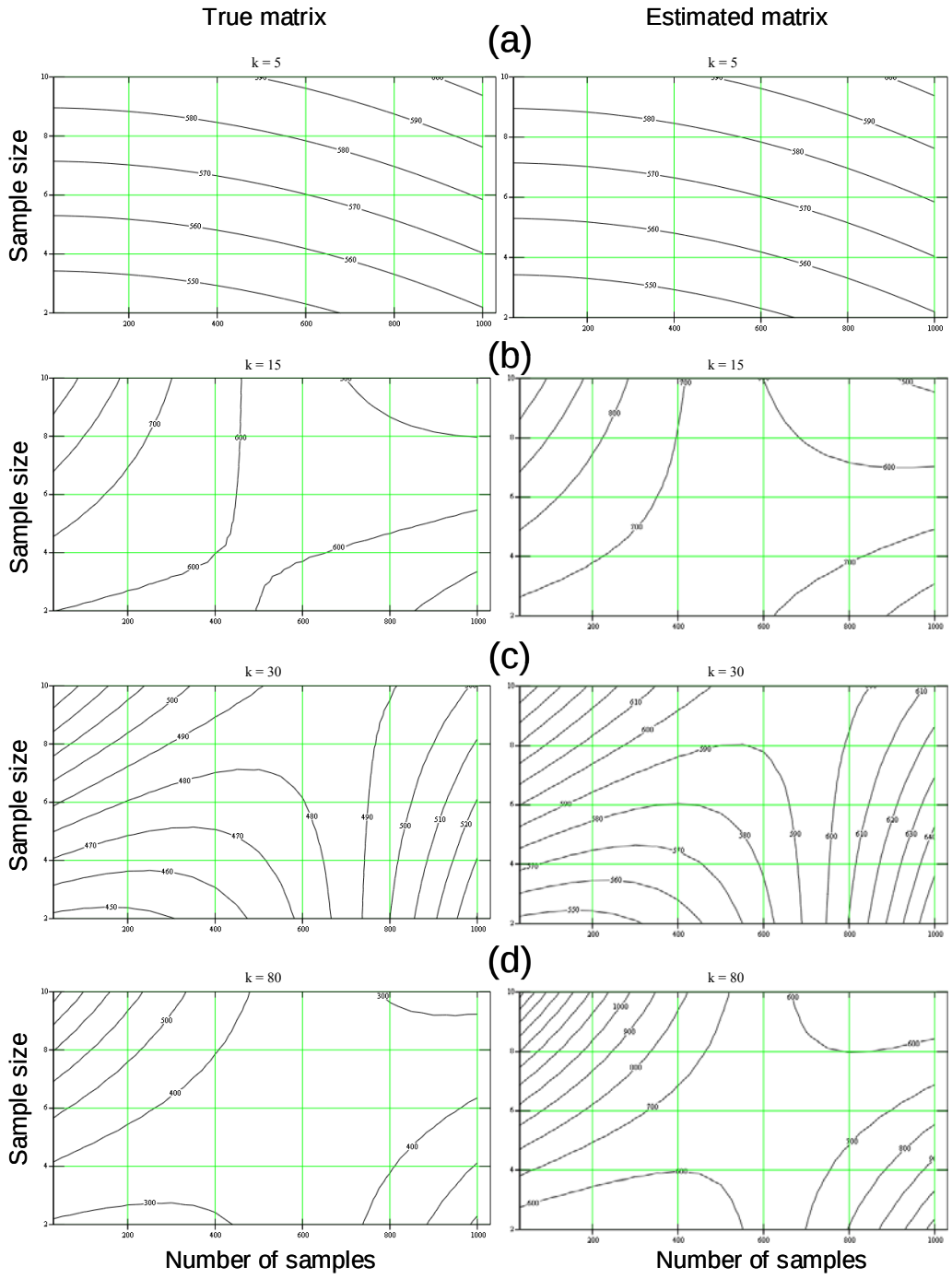


Figure 5.3-17. Effect of interaction between number of samples (*m*) and sample size (*n*) in model (a) $\hat{k} < 10$, (b) $10 \leq \hat{k} < 20$, (c) $20 \leq \hat{k} < 60$ and (d) $60 \leq \hat{k}$ for a desired ARL = 600, and $p = 4$

6 Procedure for predicting in-control ARL's

The final part of this work proposes a procedure based in the regressions developed in the previous chapter. In most of the cases, practitioners do not have large samples so the procedure can be implemented when small samples are available. In this case, there is a correction in the upper control limit (UCL), which provides more exact results than the traditional UCL.

6.1 Proposed procedure

Based in the previously described regression models, a procedure for obtaining more realistic in-control ARL's can be implemented as follows:

a. Choose the target ARL for the control chart you are designing.

b. Take *m* samples of size *n* of the *p* variables of your control chart.

c. Compute the in-control Mean vector and Covariance matrix.

d. Standardize both parameters using procedure described in Section of Chapter 3.

e. Compute the second condition number of the correlation matrix.

f. Depending of the condition number, use the corresponding model to find the value of the theoretical ARL to achieve the target ARL. If your condition number is over 60 consider to check multicollinearity and eliminate highly correlated variables.

g. With the new value of theoretical ARL, find the new UCL.

6.2 Example

The following is an illustration of how the procedure can be implemented.

Suppose you have estimated a correlation matrix of $p = 4$ with $m = 30$ and $n = 5$. also assume that the value of the condition number is 15. Using Mathcad, it is possible to find the corrected UCL to achieve a target ARL of 200:

a. The input data for this example are the target or desired in-control ARL (target_ARL), the condition number of the estimated matrix (k), sample size (n) and number of samples (m), number of variables (p) and the variable matrix is equal to 1 because it is an estimated matrix.

target_ARL := 200
k := 15
m := 30
n := 5
p := 4
matrix:= 1

b. Given the fact that the condition number is 15, the regression to be used is

$$\log(arl) = \beta_0 + \beta_1 \ln(arl_{theoretical}) + \beta_2 \ln k + \beta_3 n + \beta_4 mn + \beta_5 np + \beta_6 m^2 + \beta_7 matrix$$

where: β_0 = -0.4328219, β_1 = 0.9983577, β_2 = 0.119801, β_3 = 0.1066148, β_4 = -0.0001366, β_5 = -0.0107158, $\beta_6 = 5.17x10^{-7}$ and β_7 = 0.1140992.

c. The function ARL10k20 contains the evaluation of the in-control ARL given the input data provided in (a) or equivalently $ARL10k20 = arl = e^{\log arl}$. So one can find

the value of the theoretical in-control ARL (true_ARL or $arl_{theoretical}$) to achieve the target in-control ARL (target_ARL) using the function root in Mathcad solving the equation ARL10k20 = target_ARL.

true_ARL := root(ARL10k20(true_ARL, k, m, n, p, matrix) − target_ARL, true_ARL)
true_ARL = 148.724

d. The value obtained (true_ARL) is actually the value to achieve an in-control ARL of 200, and the UCL in Phase II (UCL) can now be estimated in the variable Corrected_UCL using the true_ARL

$$UCL(ARL) := \frac{p\cdot(m+1)\cdot(n-1)\cdot qF\left(1-\frac{1}{ARL}, p, m\cdot n-m-p+1\right)}{m\cdot(n-1)+1-p}$$

UCL(target_ARL) = 16.644
Corrected_UCL := UCL(true_ARL)
Corrected_UCL = 15.843

The corrected UCL to achieve the target ARL is 15.843. If the real matrix were used, then the case is similar as the cases presented by Champ et al (2005). Replacing matrix = 0 in the example

target_ARL := 200
k := 15
m := 30
n := 5
p := 4
matrix:= 0
true_ARL := 200
true_ARL := root(ARL10k20(true_ARL, k, m, n, p, matrix) − target_ARL, true_ARL)
true_ARL = 166.731

$$UCL(ARL) := \frac{p\cdot(m+1)\cdot(n-1)\cdot qF\left(1-\frac{1}{ARL}, p, m\cdot n-m-p+1\right)}{m\cdot(n-1)+1-p}$$

UCL(target_ARL) = 16.644

```
Corrected_UCL := UCL(true_ARL)
Corrected_UCL = 16.152
```

In this case, the corrected limit is 16.152, which compared to Champ's (16.0809), is slightly more conservative.

The following tables shows a comparison between the values obtained by Champ et al (2005) for an in-control ARL of 200 and the ones obtained by the regression models proposed in this work.

Table 23. Comparison between corrected UCLs (in-control ARL = 200)

p	*m*	*n*	Traditional UCL	Champ et al. UCL	Regressions UCL			
					Model	$\hat{k}$	True matrix	Estimated Matrix
2	30	3	12.1978	10.9763	$\hat{k} < 10$	9	12.179	12.179
					$10 \le \hat{k} < 20$	15	11.88	11.594
					$20 \le \hat{k} < 60$	50	12.333	11.809
	40	5	11.3024	10.9086	$\hat{k} < 10$	9	11.243	11.243
					$10 \le \hat{k} < 20$	15	10.68	10.429
					$20 \le \hat{k} < 60$	50	11.118	10.659
	50	3	11.5233	10.8483	$\hat{k} < 10$	9	11.506	11.506
					$10 \le \hat{k} < 20$	15	11.25	10.989
					$20 \le \hat{k} < 60$	50	11.65	11.172
	70	5	10.9940	10.775	$\hat{k} < 10$	9	10.937	10.937
					$10 \le \hat{k} < 20$	15	10.436	10.195
					$20 \le \hat{k} < 60$	50	10.827	10.385
4	30	3	18.1162	16.7850	$\hat{k} < 10$	9	18.092	18.092
					$10 \le \hat{k} < 20$	15	17.92	17.563
					$20 \le \hat{k} < 60$	50	18.285	17.63
	40	5	16.1750	15.7660	$\hat{k} < 10$	9	16.105	16.105
					$10 \le \hat{k} < 20$	15	15.72	15.424

p	*m*	*n*	Traditional UCL	Champ et al. UCL	Regressions UCL			
					Model	$\hat{k}$	True matrix	Estimated Matrix
					$20 \le \hat{k} < 60$	50	15.958	15.417
	50	3	16.7041	16.0291	$\hat{k} < 10$	9	16.683	16.683
					$10 \le \hat{k} < 20$	15	16.553	16.239
					$20 \le \hat{k} < 60$	50	16.856	16.282
	70	3	16.1455	15.6955	$\hat{k} < 10$	9	16.126	16.126
					$10 \le \hat{k} < 20$	15	16.021	15.724
					$20 \le \hat{k} < 60$	50	16.294	15.751
6	30	3	23.8820	22.3445	$\hat{k} < 10$	9	23.853	23.853
					$10 \le \hat{k} < 20$	15	23.889	23.462
					$20 \le \hat{k} < 60$	50	24.083	23.302
	40	5	20.5709	20.1584	$\hat{k} < 10$	9	20.491	20.491
					$10 \le \hat{k} < 20$	15	20.371	20.037
					$20 \le \hat{k} < 60$	50	20.326	19.715
	50	3	21.5048	19.8408	$\hat{k} < 10$	9	21.479	21.479
					$10 \le \hat{k} < 20$	15	21.534	21.173
					$20 \le \hat{k} < 60$	50	21.679	21.019
	70	5	19.6690	19.4440	$\hat{k} < 10$	9	19.595	19.595
					$10 \le \hat{k} < 20$	15	19.535	19.223
					$20 \le \hat{k} < 60$	50	19.453	18.884

It is important to note that few samples lead us to a modification in UCL. The corrections using the regression models are more conservative to the values proposed by Champ. In most of the cases, the values that the regression proposes is over Champ's value so this assure us that it is not too narrow to affect the out-of-control ARL.

7 Conclusions and Future Research

7.1 Conclusions

This work has shown the dependence of the in-control ARL to the condition number and the target in-control ARL. A model for a more exact in-control ARL has been developed and a simple procedure has been proposed. The models have been evaluated using combinations of small sample sizes and samples numbers which result in values more conservatives than those proposed by Champ et al (2005) but do not require the use of tables obtained by simulation.

As a summary of the findings, I will point out some of the conclusions observed previously:

- The experiment has confirmed the problem described in Chapter 3 also noted by Champ et al. (2005): Alt's UCL approximation to the *F* distribution is wide enough to accept the effect of estimation but does not produces exact results.
- Even thought the work of Champ et al. (2005) presents that the estimation can be skipped, the estimation is still problem for achieving targets in-control ARLs.
- The influence of the condition number in the in-control ARL is notorious when the correlation matrix is estimated. Ill-conditioned matrices affect directly in the T^2 hence the in-control ARL. This is another conclusion that has not been noted by Champ et al. (2005). Correlation matrices with condition number over 10 also influences in the probability of getting sooner false alarms.

- For estimated matrices with condition number over 10, the estimation has an effect in the in-control ARL. It tends to overestimate the ARL like in univariate charts (Quesenberry, 1993).
- The probability of getting an early false alarm when matrices are estimated is practically the same as if the true matrix were estimated when the number of variables is small (equal or less than 3) and the condition number is less than 10. As long as p increases, the probability of early false alarms increases no matter if the matrix has a small condition number.
- Large samples can reduce the effect of estimation by reducing the error of estimation up to 10% of deviation form the target value. The concept of large samples can be related to the values obtained by Champ et al. (2005) mn between 500 and 900 depending on the number of variables being monitored.

7.2 Recommendations

From this work some recommendations about when designing T^2 control charts for subgroups can be stated:

- Sample size and number of samples reduce the effect of error in the estimation of the correlation matrix. Large samples are the best option.
- When large samples are not available, the proposed procedure is an alternative to the corrected UCLs obtained, by simulation, by Champ et al. (2005).
- Finally, it is important to remark the difficulty of getting a model for large condition numbers (k). The recommendation is to eliminate correlated variables. Principal components can be an alternative to find the relationship between the variables than can be eliminated.

7.3 Future work

A natural extension of this work it is to perform a similar analysis to the out-of-control ARL. This work is important because of the modification of the UCL makes also a variation of the out-of-control ARL, specially, in the Type II error.

When large samples are not available, another alternative to the one's proposed by this work and by Champ can be the use of bootstrapping method to generate enough samples to reduce the estimation error.

Finally, it has been widely extended the use of the ARL but there is a better measure than the media. Because of the skewness of the RL, the median is always smaller than the ARL, so the probability of getting a false alarm is high. Ryan (2000) mentioned the average production length method developed by Keats, Miskulin and Runger (1995) for $\overline{X}$ control chart. This method computes the expected amount of production between the time of a parameter change and the time when the signal is received. However it has the disadvantage that it has a large standard deviation. Another option includes the development of the median run-length as a better value to design control charts.

References

Alt, F. B, Goode, J. J., and Wadsworth, H. M. (1976). "Small Sample Limits for the Mean of a Multivariate Normal Process". *ASQC Technical Conference Transactions – Toronto*, 170 – 176.

Alt, F. B., and Smith, N. D. (1988). Handbook of Statistics, Vol. 7. Elsevier Science Publishers B.V.

Atkinson, A. C., and Mulira, H. M., (1993). "The Stalactite Plot for the Detection of Multivariate Outliers". Statistics and Computing, 3.

