Universidad Nacional de Colombia, Sede Bogotá
Facultad de Ciencias - Departamento de Estadística
Maestría en Ciencias Estadística - Plan de profundización
Bogotá, julio de 2022

Carta de control multivariada sin distribución para datos funcionales y vectoriales híbridos

Propuesta de trabajo final de maestría

Andrey Duvan Rincon Torres[a]
Autor

Ruben Dario Guevara Gonzalez[b]
Director

Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

1. Introducción

El control estadístico de calidad se compone de un conjunto de herramientas tales como las cartas de control. Estas son herramientas estadísticas y gráficas que resumen la variación del un proceso a lo largo del tiempo. El investigador utiliza el conocimiento previamente adquirido del proceso para diseñar un diagrama de secuencia que capturar nueva información sobre el resultado del proceso. En general, los gráficos de control se generan a partir del cálculo de una serie de estadísticas puntuales $\boldsymbol{T}$, calculadas para diferentes muestras distinguidas en orden cronológico. Sobre estas estadísticas se definen reglas que establecen una región de alta probabilidad cuando el proceso se encuentra bajo control (Montgomery 2020). Algunos ejemplos de cartas se pueden ver en Ryan (2011).

Las tecnologías de detección avanzadas han facilitado la recogida de datos en tiempo real para la supervisión de procesos y el diagnóstico de fallos en sistemas de fabricación complejos. Los datos de perfil, donde se recogen conjuntos de m_j observaciones de una variable repetida para cada individuo $\{(y_i, x_{ij})\}_{i=1,\dots n, j=1,\dots,m_j}$, es uno de los tipos más comunes de datos recopilados por los sistemas de detección, datos que pueden ser representados por observaciones funcionales. Sin embargo, en algunas situaciones la calidad de un proceso se caracteriza mejor por la medición de múltiples características con múltiples modalidades de datos para una muestra de unidades experimentales (Jang 2021). En el trabajo final de maestría se busca proponer un marco viable de gráficos de control para el monitoreo en línea de procesos donde las mediciones corresponden a dos tipos de datos comúnmente recolectados en los procesos de producción moderna: datos funcionales y vectoriales (multivariados).

Para lograr el objetivo del trabajo, se necesita una metodología de control de calidad para monitorear las variaciones simultáneas de datos vectoriales y funcionales multivariados de un proceso con mediciones en línea, pero no existe dentro de la documentación disponible. Un posible enfoque para el control calidad con este tipo de información consiste en realizar por separado el control calidad individual de la parte funcional multivariada y la parte vectorial, ignorando las estructuras de correlación existentes entre funciones y vectores que puede proporcionar información valiosa sobre el proceso, estructura que pueden proporcionar información valiosa sobre el proceso (Jang 2021).

Como parte de la documentación del esquema de control de calidad de datos de perfiles multivariados, se han desarrollado varios métodos. Algunos bajo la perspectiva paramétrica para supervisar coeficientes

[a]Estudiante de maestría en estadística. E-mail: adrincont@unal.edu.co
[b]Profesor asociado. E-mail: rdguevarag@unal.edu.co

de modelos lineales y no lineales (Zou et al. (2007);Williams et al. (2007)) o basados en modelos de efectos mixtos (Jensen et al. (2008); Paynabar et al. (2012)). Williams et al. (2007) propone el uso de tres estadísticas T^2 (análisis de la fase I) para controlar los coeficientes resultantes de un modelo de regresión no lineal paramétrico. Se tienen enfoques para monitoreo simultáneo de vector medio y la matriz de covarianza de los perfiles lineales multivariantes (Ghashghaei et al. 2019), o enfoques que combinan en un único perfil agregado las magnitudes de las señales de los demás canales (Lei et al. 2010), pero que no captan toda la información existente en los perfiles de cada variable. Soleimani et al. (2013) proponen una carta MEWMA/χ^2 para el seguimiento de perfiles multivariantes cuando se viola el supuesto de independencia de las observaciones dentro de los perfiles.

Otros enfoques ahora bajo la perspectiva de métodos no paramétricos se pueden notar, por ejemplo, Hung et al. (2012) propone la construcción de una región de confianza simultánea para controlar perfiles reales sobre el SIDA por medio de los enfoques de SVR y método bootstrap (Fase I). Pan et al. (2019) ajustan un perfil de referencia en control mediante un modelo de regresión de vectores de soporte (SVR) y monitorean un vector de métricas entre perfiles observados y de referencia, enfoques similares (Noorossana et al. (2010), Fan et al. (2017)) construyen el perfil de referencia por medio de métodos del Spline. Zou et al. (2008) propone una carta SREWMA de arranque automático para controlar los parámetros de localización combinando los métodos de rango espacial y de media móvil ponderada exponencial.

Algunas propuestas consideran los perfiles en el marco del análisis funcional. Entre ellos, Zhang et al. (2015) plantean una metodología de monitoreo para el análisis de la Fase I con múltiples perfiles, trata las observaciones como datos funcionales multivariantes, donde los residuos ajustados por regresión son supervisados mediante el uso de una metodología de componentes principales funcionales FPCA de manera individual por componente. Por otra parte, Paynabar et al. (2016) desarrollan una carta de punto de cambio mediante el análisis de componentes principales funcionales multivariados para construir estadísticas de seguimiento. Paynabar et al. (2013) utilizaron técnicas de reducción dimensión multilineal (UMPCA) para extraer características informativas de los perfiles, para luego plantear una metodología de monitoreo de perfiles multivariados.

De forma complementaria a estos trabajos y empleando la técnica de componentes principales para datos funcionales multivariados híbridos que se describe con detalle más adelante, para este trabajo final de maestría se busca el planteamiento de una metodología de control de calidad para el monitoreo simultáneo de procesos donde las observaciones en línea estén conformadas de datos vectoriales y perfiles no lineales multivariados con estructuras de correlación entre ambas partes. Dentro de las características de este gráfico de control, se busca que esté basado en estadísticas libres de distribución, que sea factible computacionalmente en la práctica y que permita el monitoreo de mediciones tanto individuales ($n_i = 1$) como en grupos ($n_i > 1$) para cada punto en el tiempo.

2. Marco Teórico

Para el desarrollo de la carta de control y de acuerdo a la revisión de la literatura, a continuación se presentan algunos conceptos teóricos importantes en la construcción del esquema de control. Se comienza con algunos conceptos fundamentales sobre la teoría del control estadístico de calidad (SPC), continuando con una introducción al análisis de datos funcionales, para luego mostrar las bases teóricas y de implementación de métodos de componentes principales para datos vectoriales (PCA), datos funcionales multivariados (MFPCA) y datos funcionales multivariados híbridos (MFPCAH). Finalmente, se describe la prueba de rangos de wilcoxon y los motivos por los que esta sería una posible candidata para la construcción del estadístico de monitoreo para la casta de control.

