27 Congreso Nacional de Estadística e Investigación Operativa
Lleida, 8-11 de abril de 2003

CONTROL ESTADÍSTICO *Mega*VARIANTE PARA LOS PROCESOS DEL SIGLO XXI

A. Ferrer

Departamento de Estadística e Investigación Operativa Aplicadas y Calidad
Universidad Politécnica de Valencia
Camino de Vera s/n, Edificio I-3, 46022 Valencia, España
E-mail: aferrer@eio.upv.es

RESUMEN

El reto al que se enfrenta actualmente la monitorización de procesos es cómo manejar la gran cantidad de información correlacionada que puede llegar a registrarse en tiempo real mediante la informatización de estos procesos. En la ponencia se discuten las limitaciones de los enfoques clásicos uni y multivariantes del Control Estadístico de Procesos (SPC) en los modernos procesos altamente automatizados, y se presentan los métodos estadísticos de proyección sobre estructuras latentes (PCA y PLS) como alternativa para adaptar el SPC a estos entornos *mega*variantes, en lo que podría denominarse Control Estadístico *Mega*variante de Procesos. Esto se ilustra a través de varios ejemplos de procesos reales.

Palabras y frases clave: Control estadístico multivariante de procesos, Análisis de componentes principales (PCA), Regresión en mínimos cuadrados parciales (PLS).
Clasificación AMS: 62N10; 62HXX.

1. Introducción

Personalidades como Box, Fisher, Tukey y Youden han contribuido decisivamente a que el pensamiento y los métodos estadísticos se hayan introducido en un gran número de disciplinas de la ciencia y la ingeniería. Estos "estadísticos" fueron en esencia científicos (genetistas, químicos, ingenieros) que comprendieron el modo de pensar de científicos e ingenieros, e hicieron suyo sus problemas, lo que les permitió desarrollar nuevos y eficientes métodos estadísticos para resolverlos (MacGregor 1996). Se puede decir que asumieron el hecho de que "la única forma de hacer descubrimientos importantes es estudiar problemas importantes" (Medawar 1979). Y lo hicieron implicándose en esos problemas reales y trabajando en grupos multidisciplinares.

"Los mayores avances científicos, y en la ciencia estadística en particular, ocurren generalmente como resultado de la interacción entre la teoría y la práctica. ... Un

entorno escogido adecuadamente puede sugerir al investigador nuevas teorías o modelos que vale la pena investigar ” (Box 1976).
Es, por tanto, del contacto directo con la problemática real desde donde surge la necesidad de la investigación para dar respuesta a problemas sin solución aparente. Y esta actitud está en plena sintonía con la de otro gran eminente estadístico de principios del siglo XX, Karl Pearson, para quien el objetivo de la ciencia estadística fue siempre el “desarrollo de una metodología para investigar la vida real, y no el refinamiento de teorías matemáticas” (Kruskal y Tanur 1978).
El último cuarto del siglo XX ha sido testigo de una nueva revolución, que podría calificarse de tecnológica, provocada por el desarrollo explosivo de la tecnología de la electrónica y las comunicaciones, que ha traído consigo el abaratamiento de los ordenadores, sensores y otros dispositivos de medida, lo que se ha traducido en una creciente informatización de los procesos, que permite el registro en tiempo real con frecuencias hasta de milisegundos de una gran cantidad de variables. El reto al que se enfrenta actualmente la monitorización de procesos es cómo manejar la gran cantidad de información correlacionada que puede llegar a registrarse.
Este nuevo entorno ha modificado totalmente la naturaleza de los datos disponibles, lo que obliga a un cambio de paradigma: de la escasez a la sobreabundancia de datos. Esta sobrecarga de información y la falta de herramientas estadísticas apropiadas ha provocado que en la práctica se malgasten recursos y se desaprovechen oportunidades de mejora en la calidad y productividad de estos procesos.
Siguiendo el ejemplo de Box, Fisher, Pearson, Tukey y Youden es necesario salir de nuestros despachos y trabajar junto con científicos e ingenieros para dar respuesta como estadísticos a los nuevos retos que la revolución tecnológica plantea, abandonando viejos paradigmas que nada tienen que ver con los verdaderos problemas que acucian a los procesos industriales del siglo XXI, y contribuyendo al desarrollo de nuevos métodos estadísticos para abordar los nuevos retos asociados con los grandes volúmenes de datos de naturaleza multivariante que nos inundan.
En esta ponencia se discuten las limitaciones de los enfoques clásicos del Control Estadístico de Procesos (*Statistical Process Control*, SPC) en los modernos procesos altamente automatizados, y se presentan los métodos estadísticos de proyección sobre estructuras latentes como alternativa para adaptar el SPC al nuevo paradigma, fruto de la revolución tecnológica.