Champ, C. W., Jones-Farmer, L. A., and Rigdon, S. E. (2005). "Properties of the T^2 Control Chart when parameters are estimated". Technometrics, Vol .47 (4).

Djauhari, M. A. (2005). "Improved Monitoring of Multivariate Process Variability". Journal of Quality Technology, Vol. 37.

Duncan, A. J. (1974). Quality Control and Industrial Statistics. 4th Ed., Richard D. Irwin, IL.

Hauck, D. J., Runger, G. C., and Montgomery, D. C., (1999). "Multivariate Statistical Process Monitoring and Diagnosis with Grouped Regression-Adjusted Variables". Commun. Statist. – Simulation and Computation, Vol. 28 (2).

Hotelling, H. (1947). "Multivariate Quality Control." Techniques of Statistical Analysis. McGraw-Hill.

Ito, K. (1956). "Asymptotic Formulae for the Distribution of Hotelling's Generalized T^2". The Annals of Mathematical Statistics, Vol. 27 (4), 1091-1105.

Keats, J. B., Muskulin, J. D., and Runger, G. C. (1995). "Statistical Process Control Design". Journal of Quality Technology, Vol. 27 (3).

Johnson, R. A., and Wichern, D. W. (1999). Applied Multivariate Statistical Analysis. 4th Ed., Prentice Hall, NJ.

Lowry, C. A., and Montgomery, D. C., (1995). "A Review of Multivariate Control Charts". IIE Transactions, 26.

Mittag, H.-J., and Rinne, H. (1993). Statistical Methods for Quality Assurance. 1st Ed., Capman & Hall, UK.

Mason, R. L., and Young J. C. (1999). "Improving the Sensitivity of the T^2 Statistic in Multivariate Process Control". Journal of Quality Technology, Vol. 31.

Montogmery, Douglas C. (2001). Introduction to Statistical Quality Control. 4th Ed., John Wiley and Sons, NY.

Myers, Raymond H. and Montogmery, Douglas C. (1995). Response Surface Methodology. 1st Ed., John Wiley and Sons, NY.

Montogmery, Douglas C., Peck, Eliazbeth A., and Vinning, G. Geoffrey. (2001). Introduction to Linear Regression. 3rd Ed., John Wiley and Sons, NY.

Neduraman, G, and Pignatiello Jr., J. J. (2000). "On Constructing *T*-squared Control Charts for Retrospective Examination". Communications in Statistics. Part B Simulation and Computation. Vol. 29 (2), 621-632.

Quesenberry, C. P. (1993). "The Effect of Sample Size on Estimated Limits for $\overline{X}$ and X Control Charts". Journal of Quality Technology, Vol. 25 (4).

Ryan, Thomas P. (1989). Statistical Methods for Quality Improvement. 1st Ed., John Wiley & Sons, US.

Ryan, Thomas P. (2000). Statistical Methods for Quality Improvement. 2nd Ed., John Wiley & Sons, US.

Runger, G. C., Alt, F. B., and Montgomery, D. C. (1996). "Contributors to a Multivariate Statistical Process Control Signal". Communications in Statistics – Theory and Methods, Vol 25.

Siotani, M. (1959). "The Exact Value of the Generalized Distances of the Individual Points in the Multivariate Normal Sample". Annals of the Institute of Statistical Mathematics 10, pp. 183-203.

Sheo, T. (2002). "The Effects of Nonnormality on the Upper Percentiles of T2max Statistic in Elliptical Distributions". J. Japan Statist. Soc., Vol. 32 (1), 57-76.

Sullivan, J. S., and Woodall W. H. (1996). “A Comparison of Multivariate Control Charts for Individual Observations”. Journal of Quality Technology, Vol. 28.

Sullivan, J. S., and Woodall W. H. (1998). “Adapting Control Charts for the Preliminary Analysis of Multivariate Observations”. Commun. Statist. – Simulation and Computation, Vol. 27 (4).

Tracy, N. D., Young, J. C., and Mason, R. L. (1992). “Multivariate Control Charts for Individual Observations”. Journal of Quality Technology, Vol. 24.

Vargas, J. A. (2003). “Robust Estimation in Multivariate Control Charts for Individual Observations”. Journal of Quality Technology, Vol. 35.

Yai, K., and Trewn, J. (2003). Multivariate Statistical Methods in Quality Management. McGraw Hill, Inc.

Appendix A. Procedure to obtain a matrix with a pre-defined condition number

The Condition Number

The concept of condition number is related with a numerical problem of computation called well-posed problem. This problem is related with the instability of numerical results and sensitivity to little changes in the data.

In matrices, the condition number measures the sensitivity of the matrix to numerical operations. A matrix with a low condition number (the closer to one, the better) is known as well-conditioned on the contrary if the matrix has a high condition number would be an ill-conditioned matrix. Ill-conditioned matrices, hence, are not reliable.

The condition number, k, of a square matrix A is defined as

$$k_n(A) = \left\| A \right\|_n \left\| A^{-1} \right\|_n \tag{A1}$$

where $\|.\|_n$ is the underlying norm. In strict, the condition number of a matrix measures some sort of inverse distance from a singular matrix to A, normalized by $\|A\|$. When the 2-norm is used, orthogonal matrices are perfectly condition $k_2(.) = 1$. This is why this the condition number used for linear problems such as $Ax = b$, and is, usually, estimated by

$$k_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{A2}$$

For this work, when I refer to the condition number it will be assumed that I am referring to the second condition number.

Characterization of ill-conditioned matrices

In this part I will show the importance of the original matrix used as $\hat{\Sigma}_0$. As I explained in the previous part, the condition number is a measure of sensitivity to numerical operations, that is the reason that one usually chooses very low conditioned matrices and there are procedures to reduce it using matrices based in eigenvalues. However, I will show that even when the matrix is low conditioned, a slight departure from this value will lead to an ill-conditioned matrix, here is another reason that supports the importance of a good estimation of the covariance matrix.

For example, given the correlation matrix

$$M = \begin{bmatrix} 1 & r & 0.5 \\ r & 1 & 0.25 \\ 0.5 & 0.25 & 1 \end{bmatrix}$$

Figure A-1 shows how the condition number behaves to a variation in one correlation, in this example *r*, fixing the third correlation in 0.25 and making the second correlation to vary from 0.5 to 0.45 or 0.55.

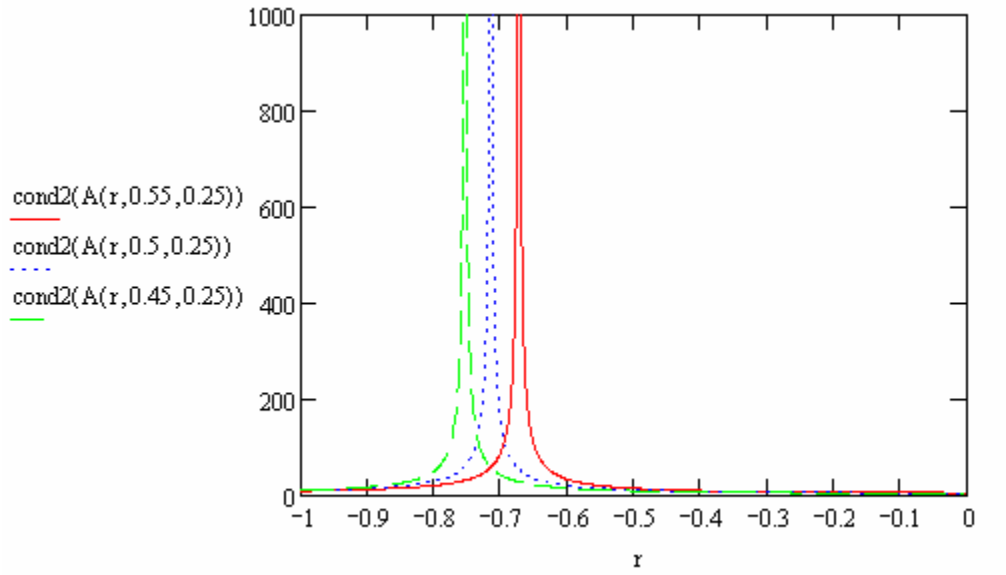


Figure A.1. Condition number behavior varying one correlation

Suppose that the real value of r is -0.6, in this case the $k_2(M) = 19.553$ but now suppose that the estimation lead us to an estimated correlation, $\hat{r} = -0.65$, in this case the condition number increases to $k_2(\hat{M}) = 35.447$. But now suppose that now two correlations were estimated and the estimated matrix is

$$\hat{M} = \begin{bmatrix} 1 & -0.65 & 0.55 \\ -0.65 & 1 & 0.25 \\ 0.55 & 0.25 & 1 \end{bmatrix}$$

For this matrix, the condition number is $k_2(\hat{M}) = 111.577$. See that, from a medium ill-conditioned matrix we have moved to an ill-conditioned matrix or in some cases to a singular matrix. Figures A-2 and A-3 shows both the contour plot and the surface plot, a bad estimation of the correlations can lead to a very conditioned matrix.

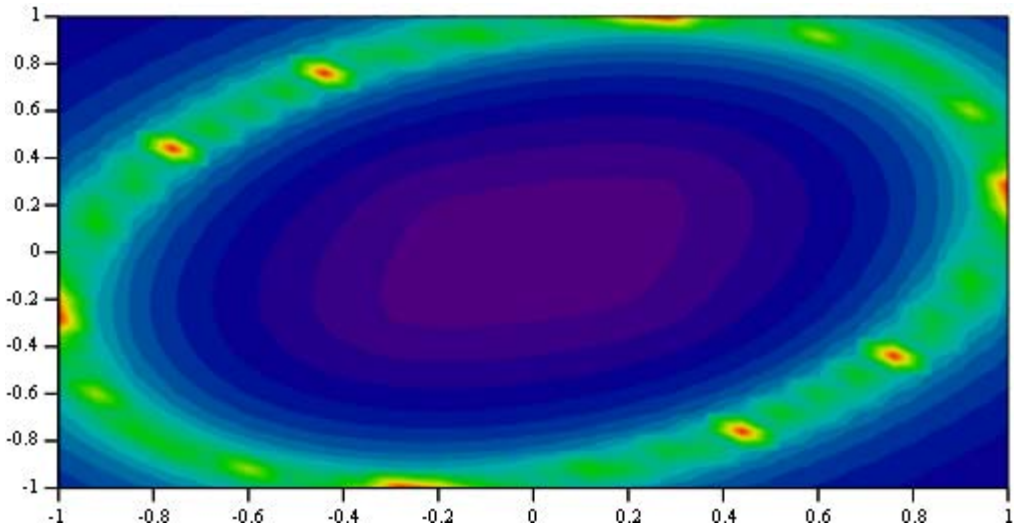


Figure A. 2. Contour plot of condition number varying two correlations (3x3 matrix)

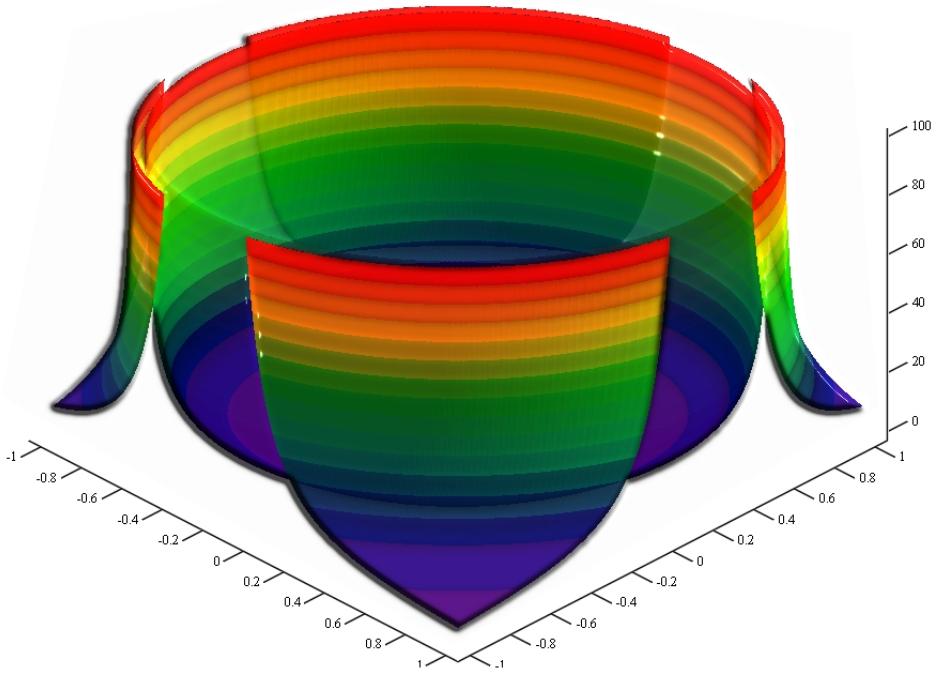


Figure A. 3. Surface Plot of the condition number varying two correlations (3x3 matrix)

To prevent singular matrices in the estimation, I have developed a procedure to get matrices that have a previously defined condition number with not singular matrices closer. This procedure is explained in the following part.

Procedure

The idea is to find matrices with a previously chosen condition number.
The procedure starts selecting an initial matrix M_{pxp}. A matrix A is positive definite if $A = MM^T$.