2.1. Cartas de control

2.1.1. Control estadístico de calidad

El control estadístico de calidad (SPC) se entiende como un conjunto de métodos estadísticos aplicados principalmente en un contexto industrial, con el fin de mejorar la calidad en el proceso de producción, mediante el control de entradas y salidas (productos) y procesos relacionados (Qiu 2013). En este contexto, la "idoneidad para el uso" se define como un conjunto de características intrínsecas del producto y su adecuación a un conjunto de especificaciones. A continuación, en función del grado de variación de dichas características, se mide la calidad en términos estadísticos, siendo mayor la calidad cuanto menor sea la varianza (Montgomery 2020).

Entre las herramientas más conocidas para mejorar la calidad son los gráficos de control, que son herramientas gráficas para supervisar el comportamiento de un proceso mientras varían sus propiedades en el tiempo. Estas características muestran una amplia gama de causas diferentes, que se dividen en dos categorías: causas comunes y causas atribuibles.

- **Causas comunes:** Reconocer lo que está oculto al proceso. Esto incluye cualquier perturbación del proceso causada por factores conocidos o desconocidos, internos o externos al proceso, cuyo control está limitado por diversos factores como el tiempo o los recursos, pero que no compromete la calidad del proceso.
- **Causas asignables:** Cualquier variación que se produzca fuera del sistema de causas comunes que introduce una variabilidad inevitable en el proceso, lo que aumenta el riesgo de que los productos no cumplan los parámetros de diseño y, en consecuencia, de que sean de mala calidad.

Cuando un proceso solo presenta pequeñas variaciones debidas a causas comunes, se considera bajo control (CI). Si hay al menos una o más causas asignables, la situación se clasifica como fuera de control (OC). Más información relacionada se puede ver en (Montgomery 2020).

2.1.2. Carta de control

Las cartas de control, como herramientas estadísticas, se basan en un fundamento teórico en el que se busca determinar cuándo un proceso está bajo control (CI) y cuándo no (CO). En una gran parte de las cartas predomina el uso de reglas definidas por límites de control que establecen una región de alta probabilidad para el estadístico de interés si se encuentre bajo control. En este tipo de cartas se utilizan casi siempre entre uno y dos límites de control. Algunos ejemplos de estas reglas, junto con sus resultados sobre la longitud de la carrera, pueden encontrarse en (Montgomery 2020).

Normalmente, las cartas de control se aplican en dos fases, denominadas fase I y fase II. La primera de ellas se refiere a un análisis retrospectivo del proceso cuyo objetivo es caracterizar la distribución estadística bajo control mediante tarjetas aplicadas a conjuntos de datos históricos con límites de prueba. Estos límites se ajustan en un proceso iterativo que identifica y elimina de los cálculos las observaciones fuera de control.

Tras la limpieza del conjunto de datos históricos y la garantía de que los datos representan un proceso en estado de control, la segunda fase consiste en la supervisión de la producción en línea mediante cartas de control con límites que se corresponden con los obtenidos en la primera fase. Esta supervisión se realiza con el objetivo de mantener la situación bajo control y detectar con rapidez y precisión cualquier cambio en alguno de los estados (Montgomery 2020)).

Riesgos del muestreo

Debido a que las gráficas de control se generan en función de la cantidad de variación entre las muestras y los límites de control que son establecidos para contener estas realizaciones con alta probabilidad, existen dos riesgos asociados con la variación en los datos estadísticos. El primero de ellos, riesgo de tipo I, que consiste en obtener valores muestrales fuera de control, aunque dicha medición provenga de

una distribución IC. Por otro lado, el riesgo de tipo II, es el valor muestral de la distribución OC que es probable que se encuentre dentro del límite de control. En la construcción de una carta, es usual dar prioridad al manejo del riesgo de tipo I o, como alternativamente se le conoce "tasa de falsas alarmas" (Montgomery 2020).

Longitud de corrida

Cuando se ha establecido la gráfica de control según la distribución IC, se genera una variable aleatoria para el número total de muestras consecutivas sin que se genere ninguna falla, a la que se le da el nombre de longitud de corrida RL. Esta variable se utiliza como medida o entrada en el cálculo de las estadísticas de rendimiento de los gráficos de control. En el caso independiente, cuando se conocen los parámetros de la distribución IC del gráfico de control y la probabilidad de obtener una señal permanece constante, en ambos casos IC y OC, debido al cambio constante, RL tiene distribución geometría de parámetros iguales a la probabilidad de la señal correspondiente, siendo α durante el estado IC y β en otro caso bajo las suposiciones anteriores (véase por ejemplo Montgomery (2020)).

Para utilizar ARL como medida de rendimiento y descripción general de una carta, se tiene en cuenta el estado del proceso que obtuvo este promedio. Para denotar partes del proceso IC, el índice 0 se agrega a la media, ARL0. Por otro lado, ARL1, se usa para el medio OC. En cuanto a ARL, una carta de fase II se considera confiable si contiene un ARL0 mayor y un ARL1 menor, de acuerdo con los requisitos del proceso.

2.2. Datos funcionales

En la actualidad, los avances en tecnología y nuevas aplicaciones de la estadística han hecho que grandes volúmenes de información para un mismo individuo estén disponibles para la investigación en diferentes áreas del conocimiento. Datos que son representados por curvas, superficies o de manera más general en forma de funciones, para el análisis de este tipo de problemas surge el campo de estudio denominado Análisis de datos funcionales (ADF) que desarrolla métodos estadísticos que tratan con estas nuevas estructuras de información.

La unidad básica para el ADF son las funciones que generalizan un conjunto de valores (Ramsay & Dalzell 1991), por lo que se puede pensar que un dato funcional es una extensión de las medidas repetidas cuando el número de mediciones es muy grande. Se define una variable aleatoria funcional χ como una variable aleatoria que toma valores en un espacio infinito dimensional (espacio funcional) (Ferraty & Vieu 2006).

En el análisis de datos funcionales (ADF) se construye la teoría sobre los espacios $L^2(\mathcal{T})$, el cual corresponde al conjunto de funciones de valor real definidas sobre un conjunto compacto $\mathcal{T} \in \mathbb{R}^p$ con medida (de Lebesgue) finita. El espacio $L^2(\mathcal{T})$ es un espacio de Hilbert separable con el producto interno definido como:

$$\begin{aligned} \langle x, y\rangle_{\mathcal{T}} &= \int_{\mathcal{T}} f(t)g(t)dt \quad \text{para }, f(t), g(t) \in L^2(\mathcal{T}) \\ L_2(\mathcal{T}) &= \left\{ f : \mathcal{T} \to \mathbb{R}, \text{ tal que } \int_{\mathcal{T}} f(t)^2 dt < \infty \right\} \end{aligned} \tag{1}$$

Una observación f de una variable aleatoria funcional χ es a lo que denominamos como dato funcional, un conjunto de datos funcionales $\{f_1, f_2, , f_n\}$ es la observación de n variables aleatorias funcionales distribuidas χ. Un dato funcional $f_i(t), t \in \mathcal{T} \subset \mathbb{R}^p$ se representa como un conjunto de M parejas $(t_i, x_{ij}) \quad t_i \in \mathcal{T}, j = 1, 2, ..., M$ observadas de la variable de interés $y_{ij} = f_i(t_j) + \varepsilon_j$ con ε_j error observacional (si existe). En el marco de AFD, normalmente se asume que se tienen elementos del espacio de funciones $L_2(\mathcal{T})$ (Horváth & Kokoszka 2012).

