Propiedades probabilísticas del gráfico T^2 con componentes principales frente a datos faltantes

J. I. Fernández[1,2] J. A. Pagura[1] M. B. Quaglino[1]

[1]Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística
Facultad de Ciencias Económicas y Estadística
Universidad Nacional de Rosario

[2]CONICET

XV Congreso Dr. Antonio Monteiro, Junio 2019

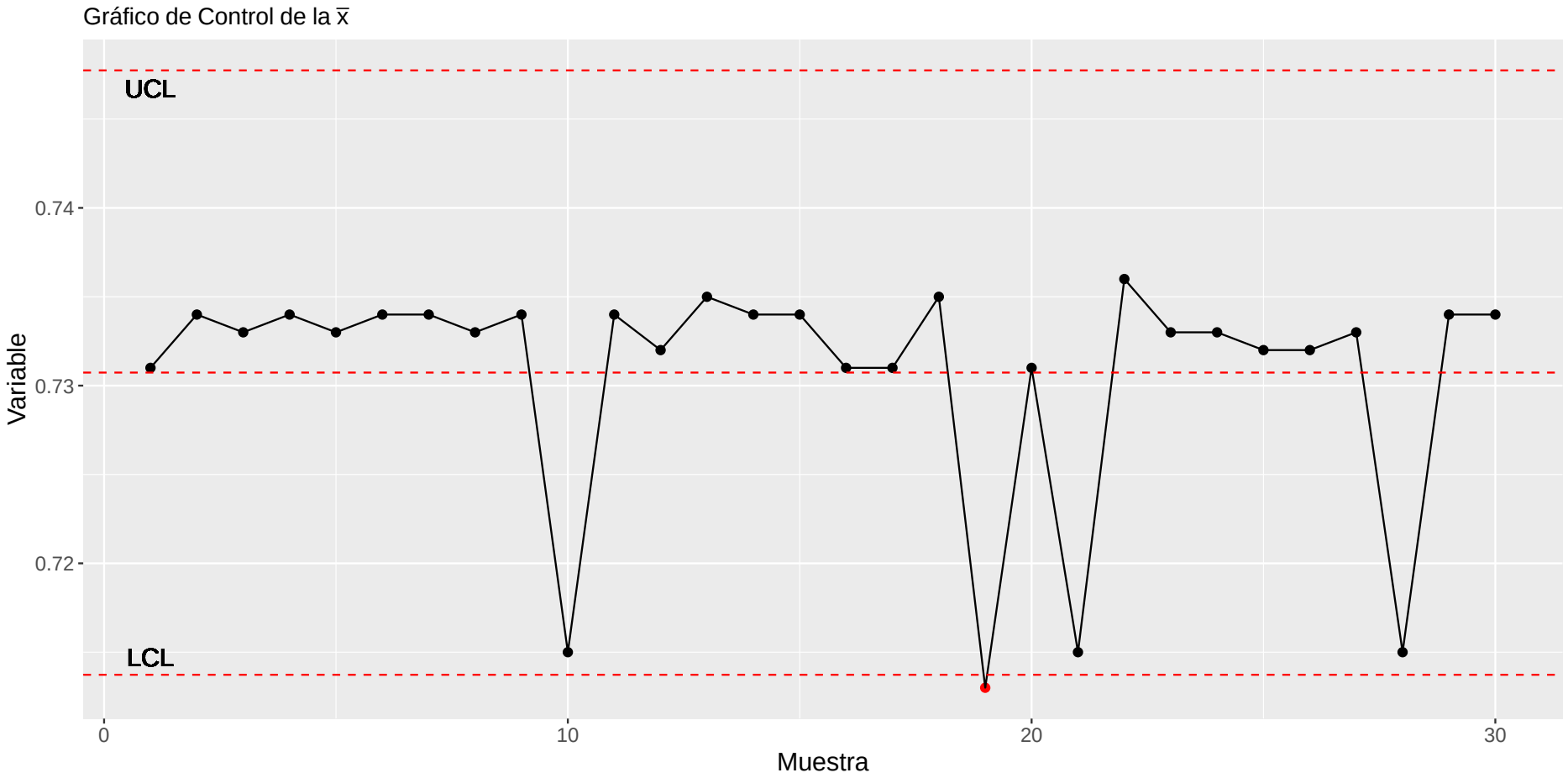
Tabla de Contenidos

Tabla de Contenidos

Control Estadístico de Procesos (SPC)

- Una de las estrategias más difundidas en el contexto de programas de mantenimiento y mejora de calidad tanto para procesos industriales como de servicios.
- Su objetivo es monitorear si el estado de un proceso a lo largo del tiempo se mantiene bajo control estadístico.
- Un ejemplo son los gráficos de control, como el gráfico de Shewhart para la media, el cual tiene límites bilaterales.

