Pagura, José Alberto
Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística.

EL USO DE COMPONENTES PRINCIPALES EN LA MONITORIZACION Y DIAGNOSTICO DE PROCESOS INDUSTRIALES

1) INTRODUCCIÓN

El control estadístico de los procesos, en su forma tradicional, analiza un pequeño conjunto de variables de calidad de forma individual y monitoriza el proceso, realizando a partir de las señales de fallo que se detecten, las acciones correctivas pertinentes. El estudio de las variables en forma separada omite el hecho de que una anormalidad en el proceso puede evidenciarse no solo por valores inusuales en cada una de las características que se estudian, sino también por distorsiones en las relaciones entre ellas. Estas variables pueden analizarse en forma conjunta lo que permite un mejoramiento en la detección de anomalías. El tratamiento puede enriquecerse con la aplicación de métodos que permiten encontrar las variables responsables de tales situaciones.

Por otra parte, en muchas industrias en la actualidad, puede disponerse de información no sólo de variables de calidad de los productos sino también de una importante cantidad de datos del proceso, posiblemente relacionados con las características de calidad. La disponibilidad de estos datos puede ser aprovechada con diferentes propósitos: controlar el sistema monitorizando las variables del proceso, predecir valores de calidad del producto a través de las variables del proceso, determinar las condiciones en las que debe funcionar un proceso para obtener mejoras en la calidad, etc.

La aplicación de métodos multivariados que permiten la reducción de las dimensiones, entre ellos el análisis en componentes principales, aparecen como herramientas adecuadas para el mayor aprovechamiento de la información disponible.

En este trabajo, se discute el uso de componentes principales para la determinación de variables responsables de una señal de salida de control y la monitorización y diagnóstico de procesos, mostrando sus posibilidades y ventajas para la mejora de la calidad.

2) EL CONTROL MULTIVARIANTE DE PROCESOS

El objetivo del control estadístico de procesos es monitorizar el comportamiento de los mismos con el fin de detectar cualquier evento inusual que desvíe el proceso de su funcionamiento normal (bajo control estadístico), pudiendo comprometer tanto la calidad del producto producido, como la seguridad del proceso de fabricación. Una vez identificada la causa que provocó la anomalía, es posible corregirla teniendo como consecuencia una mejora del proceso y por ende, de la calidad de los productos.

Los procedimientos que se utilizan en la práctica corriente, se basan en gráficos de control para un pequeño número de variables, tratándose por lo general de características de calidad del producto final. Los divulgados gráficos de Shewhart, CUSUM o EWMA son las herramientas para llevar adelante esta clase de control.

Cuando la calidad del producto está definida por varias variables, éstas pueden ser estudiadas en forma simultánea ya que una anomalía en el proceso puede estar anunciada no sólo por la detección de una salida de control en los gráficos correspondientes a algunas de las variables, sino también por una anomalía en la relación existente entre las variables que se estudian. Si se observan M características de calidad, distribuídas como una normal multivariante con vector de esperanzas $\boldsymbol{\mu}$ y matriz de covariancias $\boldsymbol{\Sigma}$, que abreviaremos con $N_M(\boldsymbol{\mu},\boldsymbol{\Sigma})$, es posible dar un tratamiento basado en las distribuciones χ^2 y T^2 de Hotelling.

Considérese un vector **z** de M componentes correspondiente a las variables que se estudian, distribuído según $N_M(\boldsymbol{\mu},\boldsymbol{\Sigma})$ en la situación "bajo control". La estadística :

$$\chi^2_o=(\mathbf{z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu}) \ ,$$

se distribuye como una chi-cuadrado con M grados de libertad, $\chi^2_{(M)}$. Con esta referencia, ante una nueva observación se puede probar la hipótesis de pertenencia a dicha población normal comparando el χ^2_o observado, con uno fijado según el nivel de significación deseado. Esto significa que se puede entonces fijar un límite superior de control igual al valor que antiacumula α en la distribución chi cuadrado con M grados de libertad, $\chi^2_{\alpha,M}$, fijando α en 0,01 o 0,05 y luego graficar las observaciones obtenidas en los diferentes momentos en los que se realizan las mediciones para el control, entendiéndose como señal de fuera de control a un valor de χ^2_o superior al límite fijado.