Dentro de esta estructura, nos referimos a datos funcionales multivariados de dimensión K, a observaciones de la forma $\boldsymbol{f} = \left[f^{(1)}, f^{(2)}, ..., f^{(K)}\right]$ de una variable aleatoria $\boldsymbol{\chi} = \left[\chi^{(1)}, \chi^{(2)}, ..., \chi^{(K)}\right] \in \mathcal{F} := \mathcal{T}_1 \times \cdots \times \mathcal{T}_k$. Es decir que cada observación funcional multivariada consiste en $K \in \mathbb{N}$ funciones

combinadas en un vector $\boldsymbol{f} \in \mathcal{F}$, donde cada función $f^{(i)} \in \mathcal{T}_i$ es una observación de la variable aleatoria funcional χ, los dominios $\mathcal{T}_i$ no necesariamente deben ser el mismo (Happ & Greven 2018).

En la práctica, contamos para cada unidad muestral con un número M_i muy grande de observaciones discretizadas $\{(t_i, x_{ij})\}_{j=1,...M_i}$ (perfil), estas observaciones se aproximan a funciones usando métodos de suavizado no paramétrico (método que depende del tipo de observaciones y el contexto del problema), por lo que el análisis de datos funcionales (ADF) tiene inherente las metodologías de estimación funcional no paramétricas, algunas de estas metodologías se pueden estudiar en Horváth & Kokoszka (2012).

Cuando vemos una curva aleatoria χ como un elemento aleatorio de $L_2(\mathcal{T})$ equipado con σ-álgebra de Borel. Si χ es integrable $E\|\chi\| = E\left[\int_{\mathcal{T}} \chi^2(t)dt\right]^{1/2} < \infty$, se garantiza que existe una única función $\mu \in L_2(\mathcal{T})$ que llamamos función media tal que $E\langle y, X\rangle_{\mathcal{T}} = \langle y, \mu\rangle_{\mathcal{T}}$ para todo $y \in L_2(\mathcal{T})$. Si además, χ es cuadrado integrable $E\|\chi\|^2 = E\left[\int_{\mathcal{T}} \chi^2(t)dt\right] < \infty$ y $E(\chi) = 0$ el operador de covarianza, se define como:

$$\begin{aligned} C_\chi y) &= E[\langle \chi, y\rangle \chi], \quad y \in L^2, \\ C_\chi(y)(t) &= E[\langle \chi, y\rangle \chi] = \int_{\mathcal{T}} \sigma(t,s)y(s)ds, \text{ donde } \sigma(t,s) = E[\chi(t)\chi(s)] \end{aligned} \tag{2}$$

Este sirve como una generalización al concepto de matriz de covarianza para el caso de vectores aleatorios. El operador C_χ cumple las propiedades de ser simétrico definido positivo y de Hilbert Schmidt como se muestra en Horváth & Kokoszka (2012). Estos conceptos fundamentales son las bases requeridas para la construcción de las metodologías de reducción de dimensionalidad ACP trabajadas sobre el análisis de datos funcionales (FPCA), detalles de estas metodologías se describen continuación.

2.3. Análisis de componentes principales para datos vectoriales

Tenemos estructura de datos, $\boldsymbol{V} = [v_1, v_2, ..., v_P]^T$ que denota un vector aleatorio p-dimensional. Se supone que $\boldsymbol{V}$ es un vector aleatorio con dos primeros momentos finitos y dotado del producto interno euclidiano usual $\langle \mathbf{v_1}, \mathbf{v_2}\rangle = \mathbf{v_1}^T\mathbf{v_2}$ y la norma $\|\mathbf{v}_1\| = \langle \mathbf{v}_1, \mathbf{v}_1\rangle^{1/2}$ para $\mathbf{v}_1, \mathbf{v}_1 \in \mathbb{R}^p$. Lo que se busca en el PCA vectorial es reducir la dimensión de $\boldsymbol{V}$ proyectándola sobre un sub-espacio de dimensión $J < P$ de base ortonormal $\{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_L\}$ tal que estas direcciones maximicen la varianza de los datos proyectados.

Asumiendo $\boldsymbol{\mu}_x = E(\mathbf{X}) = \mathbf{0}$ sin perdida de generalidad. El resultado matemático que permite la implementación del ACP es que la base del espacio sobre el que queremos proyectar, tal que las direcciones principales conserven la mayor variabilidad, corresponde a los primeros $J < P$ vectores $\{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_J\}$ ortonormales de la descomposición propia de la matriz de covarianza $\boldsymbol{C_V} = E(\boldsymbol{V}\boldsymbol{V}^T)$, descomposición dada por $\boldsymbol{C_V}\boldsymbol{w_j} = \zeta_j\boldsymbol{w_j}, j = 1, ..., P$ con $\zeta_1 \geq \zeta_2 \geq \ldots \geq \zeta_P > 0$. Aquí, ζ_j corresponde al j-ésimo valor propio de $\boldsymbol{C_V}$ para el correspondiente vector propio $\mathbf{w}_j = [w_{j1}, \ldots, w_{jp}]^T$.Este resultado conduce a la siguiente descomposición vectorial truncada de PC

$$\boldsymbol{V} \approx \boldsymbol{V}^{[J]} = \sum_{j=1}^{J} \gamma_j \boldsymbol{w}_j \text{ con } \gamma_j = \langle \mathbf{V}, \mathbf{w}_j\rangle_{\mathbb{R}^p} \tag{3}$$

Donde $J(< P)$ se elige para explicar un porcentaje δ la mayor parte de la variabilidad de $\boldsymbol{V}$, llamamos a γ_j los puntajes de la descomposición que se demuestra que cumplen $\mathrm{E}(\gamma_j) = 0$ y $\mathrm{Var}(\gamma_j) = \kappa_j$. El vector de puntajes $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_J]^T$ se utiliza como una representación de menor dimensión de los datos originales $\boldsymbol{V}$ (Wold et al. 1987). Podemos notar que los primeros vectores $\{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_L\}$ asociados a los primeros valores propios, capturan sucesivamente las direcciones con mayor cantidad de varianza en los datos.

En la práctica, las direcciones principales y los valores propios pueden estimarse a partir de la descomposición propia de la matriz de covarianza muestral de los datos vectoriales observados $\{\boldsymbol{V}_1, ...\boldsymbol{V}_n\}$ como $\hat{\mathcal{C}}_V\hat{\mathbf{w}}_j = \hat{\kappa}_j\hat{\mathbf{w}}_j, j = 1, \ldots, J$ donde $\hat{\boldsymbol{C}}_V = (n-1)^{-1}\sum_{i=1}^n (\mathbf{V}_i - \hat{\boldsymbol{\mu}}_V)(\mathbf{V}_i - \hat{\boldsymbol{\mu}}_V)^T$ con $\hat{\boldsymbol{\mu}}_V = n^{-1}\sum_{i=1}^n \mathbf{V}_i$.