2. Limitaciones del SPC Clásico

El objetivo del SPC es el de establecer un sistema permanente e inteligente de monitorización de un proceso a lo largo del tiempo con el fin de detectar precozmente cualquier causa especial de variabilidad que pueda afectarle. Para ello, construye a partir de datos del proceso recogidos en condiciones de funcionamiento normales un modelo empírico (modelo bajo control) que se utiliza como patrón de comparación del funcionamiento del proceso en el futuro. La identificación de las causas que provocan comportamientos anómalos del proceso (causas especiales) es el primer paso en la toma

de medidas pertinentes para su eliminación (o incorporación si son positivas), lo que permite una mejora continua del mismo.

En las industrias de procesos (típicas de los sectores químico y petroquímico) la revolución tecnológica ha permitido que hoy día centenares de dispositivos electrónicos (ordenadores, sensores, etc.) registren rutinariamente medidas de un gran número de variables (caudales, temperaturas, presiones, frecuencias espectrales, etc.) con frecuencias de muestreo que pueden oscilar entre milisegundos y horas, dependiendo de la dinámica de los procesos. Adicionalmente a esta gran cantidad de variables de proceso, también se registran, pero con frecuencia muy inferior, algunas pocas variables de calidad (a veces también de productividad) medidas, por lo general no en tiempo real sino *off-line*, tras determinaciones analíticas, a menudo costosas, en laboratorios de control de calidad.

La automatización también ha llegado a las industrias de piezas, donde cada vez es más frecuente la inspección al 100% de las piezas fabricadas, así como la medición de parámetros continuos de las máquinas que realizan las operaciones sobre las piezas. Por ejemplo, además de medir el diámetro y la posición de un cierto orificio en una pieza, es posible medir en continuo la presión del circuito hidráulico del brazo del robot que determina la posición y la inclinación de la cabeza taladradora que mecaniza la pieza. O también es posible medir en continuo las trayectorias de las variables de proceso (intensidad, voltaje, presión de mordaza, etc.) en procesos de soldadura de piezas, que funcionan como procesos por lotes. Esto hace que hoy día las diferencias en cuanto a la naturaleza de los datos registrados entre las industrias de procesos y las de piezas sean cada vez menores.

Las herramientas tradicionales más extendidas del SPC han consistido en la monitorización mediante gráficos de control univariante de las variables de calidad. En algunos procesos se han utilizado gráficos de control multivariante de un pequeño grupo de las variables de calidad más correlacionadas basados en el estadístico T^2-Hotelling (Jackson 1991) expresado como:

$$T^2 = (\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \qquad (1)$$

donde $\boldsymbol{y}$ es el vector de las variables de calidad con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas estimada bajo control, $\boldsymbol{S}$.

En el SPC clásico las variables de proceso son frecuentemente ignoradas en la monitorización, utilizándose sólo como ayuda en la identificación de las causas especiales de variabilidad, una vez que los gráficos de control de las variables de calidad indican una señal de falta de control estadístico.