Si se desconoce la matriz de covariancias $\boldsymbol{\Sigma}$, puede estimarse a partir de una muestra calculando la matriz de covariancias muestral S, y entonces considerar

$$T^2=(\mathbf{z}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\mathbf{z}-\boldsymbol{\mu})$$

que sigue una distribución T^2 de Hotelling con m-1 grados de libertad, donde m es el tamaño de la muestra con el que fuera calculada **S**, actuando en forma análoga al caso anterior.

Una expresión útil para fijar el límite superior de control es la que relaciona a la distribución T^2 con la F de Snedecor. Este enfoque permite, a partir de un conjunto de

$$T^2 = \frac{(N^2-1)M}{N(N-M)} F_{M,\,N-M}$$

variables de calidad del producto, detectar algún evento inusual en el proceso que afecte la calidad del producto.

Una forma de complementar la propuesta tradicional, que será tratada en el apartado siguiente, es utilizar procedimientos que permitan identificar las variables que son responsables de la señal de falta de control.

Otro aspecto a tratar, está sugerido por el hecho de que en muchos de los procesos industriales actuales se puede disponer de una importante cantidad de variables del proceso que, seguramente, están relacionadas con las variables de calidad del producto. Esta información se encuentra disponible en momentos anteriores a aquellos en los que pueden obtenerse las variables de calidad, y conociendo las condiciones operativas normales podrían monitorizar en forma anticipada al hallazgo de los fallos en las variables de calidad. Para ello, se puede establecer en las condiciones normales de operación, el comportamiento probabilístico de las mismas y establecer límites de control. Un valor observado superior al fijado como límite de control indicará una anomalía. En caso de conocer el valor poblacional de la matriz de covariancias, el límite se fijará a partir de la distribución χ^2.

3) VARIABLES RESPONSABLES DE LAS SEÑALES DE SALIDAS DE CONTROL

Al encontrar una señal de salida de control, ya sea monitorizando las variables del proceso, o las variables de calidad, debe decidirse el curso de acción a seguir, para lo cual deben conocerse los factores que provocaron dicha anomalía. Kourti y Mac Gregor (1996) analizan diferentes propuestas para detectar qué variables son las responsables de una señal de salida de control y enuncian algunos métodos que serán considerados en este trabajo.

Llamando con **z** al vector de observaciones de calidad o de observaciones del proceso, uno de los caminos posibles a seguir es calcular el error normalizado para cada variable de la observación i: $(z_{ij}-\mu_j)/\sigma_j$, i=1...N, graficar sus valores para cada variable j, con j=1...M si se trata de variables de calidad o j=1...K si corresponden al proceso. Las variables que tengan grandes errores normalizados deberán ser estudiadas. Este análisis sobre los errores normalizados sirve para evidenciar salidas de control sólo por valores inusuales, dejando de lado las vinculadas a la correlación entre las variables. Esto significa, que puede haber observaciones que tengan los valores de las variables, en forma separada, dentro de los límites de control correspondientes a los diagramas univariados, pero estos valores pueden estar desproporcionados.

Una opción que mejora la propuesta anterior es la construcción de un gráfico de los llamados "scores" normalizados.

Los "**scores" normalizados** surgen de la posibilidad de descomponer la T^2 utilizando los autovalores λ_j, j=1...J de la matriz de covariancias S y combinaciones lineales de las variables **z** a partir del siguiente razonamiento

Sea $T^2=(z-\mu)'S^{-1}(z-\mu)$

La matriz S puede diagonalizarse a partir de sus vectores propios, pudiéndose escribir como:

$S=PD_\lambda P'$, siendo P la matriz cuyas columnas son vectores propios normalizados de S, y D_λ matriz diagonal cuyos elementos $d_{ii}=\lambda_i$ valor propio de P. La matriz P verifica: $P^{-1}=P'$.