2.4. Análisis de componentes principales para datos funcionales multivariados

2.4.1. Fundamentos teóricos

Tenemos estructura datos $\boldsymbol{f} := \left(f^{(1)}, f^{(2)}, \cdots, f^{(k)}\right) \in \mathcal{F} := \mathcal{T}_1 \times \cdots \times \mathcal{T}_k$ vector compuesto por funciones con posibles diferentes dominios $\mathcal{T}_k \in \mathbb{R}^{d_j}$, en donde cada función $f^{(j)} : \mathcal{T}_j \to \mathbb{R}$ se supone que está en $L^2\left(\mathcal{T}_j\right)$. Sobre $\mathcal{F}$ se define un producto interno, $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ que define la norma $\| \cdot \|_{\mathcal{F}} = \langle \cdot, \cdot \rangle_{\mathcal{F}}^{1/2}$:

$$\langle \boldsymbol{f}_1, \boldsymbol{f}_2 \rangle_{\mathcal{F}} = \langle \boldsymbol{f}_1, \boldsymbol{f}_2 \rangle_{\mathcal{F}} = \sum_{k=1}^{K} \int_{\mathcal{T}_k} f_1^{(k)}\left(t_k\right) f_2^{(k)}\left(t_k\right) dt_k \tag{4}$$

Donde $\boldsymbol{f_1}, \boldsymbol{f_2} \in \mathcal{F}$.Se puede demostrar como en Happ & Greven (2018) que el espacio vectorial $(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mathcal{F}})$ cumple las propiedades para ser espacio de Hilbert separable con funciones cuadrado integrables, sobre el que están bien definidos la función media $\boldsymbol{\mu} := \mathrm{E}(\mathbf{f}) = 0$ y el operador de covarianza $\mathcal{C}_{\boldsymbol{f}} := \mathrm{E}(\mathbf{f} \otimes \mathbf{f})$. La forma del operador de covarianza está dada por $\boldsymbol{\mathcal{C}_f}(\boldsymbol{t}) = \left[\mathcal{C}_{\boldsymbol{f}}^{(1)}(t_1), \mathcal{C}_{\boldsymbol{f}}^{(2)}(t_2), ..., \mathcal{C}_{\boldsymbol{f}}^{(K)}(t_K)\right]^T \in \mathcal{F}$ para:

$$\mathcal{C}_{\boldsymbol{f}}^{(j)}\left(t_j\right) := \sum_{i=1}^{K} \int_{\mathcal{T}_i} \sigma_{ij}\left(s_i, t_j\right) f^{(i)}\left(s_i\right) ds_i = \langle \boldsymbol{\sigma}_{\cdot j}(., t_j), \boldsymbol{f} \rangle_{\mathcal{F}} \tag{5}$$

Donde $\sigma_{ij}\left(s_i, t_j\right) = \mathrm{Cov}\left\{f^{(i)}\left(s_i\right), f^{(j)}\left(t_j\right)\right\} \in \mathcal{T}_i \times \mathcal{T}_j$. El operador de covarianza $\boldsymbol{\mathcal{C}_f}$ cumple además las propiedades necesarias (operador compacto, autoadjunto y definido positivo de Hilbert-Schmid) que garantizan la convergencia y existencia de la representación de Karhunen-Loève en términos de las funciones y valores propios $\{\boldsymbol{\phi}(.), \eta_m\}_{m \in \mathbb{N}}$ de la descomposición del operador de covarianza $\mathcal{C}_{\boldsymbol{\phi}_m}(\mathbf{t}) = \eta_m \boldsymbol{\phi}_m(\mathbf{t})$ con $\eta_1 \geq \eta_2 \geq \ldots \geq 0$ y $\eta_m \to 0$ como $m \to \infty$. Para algún $\boldsymbol{f} \in \mathcal{F}$ esta se puede representar como:

$$\boldsymbol{f}(\mathbf{t}) - \boldsymbol{\mu_f}(\mathbf{t}) = \sum_{m=1}^{\infty} \kappa_m \boldsymbol{\phi}_m(\mathbf{t}) \approx \sum_{m=1}^{L} \kappa_m \boldsymbol{\phi}_m(\mathbf{t}), \quad \mathbf{t} \in \mathcal{F} \tag{6}$$

Con $\kappa_m = \langle \boldsymbol{f}, \boldsymbol{\eta}_m \rangle_{\mathcal{F}}$, $\mathrm{E}\left(\kappa_m\right) = 0$ y $\mathrm{E}\left(\kappa_m \kappa_n\right) = \eta_m \delta_{mn}$, y $\boldsymbol{\phi}_m(\mathbf{t}) = \left[\phi_m^{(1)}(t_1), \phi_m^{(2)}(t_2), ..., \phi_m^{(K)}(t_k)\right]^T$. En la aproximación, L es un número entero positivo finito que permite una aproximación óptima de L dimensiones (descomposición en componentes principales) que captura un porcentaje de variabilidad δ de los datos originales (Happ & Greven 2018).

2.4.2. Implementación práctica

En el artículo de Happ & Greven (2018) propone un esquema de estimación para el análisis de componentes principales para datos funcionales multivariados (MFPCA) donde las funciones no se limitan a ser observadas en el mismo intervalo ni a tener dominios unidimensionales, lo anterior lo logra al establecer una relación teórica directa entre los análisis de componentes principales univariado (FPCA) y multivariado (MFPCA), lo que resulta en una estrategia de estimación sencilla para los componentes principales funcionales multivariantes y las puntuaciones basadas en sus contrapartes univariadas.

El insumo principal para la aproximación para esquema de estimación de MFPCA es pensar en la aproximación de cada uno de los elementos del vector de funciones $\tilde{\boldsymbol{f}} := \left[\tilde{f}^{(1)}, \cdots, \tilde{f}^{(k)}\right]^T$ por medio de las representaciones de Karhunen-Loève univariadas $\tilde{f}^{(j)} = \sum_{i=1}^{L_j} \kappa_i^{(j)} \phi_h^{(j)}$ donde $\mathcal{C}_{\phi_m^{(j)}}^{(j)}(t) = \eta_m^{(j)} \phi_m^{(j)}(t)$ con $\eta_1^{(j)} \geq \eta_2^{(j)} \geq \ldots \geq 0$ y $\eta_m^{(j)} \to 0$ como $m \to \infty$ y $L \leq \sum_{j=1}^{K} L_j =: L_+$. Luego, a partir de esta representación aproximada, se logra tener una forma aproximada para el operador de covarianza $\tilde{\mathcal{C}}_{\tilde{\boldsymbol{f}}}$ y una representación de Karhunen-Loève para $\tilde{\boldsymbol{f}}$.