So, the idea is to find a matrix A that satisfies the following constraints:

a. Assuring that A be positive definite:
   - eigenvalues of $A > 0$
b. A be a correlation matrix:
   - $A_{ij} = A_{ji}$, for $i \neq j$
   - $A_{ii} = 1$
   - $A_{ij} > -1$ and $A_{ij} < 1$, for $i \neq j$

The problem is that not only that the matrix should have this condition number, but also closer matrices (A_{up} and A_{down}) should be positive semi-definite. These constraints pretend to prevent that estimated matrices cannot have inverse and have a minimum condition number:

c. Upper and lower matrices can be defined as follows:
   - $A_{up} = A + V$ and

- $A_{down} = A - V$

where, $V^{0.05}$ is a *pxp* matrix with $V_{ii} = 0$ and $V_{ij} = 0.05$ for $i \neq j$

d. Satisfy required condition numbers:

- $\|A\|_2 \|A^{-1}\|_2 > \min$ and $\|A\|_2 \|A^{-1}\|_2 < \max$
- $\|A_{up}\|_2 \|A_{up}^{-1}\|_2 > \min$
- $\|A_{down}\|_2 \|A_{down}^{-1}\|_2 > \min$

Results

The previous procedure was implemented in Mathcad v.11. The following results show the matrices that the simulation used as the true matrices:

For $k = 5$: The matrices were

- $p = 2$: $\mathrm{R} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, $k = 3$
- $p = 3$: $\mathrm{R} = \begin{bmatrix} 1 & -0.253 & -0.642 \\ -0.253 & 1 & 0.195 \\ -0.642 & 0.195 & 1 \end{bmatrix}$, $k = 4.9915$
- $p = 4$: $\mathrm{R} = \begin{bmatrix} 1 & 0.108 & 0.196 & -0.033 \\ 0.108 & 1 & -0.452 & 0.234 \\ 0.196 & -0.452 & 1 & 0.227 \\ -0.033 & 0.234 & 0.227 & 1 \end{bmatrix}$, $k = 4.638$
- $p = 5$: $\mathrm{R} = \begin{bmatrix} 1 & -0.2 & 0.535 & 0.101 & -0.514 \\ -0.2 & 1 & -0.208 & 0.027 & 0.211 \\ 0.535 & -0.208 & 1 & 0.218 & -0.513 \\ 0.101 & 0.027 & 0.218 & 1 & -0.157 \\ -0.514 & 0.211 & -0.513 & -0.157 & 1 \end{bmatrix}$, $k = 4.889$

- $p = 6$: $\mathrm{R} = \begin{bmatrix} 1 & -0.061 & -0.222 & 0.254 & 0.206 & -0.184 \\ -0.061 & 1 & 0.301 & 0.127 & -0.090 & -0.349 \\ -0.222 & 0.301 & 1 & 0.046 & -0.093 & 0.221 \\ 0.254 & 0.127 & 0.046 & 1 & -0.064 & 0.264 \\ 0.206 & -0.090 & -0.093 & -0.064 & 1 & -0.078 \\ -0.184 & -0.349 & 0.221 & 0.264 & -0.078 & 1 \end{bmatrix}$,

$k = 4.9103$

For $k = 10$: The matrices were

- $p = 2$: $\mathrm{R} = \begin{bmatrix} 1 & -0.818 \\ -0.818 & 1 \end{bmatrix}$, $k = 9.9890$
- $p = 3$: $\mathrm{R} = \begin{bmatrix} 1 & -0.308 & 0.748 \\ -0.308 & 1 & -0.428 \\ 0.784 & -0.428 & 1 \end{bmatrix}$, $k = 9.9986$
- $p = 4$: $\mathrm{R} = \begin{bmatrix} 1 & -0.058 & 0.039 & -0.720 \\ -0.058 & 1 & -0.285 & -0.117 \\ 0.039 & -0.285 & 1 & 0.354 \\ -0.720 & -0.117 & 0.354 & 1 \end{bmatrix}$, $k = 10.069$
- $p = 5$: $\mathrm{R} = \begin{bmatrix} 1 & -0.659 & 0.386 & 0.116 & -0.304 \\ -0.659 & 1 & -0.418 & 0.246 & 0.207 \\ 0.386 & -0.418 & 1 & -0.334 & -0.495 \\ 0.116 & 0.246 & -0.334 & 1 & 0.116 \\ -0.304 & 0.207 & -0.495 & 0.116 & 1 \end{bmatrix}$, $k = 9.7938$
- $p = 6$: $\mathrm{R} = \begin{bmatrix} 1 & 0.476 & -0.204 & 0.546 & 0.234 & -0.217 \\ 0.476 & 1 & -0.351 & 0.205 & -0.104 & -0.570 \\ -0.204 & -0.351 & 1 & 0.049 & -0.132 & 0.420 \\ 0.546 & 0.205 & 0.049 & 1 & 0.450 & 0.304 \\ 0.234 & -0.104 & -0.132 & 0.450 & 1 & 0.404 \\ -0.217 & -0.570 & 0.420 & 0.304 & 0.404 & 1 \end{bmatrix}$,

$k = 9.9677$

For $k = 20$: The matrices were

- $p = 2$: $\mathrm{R} = \begin{bmatrix} 1 & -0.905 \\ -0.905 & 1 \end{bmatrix}$, $k = 20.0526$

- $p = 3$: $\mathrm{R} = \begin{bmatrix} 1 & -0.65 & 0.1 \\ -0.65 & 1 & -0.69 \\ 0.1 & -0.69 & 1 \end{bmatrix}$, $k = 19.8664$

- $p = 4$: $\mathrm{R} = \begin{bmatrix} 1 & 0.367 & -0.244 & -0.741 \\ 0.367 & 1 & -0.819 & 0.101 \\ -0.244 & -0.819 & 1 & -0.288 \\ 0.741 & 0.101 & -0.288 & 1 \end{bmatrix}$, $k = 20.0477$

- $p = 5$: $\mathrm{R} = \begin{bmatrix} 1 & 0.658 & -0.242 & 0.685 & 0.198 \\ 0.658 & 1 & -0.609 & 0.239 & -0.143 \\ -0.242 & -0.609 & 1 & 0.05 & -0.155 \\ 0.685 & 0.239 & 0.05 & 1 & 0.555 \\ 0.198 & -0.143 & -0.155 & 0.555 & 1 \end{bmatrix}$, $k = 19.6578$

- $p = 6$: $\mathrm{R} = \begin{bmatrix} 1 & 0.638 & -0.199 & 0.663 & 0.241 & -0.227 \\ 0.638 & 1 & -0.567 & 0.233 & -0.108 & -0.653 \\ -0.199 & -0.567 & 1 & 0.050 & -0.145 & 0.481 \\ 0.663 & 0.233 & 0.050 & 1 & 0.638 & 0.319 \\ 0.241 & -0.108 & -0.145 & 0.638 & 1 & 0.557 \\ -0.227 & -0.653 & 0.481 & 0.319 & 0.557 & 1 \end{bmatrix}$,

$k = 20.0295$

For $k = 50$: The matrices were

- $p = 2$: $\mathrm{R} = \begin{bmatrix} 1 & -0.961 \\ -0.961 & 1 \end{bmatrix}$, $k = 50.2821$

- $p = 3$: $\mathrm{R} = \begin{bmatrix} 1 & -0.601 & 0.093 \\ -0.601 & 1 & -0.690 \\ 0.093 & -0.690 & 1 \end{bmatrix}$, $k = 49.6628$

- $p = 4$: $\mathrm{R} = \begin{bmatrix} 1 & 0.589 & 0.531 & 0.855 \\ 0.589 & 1 & 0.327 & 0.212 \\ 0.531 & 0.327 & 1 & 0.530 \\ 0.855 & 0.212 & 0.530 & 1 \end{bmatrix}$, $k = 50.0897$

- $p = 5$: $\mathrm{R} = \begin{bmatrix} 1 & 0.678 & -0.244 & 0.766 & 0.196 \\ 0.678 & 1 & -0.649 & 0.242 & -0.144 \\ -0.244 & -0.649 & 1 & 0.05 & -0.158 \\ 0.766 & 0.242 & 0.05 & 1 & 0.614 \\ 0.196 & -0.144 & -0.158 & 0.614 & 1 \end{bmatrix}$, $k = 49.5416$

- $p = 6$: $\mathrm{R} = \begin{bmatrix} 1 & 0.821 & -0.042 & 0.492 & 0.070 & -0.069 \\ 0.821 & 1 & -0.238 & 0.081 & -0.033 & -0.400 \\ -0.042 & -0.238 & 1 & 0.015 & -0.037 & 0.147 \\ 0.492 & 0.081 & 0.015 & 1 & 0.180 & 0.169 \\ 0.070 & -0.033 & -0.037 & 0.180 & 1 & 0.181 \\ -0.069 & -0.400 & 0.147 & 0.169 & 0.181 & 1 \end{bmatrix}$,

  $k = 49.3984$

For $k = 100$: The matrices were

- $p = 2$: $\mathrm{R} = \begin{bmatrix} 1 & -0.980 \\ -0.980 & 1 \end{bmatrix}$, $k = 99$

- $p = 3$: $\mathrm{R} = \begin{bmatrix} 1 & 0.675 & -0.390 \\ 0.675 & 1 & 0.388 \\ -0.390 & 0.388 & 1 \end{bmatrix}$, $k = 97.9866$

- $p = 4$: $\mathrm{R} = \begin{bmatrix} 1 & -0.186 & 0.544 & 0.298 \\ -0.186 & 1 & 0.158 & -0.967 \\ 0.544 & 0.158 & 1 & -0.011 \\ 0.298 & -0.967 & -0.011 & 1 \end{bmatrix}$, $k = 97.7144$

- $p = 5$: $\mathrm{R} = \begin{bmatrix} 1 & 0.715 & -0.242 & 0.777 & 0.196 \\ 0.715 & 1 & -0.705 & 0.244 & -0.146 \\ -0.242 & -0.705 & 1 & 0.051 & -0.159 \\ 0.777 & 0.244 & 0.051 & 1 & 0.637 \\ 0.196 & -0.146 & -0.159 & 0.637 & 1 \end{bmatrix}$, $k = 100.7881$

- $p = 6$: $\mathrm{R} = \begin{bmatrix} 1 & 0.718 & -0.195 & 0.779 & 0.243 & -0.231 \\ 0.718 & 1 & -0.656 & 0.239 & -0.109 & -0.739 \\ -0.195 & -0.656 & 1 & 0.050 & -0.148 & 0.487 \\ 0.779 & 0.239 & 0.050 & 1 & 0.650 & 0.343 \\ 0.243 & -0.109 & -0.148 & 0.650 & 1 & 0.599 \\ -0.231 & -0.739 & 0.487 & 0.343 & 0.599 & 1 \end{bmatrix}$,

$k = 101.3750$

Appendix B. Evaluation of the rotatability of the design

Rotatability

A very important characteristic of second-order response surface design is rotatability which was developed by Box and Hunter (1957). They state that "a rotatable design is one in which $N Var\hat{y}(x)/\sigma^2$ has the same value at any locations that are at the same distance from the design center. This will assure that the predicted values, for example, of two points at the same distance from the origin will be equally good by having the same variance.

As noted by Myers and Montgomery (1995), this concept provides guidelines for the selection of the center points and axial distance (ρ).

Evaluation of the design

The design (ρ = 2.00) was evaluated versus the optimal design (ρ = 1.682). The first evaluation is the orthogonality which guarantees that the estimation of one factor or interactions is free from the influence of any other factor or interaction. This can be obtained by evaluating the X^TX, where X is the matrix design. The following figure shows the results for both the optimal design (a) and the proposed design (b).

$$X^T \cdot X = \begin{pmatrix} 20 & 0 & 0 & 0 & 13.658 & 13.658 & 13.658 & 0 & 0 & 0 \\ 0 & 13.658 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 13.658 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 13.658 & 0 & 0 & 0 & 0 & 0 & 0 \\ 13.658 & 0 & 0 & 0 & 24.009 & 8 & 8 & 0 & 0 & 0 \\ 13.658 & 0 & 0 & 0 & 8 & 24.009 & 8 & 0 & 0 & 0 \\ 13.658 & 0 & 0 & 0 & 8 & 8 & 24.009 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 \end{pmatrix} \qquad X1^T \cdot X1 = \begin{pmatrix} 20 & 0 & 0 & 0 & 16 & 16 & 16 & 0 & 0 & 0 \\ 0 & 16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 16 & 0 & 0 & 0 & 0 & 0 & 0 \\ 16 & 0 & 0 & 0 & 40 & 8 & 8 & 0 & 0 & 0 \\ 16 & 0 & 0 & 0 & 8 & 40 & 8 & 0 & 0 & 0 \\ 16 & 0 & 0 & 0 & 8 & 8 & 40 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 \end{pmatrix}$$

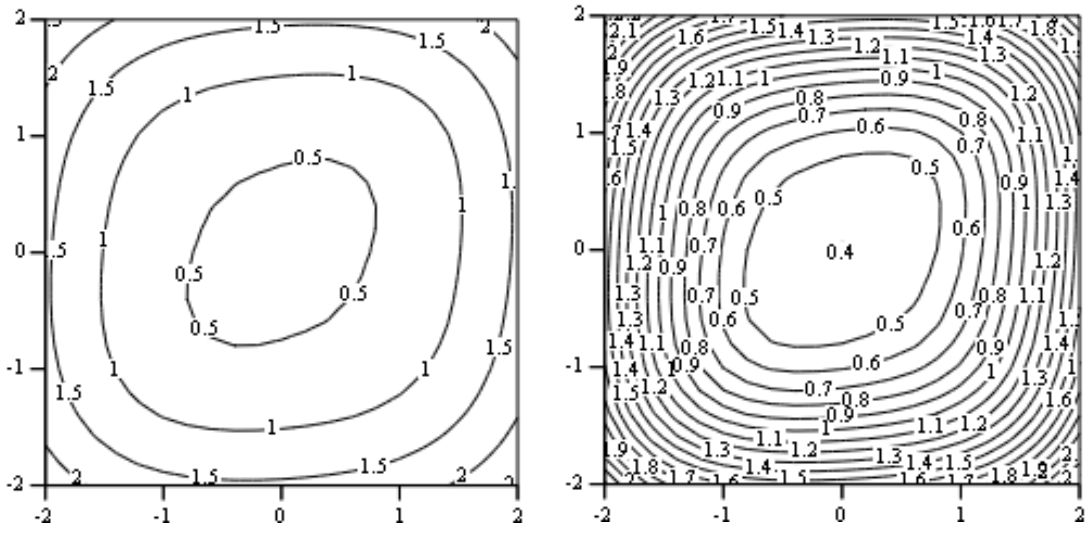
(a) ρ = 1.682 (b) ρ = 2.000

Figure B.1. X^TX matrix for (a) Optimal design and (b) Proposed design

Both designs are similar. They have the expected properties of the central composite design (CCD): the linear and the interactions columns are orthogonals but the quadratic columns not.

Now, let's evaluate the rotatability of the design used for this work. As in the previous part, the design (ρ = 2.00) was evaluated versus the optimal design (ρ = 1.682). Graphically, it can be evaluated by showing the contour plot of $\sqrt{Var\hat{y}(x)} = s\sqrt{x^{(m)'}(X'X)^{-1}x^{(m)}}$, where s is the root mean square error and $x^{(m)}$ is a vector that reflects the model.

Figure B.2 shows the contour plot without considering s. The proposed design seems to be rotable around the design region.



(a) ρ = 1.682 (b) ρ = 2.000

Figure B.2. Contours of constant response for (a) Optimal design and (b) Proposed design with factor $x_3 = 0$.

The performance of the design will now be evaluated using the variance dispersion graph as proposed by Myers and Montgomery (1995): by ploting the $\max(N Var\hat{y}(x)/\sigma^2)$ on a radios ρ against ρ. Figure B3 shows the variance dispersion graph for a CCD with 3 factors (k) and 6 center points, in this case, the maximum variance is reached at the sphere that contains the axial points.