Entonces, $S^{-1}= PD_\lambda P'$

De esta forma, T^2 se puede expresar como:

$T^2=(z-\mu)'PD_\lambda^{-1}P'(z-\mu)$

Llamando scores t a $(z-\mu)'P$, puede escribirse:

$$T^2 = \sum_{j=1}^{J} \frac{t_j^2}{\lambda_j} = \sum_{j=1}^{J} \frac{t_j^2}{s_{tj}^2}$$

Esta descomposición es llamada "descomposición en componentes principales". La estadística T^2 queda expresado como una suma de razones donde los denominadores λ_j, i=1...J son todos los autovalores no nulos de la matriz de covariancias **S** y los numeradores t_i,llamdos, scores. Se obtienen como combinaciones lineales de las **z** con coeficientes iguales a las componentes de los autovectores normalizados correspondientes a cada autovalor ($t_i=\mathbf{p'}_i\,\mathbf{z}$ donde $\mathbf{p'}_i$ es el vector propio asociado a λ_i). El número máximo de valores y vectores propios no nulos es el mínimo entre N-1 y J.

Cada score t_a puede también expresarse como sigue:

$$t_j = p'_j(z-\mu) = \sum_{k=1}^{J} p_{j,k}(z_k - \mu_k)$$

Estos scores están distribuídos normalmente, por ser combinaciones lineales de variables normales con matriz de covariancias diagonal $\mathbf{D_\lambda}$, es decir $\mathbf{D_\lambda}=\{d_{ij}\}$con $d_{ij}=\delta_{ij}\lambda_j$, j=1...M y δ_{ij}=1 si i=j, y es 0 en otro caso.

La estadística T^2 obtenido a partir de los valores de **z** es igual a la que se obtiene con los scores de las componentes principales tomando todos los términos de la suma.

Los scores normalizados, es decir divididos por su desviación típica, $\sqrt{\lambda_a}$, se representan en un gráfico de barras. Aquéllos con valores altos son tenidos en cuenta para el estudio de causas asignables. Una forma de establecer un límite indicativo para decidir qué valor es alto, es fijarlos según el criterio de Bonferroni, para mantener el nivel global de significación fijado, es decir contrastarlos con el valor de tabla que corresponde al nivel α/M, donde M es el número de scores.

Dado que los scores son combinaciones lineales de las variables originales **z**, cuando se pretenda interpretarlos con el fin de determinar que variables son responsables de la señal de anomalía, pueden presentarse problemas. Se propone entonces graficar, para aquellas observaciones con scores altos, las contribuciones de las variables en la composición de los mismos.

La **contribución de una variable z_j al score $t_{j'}$** es $p_{j'j}(z_j-\mu_j)$. Luego, las contribuciones grandes en valor absoluto y que llevan el mismo signo que el score considerado corresponderán a las variables que afectan directamente a la T^2_o que permitió detectar la anomalía.

Es frecuente encontrar más de un score con valor absoluto alto, por lo que se deben estudiar las contribuciones de cada variable en forma separada . Para simplificar esta tarea y disponer de una herramienta para la rápida identificación, se propone el cálculo de las llamadas **contribuciones totales** de cada variable a todos los scores con valores altos, pero teniendo en cuenta solo aquellas variables con contribuciones de igual signo que el score obtenido. El cálculo se puede esquematizar en los siguientes pasos:

Para los K^* scores altos:

$$cont_{j',j} = \frac{t_{j'}}{s^2_{j'}} p_{j'j}(z_j - \mu_j)$$

Calcular la contribución de una variable z_j al score normalizado $(t_{j'}/s_{tj'})^2$

$Cont_{j'j}$ tomará el valor calculado si tiene el mismo signo que el score $t_{j'}$ y será igual a 0 en caso contrario

Luego, calcular la contribución de la variable z_j como:

$$CONT_j = \sum_{j'=1}^{J} Cont_{j'j}$$

Esta última cantidad es la que permite identificar aquellas variables con altas contribuciones a la anomalía.

La utilización de las estadísticas χ^2 o de T^2 puede presentar problemas en los casos de tener muchas variables y/o alta colinealidad. Σ o S pueden no ser invertibles. Por otra parte, la alta colinealidad provocará una mayor sensibilidad en las direcciones de los menores valores propios. Si se presentan datos faltantes también se dificulta el tratamiento. Una forma de tratar con el problema de la dimensionalidad puede ser subdividir el sistema en partes para las que tengan sentido los análisis por separado, pero esto no garantiza la solución a los problemas generados por la colinealidad.