$$\tilde{\boldsymbol{f}}(\mathbf{t}) - \tilde{\boldsymbol{\mu}}_{\tilde{\boldsymbol{f}}}(\mathbf{t}) = \sum_{m=1}^{\infty} \tilde{\kappa}_m \tilde{\boldsymbol{\phi}}_m(\mathbf{t}) \approx \sum_{m=1}^{L} \tilde{\kappa}_m \tilde{\boldsymbol{\phi}}_m(\mathbf{t}), \quad \mathbf{t} \in \mathcal{F} \tag{7}$$

En la proposición 5 de Happ & Greven (2018), se establece que para la descomposición en valores y funciones propias del operador de covarianza $\tilde{\mathcal{C}}_{\tilde{\boldsymbol{\phi}_m}}(\mathbf{t}) = \tilde{\eta_m}\tilde{\boldsymbol{\phi}}_m(\mathbf{t})$ corresponden a los valores propios positivos de la descomposición $\boldsymbol{V}\mathbf{e}_m = \nu_j \mathbf{e}_m$ para la matriz $\mathbf{V} \in \mathbb{R}^{L_+ \times L_+}$ formada por bloques $\boldsymbol{V}^{(jk)} = \left[\text{Cov}\left(\kappa_m^{(j)}, \kappa_n^{(k)}\right)\right]_{\{m,n\}} \in \mathbb{R}^{L_j \times L_k}$, $m = 1, ..., L_j, n = 1, ..., L_k$, luego $\boldsymbol{V}$ caracteriza las covarianzas dentro y entre las puntuaciones PC funcionales univariadas. Las funciones propias de $\tilde{\mathcal{C}}_{\tilde{\boldsymbol{\phi}_m}}$ y los puntajes para el MFPCA están dadas por:

$$\tilde{\boldsymbol{\phi}}_m = \left[\sum_{n=1}^{L_1} [\boldsymbol{e_m}]_n^{(1)} \phi_n^{(1)}(t_j), ..., \sum_{n=1}^{L_K} [\boldsymbol{e_m}]_n^{(K)} \phi_n^{(K)}(t_j)\right]^T, \tilde{\kappa}_m = \sum_{j=1}^{K}\sum_{n=1}^{L_j} [\boldsymbol{e}_m]_n^{(j)} \kappa_n^{(j)} \tag{8}$$

Con $[\boldsymbol{e_m}]_n^{(j)} \in \mathbb{R}^{L_j}$ el j-ésimo bloque del vector propio (ortonormal) $\boldsymbol{e_m}$ de la matriz $\boldsymbol{V}$ y $\left\{\tilde{\boldsymbol{\phi}}_m, \tilde{\kappa}_m\right\}_{m\in 1,..,L}$ corresponden a las direcciones propias y puntajes (aproximados) para la descomposición en componentes principales truncada de $\tilde{\boldsymbol{f}}$. Luego, los pasos para la implementación práctica pueden ser resumidos como:

1. Para cada elemento $f^{(j)}$ de $\boldsymbol{f}$ estimar la descomposición FPCA univariadas basados en un conjunto de observaciones. Esto nos da como resultado $\left\{\hat{\phi}_m^{(j)}, \hat{\kappa}_m^{(j)}\right\}$ $m = 1, .., L_j, j = 1, .., K$ direcciones propias y puntajes de representación de las observaciones para las K descomposiciones FPCA de cada una de las variables.
2. Se define una matriz $\boldsymbol{\Xi} \in \mathbb{R}^{N \times L_+}$, donde cada fila corresponde a $\left[\hat{\kappa}_{i,1}^{(1)}, ..., \hat{\kappa}_{i,L_1}^{(1)}, ..., \hat{\kappa}_{i,1}^{(K)}, ..., \hat{\kappa}_{i,M_K}^{(K)}\right]$ la concatenación de los puntajes del FPCA de cada variable. Encontrar $\hat{\boldsymbol{V}} = (N-1)^{-1}\boldsymbol{\Xi}^\top\boldsymbol{\Xi}$ estimación dé $\hat{V}$.
3. Realice un análisis de valores y vectores propios de la matriz $\hat{\boldsymbol{V}}$ obteniendo los valores propios ν_m y vectores propios $\boldsymbol{e}_m$, donde el vector propio se puede descomponer en bloques $[\boldsymbol{e}_m]_n^{(j)}$ correspondientes a la parte del vector propio asociado a la variable j.
4. Estime las direcciones principales $\hat{\boldsymbol{\phi}}_m$ y puntajes $\hat{\kappa}_m$ asociados a la descomposición en valores y vectores propios funcional multivariada. Estas estimaciones se pueden realizar mediante las relaciones de Ecuación 8.

2.5. Análisis de componentes principales para datos funcionales híbridos

2.5.1. Fundamentos teóricos

Tenemos estructura datos $\boldsymbol{Z} = (\boldsymbol{f}, \boldsymbol{V}) \in \mathcal{H} = \mathcal{F} \times \mathbb{R}^p$ que cuentan con una parte funcional (multivariada) y otra parte vectorial. La parte vectorial es de la forma $\boldsymbol{V} = [v_1, v_2, \cdots, v_p] \in \mathbb{R}^p$ y la estructura de la parte funcional multivariada de la forma $\boldsymbol{f} := \left[f^{(1)}, f^{(2)}, \cdots, f^{(k)}\right] \in \mathcal{F} := \mathcal{T}_1 \times \cdots \times \mathcal{T}_k$ vector compuesto por funciones con posibles diferentes dimensiones. Técnicamente, $\mathcal{T}_j$ deben ser conjuntos compactos en $\mathbb{R}^{d_j}, d_j \in \mathbb{N}$ con medida (Lebesgue-) finita y $f^{(j)} : \mathcal{T}_j \to \mathbb{R}$ se supone que está en $L^2(\mathcal{T}_j)$.

Sobre la estructura $\boldsymbol{Z}$ se define un producto interno, $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ que define la norma $\| \cdot \|_{\mathcal{H}} = \langle \cdot, \cdot \rangle_{\mathcal{H}}^{1/2}$:

$$\begin{aligned} \langle \mathbf{h}_1, \mathbf{h}_2 \rangle_{\mathcal{H}} &= \langle \boldsymbol{f}_1, \boldsymbol{f}_2 \rangle_{\mathcal{F}} + \omega \langle \mathbf{v_1}, \mathbf{v_2} \rangle \\ &= \sum_{k=1}^{K} \int_{\mathcal{T}_k} f_1^{(k)}(t_k) f_2^{(k)}(t_k) \, dt_k + \omega \sum_{r=1}^{p} v_{1r} v_{2r} \end{aligned} \tag{9}$$

Donde $\boldsymbol{h_1} = (\boldsymbol{f}_1, \boldsymbol{v_1})$ y $\boldsymbol{h_2} = (\boldsymbol{f}_2, \boldsymbol{v_2}) \in \mathcal{H}$. Adicionalmente, se incluye una ponderación $\omega > 0$ para tener en cuenta la heterogeneidad entre la parte funcional y vectorial en términos de escala de medición y variación. Se logra demostrar en Jang (2021) que el espacio vectorial $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ cumple las propiedades para ser espacio de Hilbert separable con $\mathrm{E}\|\mathbf{Z}\|^2_{\mathcal{H}} < \infty$, sobre el que están definidos $\boldsymbol{\mu_Z} = \mathrm{E}(\mathbf{Z})$ y $\boldsymbol{\mathcal{C}_h} = \mathrm{E}(\mathbf{Z} \otimes \mathbf{Z})$, donde además $\boldsymbol{\mathcal{C}_h}$ pertenece al espacio de operadores de Hilbert-Schmid .

