Publicación Electrónica **ISSN 2718- 6636**

UNIVERSIDAD NACIONAL DE ROSARIO
FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA
SECRETARIA DE CIENCIA Y TECNOLOGIA E INSTITUTOS DE INVESTIGACIONES

Resumen Ampliado

Jornadas Anuales
"Investigaciones en la Facultad"
Ciencias Económicas y Estadística

Quaglino, Marta
Pagura, José
Fernández, Julia Inés
Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística.

GRÁFICOS DE CONTROL MULTIVARIADO DE PROCESOS: DESEMPEÑO EN FASE II EN PRESENCIA DE VALORES FALTANTES[1]

Resumen

La aplicación de los gráficos T^2 de Hotelling y Squared Prediction Error (SPE) construidos a partir de un subconjunto de las Componentes Principales (CP) para llevar a cabo el control estadístico multivariado de procesos es de alta efectividad cuando se considera un conjunto de objetivos definidos con indicadores continuos. Sin embargo, su uso se ve limitado por un problema frecuente en estas situaciones de control multivariado: la ocurrencia de observaciones que presentan datos faltantes. Para resolver este inconveniente, se ha propuesto el uso de métodos de imputación de scores de las Componentes Principales cuando hay observaciones faltantes en los datos originales, denominados *Known Data Regression* (KDR) y *Trimmed Score Regression* (TSR). Los métodos KDR y TSR presentan Error Cuadrático Medio bajo, pero no habían sido realizados estudios acerca de su posible impacto sobre las propiedades de los gráficos de control T^2 y SPE en Fase II, es decir, en la etapa específica de control de los procesos. Para evaluar la performance de los gráficos se realizaron estudios de simulación en distintos escenarios que consideran diferentes situaciones de aplicación. Se estimaron las curvas de *Average Run Length* (ARL) al utilizar los métodos de imputación para valores faltantes y se las compararon con las curvas de ARL correspondientes al conjunto de los datos completos. Se analizaron las propiedades distribucionales de las estadísticas T^2 y SPE en situaciones de control estadístico. También se evaluaron las discrepancias entre las estructuras de covariancia estimadas de los scores y de los residuos en comparación con aquellas correspondientes a los datos completos. Los resultados indican que hay situaciones en las cuales, al utilizar KDR y TSR para imputar scores frente a observaciones faltantes, el desempeño de los gráficos de control es deficiente en comparación con lo esperado bajo datos completos. En consecuencia, cuando se usan estos métodos de estimación de scores, dichos gráficos deben ser interpretados con precaución para lograr la eficiencia deseada en el control.

Palabras clave: Control Multivariado de Procesos, Datos Faltantes, Análisis de Componentes Principales

Abstract

Hotelling's T^2 control chart and the Squared Prediction Error (SPE) chart of selected Principal Components (PCs) are highly effective when monitoring several quantitative quality variables. However, their application is often limited by a frequent issue: the occurrence of observations with missing data. In order to solve this problem*, Known Data Regression* (KDR) and *Trimmed*

[1] Trabajo elaborado en el marco del Proyecto ECO180, titulado: "Modelo Estadísticos para el Diseño y Control de Proceso", dirigido por Marta Quaglino.

Score Regression (TSR) methods for the estimation of the scores' PCs have been proposed, and they present low Mean Squared Error. However, it had not been assessed how the use of KDR and TSR affect the performance of T^2 and SPE control charts in Phase II, when the process' state of statistical control is monitored. In order to do so, simulation studies in different scenarios that contemplate diverse situations were performed. Average Run Lengths (ARLs) of the charts were estimated when KDR and TSR are used for the imputation of observations with missing data, and compared to ARL curves corresponding to complete data. Distributional properties of the charts were analyzed in situations where the process is under control. The alterations in covariance structures of estimated scores and residuals were compared to the same structures with complete data. Results indicate that the performance of these control charts when KDR and TSR are used for the imputation of scores of observations with missing values is defective in comparison to the performance of the charts using complete data. Therefore, when these methods of score estimation are used, caution is recommended in the interpretation of the charts in order to ensure the desired effectiveness.

Keywords: Multivariate Process Control, Missing Data, Principal Components Analysis

Objetivos

El objetivo principal del trabajo es evaluar el desempeño de los gráficos de control multivariado de procesos T^2 de Hotelling construido con las Componentes Principales (CP) y el gráfico del Squared Prediction Error (SPE) al utilizar los métodos *Known Data Regression* (KDR) y *Trimmed Score Regression* (TSR) para estimar scores en base a valores faltantes.

Además se pretende analizar si las estadísticas calculadas con valores imputados mantienen las propiedades distribucionales a partir de las cuales se definen los límites de control de los gráficos, y estudiar la estructura de covariancia de los scores y de los residuos frente a datos faltantes, a fin de detectar si ocurren desviaciones de las correlaciones teóricas esperadas.