Una solución aceptable para estas situaciones, la proporcionan algunos métodos de proyección como el análisis en componentes principales (PCA).

4) ANALISIS EN COMPONENTES PRINCIPALES

Este método permite explicar la variabilidad existente en los datos de una matriz **Z**, de N observaciones y M variables, por medio de un número menor de variables latentes, combinación lineal de las primeras.

Desde el punto de vista geométrico, las filas de la matriz **Z** conforman una nube de puntos en el espacio M dimensional. Obtener las componentes principales es encontrar un subespacio de proyección sobre el cuál la nube dé la mínima deformación posible.

Si M<N-1, pueden encontrarse hasta M componentes principales, es decir tantas como número de variables originales, pero el interés estará en seleccionar un número A mucho menor que M perdiendo lo menos posible de la variabilidad existente en el conjunto original de los datos.

La primera de estas variables latentes o primera componente principal $\mathbf{t_1}$ se obtiene buscando la combinación lineal de las variables originales $\mathbf{t_1=p_1z}$, que tiene mayor variancia. Esta se expresa para el conjunto total de individuos como $\mathbf{t_1=Zp_1}$ donde el vector $\mathbf{p_1}$ es el vector propio correspondiente al mayor valor propio λ_1 de la matriz de covariancias de **z**, $\mathbf{\Sigma_z}$, y define la dirección de mayor variabilidad en el espacio de las M variables. Las componentes del vector $\mathbf{p_1}$ son denominadas cargas. El vector $\mathbf{t_1}$ de scores definido anteriormente, representa la proyección de cada observación sobre la dirección definida por $\mathbf{p_1}$. Además, la variancia del score $\mathbf{t_1}$ es igual a λ_1, ya que:

$$\sigma^2_{t_1} = \mathbf{p_1}E(\mathbf{zz'})\mathbf{p_1} = \mathbf{p_1}\Sigma_{\mathbf{z}}\mathbf{p_1} = \mathbf{p_1'}\lambda_1\mathbf{p_1} = \lambda_1$$

La segunda componente principal o variable latente es una combinación lineal de las variables originales $\mathbf{t_2=Zp_2}$, que se obtiene de forma tal que sea ortogonal a la primera, con la mayor variancia posible, inferior a la de $\mathbf{t_1}$. Siguiendo los mismos criterios pueden encontrarse las restantes componentes principales.

Dada una observación i, los scores t_{i1}, t_{i2},..., t_{iA}, ubican a dicha observación en el subespacio definido por los A primeros autovectores $\mathbf{p_1}$, $\mathbf{p_2}$,..., $\mathbf{p_A}$, que son los asociados a los A mayores valores propios.

Si el cálculo de las componentes principales se realiza con los datos originales, las diferentes escalas en que están medidas las variables puede tener influencia decisiva en las direcciones que se obtienen. Por esta razón, previamente al cálculo de las componentes principales, los datos se centran y escalan restando a cada variable su media y dividiendo por su desviación estándar (aunque puede utilizarse otra forma de escalamiento).

Si se extraen todas las componentes principales, **T**=**ZP**, donde **T** es la matriz NxM, cuyas columnas son los M vectores de scores y **P** es la matriz MxM cuyas columnas son los

M vectores propios que constituye una base ortonormal. Postmultiplicando por **P**' se tiene que: **Z**=**TP**'. Operando luego por bloques, la matriz **Z** puede expresarse a partir de los scores **t** y de los autovectores **p**, de la siguiente forma:

$$\mathbf{Z} = \sum_{j=1}^{J} \mathbf{t}_j \mathbf{p}'_j = \sum_{j=1}^{A} \mathbf{t}_j \mathbf{p}'_j + \sum_{j=A+1}^{J} \mathbf{t}_j \mathbf{p}'_j = \hat{\mathbf{Z}} + \mathbf{E}$$

donde $\hat{\mathbf{Z}}$ es la estimación de **Z** utilizando solamente A componentes principales, asociadas a los A mayores valores propios, de las M posibles. Es decir, puede utilizarse un subconjunto A de componentes para reconstruir los valores originales con un cierto error.