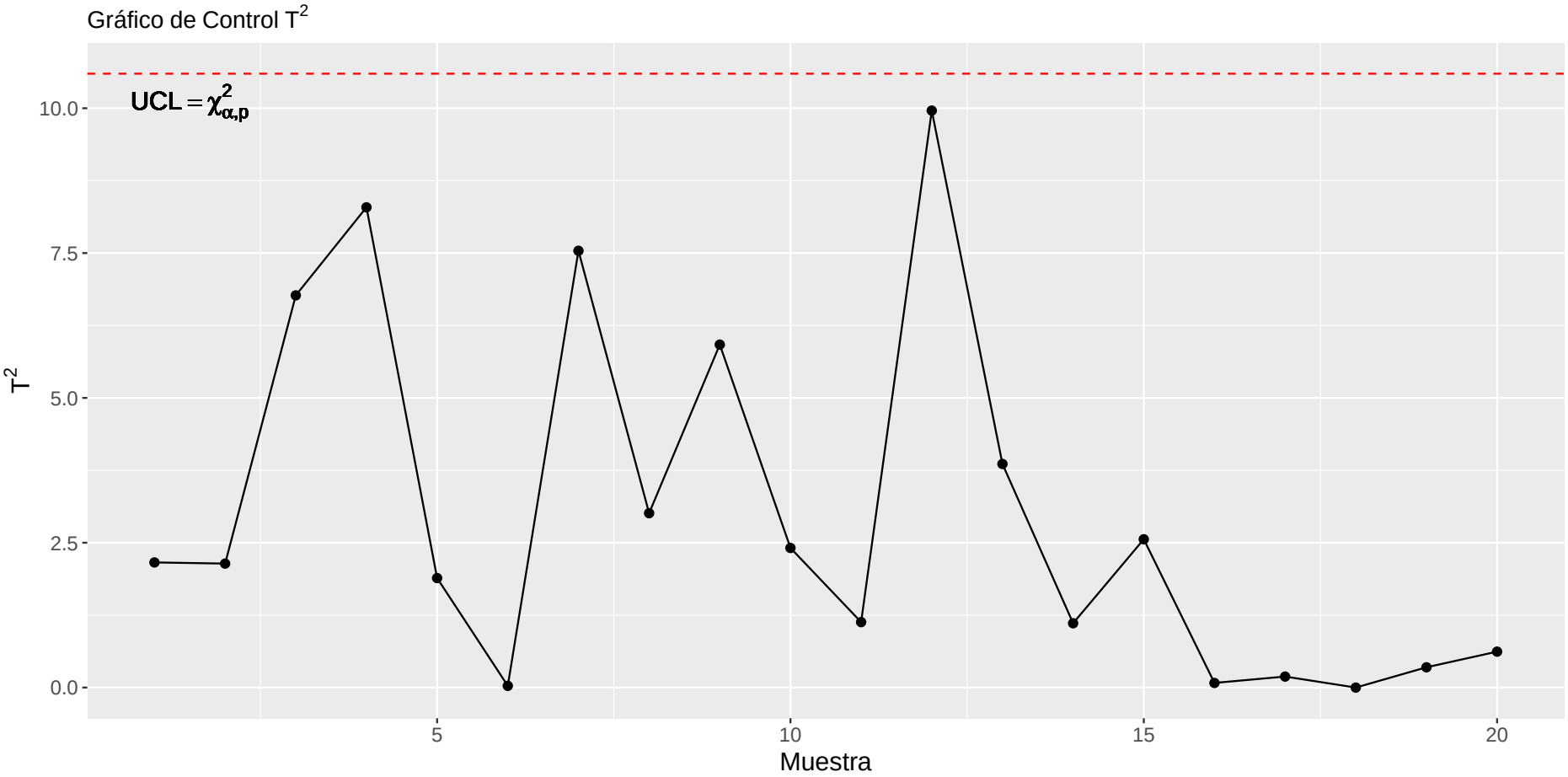
Gráfico de Shewhart para la $\bar{x}$



Control Estadístico Multivariado de Procesos (MSPC)

- Se utiliza cuando el concepto de calidad de un producto depende de un conjunto de variables.
- Un gráfico de control multivariado es el gráfico T^2 de Hotelling, el cuál tiene un sólo límite superior.
- En ciertos casos, el gráfico T^2 se construye utilizando un número de variables menor al original que retenga información relevante, aplicando Análisis de Componentes Principales (PCA) y basando el control en los scores.

