missMSPC: un paquete de herramientas gráficas para aplicar MSPC con datos faltantes

Julia Inés Fernández , Diego Marfetán Molina , José Alberto Pagura , Marta Beatriz Quaglino

Palabras clave: Control de Calidad - MSPC - Datos Faltantes - Componentes Principales

El control estadístico de procesos es una estrategia ampliamente utilizada en el contexto de control y mejora de la calidad. De acuerdo a la International Organization for Standardization (ISO), calidad es "el grado en el que un conjunto de características inherentes a un objeto (producto, servicio, proceso, persona, organización, sistema o recurso) cumple con los requisitos" (ISO 2005). La calidad constituye un factor decisivo al momento de seleccionar productos o servicios, tanto para transacciones entre organizaciones como en ventas al consumidor final. Actualmente es de suma importancia implementar técnicas de control de calidad, ya que proporcionan ventajas económicas y competitivas.

Generar productos o servicios de alta calidad no es un asunto trivial o sencillo. La constante evolución tecnológica que está transformando industrias y organizaciones, en áreas tan diversas como electrónica, metalúrgica, química, biotecnología, etc., hace posible registrar cada vez más información sobre un proceso, incluso de manera automática, y permite tener una visión más completa del mismo. Una de las consecuencias de estos avances es que la calidad no depende de un único atributo, sino que se mide a través de un **conjunto de características**.

Una de las posibles estrategias a aplicar dentro de este contexto es el Control Estadístico Multivariado de Procesos (MSPC). Entre otras cosas, el MSPC permite monitorear un proceso a lo largo del tiempo, observar si su comportamiento se ajusta al patrón esperado, y detectar eventos especiales en tiempo real que señalen que el proceso se encuentra fuera de control. La principal herramienta del MSPC son los **gráficos de control**, los cuales representan una estadística a lo largo del tiempo y la contrastan con valores límites determinados en función de su distribución de probabilidad, lo que permite determinar el estado del proceso. Existen dos etapas en la utilización de gráficos de control en MSPC, llamadas Fase I y Fase II. En la Fase I se definen los límites de control, los cuales serán utilizados para monitorear el proceso en tiempo real durante la Fase II.

La posibilidad de recolectar datos a gran escala sobre estos procesos se traduce en una baja relación de señal-ruido (*low signal-to-noise ratio*), fuertes estructuras de correlación entre las variables (multicolinealidad) y matrices de covariancia no invertibles, las cuales imposibilitan el cálculo de las estadísticas necesarias para construir gráficos de control multivariados convencionales. En estos casos se aplican métodos basados en **variables latentes**, como Análisis de Componentes Principales (PCA) y Mínimos Cuadrados Parciales (PLS). En primer lugar se proyectan los datos hacia un subespacio de menor dimensión, para luego realizar el control del proceso sobre las variables latentes identificadas como relevantes. En este trabajo se considera el uso simultáneo de los gráficos T^2 de Hotelling y SPE (*squared prediction error*, también llamada estadística Q) sobre variables latentes para el monitoreo del proceso.

Un problema común frente a grandes bases de datos es la ocurrencia de **valores faltantes** en las observaciones a controlar, los cuales pueden surgir debido a fallas en sensores, formularios incompletos, errores de registro, etc. Se han propuesto numerosas estrategias para imputar estos valores cuando el control se realiza aplicando modelos de variables latentes. Dos de las técnicas más populares son *Known Data Regression* (KDR) y *Trimmed Score Regression* (TSR) (Arteaga and Ferrer 2002), (Folch-Fortuny, Arteaga, and Ferrer 2015). El método KDR es equivalente a *Conditional Mean Replacement* (CMR) (Nelson, Taylor, and MacGregor 1996). Hasta donde los autores de este trabajo pudieron constatar, ninguna de estas técnicas se encuentra implementada actualmente en lenguaje R.