Si bien el análisis en componentes principales es un método ampliamente divulgado con aplicación en diferentes campos, el uso de esta técnica para establecer un modelo de referencia no es tan habitual en la práctica. Este enfoque, sin embargo, es de gran utilidad para la monitorización y diagnóstico de procesos.

Otra forma equivalente de expresar el modelo, una vez que los datos han sido centrados y escalados se obtiene utilizando matrices en lugar de vectores:

$$\mathbf{Z} = \mathbf{1} * \bar{\mathbf{z}}' + \mathbf{T} * \mathbf{P}' + \mathbf{E}$$

El primer término es una matriz que contiene en cada columna la media de la variable correspondiente, el producto **TP'** es la estructura del modelo, con **T** matriz de scores, **P** matriz de cargas, y **E**={e_{ij}} es la matriz de los residuales.

Cada fila de la matriz **T** contiene las coordenadas de cada observación en el nuevo espacio definido por las componentes principales. Cada columna de **P** contiene las cargas o pesos de cada una de las variables originales sobre cada variable latente. Las cargas definen la orientación del hiperplano obtenido, con respecto al espacio de las variables originales. Las cargas indican también la magnitud y sentido de la correlación de la componente considerada con cada una de las variables originales. En el caso en que las variables están tipificadas, la relación entre las cargas y el coeficiente de correlación de z_j y $t_{j'}$ está dado por : $r(z_j, t_{j'}) = p_{j'j}\sqrt{\lambda_{j'}}$

Cálculo de las componentes principales. El algoritmo NIPALS

Los algoritmos para el cálculo de valores y vectores propios, de amplia divulgación, permiten obtener las componentes principales. Se pueden encontrar algoritmos que calculan todos los valores y vectores propios en forma simultánea y hay otros que lo hacen secuencialmente. Teniendo en cuenta que sólo se requerirá un número de componentes menor al número máximo que se puede calcular, dada la estructura de correlaciones que es normal encontrar entre las variables originales, , conviene utilizar algún procedimiento que obtenga secuencialmente los vectores y valores propios desde el mayor hasta el más pequeño, indicando en qué momento no se desean más cálculos. Un algoritmo de uso habitual es el de Hotelling también conocido como "Power Method". Otro, que responde en forma adecuada, es el algoritmo NIPALS (Nonlinear Iterative Partial Least Squares). Cabe mencionar, que los algoritmos que obtienen todas las componentes en forma simultánea, son más precisos, pero pueden presentar inconvenientes si existe alta colinealidad ya que es posible que **Z**'**Z** esté mal condicionada. Los secuenciales son menos precisos pero funcionan bien en contextos de alta colinealidad y en presencia de datos faltantes.

Si bien en PCA cualquiera de los algoritmos desarrollados resultan adecuados, es de interés la utilización del NIPALS, teniendo en cuenta que, en el próximo método a estudiar, PLS, se utilizará también este algoritmo.

El algoritmo NIPALS puede sintetizarse en los siguientes pasos:

1) Elegir una columna de la matriz **Z**, por ejemplo $\mathbf{z_j}$ y hacer $\mathbf{t_{inicial}=z_j}$.
2) Obtener un vector $\mathbf{p'_j=t'_{inicial}Z/(t'_{inicial}t_{inicial})}$
3) Normalizar p
4) Calcular un nuevo $\mathbf{t_{final}=Zp_j}$. Las componentes de $\mathbf{t_{final}}$ son las pendientes de la regresión de cada una de las filas de **Z** sobre $\mathbf{p_j}$.
5) Comparar $\mathbf{t_{final}}$ con $\mathbf{t_{inicial}}$ y, estableciendo algún criterio de tolerancia, si resultan aproximadamente iguales se da por finalizada la búsqueda de la j-ésima componente. En caso contrario se debe repetir desde 2) asignando previamente el valor de $\mathbf{t_{final}}$ a $\mathbf{t_{inicial}}$ $\mathbf{(t_{inicial} \leftarrow t_{final})}$
6) Si se desea obtener otra componente, se extrae de la matriz **Z** la parte explicada por la j-ésima componente ($\mathbf{t_j p'_j}$), es decir, se asigna a **Z** el resultado de $\mathbf{Z-t_j p'_j}$, y se comienza nuevamente el proceso, es decir se vuelven a ejecutar los pasos 1 a 5.