Gráfico T^2 de Hotelling



Datos Faltantes

- En el contexto de MSPC es común encontrar datos faltantes debido a diferentes causas como errores de medición, fallas en sensores, formularios incompletos, etc.
- Arteaga y Ferrer (2002) [2] proponen y analizan un conjunto de métodos para estimar scores de nuevas observaciones cuando hay datos faltantes utilizando un modelo PCA conocido.
- El método Known Data Regression (KDR) produce estimadores con errores cuadráticos medios (ECM) pequeños y mantiene la ortogonalidad entre las componentes principales.
- La distribución condicional de los estimadores de los scores cuando son conocidos los valores observados en el caso de distribución normal de las variables originales es normal multivariada [1]. Es usual que para la construcción de la estadística T^2 en el control de procesos se utilice la matriz de covariancias de los scores en base a los datos completos.

Objetivo

Estudiar los efectos del cálculo del estadístico T^2 de Hotelling calculado estimando los scores de valores faltantes con el método KDR y utilizando la matriz poblacional de variancias y covariancias de los scores sobre las propiedades probabilísticas del gráfico de control.

Tabla de Contenidos

Modelo PCA y notación

- Modelo PCA considerando las primeras A PC's:

$$\mathbf{X} = \mathbf{T}_{1:A}\mathbf{P}'_{1:A} + \mathbf{E} \tag{1}$$

$$\hat{\mathbf{X}} = \mathbf{T}_{1:A}\mathbf{P}'_{1:A} \tag{2}$$

- Estimación de los scores de un nuevo vector de observaciones, $\mathbf{z}$, a partir del modelo PCA:

$$\boldsymbol{\tau} = \mathbf{P}'_{1:A}\mathbf{z} \tag{3}$$

Notación para datos faltantes

- Los valores incompletos en el vector de observaciones se simbolizan "#", mientras que los datos presentes se indican "$*$":

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}^{\#} \\ \mathbf{z}^{*} \end{bmatrix} \tag{4}$$

- Partición de la matriz de pesos:

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{1:A} & \mathbf{P}_{A+1:K} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^{\#}_{1:A} & \mathbf{P}^{\#}_{A+1:K} \\ \mathbf{P}^{*}_{1:A} & \mathbf{P}^{*}_{A+1:K} \end{bmatrix} \tag{5}$$

- Partición de la matriz de scores:

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{1:A} & \mathbf{T}_{A+1:K} \end{bmatrix} = \begin{bmatrix} \mathbf{T}^{\#}_{1:A} & \mathbf{T}^{\#}_{A+1:K} \\ \mathbf{T}^{*}_{1:A} & \mathbf{T}^{*}_{A+1:K} \end{bmatrix} \tag{6}$$

Método Known Data Regression (KDR) para la estimación de scores de vectores con datos faltantes

- Cuando $\mathbf{z}$ tiene observaciones faltantes, el método KDR estima el vector de scores, τ, usando las primeras A PC's como:

$$\hat{\tau}_{1:A} = \mathbf{\Theta}_{1:A}\mathbf{P}^{*'}_{1:A}(\mathbf{P}^{*}\mathbf{\Theta}\mathbf{P}^{*'})^{-1}\mathbf{z}^{*} \tag{7}$$

  donde $\mathbf{\Theta}$ es la matriz diagonal que contiene los autovalores de la matriz de covariancias a partir de la cual se construyó el modelo PCA.

- La distribución de un nuevo vector de scores condicional a los valores observados es:

$$\begin{aligned} \tau_{1:A}/\mathbf{z}^{*} \sim \quad & N\Big(\mathbf{\Theta}_{1:A}\mathbf{P}^{*T}_{1:A}(\mathbf{P}^{*}\mathbf{\Theta}\mathbf{P}^{*T})^{-1}\mathbf{z}^{*}, \\ & \left[\mathbf{I}_{1:A} - \mathbf{\Theta}_{1:A}\mathbf{P}^{*T}_{1:A}(\mathbf{P}^{*}\mathbf{\Theta}\mathbf{P}^{*T})^{-1}\mathbf{P}^{*}_{1:A}\right]\mathbf{\Theta}_{1:A}\Big) \end{aligned} \tag{8}$$

Cálculo del estadístico T^2 de Hotelling

- La propuesta de Arteaga y Ferrer para calcular el estadístico T^2 de Hotelling que se utiliza en la construcción del gráfico de control es:

$$T^2 = \hat{\boldsymbol{\tau}}'_{1:A} \boldsymbol{\Theta}^{-1}_{1:A} \hat{\boldsymbol{\tau}}_{1:A} \tag{9}$$

donde se observa que la matriz de covariancia utilizada en el cálculo es la matriz de los scores considerando observaciones sin datos faltantes.