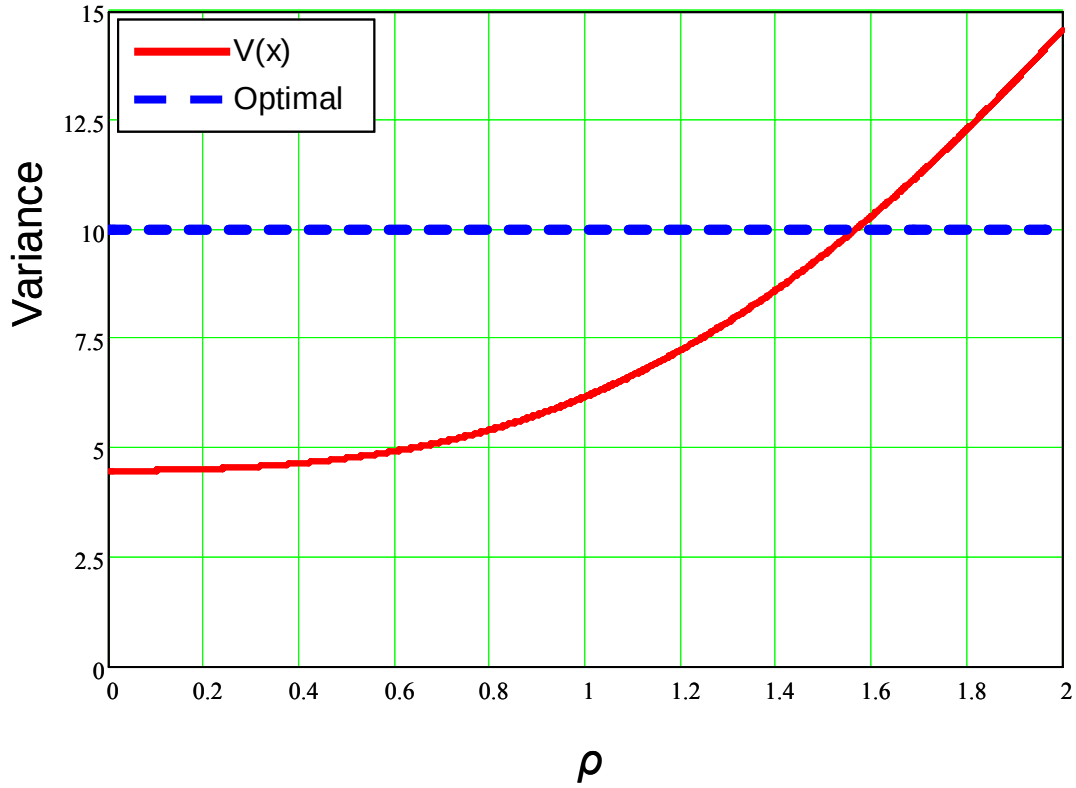


Figure B.3. Variance dispersion graph for a CCD for k = 3 and 6 central points.

The optimal value was established using the recommendation of Myers and Montgomery (1995) as the number of parameters being estimated (10). Note that the variance at ρ = 1.682 is only 10.8% of difference from the optimal value or the design is 89% efficient the proposed design instead has around 55%. One would argument that this is low, but as it was demonstrated previously it is still rotatable and the lack of efficiency can be sacrificed to expand the design region.

Appendix C. Fractional Factorial 2^{k-1} designs for $p \geq 4$

The designs for 4, 5 and 6 variables are

Table 24. Fractional Factorial 2^{6-1} design for p=4

	Correlations					
Run	A	B	C	D	E	F
1	1	-1	1	-1	1	1
2	1	1	-1	1	1	-1
3	1	-1	-1	-1	1	-1
4	1	1	-1	-1	-1	-1
5	-1	1	-1	1	1	1
6	-1	-1	-1	1	1	-1
7	-1	-1	1	1	-1	-1
8	-1	1	-1	-1	1	-1
9	-1	1	1	-1	1	1
10	1	1	1	-1	1	-1
11	-1	1	1	-1	-1	-1
12	-1	1	-1	-1	-1	1
13	1	-1	-1	1	1	1
14	1	1	-1	-1	1	1
15	1	-1	1	1	1	-1
16	-1	1	1	1	1	-1
17	-1	-1	-1	-1	1	1
18	1	-1	-1	1	-1	-1
19	1	1	1	1	-1	-1
20	1	-1	-1	-1	-1	1
21	1	-1	1	1	-1	1
22	-1	-1	-1	-1	-1	-1
23	-1	-1	-1	1	-1	1
24	1	1	1	-1	-1	1
25	-1	-1	1	1	1	1
26	-1	-1	1	-1	1	-1
27	1	1	-1	1	-1	1
28	1	-1	1	-1	-1	-1
29	1	1	1	1	1	1
30	-1	1	-1	1	-1	-1
31	-1	1	1	1	-1	1
32	-1	-1	1	-1	-1	1

+1: Over estimated
-1: Under estimated

Table 25. Fractional Factorial 2^{10-4} design for p=5

	Correlations									
Run	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**J**	**K**
1	1	-1	-1	1	1	1	1	-1	-1	1
2	-1	-1	1	-1	1	-1	1	-1	1	1
3	-1	-1	1	-1	-1	-1	-1	1	-1	-1
4	1	-1	1	-1	1	1	-1	1	-1	1
5	-1	1	-1	-1	1	-1	1	1	-1	1
6	-1	1	1	1	1	-1	-1	-1	-1	1
7	1	1	-1	-1	1	1	-1	-1	1	1
8	1	-1	1	1	1	-1	-1	-1	1	-1
9	-1	1	-1	1	-1	1	-1	1	-1	1
10	-1	1	1	1	-1	-1	1	1	1	-1
11	1	1	1	-1	-1	-1	-1	1	1	1
12	1	1	1	1	1	1	1	1	1	1
13	1	1	-1	1	-1	-1	1	-1	1	1
14	-1	1	-1	-1	-1	-1	-1	-1	1	-1
15	-1	-1	-1	1	1	-1	-1	1	1	1
16	1	-1	-1	-1	-1	-1	-1	-1	-1	1
17	-1	1	-1	1	1	1	1	-1	1	-1
18	-1	-1	1	1	1	1	1	1	-1	-1
19	1	-1	1	-1	-1	1	1	-1	1	-1
20	1	1	-1	-1	-1	1	1	1	-1	-1
21	1	1	1	-1	1	-1	1	-1	-1	-1
22	-1	-1	1	1	-1	1	-1	-1	1	1
23	1	-1	-1	-1	1	-1	1	1	1	-1
24	-1	-1	-1	1	-1	-1	1	-1	-1	-1
25	-1	1	1	-1	-1	1	1	-1	-1	1
26	1	1	-1	1	1	-1	-1	1	-1	-1
27	1	-1	1	1	-1	-1	1	1	-1	1
28	1	-1	-1	1	-1	1	-1	1	1	-1
29	-1	-1	-1	-1	-1	1	1	1	1	1
30	-1	-1	-1	-1	1	1	-1	-1	-1	-1
31	1	1	1	1	-1	1	-1	-1	-1	-1
32	-1	1	1	-1	1	1	-1	1	1	-1

+1: Over estimate

-1: Under estimated

Table 26. Fractional Factorial 2^{15-9} design for p=6

	Correlations														
Run	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P
1	1	1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	-1
2	-1	1	1	-1	-1	-1	1	1	1	1	-1	-1	-1	1	1
3	1	-1	1	-1	-1	-1	1	1	-1	-1	1	1	1	-1	1
4	-1	1	-1	-1	-1	1	1	1	-1	-1	-1	1	1	1	-1
5	-1	1	1	1	1	-1	-1	-1	-1	-1	-1	1	1	1	1
6	-1	-1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	1
7	1	-1	-1	-1	1	1	1	-1	1	-1	-1	-1	1	1	1
8	-1	-1	1	-1	-1	1	-1	-1	1	1	-1	1	1	-1	1
9	1	1	1	-1	1	1	-1	1	-1	1	-1	-1	1	-1	-1
10	-1	1	1	1	-1	-1	-1	1	-1	1	1	1	-1	-1	-1
11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12	-1	1	-1	-1	1	1	1	-1	-1	1	1	1	-1	-1	1
13	1	1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1
14	1	1	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1
15	1	1	-1	1	1	-1	1	1	-1	-1	1	-1	-1	1	-1
16	-1	-1	1	1	-1	1	1	-1	-1	1	1	-1	1	1	-1
17	1	1	-1	-1	1	-1	-1	1	1	-1	-1	1	-1	-1	1
18	1	1	1	-1	-1	1	-1	-1	-1	-1	1	-1	-1	1	1
19	-1	-1	-1	1	-1	-1	1	-1	1	-1	1	1	-1	1	1
20	-1	-1	-1	-1	1	-1	-1	1	-1	1	1	-1	1	1	1
21	1	-1	1	1	-1	-1	-1	1	1	-1	-1	-1	1	1	-1
22	1	-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1
23	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
24	-1	1	-1	1	1	1	-1	-1	1	1	-1	-1	-1	1	-1
25	-1	1	-1	1	-1	1	-1	1	1	-1	1	-1	1	-1	1
26	1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1
27	1	-1	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	-1
28	-1	-1	1	-1	1	1	-1	1	1	-1	1	1	-1	1	-1
29	-1	1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	-1
30	-1	-1	-1	1	1	-1	1	1	1	1	-1	1	1	-1	-1
31	1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1	-1
32	1	-1	1	1	1	-1	-1	-1	1	1	1	-1	-1	-1	1

+1: Over estimate

-1: Under estimated

Appendix D. Sample size using correlation confidence interval

Based in the following test of hypothesis

$$H_0: \quad \rho = \rho_0$$

it is possible to construct the following 100(1- α) % confidence interval

$$\tanh(a\tanh r - \frac{Z_{\alpha/2}}{\sqrt{n-3}}) \leq \rho \leq \tanh(a\tanh r + \frac{Z_{\alpha/2}}{\sqrt{n-3}}) \qquad \text{(C1)}$$

This interval uses the statistic

$$Z_0 = (a\tanh r - a\tanh \rho_0)(n-3)^{1/2} \qquad \text{(C2)}$$

which is based in that, for large samples ($n \geq 25$), the statistic

$$Z = a\tanh r \sim N(a\tanh \rho, \frac{1}{n-3})$$

See Montgomery, Peck and Vinning (2001) for more details. So, it is possible to find a sample size given determined interval, in this case Table 27 shows the results. Based on this results it is clear that a big amount of samples are necessary if we need that the correlation would not be far from its real value specially for lower correlations.

Table 27. Sample size to achieve a determined error using the correlation confidence interval

Correlation	Error = +/- 0.01			Error = +/- 0.05		
	α =0.01	α =0.05	α =0.10	α =0.01	α =0.05	α =0.10
0.10	98,334	49,182	49,182	3,102	1,566	1,566
0.20	98,334	49,182	49,182	3,102	1,566	1,566
0.30	98,334	49,182	24,606	3,102	1,566	1,566
0.40	49,182	49,182	24,606	3,102	1,566	1,566
0.50	49,182	24,606	24,606	3,102	1,566	798
0.60	49,182	24,606	12,318	1,566	798	798
0.70	24,606	12,318	12,318	798	798	414
0.80	12,318	6,174	6,174	798	414	222
0.90	3,102	1,566	1,566	222	126	78

Appendix E. Percentiles and distribution of the run-length

Table 28. In-control Run-Length summary for $p = 2$

		Estimated Matrix						True Matrix ($mn \to \infty$)					
	k	#obs	Mean	SDRL	Q10	Q50	Q90	#obs	Mean	SDRL	Q10	Q50	Q90
ARL = 200	3	1000	163.64	171.29	12	110	388	500	163.98	172.00	12	110	386.5
	10	1000	178.36	199.54	15	111	410.5	500	180.41	183.11	15	113.5	414.5
	20	1000	210.44	222.91	19	138	490	500	213.10	232.79	19	141	490
	50	1000	210.44	222.91	19	138	490	500	211.23	231.93	19	139.5	486
	100	1000	210.44	222.91	19	138	490	500	210.63	232.07	18.5	137.5	486
	Avg	1000	194.66	209.77	17	126	454	500	195.87	212.88	16	126.5	454
ARL = 400	3	1000	327.05	326.51	34	209	759	500	327.86	326.59	34	210	759.5
	10	1000	375.45	390.47	41	234	893	500	351.45	343.14	45	222.5	832.5
	20	1000	427.82	411.65	47.5	290	1001	500	427.14	410.00	47.5	290	999
	50	1000	426.65	409.99	47.5	290	1001	500	422.37	406.88	46	290	977.5
	100	1000	425.54	409.15	47.5	290	999	500	422.06	406.16	46	290	977.5
	Avg	1000	396.50	392.79	45	258	939	500	390.18	382.28	45	257.5	924
ARL = 600	3	1000	490.13	485.41	54	360	1127	500	491.97	489.49	53.5	360	1133
	10	1000	544.01	558.02	61	378	1258.5	500	553.19	539.25	65	393	1254.5
	20	1000	660.57	630.30	74	478	1452	500	660.28	632.17	72	477.5	1453
	50	1000	654.68	623.89	74	475	1448	500	656.05	627.33	72	477.5	1447.5
	100	1000	652.46	624.04	72	469.5	1448	500	653.41	625.34	72	475	1446
	Avg	1000	600.37	590.96	68	421	1360	500	602.98	589.12	68	422	1359
ARL = 800	3	1000	626.62	609.26	64	416	1378.5	500	621.43	603.20	63.5	415.5	1374.5
	10	1000	690.10	806.18	65	432	1501	500	674.15	702.30	65.5	456	1473.5
	20	1000	851.26	882.42	75.5	595	1967.5	500	848.07	878.95	75.5	595	1943.5
	50	1000	846.88	875.56	75.5	594	1946	500	845.10	879.08	75.5	590	1943.5
	100	1000	843.51	872.43	70.5	590	1930.5	500	843.67	876.90	75.5	590	1943.5
	Avg	1000	771.67	820.94	67	503	1797	500	766.49	801.83	67	515	1780
ARL = 1000	3	1000	785.45	798.01	75	551	1727	500	781.35	779.22	79.5	550.5	1716.5
	10	1000	917.52	1031.63	72	594.5	2048.5	500	895.39	945.46	71	605.5	2022
	20	1000	1118.16	1176.72	105	770	2759	500	1109.39	1174.90	104	767	2686
	50	1000	1110.03	1170.90	104	767	2756	500	1103.81	1173.58	104	766	2686
	100	1000	1109.44	1171.17	104	766	2756	500	1103.74	1173.58	104	766	2686
	Avg	1000	1008.12	1087.60	90	680.5	2390	500	998.73	1069.52	90	680.5	2347
ARL = 1200	3	1000	1026.03	1000.72	107	704	2429	500	1024.64	1012.72	103	686.5	2429
	10	1000	1065.96	1159.69	97	685	2515	500	1059.69	1053.83	108	736.5	2414
	20	1000	1319.51	1277.63	124	928	2961	500	1327.65	1287.78	126	929	2993
	50	1000	1309.85	1267.11	124	922	2948	500	1323.42	1276.12	126	929	2993
	100	1000	1308.50	1267.57	124	921	2948	500	1318.16	1269.25	126	929	2954.5
	Avg	1000	1205.97	1205.93	115.5	820	2795	500	1210.71	1193.16	116.5	832	2795
ARL = 1400	3	1000	1062.04	1083.38	114	692	2612.5	500	1063.77	1100.55	113	691.5	2589.5
	10	1000	1220.54	1275.22	110	864	2832	500	1158.64	1122.15	113	837.5	2615
	20	1000	1455.52	1466.09	149.5	1002	3403	500	1435.98	1419.62	149.5	1007	3285
	50	1000	1455.31	1466.25	149.5	1002	3403	500	1430.10	1419.71	143	1004.5	3285
	100	1000	1446.88	1463.26	139	1000	3352.5	500	1427.79	1420.47	143	999.5	3285
	Avg	1000	1328.06	1368.34	131	895.5	3083	500	1303.26	1314.00	131	893	3046.5