En la convergencia se tiene que:

$$\mathbf{p_j=Z't_j/t_j't_j = Z'Zp_j/t_j't_j} \text{ ; luego, } \mathbf{Z'Zp_j= t_j't_j\, p_j} \text{ ,}$$

luego $\mathbf{p_j}$ converge al vector propio correspondiente al mayor valor propio de **Z'Z** . El valor propio será $\mathbf{t'_j \,.\, t_j}$.

La variancia del score $\mathbf{t_j}$ encontrado, es el mayor autovalor de **Z'Z/(**N-1), si se trata de una muestra, debiendo dividir entonces $\mathbf{t'_j \,.\, t_j}$ por (N-1) para tener el valor propio buscado.

HERRAMIENTAS DE DIAGNÓSTICO

PCA presenta interesantes herramientas de diagnóstico. Las mismas permiten detectar agrupamientos en las observaciones y en las variables así como tendencias, datos anómalos, etc.

Graficar los scores en los planos definidos por las variables latentes, permitirá encontrar agrupamientos y detectar datos con anomalías severas ("outliers" severos). Al representar graficamente los scores junto con la elipse definida por la T^2 de Hotelling, estos datos anómalos aparecen a una distancia del centro de la elipse mayor que cualquier punto de ella. Esta última comparación se propone a partir de la descomposición de la T^2 en la suma de los scores. Además una observación con aporte t^2_{ij} elevado puede ser analizada observando a qué componente se debe ese aporte y adicionalmente, qué variables están relacionadas con tal componente.

Graficar las cargas permitirá visualizar las asociaciones entre las variables. La representación conjunta de observaciones y variables permitirá encontrar explicaciones sobre qué variables caracterizan los agrupamientos y tendencias hallados en las observaciones.

Es posible que, además de existir datos con anomalías severas, haya datos con anomalías moderadas. Una observación con estas características podrá tener coordenadas

interiores a la elipse definida por la T^2 de Hotelling para un nivel de confianza dado, pero no estará bien ajustada por el modelo, es decir estará alejada del plano de ajuste de la nube. Esta situación estará reflejada en los residuales expresados en la matriz **E** de la ecuación del modelo. Una observación en estas condiciones tendrá, en su correspondiente fila de la matriz **E** valores absolutos grandes. Para la detección de "outliers" moderados, se puede utilizar la medida $DmodZ_i$ que representa la desviación típica residual del individuo i al modelo PCA ajustado ($\sqrt{\frac{SCR}{gl}}$) y que, como se observa, es proporcional a la raíz cuadrada de la distancia euclídea del individuo i a su proyección en el modelo.:

$$DModZ_i = \sqrt{\frac{\sum_{j=1}^{J} e_{ij}^2}{J - A}}$$

donde e_{ij} es la diferencia entre la observación z_{ij} y su estimación por el modelo.

Si se define una distancia media de una observación al modelo como:

$$s_0 = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{A} e_{ij}^2}{(N-A-1)(J-A)}}$$

Puede calcularse la estadística $(DmodX/s_0)^2$, llamado también $DmodZ_{norm}$ (distancia la modelo normalizada) que sigue una distribución $F_{J\text{-}A,(N\text{-}A\text{-}1)(J\text{-}A)}$. De esa manera, comparando para una observación dada, el valor de $DmodZ_{norm}$ con la raíz cuadrada del percentil $(1\text{-}\alpha)\%$ de una distribución $F_{J\text{-}A,(N\text{-}A\text{-}1)(J\text{-}A)}$, llamada distancia crítica normalizada, se puede probar si la distancia al modelo de una observación puede considerarse o no anómala. Si $DmodZ_{norm} > \sqrt{F^{\alpha}_{J-A,(N-A-1)(J-A)}}$ la observación i, es anómala moderada, es decir, no se ajusta bien al modelo PCA.