Escenarios para el estudio por simulación

- Se usan dos matrices de correlación (generadas por el Método de Proyecciones Alternativas implementado por Waller [4]). A partir de éstas se construyen dos modelos PCA centrados reteniendo las componentes que explican al menos el 80 % de la variabilidad.
- Se generan observaciones normales multivariadas con vector de medias $\boldsymbol{\mu} = \mathbf{0}$ y se provocan faltantes según dos esquemas de pérdidas: completamente al azar (MCAR) y al azar en las variables más influyentes en la primera componente principal (MRIV).
- Los porcentajes de pérdidas considerados son: 5, 10, 15 y 20 %.
- El estudio se llevó a cabo usando el software libre R [3].

Primer estudio por simulación

- En cada uno de los escenarios propuestos, combinación de las alternativas de matrices de correlación, mecanismos de generación de pérdidas, porcentajes de datos faltantes, bajo control, se generan 10000 observaciones. Los scores de los valores faltantes se estiman con el método KDR y con ellos se calcula el estadístico T^2.
- Se realiza una prueba de bondad de ajuste para evaluar si la distribución puede ajustarse a una χ^2, con la que se determinaron los límites de control para los gráficos.

Segundo estudio por simulación

- Obtener los ARL (Average Run Length) estimados y sus ECM en situaciones bajo y fuera de control para analizar las propiedades del gráfico de control.
- ARL es el promedio del número de puntos en el gráfico de control hasta que se detecta un valor del estadístico T^2 superior al límite de control. Bajo control, el valor de ARL debe ser grande para evitar falsas alarmas. Fuera de control, es deseable que ARL sea pequeño para detectar rápidamente esta situación.

Segundo estudio por simulación

- Para reproducir situaciones fuera de control se producen corrimientos constantes en el vector de medias: $\boldsymbol{\mu} = \mathbf{0} + \boldsymbol{\Delta}$, con valores de $\boldsymbol{\Delta}$ entre $\mathbf{0,01}$ y $\mathbf{1}$. Estos desvíos se expresan como la distancia de Mahalanobis entre el vector original, $\mathbf{0}$, y $\mathbf{0} + \boldsymbol{\Delta}$
- Se simulan situaciones bajo control y fuera de control cambiando los parámetros de la distribución. En cada caso se estima el ARL con 10000 repeticiones del proceso hasta la salida de control (RL).