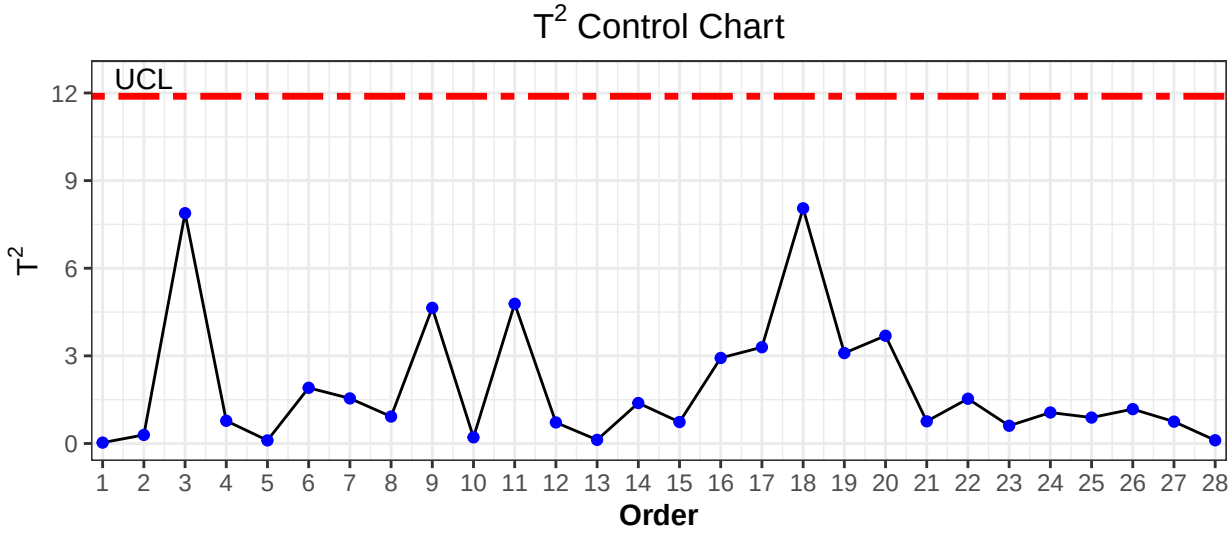
El paquete propuesto `missMSPC` permite construir los gráficos de control multivariados T^2 y SPE a partir de un modelo PCA en Fase II en ambas situaciones: datos completos o incompletos. Ante la presencia de valores faltantes, se permite realizar la imputación a través de alguno de los métodos propuestos (KDR/CMR o bien TSR) mediante la función `score_imp()`. Consideramos que la importancia de este paquete reside en el hecho de permitir que cualquier interesado/a en realizar control estadístico de procesos tenga acceso a herramientas

avanzadas mediante un software libre. A futuro se planifica implementar el método de control en base a PLS, la estimación del modelo PCA en Fase I cuando se trabaja con datos incompletos, y la utilización de técnicas de validación cruzada para elegir el número óptimo de componentes a incorporar en el modelo PCA. Además, se pretenden incorporar nuevos gráficos de control de mejor performance en presencia de datos faltantes.

A continuación se presenta un ejemplo de aplicación, donde se utilizan los conjuntos de datos *bimetal1* y *bimetal2* del paquete `MSQC` (Santos-Fernández 2013). El modelo PCA se ajusta sobre los datos *bimetal1*, y luego se aplica el método CMR a una versión de *bimetal2* en la que se generaron aleatoriamente un 25 % de valores faltantes. Finalmente, se construye el gráfico T^2 de Hotelling:

```
bm1 <- scale(bimetal1) #data(bimetal1) del paquete MSQC
bm2 <- scale(bimetal2) #data(bimetal1) del paquete MSQC
set.seed(1974) #Semilla
datos <- mice::ampute(bm2, mech = "MCAR", prop = 0.25)$amp #Agrego valores faltantes
acp <- princomp(bm1) #Ajuste modelo CPA
autov <- acp$sdev^2 #Vector de autovalores
pesos <- matrix(as.numeric(acp$loadings), ncol = length(autov)) #Matriz de Pesos

#Imputación de Scores y Gráfico T2
ajuste <- score_imp(datos = datos, pesos = pesos, autov = autov, A = 2, metodo = "CMR")
graficoT2(x = ajuste, alfa = 0.01)
```



Referencias

10 Arteaga, F., and A. Ferrer. 2002. "Dealing with Missing Data in MSPC: Several Methods, Different Interpretations, Some Examples." *Journal of Chemometrics* 16 (810): 408–18. https://doi.org/10.1002/cem.750.

Folch-Fortuny, A., F. Arteaga, and A. Ferrer. 2015. "PCA Model Building with Missing Data: New Proposals and a Comparative Study." *Chemometrics and Intelligent Laboratory Systems* 146: 77–88. https://doi.org/https://doi.org/10.1016/j.chemolab.2015.05.006.

ISO. 2005. "ISO 9000:2005, Quality Management Systems - Fundamentals and Vocabulary." International Organization for Standardization.

Nelson, Philip R. C., Paul A. Taylor, and John F. MacGregor. 1996. "Missing Data Methods in PCA and PLS: Score Calculations with Incomplete Observations." *Chemometrics and Intelligent Laboratory Systems* 35 (1): 45–65. https://doi.org/https://doi.org/10.1016/S0169-7439(96)00007-X.

Santos-Fernández, Edgar. 2013. *Multivariate Statistical Quality Control Using r*. Vol. 14. Springer. http://www.springer.com/statistics/computational+statistics/book/978-1-4614-5452-6.