Es también importante diagnosticar qué variables están bien explicadas por el modelo. Para cada variable z_j , puede calcularse el % de variabilidad de esa variable, explicada por el modelo con A componentes, R^2_j ,

$$R^2 = 1 - \frac{SCR_j}{SCT}$$

donde $SCR_j = \sum_{i=1}^{N} e_{ij}^2$, y $SCT = \sum_{i=1}^{N} z_{ij}^2$ y e_{ij} son los residuos obtenidos al extraer A componentes.

Este valor puede variar entre 0, que corresponde al caso en el que el modelo no explique nada de la variable, y 1 para la situación contraria en la que toda la variabilidad de z_j está recogida por el modelo. Al aumentar el número de componentes principales, los valores de R^2_j irán aumentando hasta llegar a 1 cuando todas las variables latentes se incluyen en el modelo. Puede calcularse también el porcentaje de variabilidad total de **Z** explicada por el modelo, R^2, con A componentes, como:

$$R^2 = 1 - \frac{SCR}{SCT}$$

donde $SCR = \sum_{i=1}^{N}\sum_{j=1}^{J} e_{ij}^2$ y $SCT = \sum_{i=1}^{N}\sum_{j=1}^{J} z_{ij}^2$

Estos coeficientes R^2 pueden ajustarse dividiendo cada suma de cuadrados por sus correspondientes grados de libertad obteniendo R^2_{aj} que mide el porcentaje de variancia explicada.

Cabe también analizar el número de componentes principales que debe incluirse en el modelo. La discusión se genera porque no es lo mismo considerar la capacidad del modelo para la representación simplificada de los datos reales o bondad de ajuste, que su capacidad de predecir nuevos datos. La primera de las características está medida por el valor R^2 y a medida que se complica el modelo agregando variables latentes, su valor aumenta. Sin embargo, en la medida en que el modelo esté sobreparametrizado, su capacidad de predicción disminuirá. La capacidad depredicción puede ser evaluada mediante la técnica de validación cruzada. Esto es:

1) Dividir la muestra de N observaciones en L grupos de I unidades cada uno
2) Eliminar cada uno de los grupos por turno
3) Obtener el modelo con las N-I unidades restantes y luego predecir los valores de **z** para las unidades eliminadas.

Para cada unidad i se podrá calcular el error cuadrático de predicción: $\sum_{j=1}^{J} (z_{ij} - \hat{z}_{ij})^2$

donde $\hat{z}_{ij}$ es el valor de la j-ésima variable para el i-ésimo individuo según la estimación provista por el modelo. La suma de cuadrados de los residuos de las predicciones para todas las observaciones se denomina PRESS (Predictive Residual Sum of Squares). A partir del PRESS se pueden construir diferentes indicadores de la bondad de predicción.

Si bien se cumple siempre que SCR(j) < SCR(j-1), sin embargo, PRESS(j) será menor que SCR(j-1) sólo si la componente j-ésimamejora la capacidad predictiva del modelo, y en ese caso $Q^2 > 0$. En caso contrario PRESS(j) > SCR(j-1) y $Q^2 < 0$.

$$Q_j^2 = 1 - \frac{PRESS(j)}{SCR_{j-1}}$$

SC_{j-1} es la suma de cuadrados de los residuos del modelo con j-1 componentes y PRESS(j) es la suma de los cuadrados de los residuos obtenidos por validación cruzada en el modelo con j componentes. Q^2_j mide la variación de **Z** predicha por la componente j-ésima.

$Q^2(cum)_j$=1-PRESS/SCT mide la bondad de predicción del modelo PCA con j componentes. Q^2 indica la fracción de la variación total que puede ser predicha por una componente. Q^2 no varía entre 0 y 1 como el caso de R^2; el indicador de bondad de predicción crece desde 0 hasta un punto máximo donde comienza a decrecer aunque aumente el número de componentes, indicando en este caso un sobreajuste del modelo.

Q^2 puede ser calculado también para cada variable , lo que permite evaluar cuan bueno es el modelo para predecir los valores de cada una de las variables por separado. Un criterio práctico para decidir el número de variables latentes a utilizar, sugerido por Wold establece que si $Q^2>0,5$ es buena, y $Q^2>0,9$ es excelente. Además, la diferencia entre R^2 y Q^2 no debe ser muy grande.