Metodología y análisis de datos considerados en la investigación

Se realizaron estudios de simulación en distintos escenarios que consideran diferentes situaciones de aplicación de los gráficos de control. Para cumplir con el objetivo principal, se estiman las curvas de Average Run Length (ARL) al utilizar métodos de estimación de scores frente a valores faltantes y se las compara con las curvas de ARL correspondientes al conjunto de los datos completos. Esto se realizó tanto en situaciones en que el proceso está bajo control estadístico, como en escenarios en los que el proceso se encuentra fuera de control.

En relación a los objetivos secundarios del estudio, se analizaron las propiedades distribuciones de las estadísticas T^2 y SPE en situaciones de control estadístico para evaluar la adecuación de los límites usuales. Se realizaron pruebas de bondad de ajuste de las estadísticas a las distribuciones utilizadas para definir los límites de control de cada gráfico, y se calcularon las proporciones de tests en los que se rechaza la hipótesis nula. Si los límites de los gráficos están definidos correctamente, se espera que la proporción de pruebas rechazadas sea aproximadamente igual al nivel de significación utilizado en su identificación.

También se evaluaron las discrepancias entre las estructuras de covariancia estimadas de los scores y de los residuos en comparación con aquellas correspondientes a los datos completos. Se midieron la distancia euclídea entre las diagonales de las matrices de covariancia para comparar las diferencias en variancia, y para contrastar las covariancias se calculó la suma de las diferencias entre los elementos del triángulo superior de las matrices.

Los escenarios para los estudios de simulación se construyeron considerando cuatro estructuras de covariancia, denominadas I, II, III y IV, dos patrones de generación de datos faltantes, llamados A y B, cuatro porcentajes de valores faltantes: 5, 10, 15 y 20%. Con cada estructura de covariancias se construye un modelo de componentes principales.

Se consideraron 8 variables de calidad, con las que se construyen los modelos de componentes principales:

$$\mathbf{T}_{1:\mathrm{A}} = \mathbf{X}\mathbf{P}_{1:\mathrm{A}} + \mathbf{E}$$

donde $\mathbf{X}$ es la matriz de datos centrada y estandarizada, $\mathbf{P}_{1:\mathrm{A}}$ es la matriz de cargas de las A primeras CP, $\mathbf{E}$ es la matriz de residuales, y $\mathbf{T}_{1:\mathrm{A}}$ son los scores de las unidades en las A primeras CP. En cada modelo, se retuvieron las A primeras CP que conservan al menos el 80% de la variabilidad total. Las estructuras presentan niveles decrecientes de correlación entre las variables, desde la que presenta mayor proporción de correlaciones altas, I, a la que presenta menor proporción, IV.

El patrón A de generación de datos faltantes provoca valores faltantes completamente al azar en todas las variables de calidad, mientras que el B sólo produce datos faltantes en las variables que presentan una correlación con la primera CP de 0,5 o mayor en valor absoluto.

La estadística para el gráfico T^2 de Hotelling es: $T^2 = \boldsymbol{\tau}_{1:\mathrm{A}}{}'\boldsymbol{\Theta}_{1:\mathrm{A}}\boldsymbol{\tau}_{1:\mathrm{A}}$, siendo $\boldsymbol{\tau}_{1:\mathrm{A}} = \mathbf{P}'_{1:\mathrm{A}}\mathbf{z}$ y $\boldsymbol{\Theta}_{1:\mathrm{A}}$ la matriz de variancias y covariancias de los scores del modelo. El límite de control del gráfico fue definido como el percentil del $(1-\alpha)\times 100\%$ de la distribución χ^2 con A grados de libertad.

Para el gráfico SPE, la estadística se calcula como $SPE = \boldsymbol{e}'\boldsymbol{e}$, y el límite de control se calculó utilizando la aproximación de Box como el percentil del $(1-\alpha)\times 100\%$ de una distribución Gamma con parámetro de forma $\sum_{j=A+1}^{p}\lambda_j^2 / \sum_{j=A+1}^{p}\lambda_j$ y parámetro de escala $\left(\sum_{j=A+1}^{p}\lambda_j\right)^2 / \sum_{j=A+1}^{p}\lambda_j^2$. La estrategia de control se diseñó considerando un nivel de significación de $\alpha = 0{,}5\%$, de forma tal que el ARL esperado para los gráficos es de 200.

Para estimar los vectores de scores de observaciones que presentan valores faltantes, se emplearon los métodos *Known Data Regression* (KDR) y *Trimmed Score Regression* (TSR), los cuales producen estimadores con bajo Error Cuadrático Medio y que no afectan la ortogonalidad de las componentes principales. El estimador KDR de los scores se obtiene como: $\hat{\boldsymbol{\tau}}_{1:\mathrm{A}} = \boldsymbol{\Theta}_{1:\mathrm{A}}\mathbf{P}^*_{1:\mathrm{A}}(\mathbf{P}^*\boldsymbol{\Theta}\mathbf{P}^{*\prime})^{-1}\mathbf{z}^*$, y el del método TSR es: $\hat{\boldsymbol{\tau}}_{1:\mathrm{A}} = \boldsymbol{\Theta}_{1:\mathrm{A}}\mathbf{P}^{*\prime}_{1:\mathrm{A}}\mathbf{P}^*_{1:\mathrm{A}}(\mathbf{P}^{*\prime}_{1:\mathrm{A}}\mathbf{P}^*\boldsymbol{\Theta}\mathbf{P}^{*\prime}\mathbf{P}^*_{1:\mathrm{A}})^{-1}\mathbf{P}^{*\prime}_{1:\mathrm{A}}\mathbf{z}^*$, donde el símbolo "*" indica las variables que presentan datos observados. Los autores de estos métodos de imputación sugieren reemplazar el vector de scores por el vector estimado directamente en el cálculo de las estadísticas T^2 y SPE, y utilizar los límites de control usuales.