$$ECM_{\hat{ARL}} = E[\hat{ARL} - ARL]^2 \tag{10}$$

Tabla de Contenidos

Pruebas de bondad de ajuste

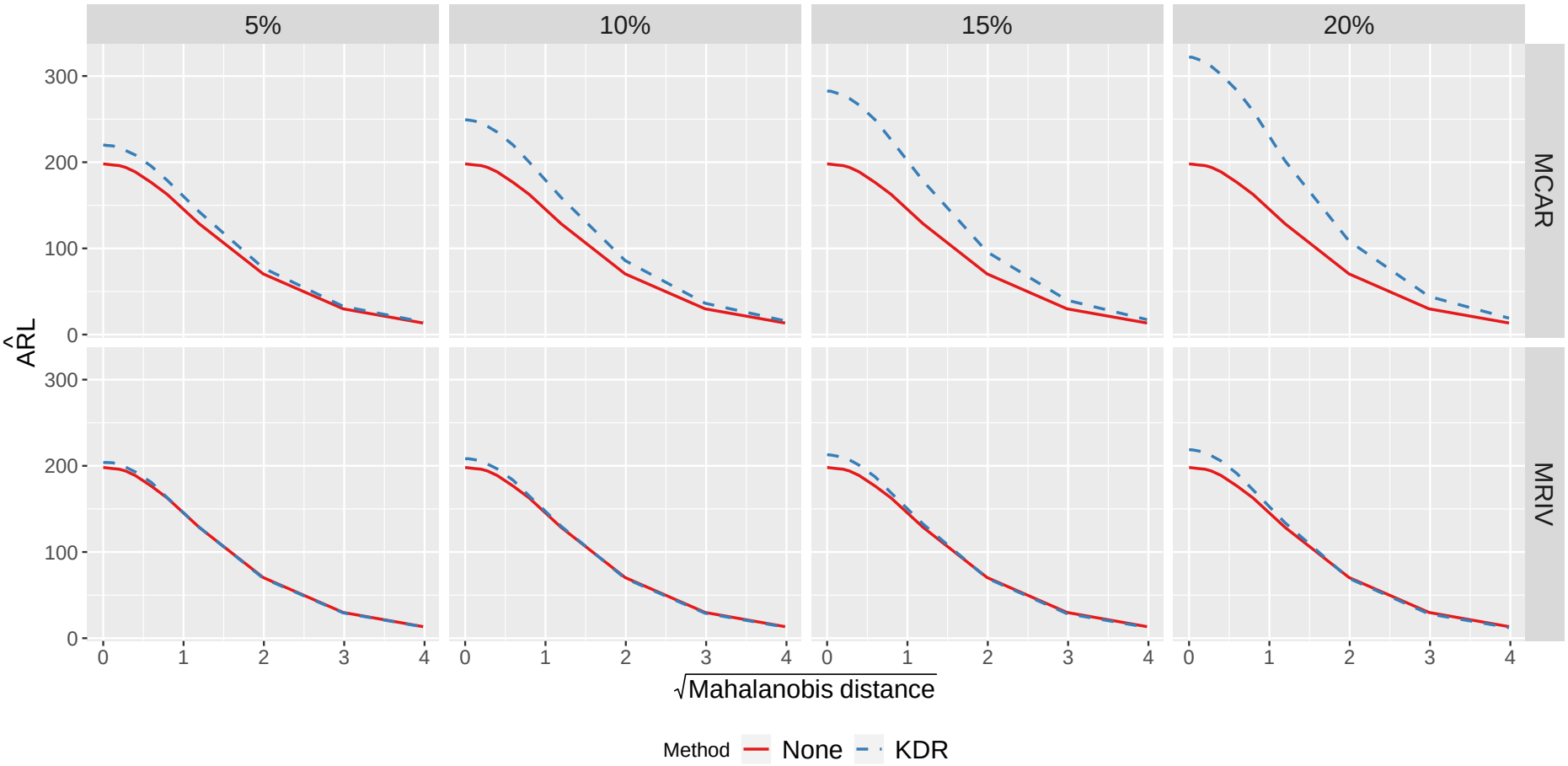
Mecanismo de pérdida	Porcentaje de datos faltantes	Matriz de correlación 1	Matriz de correlación 2
MCAR	5	0.0132	0.5409
	10	0.0000	0.0023
	15	0.0000	0.0000
	20	0.0000	0.0000
MRIV	5	0.0472	0.4649
	10	0.1712	0.4788
	15	0.2036	0.6925
	20	0.3947	0.7694

Cuadro 1: p-values de las pruebas de Kolmogorov-Smirnov del estadístico T^2 para las matrices de correlación 1 y 2.

- Cuando las pérdidas se produjeron en las variables altamente correlacionadas con la primera PC, en los escenarios bajo control el estadístico T^2 sigue una distribución χ^2 (Tabla 1).