Este enfoque tradicional, desarrollado inicialmente por Shewhart (1931) y que ha dado excelentes resultados en contextos caracterizados por el registro de pocas variables de naturaleza independiente, típicos de buena parte de los procesos industriales de décadas pasadas, resulta del todo inadecuado en los modernos contextos de sobreabundancia de datos, fundamentalmente por dos motivos:

a) La naturaleza multivariante de la calidad

En realidad la calidad de un producto es una propiedad multivariante y debe ser tratada como tal. Para que un producto satisfaga las necesidades de los clientes, el modelo bajo

control asume que, no sólo las distribuciones marginales deben estar bajo control (lo que podría monitorizarse con gráficos de control univariantes), sino que también debe estarlo la estructura de correlación entre las variables, caracterizada por la matriz de covarianzas. Esto provoca que el esquema de monitorización clásico univariante sea completamente desaconsejable en procesos con variables correlacionadas.

Podría decirse que la presencia de correlación entre las variables provoca la misma ineficiencia en la utilización del SPC univariante, como la presencia de interacciones entre los factores provoca en el diseño de experimentos desarrollado bajo la inoperante estrategia de ir modificando cada vez los niveles de un factor, dejando fijos los de los restantes (MacGregor 1996).

Como alternativa, los gráficos de control multivariantes clásicos, basados en el estadístico T^2-Hotelling construido a partir de las variables de calidad registradas en los procesos (ecuación (1)), pueden considerarse herramientas útiles para la monitorización de estos siempre que se cumpla que: haya más observaciones que variables, la matriz de covarianzas de las variables no sea de grandes dimensiones, no esté mal condicionada debido a problemas de colinealidad, y no existan datos faltantes. Sin embargo, estos supuestos son poco razonables en los modernos contextos automatizados, como se discute más adelante.

b) El desaprovechamiento de gran parte de la información registrada en el proceso

Utilizar sólo las variables de calidad para monitorizar los procesos es, sin duda, el motivo más importante de la ineficiencia del SPC tradicional en los modernos procesos industriales del siglo XXI, donde el ratio de variables de proceso respecto a variables de calidad registradas puede ser del orden de 100, o con frecuencia superior. Por ejemplo, en un proceso por lotes de polimerización de óxido de polipropileno el ratio llegó a ser de 4500 (Zarzo *et al* 2002a, 2002b). Esto supone desaprovechar para la monitorización la práctica totalidad de la información registrada del proceso y prácticamente el 100% de la información registrada en tiempo real, pues, a diferencia de las variables de proceso, la información de las variables de calidad suele venir desfasada a menudo varias horas, una vez se obtienen los resultados de los análisis de laboratorio.

Otro problema asociado es el del tiempo de reacción desde que se produce una anomalía en el proceso hasta que el sistema de monitorización la detecta. Si sólo se utilizan las variables de calidad, este tiempo en los procesos continuos puede ser en el mejor de los casos del orden de varias horas (asumiendo una frecuencia de muestreo de varias horas y que el *Average Run Length* -ARL- fuera de control de los gráficos de control de las variables de calidad es de una unidad). Es posible, sin embargo, que estas anomalías, aunque sí afecten a la calidad percibida por el cliente, no se reflejen en las variables de calidad medidas (pues éstas no definen por completo la calidad del producto), por lo que en ese caso no serían detectables en el proceso.

En el caso de los procesos por lotes, la utilización sólo de las medidas de calidad del producto, obtenidas una vez finalizado la producción del lote y tras la realización de las pruebas analíticas en laboratorio, provoca que los tiempos de reacción puedan llegar a ser hasta de varios días, monitorizándose únicamente la variación entre lotes, pero no la existente dentro de los mismos.