Bajo control, se generaron observaciones normales multivariadas con vector de medias nulo y usando como matriz de correlación las estructuras I, II, III y IV. Fuera de control, se alteró el vector de medias con incrementos constantes en todos sus elementos de $\pm 0.01,\ 0.03,\ 0.05,\ 0.07,\ 0.1,\ 0.15,\ 0.2,\ 0.3,\ 0.5,\ 0.75,\ 1$.

Problemas planteados, principales hipótesis y resultados

Al realizar la comparación de las curvas de ARL estimadas en presencia de observaciones incompletas con las curvas de ARL correspondientes a los datos completos que se obtuvieron en los estudios de simulación realizados, se encontró que, en general, el desempeño de los gráficos T^2 de Hotelling y SPE no se ven severamente afectado con ninguno de los dos métodos de imputación cuando el porcentaje de valores faltantes es del 5% o inferior. Cuando

los porcentajes de datos faltantes son superiores, las curvas estimadas indican que hay pérdida de eficiencia de la estrategia de control, ya que los valores de ARL estimados son mayores a los ARL que se obtienen con los datos completos. Los comportamientos observados difieren para las cuatro estructuras de covariancia estudiadas y para los dos patrones de generación de pérdidas. En términos prácticos, esto significa que cuando el porcentaje de observaciones incompletas es superior al 5% se requiere mayor cantidad de tiempo para detectar una salida de control del proceso, con consecuencias negativas para la calidad de los productos. Si bien el desempeño de los gráficos no es uniforme para todas las estructuras de covariancia consideradas, en todos los casos se encontró que la eficiencia de los gráficos de control T^2 y SPE se ve más afectada cuando las pérdidas ocurren al azar en todas las variables de calidad medidas que cuando sólo ocurren en las variables más correlacionadas con la primera CP.

Respecto de las posibles causas de las alteraciones observadas en las curvas de ARL estimadas en presencia de datos faltantes para los gráficos de control T^2 y SPE, se analizaron los resultados de las pruebas de bondad de ajuste realizadas cuando el proceso está bajo control. Para el gráfico T^2 se encontró que en la mayoría de los escenarios considerados en el estudio de simulación el límite control establecido no fue adecuado. El límite de control del gráfico SPE no fue apropiado en ninguna de las situaciones estudiadas.

Los resultados del estudio de simulación muestran que las estructuras de covariancia de los scores y de los residuos se alteran cuando se utilizan los métodos de imputación KDR y TSR en comparación con las estructuras de covariancia con datos completos. En relación a los scores, la medida más afectada es la variancia, mientras que las covariancias presentan un cambio de menor magnitud. Con respecto al SPE también se observan modificaciones en las variancias.

Descripción de la novedad y relevancia del trabajo

La evaluación de los métodos de imputación KDR y TSR ha sido realizada en su calidad de estimadores, evaluando sus errores cuadráticos medios. Sin embargo, no se había realizado previamente un estudio del impacto de los métodos de estimación sobre las propiedades de los gráficos de control T^2 de Hotelling y SPE. Conocer las alteraciones que se producen en los gráficos ayuda a prevenir situaciones de retraso en la detección de la salida de control de un proceso, evitando la producción de unidades de menor calidad o que deban ser desechadas.

REFERENCIAS BIBLIOGRÁFICAS

Arteaga, F., & Ferrer, A. (2002). Dealing with missing data in MSPC: several methods, different interpretations, some examples. Journal of Chemometrics: A Journal of the Chemometrics Society, 16(8-10), 408-418.

Box, G. E. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. The annals of mathematical statistics, 25(2), 290-302.

Fernández, J. I., Pagura, J. A., & Quaglino, M. B. (2020). Assessment of the effect of imputation of missing values on the performance of Phase II multivariate control charts. Quality and Reliability Engineering International.

Jackson, J. E., & Mudholkar, G. S. (1979). Control procedures for residuals associated with principal component analysis. Technometrics, 21(3), 341-349.

MacGregor, J. F., & Kourti, T. (1995). Statistical process control of multivariate processes. Control engineering practice, 3(3), 403-414.

Nelson, P. R., Taylor, P. A., & MacGregor, J. F. (1996). Missing data methods in PCA and PLS: Score calculations with incomplete observations. Chemometrics and intelligent laboratory systems, 35(1), 45-65.

Agradecimientos

Consejo Nacional de Investigaciones Científicas y Técnicas