Table 29. In-control Run-Length summary for $p = 3$

		Estimated Matrix						True Matrix ($mn \to \infty$)					
	k	#obs	Mean	SDRL	Q10	Q50	Q90	#obs	Mean	SDRL	Q10	Q50	Q90
	5	16000	187.62	182.26	22	134	427	2000	187.28	183.80	22	133	431
	10	16000	195.19	188.94	25	137	445	2000	200.45	194.22	26	144	450
ARL = 200	20	16000	187.51	201.24	19	125	439	2000	191.57	187.37	22	135	445
	50	16000	229.64	356.67	10	113	540	2000	173.73	170.84	21	120	392
	100	16000	300.14	650.50	6	116	674	2000	194.31	191.73	22	135	455
	Avg	16000	220.02	365.78	15	127	486	2000	189.47	185.95	22	133	436
	5	16000	407.83	411.94	39	279.5	938	2000	410.86	418.19	40	280	935
	10	16000	420.25	433.50	38	286	973	2000	433.41	440.38	39	302	1007
ARL = 400	20	16000	397.22	437.87	29	250	947	2000	412.21	427.60	36	280	955
	50	16000	471.64	765.08	14	225	1166	2000	360.82	365.22	38	240	835
	100	16000	635.74	1465.19	7	231	1443	2000	386.72	386.11	43	263	872
	Avg	16000	466.54	814.91	23	254	1065	2000	400.80	409.12	39	273	931
	5	16000	567.39	580.35	51.5	388	1315	2000	571.26	585.36	52	389	1336
	10	16000	602.98	602.74	57	426	1372.5	2000	614.92	598.35	57	439	1420
ARL = 600	20	16000	559.88	630.77	40	353	1326	2000	560.96	558.55	54	391	1314
	50	16000	683.73	1135.51	18	312	1659.5	2000	499.68	500.15	42	333	1142
	100	16000	944.75	2105.50	9	328	2205	2000	542.47	544.66	54	365	1248
	Avg	16000	671.75	1176.69	29	364	1523	2000	557.86	559.62	53	385	1292
	5	16000	746.46	748.22	89	503	1763	2000	749.77	740.41	91	508	1769
	10	16000	789.54	796.47	86	536	1881	2000	820.66	808.38	95	573	1922
ARL = 800	20	16000	747.91	855.56	62	475	1765	2000	753.29	779.09	86	498	1717
	50	16000	931.14	1607.41	24	413	2281	2000	653.63	668.35	71	427	1577
	100	16000	1263.52	2756.61	10	447	2922	2000	747.89	775.63	80	505	1732
	Avg	16000	895.72	1568.45	42	481	2041	2000	745.05	757.63	83	498	1750
	5	16000	974.11	990.84	103	686	2156	2000	980.20	1017.39	103	691	2159
	10	16000	997.73	1056.32	101	687	2243	2000	1033.44	1096.50	105	737	2344
ARL = 1000	20	16000	943.51	1095.24	74	588	2218	2000	967.79	1033.34	103	673	2146
	50	16000	1178.62	2073.57	26	552	2776.5	2000	849.90	925.96	85	579	1894
	100	16000	1565.56	3561.94	11	555	3591.5	2000	919.10	926.70	105	611	2147
	Avg	16000	1131.91	2027.41	50	619	2543	2000	950.09	1003.84	101	652	2152
	5	16000	1156.45	1174.03	115	778	2688	2000	1166.29	1186.00	115	789	2737
	10	16000	1177.31	1224.91	116.5	795	2750	2000	1229.96	1254.94	130	822	2912
ARL = 1200	20	16000	1111.45	1271.28	78	691	2689	2000	1140.35	1195.36	113	756	2676
	50	16000	1391.86	2406.78	26	607	3429	2000	967.91	1030.26	98	663	2305
	100	16000	1928.07	4368.84	11	627	4476	2000	1115.07	1194.76	110	718	2584
	Avg	16000	1353.03	2442.66	54	708	3023	2000	1123.91	1177.64	111	746	2614
	5	16000	1368.95	1331.38	151	961	3164	2000	1381.62	1319.09	152	975	3228
	10	16000	1410.13	1415.39	144	980	3261	2000	1476.88	1479.26	152	1027	3421
ARL = 1400	20	16000	1345.50	1517.29	100	846	3223	2000	1373.31	1377.01	151	962	3170
	50	16000	1684.16	2880.51	37	740	4000	2000	1195.18	1175.06	144	816	2812
	100	16000	2308.20	5231.65	13	799.5	5201.5	2000	1303.00	1254.14	158	909	2993
	Avg	16000	1623.39	2912.17	69	869	3644	2000	1346.00	1328.02	152	938	3103

Table 30. In-control Run-Length summary for $p = 4$

		Estimated Matrix						True Matrix ($mn \to \infty$)					
	k	#obs	Mean	SDRL	Q10	Q50	Q90	#obs	Mean	SDRL	Q10	Q50	Q90
	5	170000	184.44	191.82	17	124	433	5000	189.05	190.79	20	130	439
	10	170000	249.87	346.51	21	146	565	5000	261.04	347.70	23	150	589.5
ARL = 200	20	170000	175.27	197.15	16	113	417	5000	199.24	205.29	20.5	132	475
	50	170000	238.34	746.58	13	113	453	5000	200.12	216.80	20	131	464
	100	170000	258.11	995.26	9	89	423	5000	187.26	192.30	20	127	433
	Avg	170000	221.21	591.54	15	117	460	5000	207.34	239.63	21	132	481
	5	170000	385.96	406.20	38	264	920	5000	403.10	413.84	42	278	943.5
	10	170000	531.33	803.22	46	307	1161	5000	558.39	808.17	48	323.5	1217
ARL = 400	20	170000	347.66	395.60	30	214	856	5000	411.07	420.83	42	273	979.5
	50	170000	507.95	1777.41	24	221	940	5000	417.67	456.93	39	273	999
	100	170000	542.32	2281.13	15	171	863	5000	389.50	402.65	37	265	925
	Avg	170000	463.04	1368.32	27	234	947	5000	435.95	527.52	41.5	281	1006
	5	170000	542.49	566.33	50	363	1289	5000	568.62	576.34	59	391	1335
	10	170000	798.97	1278.36	67	448	1707	5000	852.63	1341.44	75	480	1815
ARL = 600	20	170000	505.82	568.64	36	319	1218	5000	603.66	615.86	60	410	1423
	50	170000	754.68	2891.04	33	318	1340	5000	615.08	691.72	62	397	1430
	100	170000	818.98	3704.47	19	242	1203	5000	553.68	563.02	54	386.5	1288.5
	Avg	170000	684.19	2211.22	35	332	1355	5000	638.73	820.43	62	409	1464
	5	170000	757.38	784.99	67	499	1781	5000	796.58	796.70	81	535	1847
	10	170000	1107.14	1916.18	97	597	2312	5000	1196.87	1949.13	103	637	2566
ARL = 800	20	170000	683.33	777.82	56	425	1623	5000	814.92	834.74	87	561	1832
	50	170000	1042.79	4018.50	40	424	1797	5000	833.93	906.15	81	543	1907.5
	100	170000	1170.02	5429.28	23	326	1673	5000	768.28	790.62	77	519	1766.5
	Avg	170000	952.13	3184.59	50	445	1832	5000	882.12	1157.75	84	559	1980
	5	170000	976.68	1061.49	96	623	2349	5000	1029.50	1090.09	109	669	2415
	10	170000	1479.39	2572.46	112	750	3199	5000	1580.07	2503.78	127	833	3445
ARL = 1000	20	170000	885.77	1066.72	68	500	2173	5000	1087.52	1161.36	108	697	2579
	50	170000	1352.36	5591.48	56	502	2403	5000	1079.54	1223.65	107	688.5	2532.5
	100	170000	1477.43	7143.46	26	388	2135	5000	999.87	1060.12	96	651	2405
	Avg	170000	1234.33	4277.77	62	546	2441	5000	1155.30	1526.83	109	699	2659
	5	170000	1105.38	1169.48	101	714	2663	5000	1162.88	1191.45	119	786	2755.5
	10	170000	1760.27	3426.41	134	878	3677	5000	1894.91	3412.79	145	936.5	4010.5
ARL = 1200	20	170000	985.09	1140.21	79	595	2341	5000	1193.26	1236.90	128	782	2839
	50	170000	1559.22	6298.73	59	595	2684	5000	1252.30	1410.91	132.5	825	2875
	100	170000	1765.91	8972.59	31	444	2277	5000	1115.52	1136.84	119.5	755	2549.5
	Avg	170000	1435.17	5198.67	71	643	2755	5000	1323.77	1912.82	128	806	2924.5
	5	170000	1319.81	1405.03	108	858	3101	5000	1396.84	1431.39	132	939	3197.5
	10	170000	2115.80	4140.71	134	1069	4343	5000	2242.44	3867.41	148.5	1172	4796
ARL = 1400	20	170000	1210.27	1451.30	75	707	2955	5000	1467.20	1592.65	124	938	3440
	50	170000	1901.16	7874.53	64	717	3218	5000	1498.10	1694.50	119	960	3462
	100	170000	2152.53	11002.71	35	523	2844	5000	1326.12	1389.76	116	881	3100
	Avg	170000	1739.92	6404.49	71	763	3308	5000	1586.14	2231.47	128	973	3499

Table 31. In-control Run-Length summary for $p = 5$

		Estimated Matrix						True Matrix ($mn \to \infty$)					
	k	#obs	Mean	SDRL	Q10	Q50	Q90	#obs	Mean	SDRL	Q10	Q50	Q90
ARL = 200	5	68000	186.92	183.02	21	125	426	2000	192.23	187.83	22	131	435
	10	68000	190.31	191.69	21	125	443	2000	194.46	192.22	22	131	450
	20	68000	175.83	192.36	16	112	413	2000	211.47	212.23	23	143.5	475
	50	68000	194.70	517.23	9	80	410	2000	227.84	230.24	26	158	509.5
	100	68000	509.04	1289.05	4	93	1256	2000	215.31	217.38	25	147	498
	Avg	68000	251.36	651.09	10	110.5	485	2000	208.26	208.97	24	141	472
ARL = 400	5	68000	362.57	355.44	41	254	844	2000	375.95	361.39	42	267.5	856
	10	68000	391.33	388.41	42	279	903	2000	401.88	391.18	42	287.5	922
	20	68000	348.20	386.73	27	219	843	2000	422.66	423.24	46	293.5	991.5
	50	68000	391.93	1189.22	13	140	827	2000	455.15	466.03	46	308	1092
	100	68000	1067.16	2745.46	4	161	2817	2000	431.03	439.59	44	295.5	1012.5
	Avg	68000	512.24	1397.48	18	212	997	2000	417.33	418.67	44	290	969
ARL = 600	5	68000	564.60	562.19	60	395	1311	2000	591.47	585.69	63.5	417.5	1379
	10	68000	611.15	626.97	63	420	1415	2000	647.44	640.36	67	443	1521
	20	68000	529.17	607.01	39	335	1256	2000	670.28	674.32	70.5	475	1536.5
	50	68000	597.55	1882.07	17	205	1249	2000	709.78	731.17	68	485	1695
	100	68000	1711.49	4542.86	5	235	4440	2000	668.22	677.15	72.5	449	1556
	Avg	68000	802.79	2293.18	25	329	1521	2000	657.44	664.46	67	455	1523
ARL = 800	5	68000	750.09	769.48	73	504	1784	2000	797.02	815.56	76	543	1871
	10	68000	808.63	833.18	80	552	1898	2000	850.77	875.92	83.5	578	1975
	20	68000	693.96	810.79	48	411	1713	2000	887.16	934.04	86	595.5	2008
	50	68000	794.08	2545.52	19	262	1599	2000	924.14	1015.48	92	592	2133
	100	68000	2302.24	6194.36	6	301	5962	2000	853.10	888.53	93	560	1943.5
	Avg	68000	1069.80	3120.88	29	411	1995	2000	862.43	909.15	85	572	1977
ARL = 1000	5	68000	894.48	900.52	96	618	2028	2000	948.20	953.85	102	648	2139.5
	10	68000	958.35	981.49	102	661	2129	2000	995.72	1005.86	109	699	2193
	20	68000	831.24	960.65	62	527	1955	2000	1036.57	1009.51	115.5	752	2388.5
	50	68000	985.06	3262.48	23	314	1995	2000	1168.83	1206.59	118	806	2757
	100	68000	2914.89	7736.70	6	362	7593	2000	1100.72	1134.87	102	761	2556
	Avg	68000	1316.80	3909.05	35	510	2404	2000	1050.01	1068.87	108	727	2403
ARL = 1200	5	68000	1080.97	1109.02	111	712	2490	2000	1135.23	1150.40	111	752	2635
	10	68000	1197.94	1211.05	124	800	2820	2000	1276.04	1275.86	131	863	3023.5
	20	68000	1012.75	1156.06	73	618	2465.5	2000	1294.09	1261.35	136	873	3056.5
	50	68000	1216.07	4180.18	26	402	2446	2000	1444.10	1418.43	144	1000.5	3271.5
	100	68000	3610.91	9784.74	6	455	9412	2000	1348.95	1330.94	139	955.5	3059
	Avg	68000	1623.73	4943.91	39	604	3033	2000	1299.68	1294.08	130	882	3023.5
ARL = 1400	5	68000	1305.40	1306.09	134	915	3006	2000	1407.29	1377.84	153.5	996	3226
	10	68000	1385.32	1423.98	138	926	3230	2000	1463.75	1501.40	141	986.5	3429
	20	68000	1225.23	1412.37	88	767	2923	2000	1609.35	1632.07	187.5	1121	3795
	50	68000	1479.52	5242.38	28	446	2842	2000	1687.07	1697.60	191.5	1163	3918.5
	100	68000	4185.28	11420.53	7	475	10875	2000	1497.18	1515.89	161.5	1047	3396.5
	Avg	68000	1916.15	5832.83	47	725	3561	2000	1532.93	1551.95	163	1067	3563