Ahora, Jang (2021) determinan una forma explícita para el operador de covarianza, $\boldsymbol{\mathcal{C}_h}$, bajo la perspectiva de que $\{\mathbf{Z}(\mathbf{t}) : \mathbf{t} \in \mathcal{F}\}$ se puede tratar como un proceso estocástico $(K+p)$-dimensional con $\mathbf{Z}(\mathbf{t}) = [f(\mathbf{t}), \mathbf{V}]^T = \left[f^{(1)}(t_1), \ldots, f^{(k)}(t_K), v_1, \ldots, v_p\right]^T \in \mathbb{R}^{K+p}$ respecto a $\mathbf{t} = [t_1, \ldots t_k]^T$. La forma explícita del operador de covarianza $\mathcal{C}_{\mathbf{h}}(\mathbf{t}) = \left[\mathcal{C}^{(1)}_{\mathbf{h}}(t_1), \mathcal{C}^{(K)}_{\mathbf{h}}(t_K), (\mathcal{C}_{\mathbf{h}})_1, \ldots, (\mathcal{C}_{\mathbf{h}})_p\right]^T \in \mathbb{R}^{K+p}$ donde sus componentes están dados por:

$$
\begin{aligned}
\mathcal{C}^{(k)}_{\mathbf{h}}(t_k) &= \left\langle \boldsymbol{\Sigma}^{(k)}(t_k, \cdot), \mathbf{h} \right\rangle_{\mathcal{H}} = \sum_{u=1}^{K} \left\langle \sigma^{(uk)}_{y}(., t_k), f^{(u)} \right\rangle_{\mathcal{T}_k} + \left\langle \boldsymbol{\sigma}^{(k)}_{yx}(t_k, .), \boldsymbol{v} \right\rangle \\
&= \sum_{u=1}^{K} \int_{\mathcal{T}_u} \sigma^{(uk)}_{y}(s_u, t_k) f^{(u)}(s_u)\, ds_u + \sum_{r=1}^{p} \sigma^{(k)}_{yx}(t_k, r)\, v_r \\
(\mathcal{C}_{\mathbf{h}})_r &= \langle \boldsymbol{\Sigma}(\cdot, r), \mathbf{h} \rangle_{\mathcal{H}} = \sum_{u=1}^{K} \left\langle \sigma^{(u)}_{yx}(., r), f^{(u)} \right\rangle_{\mathcal{T}_k} + \langle \boldsymbol{\sigma}_x(., r), \boldsymbol{v} \rangle \\
&= \sum_{u=1}^{K} \int_{\mathcal{T}_u} \sigma^{(u)}_{yx}(s_u, r) f^{(u)}(s_u)\, ds_u + \sum_{q=1}^{p} \sigma_x(q, r) v_q
\end{aligned}
\tag{10}
$$

Donde $\sigma^{(uk)}_{y}(s_u, t_k) = \mathrm{Cov}\left\{Y^{(u)}(s_u), Y^{(k)}(t_k)\right\} \in \mathcal{T}_u \times \mathcal{T}_k$ (covarianza entre dos funciones), $\sigma_x(q, r) = \mathrm{Cov}(X_q, X_r) \in \mathbb{R}$ (covarianza entre elementos vectoriales) y $\sigma^{(k)}_{yx}(t_k, r) = \mathrm{Cov}\left\{Y^{(k)}(t_k), X_r\right\} \in \mathcal{T}_k$ (covarianza cruzada entre vector y función) para $s_u \in \mathcal{T}_u, t_k \in \mathcal{T}_k, u, k = 1, \ldots, K$.

Dado que Jang (2021) muestran que se cumplen las condiciones sobre el operador de covarianza $\mathcal{C}_{\mathbf{h}}(\mathbf{t})$ (operador compacto, autoadjunto y definido positivo de Hilbert-Schmid) se demuestra que se puede garantizar en este caso la convergencia y la existencia de representación de Karhunen-Loève en términos de las funciones y valores propios $\{\boldsymbol{\xi}_m, \lambda_m\}_{m \in \mathbb{N}}$ de la descomposición del operador de covarianza $\boldsymbol{\mathcal{C}_{\xi_m}}(\mathbf{t}) = \lambda_m \boldsymbol{\xi}_m(\mathbf{t})$ con $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ y $\lambda_m \to 0$ como $m \to \infty$.

$$
\mathbf{Z}(\mathbf{t}) - \boldsymbol{\mu_Z}(\mathbf{t}) = \sum_{m=1}^{\infty} \rho_m \boldsymbol{\xi}_m(\mathbf{t}) \approx \sum_{m=1}^{M} \rho_m \boldsymbol{\xi}_m(\mathbf{t}), \quad \mathbf{t} \in \mathcal{F} \tag{11}
$$

Con $\rho_m = \langle \mathbf{Z}, \boldsymbol{\xi}_m \rangle_{\mathcal{H}}$, $\mathrm{E}(\rho_m) = 0$ y $\mathrm{E}(\rho_m \rho_n) = \lambda_m \delta_{mn}$, y $\boldsymbol{\xi}_m(\mathbf{t}) = [\boldsymbol{\psi}_m(\mathbf{t}), \boldsymbol{\theta}_m]^T$. En la aproximación M es un número entero positivo finito que permite una aproximación óptima de M dimensiones (descomposición en componentes principales para datos híbridos) (Jang 2021).

2.5.2. Implementación práctica

En el artículo de Jang (2021) se propone un esquema de estimación para el análisis de componentes principales para datos funcionales multivariados híbridos (MFPCAH), siguiendo un esquema similar al planteado anteriormente por Happ & Greven (2018). El esquema se basa en la existencia de una relación teórica directa entre los análisis de componentes principales para la parte funcional multivariada (MFPCA) y para la parte vectorial (PCA), lo que resulta en una estrategia de estimación sencilla para los componentes principales funcionales multivariantes híbridos y las puntuaciones basadas en los análisis individuales MFPCA y PCA.