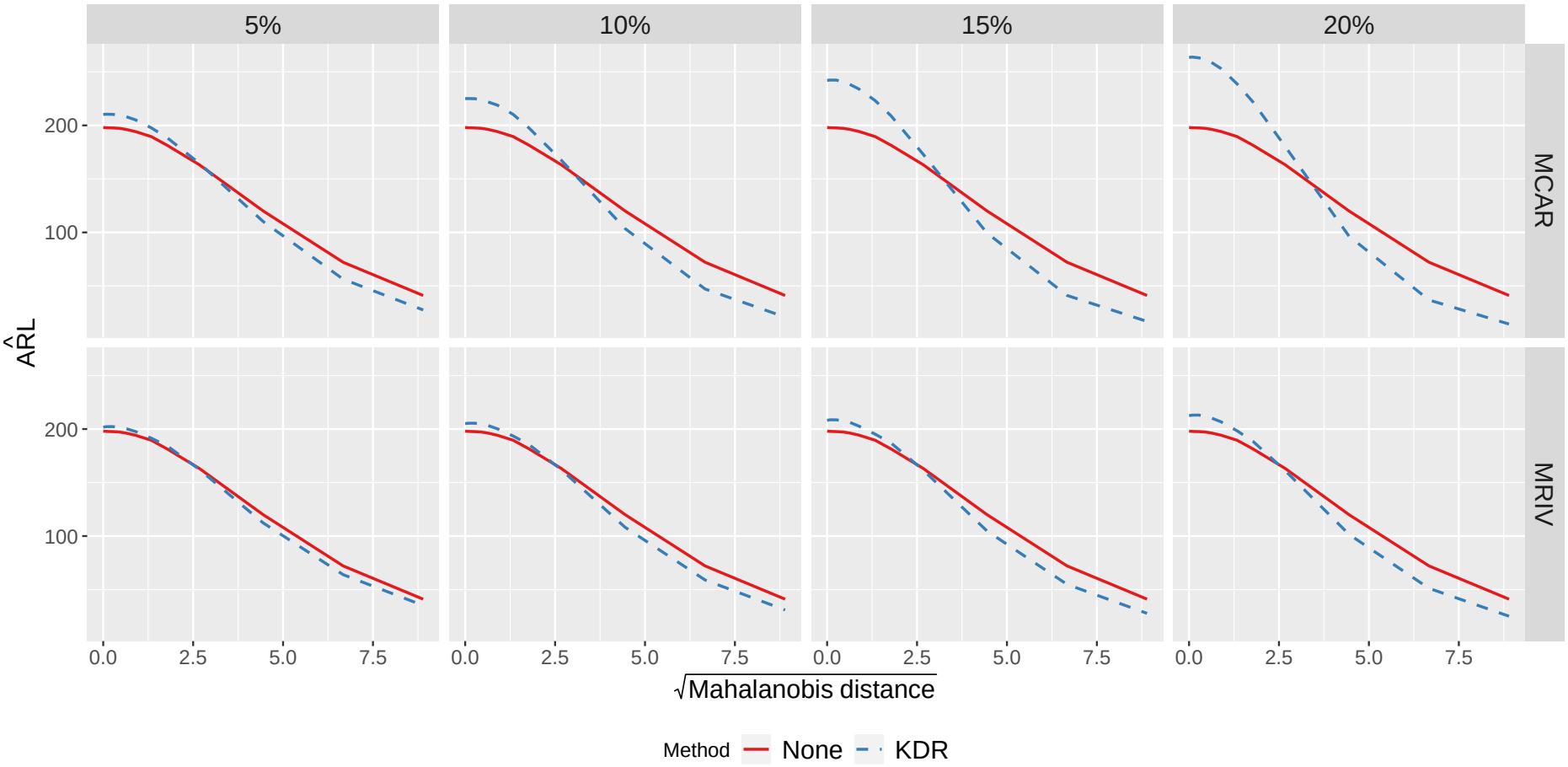
Curvas de ARL de la matriz de correlación 1

Figura 1: Curvas de $\hat{ARL}$ del estadístico T^2 de Hotelling con la estructura de correlación 1



Curvas de ARL de la matriz de correlación 2

Figura 2: Curvas de $\hat{ARL}$ del estadístico T^2 de Hotelling con la estructura de correlación 2