Table 32. In-control Run-Length summary for $p = 6$

		Estimated Matrix						True Matrix ($mn \to \infty$)					
	k		Mean	SDRL	Q10	Q50	Q90		Mean	SDRL	Q10	Q50	Q90
	5	17000	189.15	200.06	15	127	432	500	200.44	217.20	17	128	480
	10	17000	135.36	146.81	12	89	312	500	140.44	158.22	12	90	311.5
ARL = 200	20	17000	171.22	189.92	14	103.5	406	500	211.00	211.94	21.5	147.5	475
	50	17000	139.60	172.12	9	79	347	500	168.22	172.49	15	109	398
	100	17000	200.53	216.00	16	129	476	500	242.67	242.83	24.5	158	565
	Avg	17000	167.17	188.31	12	100	405	500	192.55	205.79	17	127	444.5
	5	17000	364.22	352.61	49	259	820	500	390.32	383.07	54	272	855
	10	17000	275.84	267.07	34	187	640	500	298.88	285.03	36	196.5	679.5
ARL = 400	20	17000	330.48	360.39	31	214	775	500	424.83	428.01	55	291	934
	50	17000	266.28	317.84	17	158	683	500	337.60	318.69	40.5	237.5	763
	100	17000	409.79	443.82	40	266	968	500	490.09	510.04	56	335.5	1131.5
	Avg	17000	329.32	357.22	31	209	764	500	388.35	398.48	48	267	848.5
	5	17000	523.31	506.82	49	380	1169	500	570.96	542.84	55	433.5	1268
	10	17000	376.42	391.66	29	245	912	500	407.30	407.11	31.5	270.5	960
ARL = 600	20	17000	466.95	539.77	37	297	1095	500	608.05	618.12	55	434	1362.5
	50	17000	373.23	457.89	21	208	940	500	480.75	453.93	47.5	363	1075.5
	100	17000	572.86	632.18	40	382	1329	500	700.57	686.00	58.5	516.5	1539
	Avg	17000	462.55	518.10	33	299	1092	500	553.53	560.05	49	392	1228
	5	17000	720.56	725.32	74	470	1682	500	763.44	749.07	76.5	509	1820.5
	10	17000	508.95	533.77	52	339	1227	500	553.03	544.47	56.5	383	1311.5
ARL = 800	20	17000	625.98	741.62	52	380	1526	500	854.42	915.40	96.5	572.5	1984
	50	17000	496.95	626.98	30	271	1294	500	636.53	649.86	73	421	1460
	100	17000	772.78	856.80	73	470	1832	500	921.54	916.07	117	623	2076.5
	Avg	17000	625.04	713.99	52	388	1497	500	745.79	780.31	77	477.5	1729.5
	5	17000	950.42	967.94	100	649	2130	500	1023.82	999.96	107	719.5	2371.5
	10	17000	679.20	700.13	71	469	1517	500	752.60	755.40	75.5	560	1643
ARL = 1000	20	17000	802.67	944.56	66	474	1959	500	1068.05	1073.95	104.5	736	2447.5
	50	17000	632.95	820.53	32	343	1646	500	806.88	800.52	84.5	586	1885.5
	100	17000	1011.84	1109.03	95	653	2382	500	1220.26	1154.90	134	883.5	2795.5
	Avg	17000	815.42	930.71	67	517	1958.5	500	974.32	983.87	100	682	2223
	5	17000	1082.38	1154.91	93	724	2448	500	1173.65	1292.78	103.5	798	2583
	10	17000	742.74	788.82	59	477	1719	500	827.62	869.43	73.5	533.5	1885.5
ARL = 1200	20	17000	918.59	1063.51	62	570	2259	500	1267.45	1291.31	139	918	2855.5
	50	17000	694.50	901.43	35	367	1831	500	874.41	913.40	94	606	2060.5
	100	17000	1120.52	1284.33	84	698	2682	500	1384.72	1429.68	134	973.5	3102.5
	Avg	17000	911.75	1067.47	60	567	2220	500	1105.57	1200.11	104	725.5	2564
	5	17000	1316.86	1347.69	145	917	2895	500	1471.86	1488.28	161.5	1077	3118
	10	17000	849.90	871.02	89	585	1983.5	500	973.20	1014.63	107.5	655.5	2237.5
ARL = 1400	20	17000	1127.92	1295.10	94	654	2721	500	1626.49	1580.33	201	1166	3779
	50	17000	861.09	1141.84	45	440	2171	500	1152.79	1316.01	136.5	740	2480.5
	100	17000	1367.86	1531.14	117	836	3222	500	1718.13	1742.35	177	1231.5	3876.5
	Avg	17000	1104.73	1275.86	86	665	2594.5	500	1388.50	1476.02	154	937	3064

Table 33. Run-length summary for condition number ($\hat{k}$ < = 5) by theoretical ARL and number of variables (p)

Theoretical ARL		Number of variables (p)				
		2	3	4	5	6
200	ARL	163.75	188.27	193.11	189.25	202.73
	SDRL	171.47	182.55	196.47	184.49	214.17
	Q10	12	22	20	22	17
	Q50	110	135	131	127	135
	Q90	388	431	452	430	476
400	ARL	327.32	414.60	406.03	372.34	391.82
	SDRL	326.43	418.55	417.04	364.97	376.44
	Q10	34	39	42	41	52
	Q50	209.5	285	280	263.5	279
	Q90	759	947	950	856	892
600	ARL	490.75	577.11	572.00	581.04	560.59
	SDRL	486.61	593.73	580.88	577.82	535.09
	Q10	54	53	59	62	52
	Q50	360	389	393	405	414.5
	Q90	1127	1340	1337	1372.5	1247
800	ARL	624.89	758.95	799.48	769.06	784.95
	SDRL	607.05	757.68	805.70	791.46	787.84
	Q10	64	91	82	75	77
	Q50	416	514	535	514	533
	Q90	1376	1791	1852	1850	1835
1000	ARL	784.08	990.08	1032.59	920.74	1028.78
	SDRL	791.54	1025.55	1091.31	926.30	1034.73
	Q10	77	103	108	102	106
	Q50	551	691	672	638	714
	Q90	1727	2160	2418	2080	2351
1200	ARL	1025.56	1177.31	1166.61	1113.12	1189.03
	SDRL	1004.40	1199.07	1198.46	1150.62	1257.75
	Q10	103	115	119	111	104
	Q50	696	802	786	724	842
	Q90	2429	2781	2787	2578	2664
1400	ARL	1062.62	1393.15	1396.98	1339.57	1441.02
	SDRL	1088.76	1350.80	1444.72	1334.06	1452.24
	Q10	114	153	128	136	164.5
	Q50	692	976	937	933	1026
	Q90	2612	3228	3253.5	3104	3064

Table 34. Run-length summary for condition number (5< *k* <=10) by theoretical ARL and number of variables (p)

Theoretical ARL		Number of variables (*p*)				
		2	3	4	5	6
200	ARL	199.00	197.18	254.55	194.15	158.91
	SDRL	210.25	191.29	387.27	193.58	170.09
	Q10	18	24	19	22	13
	Q50	124.5	139	137	130	100
	Q90	477	448	570	446	367
400	ARL	409.13	422.86	546.61	387.98	313.95
	SDRL	408.21	432.75	903.65	383.35	302.37
	Q10	49	39	40	42	39
	Q50	260.5	291.5	295	278	214
	Q90	977.5	984	1168	900	717
600	ARL	617.05	598.92	834.27	611.44	446.27
	SDRL	599.64	599.38	1477.21	622.03	446.68
	Q10	70	55	57	64	39
	Q50	426.5	422	417	422	312
	Q90	1390.5	1364.5	1767	1414	1037
800	ARL	783.41	791.13	1162.92	808.21	600.40
	SDRL	861.28	794.34	2226.74	827.15	609.39
	Q10	77	93	77	79	59
	Q50	502	542	565	555	402
	Q90	1796	1868	2389	1894	1424
1000	ARL	1050.67	1013.46	1565.17	960.51	796.62
	SDRL	1126.61	1049.90	3007.63	976.37	814.18
	Q10	85	105	103	102	86
	Q50	752.5	714	709	664	569
	Q90	2499.5	2279	3331	2139	1790
1200	ARL	1210.19	1202.41	1889.94	1186.38	883.75
	SDRL	1245.17	1236.11	4034.07	1197.52	944.46
	Q10	120	124	113	125	73
	Q50	812	805	838	798	580
	Q90	2716	2833.5	3837	2753	2107
1400	ARL	1363.85	1431.23	2263.04	1399.75	1049.83
	SDRL	1357.72	1418.90	4907.90	1423.34	1088.95
	Q10	135	152	114	144	117
	Q50	965	992	1003	953	693
	Q90	3223	3334	4495	3238	2329

Table 35. Run-length summary for condition number (10< $\hat{k}$ <=20) by theoretical ARL and number of variables (*p*)

Theoretical ARL		Number of variables (*p*) 2	3	4	5	6
200	ARL	176.56	210.63	211.18	198.60	161.31
	SDRL	199.90	214.56	237.00	205.29	180.95
	Q10	15	25	20	22	14
	Q50	110	147	134	131	99
	Q90	404	477	493	451	376
400	ARL	359.07	451.78	434.65	401.61	321.94
	SDRL	362.88	465.52	508.97	410.26	337.50
	Q10	36	40	40	42	34
	Q50	228.5	309	278	279	209
	Q90	857.5	1060	1022	935	723
600	ARL	538.86	647.89	630.20	619.00	445.77
	SDRL	550.20	675.30	726.01	652.32	502.36
	Q10	60.5	59	59	63	35
	Q50	376.5	447	404	421	280
	Q90	1269.5	1503	1491	1421	1045
800	ARL	674.81	858.44	865.67	829.37	603.28
	SDRL	745.20	909.68	1034.15	882.93	683.23
	Q10	61	93	81	79	58
	Q50	451	577	554	547	383
	Q90	1493	1994	1987	1949	1422
1000	ARL	886.39	1083.76	1132.60	983.35	794.03
	SDRL	978.40	1168.71	1371.54	1046.96	885.39
	Q10	80	107	105	103	81
	Q50	587	739	687	671	523
	Q90	1995	2473	2654.5	2200	1829
1200	ARL	1052.13	1280.62	1286.64	1217.16	886.04
	SDRL	1093.47	1354.42	1645.96	1263.82	982.50
	Q10	99.5	124	121	126.5	69
	Q50	697	856	788	799	572
	Q90	2455.5	2978	2945	2852	2166
1400	ARL	1184.07	1551.35	1571.27	1447.94	1054.96
	SDRL	1256.55	1603.67	1969.11	1531.86	1179.57
	Q10	110	159	117	144	97.5
	Q50	824.5	1051	962	965	658
	Q90	2740.5	3547	3603	3359	2441

Table 36. Run-length summary for condition number (20< $\hat{k}$ <=50) by theoretical ARL and number of variables (*p*)

Theoretical ARL		Number of variables (*p*) 2	3	4	5	6
200	ARL	211.33	201.77	184.78	184.75	170.17
	SDRL	226.17	238.08	209.47	210.02	186.25
	Q10	19	18	17	18	14
	Q50	139.5	126.5	117	114	103
	Q90	490	469	437	430	417
400	ARL	427.66	421.33	375.06	366.30	327.40
	SDRL	410.95	508.33	442.29	429.41	348.23
	Q10	47.5	31	34	31	34
	Q50	290	248	232	221	213
	Q90	1001	1015	913	876	761
600	ARL	660.47	605.44	543.25	554.80	463.20
	SDRL	630.71	781.58	635.08	667.32	510.74
	Q10	73	39	44	45	38
	Q50	478	342	339	341	306
	Q90	1452	1429	1299.5	1315	1075
800	ARL	850.20	815.73	740.20	722.70	624.93
	SDRL	880.97	1055.03	885.31	899.62	705.54
	Q10	75.5	60	63	61	58
	Q50	595	462	452	426	388
	Q90	1959	1980	1744	1732	1471
1000	ARL	1115.24	1030.34	945.50	883.49	798.71
	SDRL	1175.73	1350.40	1160.31	1094.87	902.76
	Q10	105	70	81	73	73
	Q50	768.5	590	539	532	513
	Q90	2756	2491	2349	2054.5	1912.5
1200	ARL	1322.22	1220.73	1078.14	1074.13	897.79
	SDRL	1280.60	1660.96	1331.89	1306.99	1014.96
	Q10	124	77	89	84	67
	Q50	928	676	645	642	573
	Q90	2961	2945	2577	2565	2192
1400	ARL	1449.01	1456.84	1305.69	1279.68	1115.90
	SDRL	1450.32	1891.14	1611.30	1600.84	1264.00
	Q10	149.5	105	88	101	96
	Q50	1004	819	759	766	676
	Q90	3352.5	3503	3101	2993	2622

Table 37. Run-length summary for condition number (50< $\hat{k}$ <=100) by theoretical ARL and number of variables (*p*)

Theoretical ARL		Number of variables (*p*) 2	3	4	5	6
200	ARL	209.26	240.46	209.02	208.70	175.13
	SDRL	219.01	319.32	555.25	408.85	206.04
	Q10	19	20	15	12	12
	Q50	137.5	136	108	94	99
	Q90	489	571	416	468	429
400	ARL	423.37	489.53	433.14	411.82	354.12
	SDRL	407.40	678.30	1344.61	834.69	423.61
	Q10	46	36	26	19	29
	Q50	290	266	208	164.5	206
	Q90	978	1185	842	950	853
600	ARL	647.90	727.75	633.56	622.92	493.23
	SDRL	619.31	1043.34	2119.65	1324.16	608.50
	Q10	72	43	35	26	33
	Q50	465	385	298	238	285
	Q90	1447	1742	1186	1457	1186
800	ARL	839.56	993.64	894.12	820.49	665.24
	SDRL	868.84	1451.16	3254.54	1801.23	829.84
	Q10	74	64	49	32	53
	Q50	582.5	516	403	306	375
	Q90	1928	2357.5	1611	1853.5	1659.5
1000	ARL	1098.42	1233.98	1122.33	1049.11	859.81
	SDRL	1164.07	1848.41	4101.94	2506.20	1075.09
	Q10	103	78	61	37	63
	Q50	764	657	469	370	466
	Q90	2693	2879.5	2056	2345	2159
1200	ARL	1301.22	1471.00	1316.09	1252.94	961.68
	SDRL	1259.45	2190.03	5233.80	2852.13	1245.28
	Q10	124	89	69	43	63
	Q50	920	747	558	462	530
	Q90	2942.5	3578	2232	2860	2452
1400	ARL	1436.08	1801.40	1590.26	1466.39	1167.53
	SDRL	1446.72	2631.91	6488.90	3502.44	1487.35
	Q10	139	117.5	71	52	86
	Q50	997	948	645	508	631
	Q90	3302	4217	2782	3281	2928

Table 38. Run-length summary for condition number ($100< \hat{k}$) by theoretical ARL and number of variables (p)

Theoretical ARL		Number od variables (p) 3	4	5	6
200	ARL	302.46	336.15	518.13	142.83
	SDRL	858.56	1489.44	1424.74	175.37
	Q10	2	5	2	8
	Q50	29	53	43	79
	Q90	749	362	1472	383
400	ARL	648.92	729.20	1100.35	285.60
	SDRL	1944.21	3429.40	3080.04	356.05
	Q10	2	7	3	14
	Q50	46	95	63	158
	Q90	1562.5	704	3268	733
600	ARL	939.34	1144.95	1774.54	392.61
	SDRL	2753.87	5629.24	5074.53	495.43
	Q10	3	9	3	18
	Q50	59	128	85	209
	Q90	2317.5	987.5	5256	1003
800	ARL	1234.61	1612.52	2397.55	513.91
	SDRL	3633.18	7945.23	6909.10	662.43
	Q10	3	10	4	23
	Q50	75	164	105	272.5
	Q90	2974	1364	7006	1355
1000	ARL	1544.51	2108.77	3012.55	678.49
	SDRL	4715.09	10869.88	8605.37	873.54
	Q10	3	10	4	28
	Q50	87	193	117	352
	Q90	3796.5	1714	8797.5	1823
1200	ARL	1915.29	2496.04	3780.26	731.64
	SDRL	5717.11	12941.03	10984.93	974.59
	Q10	3	11	4	30
	Q50	99	215	137	370
	Q90	4902.5	1852	11075	1910
1400	ARL	2275.29	3078.19	4428.90	894.99
	SDRL	6889.37	15956.64	12922.38	1179.76
	Q10	3	12	4	33
	Q50	115	262	148	431
	Q90	5737	2210	13067.5	2306

Appendix F. Box-Cox transformation procedure

Trying to improve the model, a Box-Cox transformation was also tested. The Box-Cox procedure followed was

a. Find a Box-Cox transformation for a single independent variable fixing the other ones.

b. Find a Box-Cox transformation for the following independent variable, maintaining the previously transformed variable and keeping the other ones without transformation.

c. Repeat the process until all independent variables have been transformed (not including the interactions).

d. Finally, with the independents transformed variables, transform the response variable.