La herramienta principal para la aproximación de un esquema de estimación para el MFPCAH es pensar en aproximar los elementos híbridos $\boldsymbol{Z} = (Y, \boldsymbol{X})$ a su versión $\tilde{\boldsymbol{Z}} = (\tilde{\boldsymbol{f}}, \tilde{\boldsymbol{V}}) = \left(\sum_{h=1}^{L} \kappa_h \boldsymbol{\phi}_h, \sum_{j=1}^{J} \gamma_j \mathbf{w}_j\right)$, en esta representación $\{\kappa_m, \boldsymbol{\phi}_m\}_{m\in\mathbb{N}}$ corresponden a los puntajes y direcciones propias de la descomposición en componentes principales truncada de la parte funcional (MFPCA) y $\{\gamma_m, \mathbf{w}_m\}_{m\in\mathbb{N}}$ de la parte vectorial (PCA). Luego, a partir de esta representación aproximada, se logra obtener la forma aproximada para el operador de covarianza $\tilde{\mathcal{C}_{\mathbf{h}}}$ y la representación de Karhunen-Loève para $\tilde{\boldsymbol{Z}}$ (para más detalles, ver Jang (2021)):

$$\tilde{\mathbf{Z}}^{[M]}(\mathbf{t}) = \sum_{m=1}^{M} \tilde{\rho}_m \tilde{\boldsymbol{\xi}}_m(\mathbf{t}), \quad \mathbf{t} \in \mathcal{F} \tag{12}$$

En el Teorema 4 de Jang (2021), se establece que para la descomposición en valores y funciones propias del operador de covarianza $\tilde{\mathcal{C}}_{\tilde{\boldsymbol{\xi}_m}}(\mathbf{t}) = \tilde{\lambda_m} \tilde{\boldsymbol{\xi}}_m(\mathbf{t})$ corresponden a los valores propios positivos de la descomposición $\boldsymbol{V}\mathbf{e}_m = \nu_j \mathbf{e}_m$, donde $\tilde{\boldsymbol{V}}$ es una matriz que caracteriza las covarianzas dentro y entre las puntuaciones PC funcionales y vectoriales. Las funciones propias de $\tilde{\mathcal{C}}_{\tilde{\boldsymbol{\xi}_m}}$ y los puntajes para el MFPCA están dadas por:

$$\tilde{\boldsymbol{\xi}}_m = \left(\sum_{h=1}^{L} c_{mh} \boldsymbol{\phi}_h(\mathbf{t}), \sum_{j=1}^{J} d_{mj} \mathbf{w}_j\right) \quad , \tilde{\rho}_m = \sum_{h=1}^{L} \kappa_h c_{mh} + \sum_{j=1}^{J} \gamma_j d_{mj} \tag{13}$$

Con $\mathbf{e}_m = \left[\mathbf{c}_m^T, \mathbf{d}_m^T\right]^T$ vector propio de la matriz $\boldsymbol{V}$, donde $\mathbf{c}_m = [c_{m1}, \ldots, c_{mL}]^T$ denota los primeros L elementos de $\mathbf{e}_m$, $\mathbf{d}_m = [d_{m1}, \ldots, d_{mJ}]^T$ denota los siguientes J elementos de $\mathbf{e}_m$ y $\left\{\tilde{\boldsymbol{\xi}}_m, \tilde{\rho}_m\right\}_{m\in 1,..,M}$ corresponden a las M direcciones propias y puntajes aproximados para la descomposición en componentes principales truncada aproximada para $\tilde{\boldsymbol{Z}}$. Luego, los pasos para la implementación práctica están dados por:

1. Se realiza la estimación de un ACPF para obtener los puntajes para cada observación y cada función, teniendo en cuenta los diferentes tipos de funciones.
2. Obtenga la estimación del ACP para las componentes del vector multidimensional, y extraiga los puntajes de cada componente y para cada individuo.
3. Se define una matriz $\boldsymbol{\Xi} \in \mathbb{R}^{N\times(L+J)}$, donde cada fila concatena los puntajes CP de la parte funcional y los puntajes de CP vectoriales. Encontrar $\hat{\boldsymbol{V}} = (N-1)^{-1}\boldsymbol{\Xi}^\top\boldsymbol{\Xi}$ estimación dé $\hat{V}$ del Teorema 4 anteriormente presentado.
4. Realice un análisis de valores y vectores propios de la matriz $\boldsymbol{V}$ para obtener los valores propios ν_m y vectores propios, $\boldsymbol{e}_m = [\hat{c_m}, \hat{d_m}]$ donde el vector propio se descompone en las L componentes del vector e_m correspondiente a los vectores propios de la parte funcional y las J componentes correspondientes a la parte multidimensional.
5. Estime la parte funcional y vectorial del vector propio híbrido $\boldsymbol{\xi}_m = (\boldsymbol{\psi}_m, \boldsymbol{\theta}_m)$, asociados a la descomposición en valores y vectores propios funcional multivariada hibrida. Estas estimaciones se pueden realizar mediante las relaciones dadas en Ecuación 13.

2.6. Rango de Wilcoxon

Es de interés alguna prueba para comparar localización libre de distribución, en el sentido que la distribución de la estadística de la prueba no dependa de F_0 (distribución de los datos bajo la hipótesis nula) en el caso de datos multivariados. Existen procedimientos que se construyen con base en el signo o rango espacial (Oja 2010) pero solo son libres de distribución bajo ciertas ciertos modelos y supuestos, como se discute en Zou et al. (2011). Otros conjuntos de pruebas tratan el problema por medio de

medidas de profundidad sobre las que se aplican procedimientos uni variados conocidos (Wilcox 2005) sin embargo, dependen de las funciones de profundidad seleccionadas con base en características sobre la distribución de los datos. Por ello, en un primer momento nos interesamos en un tipo de prueba basada en la combinación de las pruebas de rango de Wilcoxon para cada componente mencionada en Chen et al. (2016).

En el contexto de la prueba, contamos con dos muestras independientes $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\}, \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n\}$, $\boldsymbol{y}_i, \boldsymbol{x}_i \in \mathbb{R}^p$ que se distribuyen como $F_0(\mathbf{t}; \boldsymbol{\mu}_0)$ y $F_0(\mathbf{t}; \boldsymbol{\mu}_1)$, estamos interesados en juzgar la hipótesis $H_0 : \boldsymbol{\mu}_0 = \boldsymbol{\mu}_1$ vs $H_1 : \boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$.Luego se plantea el siguiente estadístico de prueba basado en los rangos de las observaciones dentro de cada uno de los componentes:

$$T_j = \frac{\sum_{i=1}^{m} R_{ji} - m(m+n+1)/2}{\sqrt{mn(m+n+1)/12}}, j = 1, ..., p \tag{14}$$

Donde R_{ij} es el rango de X_{ij} sobre la muestra conjunta $\{x_{j1}, x_{j2}, \cdots, x_{jm}, y_{j1}, y_{j2}, \cdots, y_{jn}\}$. A continuación, combinamos toda la información de T_j para $j = 1, ..., p$, mediante el uso de 'max' o $\sum$. Por ejemplo, considerar el estadístico $T = \sum_j T_j^2$ y se esperaría que bajo H_0 resulte un valor de T grande. Luego, rechazamos H_0 para valores $T > \omega$ tal que $P_{H_0}(T > \omega) = \alpha$ con α determinado para un cierto nivel de confianza, dicho valor ω se puede determinar por procedimientos de permutación.