APLICACIÓN A LA MONITORIZACIÓN

Para la monitorización del proceso el procedimiento consiste en registrar, datos en condiciones normales de operación, y analizar los mismos decidiendo el número A de componentes principales a utilizar, obteniendo así la matriz de cargas **P**. Ante una nueva observación $\mathbf{z_i}$ se calculará su valor de T^2

$$T^2_{i,A} = \sum_{j=1}^{A} \frac{t^2_{i,j}}{S^2_{t_j}} = \sum_{j=1}^{A} \left[\frac{\mathbf{p}'_j(\mathbf{z_n} - \boldsymbol{\mu})}{S_j} \right]^2$$

$$T^2_A \sim \frac{A(N^2-1)}{N(N-A)} F_{A,N-A}$$

Ese valor se compara con el límite superior establecido, que corresponde al percentil (1-α) % de una distribución $F_{A,(N-A)}$ por la constante de la expresión anterior .

Debe también controlarse la distancia al modelo, ya que aunque la T^2 calculada para una nueva observación, esté dentro de la región de aceptación, la misma puede estar mal ajustada por el modelo. Para ello, se puede calcular la distancia al modelo normalizada de la nueva observación y, compararla con distancia crítica normalizada. Para su aplicación práctica, se pueden construir gráficos de control de T^2 y de ($DmodZ/S_0$), cada uno con su límite superior de control. Ante una anomalía proceder al diagnóstico utilizando alguna de las herramientas propuestas.

5) COMENTARIOS FINALES

La gran cantidad de información que generan muchos de los procesos actuales, orienta hacia el desarrollo de métodos que permitan un mayor aprovechamiento de la misma. En primer lugar, los diagramas de control multivariados resultan una herramienta útil para la detección de salidas de control. En segundo lugar, para tomar las acciones correctivas adecuadas se deben encontrar las variables que provocaron dicha señal, lo cual no es tan evidente como en el control estadístico univariado. Esta tarea no es sencilla y se pueden encontrar en publicaciones recientes, diferentes propuestas, cuya implementación de manera inmediata, no parece fácil. No se cuenta aún, en el software estadístico de mayor divulgación la posibilidad de un fácil acceso a estas herramientas, sino que deben ser programadas.

Por otra parte, situaciones como lel estudio de un gran número de variables, con alta colinealidad y datos faltantes, debilitan los instrumentos mencionados. La alternativa de la utilización del análisis de componentes principales, se constituye en un método apto para el tratamiento de datos con estas características, permitiendo el desarrollo de herramientas gráficas y medidas adecuadas, para la monitorización y diagnóstico de procesos. Nuevamente la mayor dificultad actual, es la posibilidad de una implementación rápida y

accesible, ya que la obtención de gráficos y medidas necesarias, no es inmediata en el software estadístico corriente, sino que deben desarrollarse los programas, encontrando solo en software más específicos como SIMCA[1] y PLS_Toolbox[2] la posibilidad de una instrumentación sencilla en los procesos de producción.

Bibliografía

- Fuchs, Camil Kenett, Ron S. "Multivariate Quality Control. Theory and Applications", Marcel Dekker Inc. , 1998

- Geladi, P. Y Kowalski, B. R. "Partial Least Squares Regression: A Tutorial". Analytica Chimica Acta 185, pág. 1-17, 1986

- Hodouin D, MacGregor J.F. Hou M. Franklin M. "Multivariate statistical analysis of mineral processing plant data". CIM Bulletin,Vol 85, Nº 975, pág. 23-34, 1993

- Kourti T. and MacGregor J. F. "Multivariate SPC Methods for Process and Product Monitoring Journal of Quality Technology" Vol 28 Nº 4, 1996

- Wold S. Multi and Megavariate Data Anlysis using Projection Methods

- SIMCA –P 8.0 User Guide and Tutorial UMETRICS, 1999

[1] SIMCA-P, UMETRICS AB, Box 7960, S-907 19 Umea, Sweden

[2] PLS_Toolbox, Barry M. Wise-Neal B. Gallagher. Eigenvector Technologies, P.O. Box 483, 196 Hyacinth Avenue, Manson, WA, USA