The model obtained with this procedure is like the following Equation

$$y^{\theta} = x_1^{\lambda_1} + x_2^{\lambda_2} + \ldots + x_n^{\lambda_n} \tag{7.1}$$

where

$$y^{\theta} = \begin{cases} \dfrac{y^{\theta} - 1}{\theta} & if \quad \theta \neq 0 \\ \ln(y) & if \quad \theta = 0 \end{cases} \quad \text{and} \quad x^{\lambda} = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & if \quad \lambda \neq 0 \\ \ln(x) & if \quad \lambda = 0 \end{cases} \tag{7.2}$$

The exponents for the independent transformation are

Table 39. Box-Cox independent lambdas for the independent variables

Variable	**Variable (before transformation)**	**Variable (after transformation)**	**Theta/ Lambda**
Log (Theoretical ARL)	Larl_theo	Larl_theo_bc	2.646475
Log (k)	Logk	Logk_bc	1.894422
Sample size or *n*	N_u	n_u_bc	1.318188
Interaction number of samples and sample size	mn1	mn1_bc	1.314768
Interaction samples size and number of samples	np1	np1_bc	1.314239
Square of number of samples	m2	m2_bc	1.313765

The transformation for the dependent variable (larl_sim or the logarithm of the simulated ARL) is 1.629404.

The regression was evaluated with the effect without being transformed. The results show an slightly improvement in the multiple determination coefficient but the root of the MSE deteriorates up to 3.34. The residuals show practically the same pattern of the initial model (see Figures E-1 and E-2) so the initial model can be used instead of the Box-Cox model.

Table 40. Regression model using Box-Cox transformations

```
      Source |       SS       df       MS              Number of obs =  197400
-------------+------------------------------           F(  7,197392) =17057.37
       Model |  830938.012     7   118705.43           Prob > F      =  0.0000
    Residual |  1373687.53197392  6.95918542           R-squared     =  0.3769
-------------+------------------------------           Adj R-squared =  0.3769
       Total |  2204625.54197399  11.1683724           Root MSE      =   2.638

------------------------------------------------------------------------------
 larl_sim_bc        Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------------------------------------------------------------------
_cons            4.586773   .0320252   143.22   0.000     4.524005    4.649542
larl_theo~c      .1429793   .0004514   316.74   0.000     .1420946    .1438641
logk_bc         -.2041017   .0015615  -130.71   0.000    -.2071621   -.2010412
n_u_bc           .1672123   .0046623    35.86   0.000     .1580742    .1763504
mn1_bc          -.0000258   8.32e-07   -30.98   0.000    -.0000274   -.0000241
np1_bc          -.0088093   .0004763   -18.49   0.000    -.0097428   -.0078757
m2_bc            2.60e-08   9.62e-10    27.00   0.000     2.41e-08    2.79e-08
effect
           1     .5902392   .0321951    18.33   0.000     .5271376    .6533408
           2    (dropped)
------------------------------------------------------------------------------
```

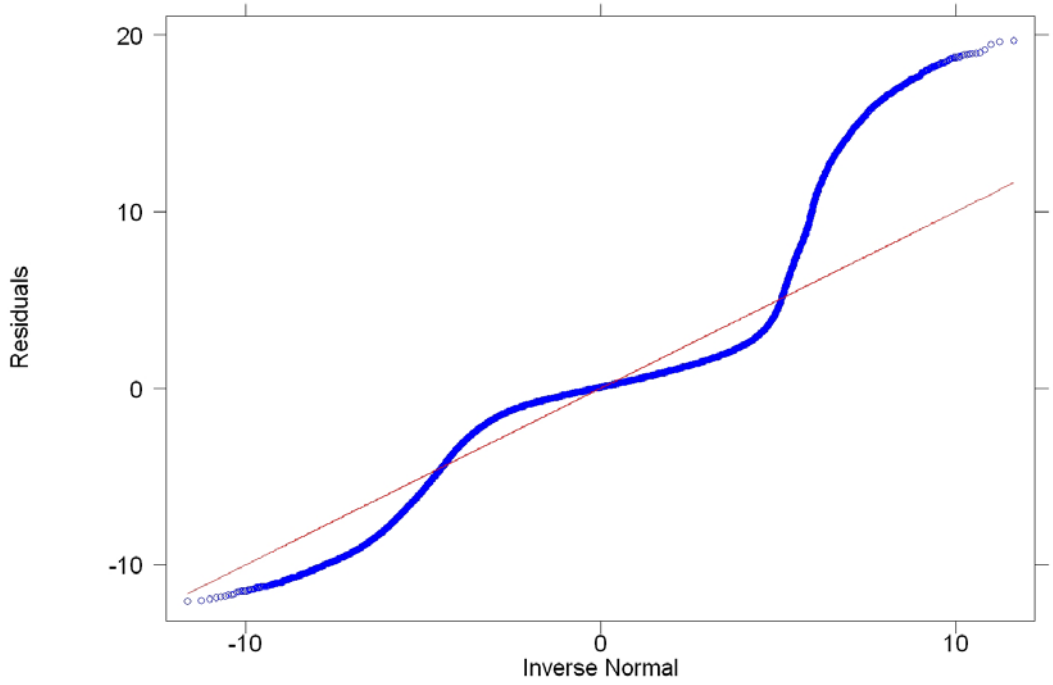
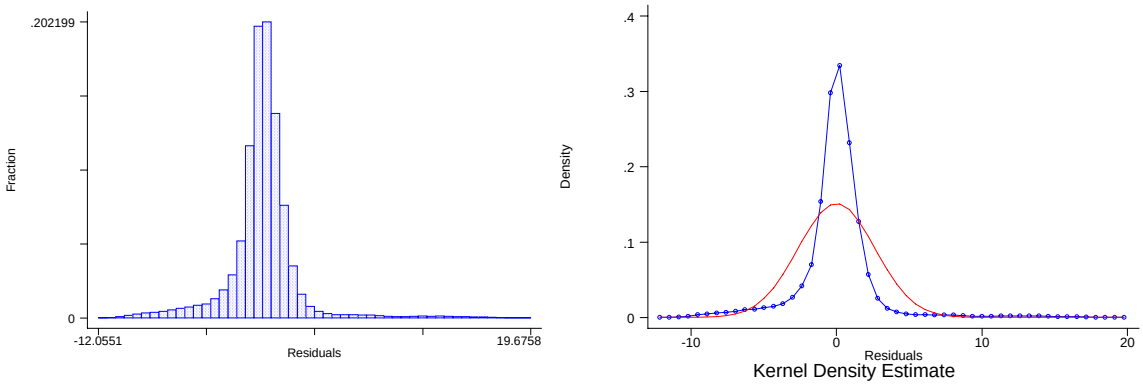


Figure F.1. Normal Q-Q plot for the residuals after Box-Cox transformation



(a) (b)

Figure F. 2. (a) Histogram (b) Kernel density for residuals in the Box-Cox model

Appendix G. Regression models

Table 41. Regression model when $\hat{k} < 10$

```
      Source |       SS       df       MS              Number of obs =   59010
-------------+------------------------------           F(  4, 59005) =41014.75
       Model |  24306.0699     4  6076.51748           Prob > F      =  0.0000
    Residual |  8741.85204 59005  .148154428           R-squared     =  0.7355
-------------+------------------------------           Adj R-squared =  0.7355
       Total |   33047.922 59009  .560048839           Root MSE      =  .38491

------------------------------------------------------------------------------
    larl_sim |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -.4592629   .0192468   -23.86   0.000    -.4969867   -.4215391
   larl_theo |   1.007269   .0024992   403.04   0.000     1.002371    1.012168
        logk |   .1819327   .0047722    38.12   0.000     .1725793    .1912862
         n_u |   .0096051   .0008759    10.97   0.000     .0078884    .0113218
          m2 |   2.99e-08   6.24e-09     4.79   0.000     1.77e-08    4.21e-08
------------------------------------------------------------------------------
```

Table 42. Regression model when $10 \leq \hat{k} < 20$

```
      Source |       SS       df       MS              Number of obs =   37170
-------------+------------------------------           F(  7, 37162) =16871.62
       Model |   15294.102     7  2184.87171           Prob > F      =  0.0000
    Residual |  4812.47278 37162  .129499833           R-squared     =  0.7607
-------------+------------------------------           Adj R-squared =  0.7606
       Total |  20106.5748 37169  .540950114           Root MSE      =  .35986

------------------------------------------------------------------------------
    larl_sim |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -.4328219   .0271763   -15.93   0.000    -.4860882   -.3795556
   larl_theo |   .9983577    .002944   339.12   0.000     .9925874    1.004128
        logk |    .119801    .006834    17.53   0.000     .1064063    .1331957
         n_u |   .1066148   .0027878    38.24   0.000     .1011507    .1120789
         mn1 |  -.0001366   3.68e-06   -37.07   0.000    -.0001438   -.0001294
         np1 |  -.0107158   .0004426   -24.21   0.000    -.0115834   -.0098483
          m2 |   5.17e-07   2.00e-08    25.78   0.000     4.77e-07    5.56e-07
      effect |
           1 |   .1140992   .0103353    11.04   0.000     .0938418    .1343566
           2 |  (dropped)
------------------------------------------------------------------------------
```

Table 43. Regression model when $20 \leq \hat{k} < 60$

```
      Source |       SS       df       MS              Number of obs =   50400
-------------+------------------------------           F(  7, 50392) =10375.69
       Model |  18829.3112     7  2689.90161           Prob > F      =  0.0000
    Residual |  13064.1395 50392  .259250268           R-squared     =  0.5904
-------------+------------------------------           Adj R-squared =  0.5903
       Total |  31893.4508 50399  .632819119           Root MSE      =  .50917

------------------------------------------------------------------------------
    larl_sim |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_cons        |  -.3269787   .0343999    -9.51   0.000    -.3944029   -.2595544
larl_theo    |   .9558468   .0035772   267.21   0.000     .9488355    .9628582
logk         |   .0793531   .0062001    12.80   0.000     .0672009    .0915054
n_u          |   .0673399   .0033923    19.85   0.000      .060691    .0739888
mn1          |  -.0000337   4.65e-06    -7.24   0.000    -.0000428   -.0000246
np1          |  -.0107668   .0005265   -20.45   0.000    -.0117988   -.0097349
m2           |   2.54e-07   2.43e-08    10.42   0.000     2.06e-07    3.01e-07
effect       |
           1 |   .1996949   .0110428    18.08   0.000     .1780509    .2213388
           2 |  (dropped)
------------------------------------------------------------------------------
```

Table 44. Regression model when $60 \leq \hat{k}$

```
      Source |       SS       df       MS              Number of obs =   50820
-------------+------------------------------           F(  6, 50813) = 1761.03
       Model |   24164.146     6  4027.35767           Prob > F      =  0.0000
    Residual |  116205.927 50813  2.28693301           R-squared     =  0.1721
-------------+------------------------------           Adj R-squared =  0.1720
       Total |  140370.073 50819  2.76215733           Root MSE      =  1.5123

------------------------------------------------------------------------------
    larl_sim |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_cons        |   2.913661   .0899396    32.40   0.000     2.737378    3.089944
larl_theo    |   .8278754   .0105805    78.25   0.000     .8071374    .8486134
logk         |  -.6466551   .0104033   -62.16   0.000    -.6670457   -.6262645
n_u          |   .1547571   .0073366    21.09   0.000     .1403774    .1691368
mn1          |  -.0002547   .0000129   -19.78   0.000      -.00028   -.0002295
m2           |   1.25e-06   7.30e-08    17.08   0.000     1.10e-06    1.39e-06
effect       |
           1 |   .6121385   .0410418    14.91   0.000     .5316961    .6925809
           2 |  (dropped)
------------------------------------------------------------------------------
```

Appendix H. Regression models with true matrices only

Table 45. Regression model when $\hat{k}$ < 10 with true matrices only

```
      Source |       SS       df       MS              Number of obs =    2100
-------------+------------------------------           F(  8,  2091) = 3653.66
       Model |  878.329207     8  109.791151           Prob > F      =  0.0000
    Residual |  62.8338262  2091  .030049654           R-squared     =  0.9332
-------------+------------------------------           Adj R-squared =  0.9330
       Total |  941.163034  2099  .448386391           Root MSE      =  .17335

------------------------------------------------------------------------------
    larl_sim |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -1.097045   .085641   -12.81   0.000    -1.264995   -.9290943
   larl_theo |   1.012063  .0059664   169.63   0.000     1.000363    1.023764
        logk |    .077417  .0106969     7.24   0.000     .0564393    .0983947
         m_u |   .0004837  .0001082     4.47   0.000     .0002715    .0006959
         p_u |   .3705704  .0283494    13.07   0.000     .3149744    .4261664
         mp1 |   .0000939  .0000207     4.53   0.000     .0000533    .0001344
          m2 |  -7.80e-07  6.45e-08   -12.10   0.000    -9.06e-07   -6.53e-07
          n2 |   .0014918  .0001737     8.59   0.000     .0011512    .0018324
          p2 |  -.0527516  .0032686   -16.14   0.000    -.0591617   -.0463415
------------------------------------------------------------------------------
```

Table 46. Regression model when $10 \leq \hat{k} < 20$ with true matrices only

```
      Source |       SS       df       MS              Number of obs =    1260
-------------+------------------------------           F(  7,  1252) = 1643.41
       Model |  685.764531     7  97.9663616           Prob > F      =  0.0000
    Residual |  74.6337542  1252  .059611625           R-squared     =  0.9018
-------------+------------------------------           Adj R-squared =  0.9013
       Total |  760.398286  1259  .603970044           Root MSE      =  .24415

------------------------------------------------------------------------------
    larl_sim |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   1.290768   .108404    11.91   0.000     1.078094    1.503441
   larl_theo |   1.025502  .0108487    94.53   0.000     1.004218    1.046786
        logk |  -.3563605  .0204785   -17.40   0.000    -.3965366   -.3161845
         m_u |  -.0046875  .0001037   -45.21   0.000    -.0048909   -.0044841
         p_u |   .3372907  .0229407    14.70   0.000     .2922842    .3822972
         np1 |  -.0463295  .0034163   -13.56   0.000    -.0530318   -.0396272
          m2 |   3.67e-06  9.42e-08    39.02   0.000     3.49e-06    3.86e-06
          n2 |   .0145395  .0011331    12.83   0.000     .0123166    .0167624
          p2 |  (dropped)
------------------------------------------------------------------------------
```