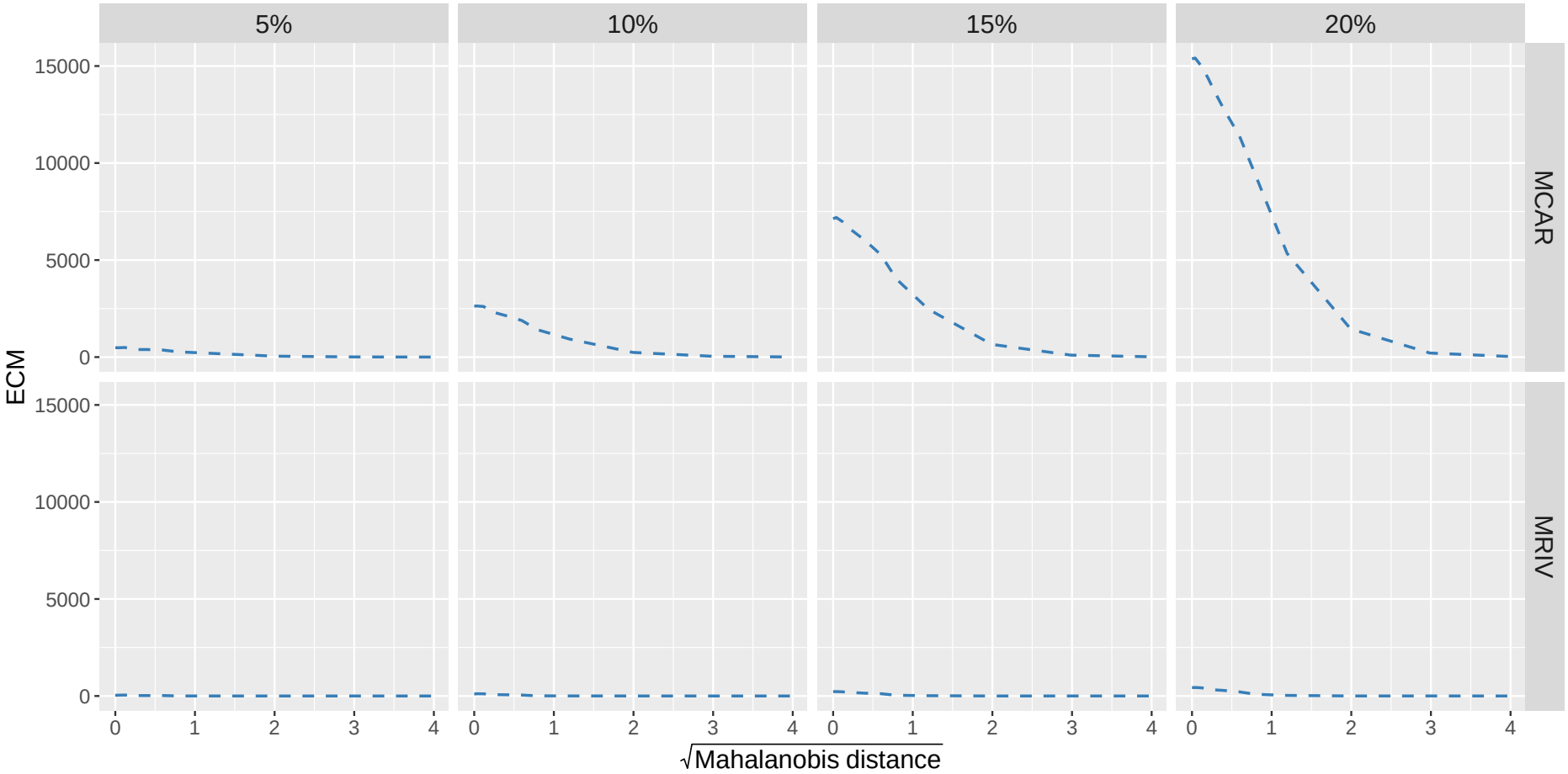


Curvas de ARL

- Las curvas de ARL estimado son más parecidas a las curvas de ARL calculadas sin datos faltantes cuando la pérdida se produce en las variables influyentes sobre la primera componente principal, y las aproximaciones empeoran a medida que aumenta el porcentaje de valores faltantes.
- Con la matriz de correlación 1, cuando las pérdidas se producen sobre las variables más influyentes sobre la primera PC prácticamente no se distinguen diferencias entre las curvas de ARL estimadas con y sin valores faltantes. En cambio cuando las pérdidas son completamente al azar y el proceso está bajo control o los desvíos en el vector de medias son pequeños o medianamente grandes el ARL estimado con KDR es superior al valor esperado (Figura 1).
- Con la matriz de correlación 2, para los desvíos más grandes de la situación de control se encuentra que el valor estimado de ARL obtenido con el método KDR es menor que los valores obtenidos sin datos faltantes (Figura 2).

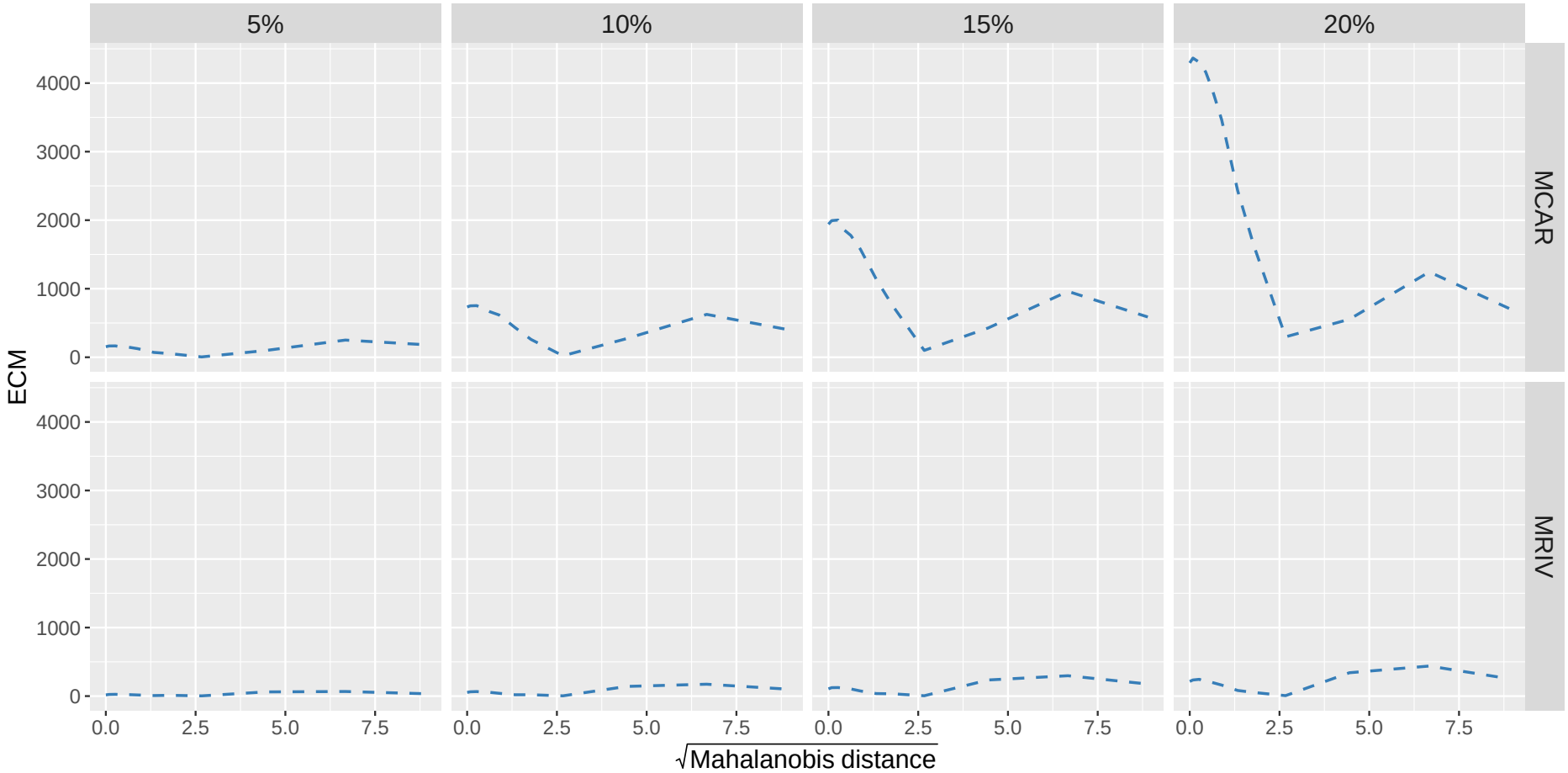
ECM del ARL de la matriz de correlación 1