Este procedimiento resulta ser robusto a valores extremos y cumple la propiedad de ser incondicionalmente libres de distribución. Es libre de distribución en el sentido de que $P_{H_0}(T > \omega) = \alpha$ es válida independientemente de la elección de F_0, utilizando un umbral dependiente de los datos (procedimiento de permutación). Es una prueba que cumple la propiedad de consistencia, como se ve en (Chen et al. 2016). En el contexto del control de calidad, para tamaños m, n pequeños, la prueba es capaz de ofrecer un buen poder de detección comparándola con T^2 de Hotelling y otras pruebas de localización no paramétricas multivariantes construidas en el contexto de normalidad, como se puede notar en Chen et al. (2016).

3. Objetivos

3.1. Objetivo General

Desarrollar una carta de control libre de distribución para el monitoreo y supervisión en línea de procesos donde las observaciones estén conformadas de datos vectoriales y conjuntos de perfiles no lineales multivariados con estructuras de correlación entre ambas partes (datos híbridos).

3.2. Objetivos Específicos

1. Definir y justificar la parte teórica de un esquema de control de calidad para datos funcionales multivariados híbridos con base en los puntajes obtenidos del análisis de componentes principales para datos híbridos.

2. Construir un algoritmo computacionalmente efectivo para el cálculo de la carta de control en datos funcionales multivariados híbridos y sus límites de control.

3. Comparar el rendimiento del nuevo gráfico de control contra otros esquemas de control para datos funcionales multivariados y datos vectoriales, lo anterior en términos de ARL_0 y longitud de corrida.

4. Realizar la aplicación de la metodología planteada sobre un conjunto de datos reales.

Actividades	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov
Revisión bibliográfica	x	x	x	x					
Desarrollo teórico					x	x			
Análisis computacional							x	x	
Redacción del documento					x	x	x	x	
Socialización del trabajo									x

4. Cronograma

Referencias

Chen, N., Zi, X. & Zou, C. (2016), 'A distribution-free multivariate control chart', *Technometrics* **58**(4), 448–459.

Fan, S.-K. S., Jen, C.-H. & Lee, T.-Y. (2017), 'Modeling and monitoring the nonlinear profile of heat treatment process data by using an approach based on a hyperbolic tangent function', *Quality Engineering* **29**(2), 226–243.

Ferraty, F. & Vieu, P. (2006), *Nonparametric functional data analysis: theory and practice*, Vol. 76, Springer.

Ghashghaei, R., Amiri, A. & Khosravi, P. (2019), 'New control charts for simultaneous monitoring of the mean vector and covariance matrix of multivariate multiple linear profiles', *Communications in Statistics-Simulation and Computation* **48**(5), 1382–1405.

Happ, C. & Greven, S. (2018), 'Multivariate functional principal component analysis for data observed on different (dimensional) domains', *Journal of the American Statistical Association* **113**(522), 649–659.

Horváth, L. & Kokoszka, P. (2012), *Inference for functional data with applications*, Vol. 200, Springer Science & Business Media.

Hung, Y.-C., Tsai, W.-C., Yang, S.-F., Chuang, S.-C. & Tseng, Y.-K. (2012), 'Nonparametric profile monitoring in multi-dimensional data spaces', *Journal of Process Control* **22**(2), 397–403.

Jang, J. H. (2021), 'Principal component analysis of hybrid functional and vector data', *Statistics in medicine* **40**(24), 5152–5173.

Jensen, W. A., Birch, J. B. & Woodall, W. H. (2008), 'Monitoring correlation within linear profiles using mixed models', *Journal of Quality Technology* **40**(2), 167–183.

Lei, Y., Zhang, Z. & Jin, J. (2010), 'Automatic tonnage monitoring for missing part detection in multi-operation forging processes', *Journal of manufacturing science and engineering* **132**(5).

Montgomery, D. C. (2020), *Introduction to statistical quality control*, John Wiley & Sons.

Noorossana, R., Eyvazian, M. & Vaghefi, A. (2010), 'Phase ii monitoring of multivariate simple linear profiles', *Computers & Industrial Engineering* **58**(4), 563–570.

Oja, H. (2010), *Multivariate nonparametric methods with R: an approach based on spatial signs and ranks*, Springer Science & Business Media.

Pan, J.-N., Li, C.-I. & Lu, M. Z. (2019), 'Detecting the process changes for multivariate nonlinear profile data', *Quality and Reliability Engineering International* **35**(6), 1890–1910.

Paynabar, K., Jin, J., Agapiou, J. & Deeds, P. (2012), 'Robust leak tests for transmission systems using nonlinear mixed-effect models', *Journal of quality technology* **44**(3), 265–278.

Paynabar, K., Jin, J. & Pacella, M. (2013), 'Monitoring and diagnosis of multichannel nonlinear profile variations using uncorrelated multilinear principal component analysis', *Iie transactions* **45**(11), 1235–1247.

Paynabar, K., Zou, C. & Qiu, P. (2016), 'A change-point approach for phase-i analysis in multivariate profile monitoring and diagnosis', *Technometrics* **58**(2), 191–204.

Qiu, P. (2013), *Introduction to statistical process control*, CRC press.

Ramsay, J. O. & Dalzell, C. (1991), 'Some tools for functional data analysis', *Journal of the Royal Statistical Society: Series B (Methodological)* **53**(3), 539–561.

Ryan, T. P. (2011), *Statistical methods for quality improvement*, John Wiley & Sons.

Soleimani, P., Noorossana, R. & Niaki, S. (2013), 'Monitoring autocorrelated multivariate simple linear profiles', *The International Journal of Advanced Manufacturing Technology* **67**(5), 1857–1865.

Wilcox, R. R. (2005), 'Depth and a multivariate generalization of the wilcoxon-mann-whitney test', *American Journal of Mathematical and Management Sciences* **25**(3-4), 343–363.

Williams, J. D., Woodall, W. H. & Birch, J. B. (2007), 'Statistical monitoring of nonlinear product and process quality profiles', *Quality and Reliability Engineering International* **23**(8), 925–941.

Wold, S., Esbensen, K. & Geladi, P. (1987), 'Principal component analysis', *Chemometrics and intelligent laboratory systems* **2**(1-3), 37–52.

Zhang, J., Ren, H., Yao, R., Zou, C. & Wang, Z. (2015), 'Phase i analysis of multivariate profiles based on regression adjustment', *Computers & Industrial Engineering* **85**, 132–144.

Zou, C., Jiang, W. & Tsung, F. (2011), 'A lasso-based diagnostic framework for multivariate statistical process control', *Technometrics* **53**(3), 297–309.

Zou, C., Tsung, F. & Wang, Z. (2007), 'Monitoring general linear profiles using multivariate exponentially weighted moving average schemes', *Technometrics* **49**(4), 395–408.

Zou, C., Tsung, F. & Wang, Z. (2008), 'Monitoring profiles based on nonparametric regression methods', *Technometrics* **50**(4), 512–526.