Table 47. Regression model when $20 \leq \hat{k} < 60$ with true matrices only

```
      Source |       SS       df       MS              Number of obs =    2240
-------------+------------------------------           F(  8,  2231) = 2214.89
       Model |  946.808786     8  118.351098           Prob > F      =  0.0000
    Residual |  119.211724  2231  .053434211           R-squared     =  0.8882
-------------+------------------------------           Adj R-squared =  0.8878
       Total |  1066.02051  2239  .476114565           Root MSE      =  .23116

------------------------------------------------------------------------------
    larl_sim |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -2.367233   .1669656   -14.18   0.000    -2.694657   -2.039809
   larl_theo |   1.006397   .0077034   130.64   0.000     .9912904    1.021504
         m_u |   .0019492   .0001964     9.92   0.000      .001564    .0023344
         p_u |   .6217728   .0357629    17.39   0.000     .5516408    .6919048
         n_u |   .2199887   .0247782     8.88   0.000      .171398    .2685794
         mn1 |  -.0001148   .0000195    -5.88   0.000     -.000153   -.0000765
         mp1 |  -.0003408    .000039    -8.73   0.000    -.0004173   -.0002643
         np1 |  -.0701145   .0048841   -14.36   0.000    -.0796924   -.0605367
          n2 |   .0062013     .00094     6.60   0.000      .004358    .0080446
------------------------------------------------------------------------------
```

Table 48. Regression model when $60 \leq \hat{k}$ with true matrices only

```
      Source |       SS       df       MS              Number of obs =    1400
-------------+------------------------------           F(  9,  1390) = 2232.96
       Model |  583.161644     9  64.7957382           Prob > F      =  0.0000
    Residual |  40.3348048  1390  .029017845           R-squared     =  0.9353
-------------+------------------------------           Adj R-squared =  0.9349
       Total |  623.496449  1399  .445672944           Root MSE      =  .17035

------------------------------------------------------------------------------
    larl_sim |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .8641143   .1377138     6.27   0.000     .5939649    1.134264
   larl_theo |   .9942502   .0071807   138.46   0.000      .980164    1.008336
         m_u |   .0019694   .0001123    17.54   0.000     .0017492    .0021896
         p_u |  -.5156502   .0393393   -13.11   0.000    -.5928211   -.4384793
         n_u |  -.1816115   .0203823    -8.91   0.000     -.221595   -.1416281
         mn1 |  -.0001848   .0000144   -12.85   0.000     -.000213   -.0001566
         np1 |   .0263233   .0035992     7.31   0.000     .0192628    .0333838
          m2 |  -6.56e-07   6.51e-08   -10.07   0.000    -7.83e-07   -5.28e-07
          n2 |     .01461   .0010152    14.39   0.000     .0126186    .0166015
          p2 |   .0510166   .0040607    12.56   0.000     .0430508    .0589823
------------------------------------------------------------------------------
```

Appendix I. Programs

Appendix I1. Main Program

```
clear
clc
%-------------------------------------------------------------------------
% The input data comes from the file: info_p2.txt
% For this example, the first three columns are p, m and n, the fourth
% column is the %theoretical k of the correlation matrix used
% Mu0 = columns 5-6
% Sigma0 = columns 7-10
load info_p2.txt
final =[];
for i=1:size(info_p2,1) %goes thruogh all exp conditions to assign
parameters
    k_theo = info_p2(i,5);
    p = info_p2(i,6);
    m = info_p2(i,7);
    n = info_p2(i,8);
    Mu0 = info_p2(i,9:10)';
    Sigma0 = reshape(info_p2(i,11:size(info_p2,2)),p,p);

    indicadores =info_p2(i,1:8);

    con=[]; nor=[];
    ISigma0 = inv(Sigma0); %inverse of the current Sigma0
    [p k1] = size(Mu0);

    cov_size = size(Sigma0);
%Both, MuHat and SigmaHat are non standarized values

%Creating n samples of random data with parameters defined above (p=3)
[MuHat,SigmaHat,sample,vec_medias]=multiv_parameter(Mu0,Sigma0,m,n);
    MuHat_NS = MuHat;
    SigmaHat_NS = SigmaHat;
    ISigmaHat_NS = inv(SigmaHat_NS);

%Standardization
    [MuHat_St,Z,standard] = Mu_stand(MuHat,SigmaHat,m,vec_medias);
    [SigmaHat_St] = Sigma_stand(SigmaHat);
    MuHat_St;
    SigmaHat_St;
    ISigmaHat_St = inv(SigmaHat_St);

%Deviation from true correlation matrix
    [cambios] = variaciones(SigmaHat_St,Sigma0);
```

```
    St_8=reshape(cambios,cov_size(1),cov_size(1));
    IS_St_8=inv(St_8);
    S = reshape(SigmaHat_St,1,cov_size(1)*cov_size(2));

%Correlation matrices estimation
    [combinaciones] = CovMatr_combs(SigmaHat_St,cambios,S);
%Full factorial for p<=3, if p>=4 the function CovMatr_combs is
%replaced by the orthogonal function

    mat = [S;cambios;combinaciones]; %mat = all considered matrices

%god = true correlation matrix
    god= reshape(Sigma0,1,cov_size(1)*cov_size(2));
    mat_cov=[mat;god]; %first row is the estimated mat/last true mat

%estimating condition number and determinant of all estimated matrices
    [matrices,condicion] = condition_determinantes(mat_cov,Sigma0);
    matrices; %this variable contains all non singular matrices
    condicion; %includes all matrices with condition number/determinant

%estimated the norm of the matrices and the deviation from the true
%matrix
    diff_norm = normas(condicion,p);
    [pa1 pa2] = size(condicion);
    [ca1 ca2] = size(matrices);

%the following loop calulates matrices' inverses and save it in
%matrices2
    matrices2=[];
    for i=1:ca1
        tempo2 = inv(reshape(matrices(i,:),p,p));
        tempo3 = reshape(tempo2,1,p*p);
        matrices2 = [matrices2;tempo3];
    end


%Phase II of T2 Control Charts and ARL estimation
%This procedure evaluates the ARL for tmax replicates
    tmax = 50;
    [f1 c1] = size(mat_cov);
%The ARL will be evaluated using standarized values
%ARL varying SigmaHat
    state1 = 0;
    V1=0;
    results2 =[];
    for j=1:10 %10 replicates
        ARL= []; Run1 =[]; Runs=[]; Runs2=[];
        state1 = state1 + V1;
        state=0+state1; V =0; results=[];
        for d=200:200:1400
```

```
            state =state+V;
            rand('state', state);
            randn('state', state);
            UCL2 = (p*(m+1)*(n-1))* finv(1-(1/d),p,m*n-m-p+1)/(m*n-m-
p+1);
            Run1=[];
            RunL=1:f1; flags=1:f1;
            for t =1:tmax
                flag = 0;
                RunL=zeros(1,f1);
                flags = zeros (1,f1);
                while flag ==0
                    sample1 = mvnrnd(Mu0,Sigma0,n);
                    promedio1 = mean(sample1);
                    Z1 = (promedio1 - MuHat_NS)./standard;
                    for i = 1:f1
                        if flags(i) ==0
                            RunL(i) = RunL(i) + 1;
                            T = n*Z1*reshape(matrices2(i,:),p,p)*Z1';
                            if T > UCL2
                                flags(i) = 1;
                            end
                        else
                        RunL(i) = RunL(i);
                        end
                    end
                    if (flags==1)
                        flag = 1;
                    end
                end
                Run1 = [Run1;RunL];
            end
            Runs2 = [Runs2;Run1'];
            Runs=[Runs,Run1'];
            ARLsim= sum(Run1)/tmax;
            ARL=[ARL;ARLsim];
            V=5;
            for y=1:size(ARLsim,1)
                results = [results;(zeros(1,size(ARL,2))+d)' ARLsim'];
            end
        V1=3;
        end
        results = [results Runs2];
        results2 = [results2; results];

        for zz=1:size(results2,1)
            con = [con; condicion];
            nor = [nor; diff_norm];
        end

    end % this is the end of the first for
```

```
    for z=1:size(results2,1)
        final = [final ; indicadores con(z,pa2-1) nor(z,:)
results2(z,:)] ;
    end
    state1 = 2;

end
```

Appendix I2. Parameters and Random number generation function

```
function
[MuHat,SigmaHat,sample,vec_medias]=multiv_parameter(Mu0,Sigma0,m,n)
sampl = []; vec_med=[]; covarianzas =[];
cov_size=size(Sigma0);
%Multivarate Random Sample Generation
%Sample contains the values of the n multivariate random number
generated
%every 3 columns
for i=1:m
    B=mvnrnd(Mu0,Sigma0,n);
    sampl = [sampl;B];
    vec_med = [vec_med;mean(B)];
    paso=cov(B);
    covarianzas=[covarianzas;paso(1:1:end)];
end
MuHat=mean(vec_med);
SigmaHat_vec=mean(covarianzas);
SigmaHat=reshape(SigmaHat_vec,cov_size(1),cov_size(1));
sample = sampl;
vec_medias= vec_med;
```

Appendix I3. Standardization function

```
function
[MuHat_St,SigmaHat_St,Z]=standarization(MuHat,SigmaHat,m,vec_medias)
temp_Mu = [];
for i =1:m
    temp_Mu = [temp_Mu; MuHat];
end
temp_St = [];
for i =1:m
    temp_St = [temp_St; sqrt(diag(SigmaHat)')];
end
Zi = (vec_medias - temp_Mu)./temp_St;
%Zs the standarized values of sample in the same column order.
%The parameters estimated are
Mu = mean(Zi);
[row col] = size(SigmaHat);
S = []; S1 =[]
```

```
for i = 1:row
    for j = 1:col
        s = SigmaHat(i,j)/sqrt(SigmaHat(i,i)*SigmaHat(j,j));
        S =[S s];
    end
    S = [S S1];
end
Sigma = reshape(S,row,row);
MuHat_St = Mu;
SigmaHat_St = Sigma;
Z =Zi;
```

Appendix I4. Deviations computation from true correlations function

```
function [cambios] = variaciones(SigmaHat,Sigma0)
[rows cols] = size (SigmaHat);
[row col] = size(SigmaHat);
S = []; S1 =[]
for i = 1:row
    for j = 1:col
        if ~(i==j)
            s = tanh(2*atanh(Sigma0(i,j))-atanh(SigmaHat(i,j)));
            S =[S s];
        else
            S = [S SigmaHat(i,j)];
        end
    end
    S = [S S1];
end
cambios = S;
```

Appendix I5. Correlation matrices function

```
function [combinaciones] = Cov_Matrices_comb(SigmaHat_St,cambios,S)
[r w] = size(SigmaHat_St);
sigma = reshape(SigmaHat_St,1,r*w);
vars = reshape(triu(SigmaHat_St,1),1,r*w);
a = find(~(vars==0));
[row col] = size(a);
[r1 w1] = size(vars);
V=[];
for i = 1:(col-1)
    c = combntns(a,i);
    [r2 w2] = size(c);
    for j = 1:r2
        temp = 0;
        temp = reshape(eye(r),1,r*w);
        for k=1:w2
```

```
            temp(c(j,k)) = cambios(c(j,k));
        end
        d = reshape(temp,r,w);
        e = d + d'-eye(r);
        temp1 = reshape(e,1,r*w);
        V = [V; temp1];
    end
end
[r3 w3] = size(V);
temp2 = [];
for i=1:r3
    temp2 = [temp2;sigma];
end
posit = find(V==0);
v = V;
v(posit) = temp2(posit);
combinaciones = v;
```

Appendix I6. Orthogonal designs function

```
function [combinaciones] = Ortogonal(SigmaHat,Sigma0,cambios,p)

[k1 l1] = size(Sigma0);
if p==5
    load disen_frac_p5.txt;
    datos = disen_frac_p5;
end
if p==6
    load disen_frac_p6.txt;
    datos = disen_frac_p6;
end

[k2 l2] = size(datos);
original = reshape(Sigma0,1,k1*l1);
estimado = reshape(SigmaHat,1,k1*l1);

m_dis = ones(k1,l1) - eye(k1,l1);
m_dis = reshape(triu(m_dis),1,l1*k1);

posiciones = find(~(m_dis==0));
m_diseno=[];
for i = 1:k2
    m_dis(posiciones) = datos(i,:);
    m_diseno =[m_diseno;m_dis];
end

matri=[];
for i = 1:k2
```

```
    tempo = reshape(m_diseno(i,:),k1,l1);
    tempo2 = tempo + tempo';
    matri = [matri; reshape(tempo2,1,k1*l1)];
end

pos_up = find((estimado>original));
pos_lo = find((estimado<original));

hola=matri;
for i = 1:k2
    for j=1:size(pos_up,2)
        if hola(i,pos_up(j))==1
            hola(i,pos_up(j)) = estimado(1,pos_up(j));
        else
            hola(i,pos_up(j)) = cambios(1,pos_up(j));
        end
    end
    for k=1:size(pos_lo,2)
        if hola(i,pos_lo(k))==1
            hola(i,pos_lo(k)) = cambios(1,pos_lo(k));
        else
            hola(i,pos_lo(k)) = estimado(1,pos_lo(k));
        end
    end

end

diseno =[];
for i = 1:k2
    temp = reshape(hola(i,:),k1,l1);
    temp2 = temp + diag(ones(1,k1));
    diseno = [diseno; reshape(temp2,1,k1*l1)];
end
combinaciones = diseno;
```

Appendix I7. Determinant and condition number function

```
%mat_cov are all the combinations of matrices
%This function uses mat_cov to evaluate the condition number and the
%determinant.
%matrices = The answer of this function is a mtrix that contains only
the
%nonsingular combinations
%condition has the matrices in one line and the condition number and
the
%determinant of this matrix
function [matrices,condition] = condition_determinantes(mat_cov,Sigma0)
[fil col] = size(Sigma0);
[filas columnas] = size(mat_cov);
matrices =[];
```

```
condition =[];
for i=1:filas
    A = reshape (mat_cov(i,:),fil,col);
    k = cond(A,2); %k = condition number
    t = det(A);    %t = determinant
    if ~(t==0) %only matrices with det<>0 --> matrices
        matrices = [matrices ;mat_cov(i,:)];
    end
   condition = [condition ;mat_cov(i,:) k t]; %information of all
matrices
end
```

Appendix I8. Deviation from true matrix and norm function

```
% Calculates the difference between estimated matrix and the
% real matrix (god's matrix). In this case only considers the upper
%diag matrix
function [diffnorms] = normas(condicion,p)
diffnorms=[];
dios = reshape(condicion(size(condicion,1),1:size(condicion,2)-2),p,p);
for i=1:size(condicion,1)-1
    mattemp = reshape(condicion(i,1:size(condicion,2)-2),p,p);
    %norma 2 is the second norm of the difference matrix
    norma2 = norm(mattemp - dios);
    triansup = triu(mattemp)- triu(dios);
    %sumaup is the absolute value of the sum of the correlation in the
    %upper triag
    sumaup = abs(sum(sum(triansup)));
    %sumatot is the absolute value of the sum of all differences
between
    %correlations
    sumatot = abs(sum(sum(mattemp-dios)));
    %proup is the average of the absolute value over the correlation
used in
    %the difference
    proup = sumaup/(size(find(triansup),1));
    %protot is the averaga of the absolute value over the total
    %correlations
    protot = sumatot/(p*p);
    diffnorms = [diffnorms; sumaup sumatot proup protot norma2];
end

diffnorms = [diffnorms;zeros(1,size(diffnorms,2))];
```