Figura 3: Curvas de ECM del $A\hat{R}L$ con la estructura de correlación 1



ECM del ARL de la matriz de correlación 2

Figura 4: Curvas de ECM del $A\hat{R}L$ con la estructura de correlación 2



ECM de las curvas de ARL

- En ambas estructuras de correlación se observa que los mayores valores de ECM se encontraron en escenarios en los que el mecanismo de generación de datos faltantes es completamente al azar (Figuras 4 y 3).
- Con la matriz de correlación 1 los valores de ECM aumentan cuando crece el porcentaje de valores faltantes y para los casos bajo control o con desviaciones leves a moderadas de esta situación (Figura 3).
- Con la matriz de correlación 2, los valores de ECM aumentan cuando crece el porcentaje de valores faltantes. Con ambos mecanismos de pérdida se observa que los mayores ECM se dan en las situaciones bajo control o con pequeñas desviaciones del control, pero también hay un crecimiento del ECM estimado cuando el vector de medias se aleja mucho de $\mathbf{0}$ (Figura 4).

Tabla de Contenidos

Discusión

- Los mayores errores en la estimación de las curvas de ARL se producen cuando la pérdida de información es completamente al azar y crecen en relación al porcentaje de datos faltantes.
- Si la pérdida es completamente al azar, las pruebas de bondad de ajuste indican que la distribución del estadístico T^2 calculado con scores estimados mediante el método KDR no siguen la distribución χ^2 con la que se fijan los límites de control en muchos de los escenarios bajo control.

Discusión

- Cuando se realizaron las simulaciones con la matriz de correlación 1 y pérdidas completamente al azar, para porcentajes de pérdida de 15 y 20 % las salidas de control del vector de medias no son detectadas con la misma rapidez que cuando no ocurren datos faltantes. Esto puede ser especialmente perjudicial en procesos donde el control se realiza a intervalos largos de tiempo.
- En los casos en los que se utilizó la matriz de correlación 2 y se presentan grandes desviaciones en el vector de medias de las variables del proceso, las salidas de control se detectaron más rápidamente en presencia de información incompleta.

Tabla de Contenidos

Bibliografía

[1] Arteaga, F. (2003). *Control Estadístico Multivariante de Procesos con datos faltantes mediante Análisis de Componentes Principales* (tesis doctoral). Universidad Politécnica de Valencia, Valencia, España.

[2] Arteaga, F. y Ferrer, A. (2002). Dealing with missing data in MSPC: several methods, different interpretations, some examples. *Journal of Chemometrics*, 16, 408-418.

[3] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[4] Waller, N.G. (2018). Generating correlation matrices with specified eigenvalues using the method of alternating projections. *The American Statistician*, DOI: 10.1080/00031305.2017.1401960.