Un tercer inconveniente es el del diagnóstico de los problemas. Si sólo se monitorizan variables de calidad y, en el mejor de los casos, se detecta la/s variable/s causante/s de la señal de falta de control, la tarea de identificar las condiciones del proceso causantes de la anomalía puede resultar muy costosa.
La incorporación de las variables de proceso en el esquema de monitorización del mismo es una de las claves de la adaptación del SPC a los modernos entornos altamente automatizados. Su medición en tiempo real con alta frecuencia de muestreo, su reducido coste (comparado con el coste de análisis de las variables de calidad) y su gran sensibilidad a anomalías en el proceso (que suelen dejar su "huella" en las variables de proceso), contribuyen a disminuir sensiblemente el tiempo de reacción ante problemas en el proceso (reducción que puede llegar a ser de días u horas, a minutos o segundos), a la vez que facilitan la tarea de diagnóstico de las causas especiales que han provocado los problemas, con un incremento de coste despreciable ante los beneficios que supone su utilización.

3. Dificultades en el manejo de los datos de proceso

Sin embargo, la utilización de todos los datos obtenidos de los procesos (variables de proceso y de calidad) genera una serie de dificultades añadidas en su procesamiento:

a) Dimensionalidad

El primer problema es que la dimensión de las matrices de datos generadas: (tiempo x variables) en procesos continuos, o (lotes x variables x tiempo) en procesos por lotes es muy elevada. En procesos continuos pueden llegar a medirse cientos o incluso miles de variables de proceso cada pocos segundos, y una decena o más de variables de calidad cada pocas horas. Además, los avances en la tecnología de la instrumentación de procesos están permitiendo también obtener medidas de variables de calidad en tiempo real (*on-line*) cada pocos minutos. Por otra parte, en los procesos por lotes, además de las variables de calidad medidas al finalizar la producción de lote, la tecnología actual permite el registro de una gran cantidad de medidas de variables de proceso (hasta 50 o incluso más) cada muy poco tiempo (que puede ser hasta de milisegundos) a lo largo del periodo de procesado del lote (que puede variar entre unos pocos minutos en procesos de soldadura, hasta varios días en procesos de polimerización). Por tanto, de cada lote pueden llegar a registrarse centenares de miles de variables de proceso, que describen las trayectorias temporales de estas variables a lo largo de toda la duración del procesado del lote.
Con el fin de enfatizar la elevada dimensionalidad asociada a estas bases de datos, se acuña el término *mega*variante para distinguir estos contextos altamente automatizados, de los enfoques *multi*variantes clásicos donde se suele trabajar con unas pocas variables de calidad.
Ante esta sobrecarga de datos, muchos operarios y técnicos se sienten abrumados, y acaban utilizando para la toma de decisiones sólo unas pocas de las variables consideradas como claves según su experiencia, obviando el resto.

b) Colinealidad

La dimensión real de sucesos independientes que afectan a los procesos es muy inferior a la dimensión aparente de las matrices de datos. El sistema de causas comunes de variabilidad está regulado por unas pocas variables latentes (no medibles explícitamente) independientes que se expresan a través de los cientos o miles de variables medidas, generando una fuerte estructura de correlación entre las últimas.

c) Ruido

Todas las variables (de proceso y de calidad) están medidas con error (error de muestreo, error de medida, etc.). Debido a que están registradas durante el funcionamiento normal del proceso, es lógico que el ratio señal-ruido de cada variable sea pequeño, pues el objetivo de los operarios y técnicos es que el proceso se desvíe lo menos posible de su trayectoria objetivo.

d) Valores faltantes

La elevada automatización hace que las bases de datos registradas de los procesos contengan, con frecuencia, valores faltantes (a veces hasta un 20%) debido a fallos en el registro de los sensores, mantenimiento de los mismos, retraso en los análisis de laboratorio, etc. Por su parte, en la monitorización en tiempo real de los procesos por lotes, en cada instante t los valores futuros de las trayectorias de las variables de proceso correspondientes al periodo existente entre el instante t y el final del lote son desconocidos y, por tanto, se han de tratar como datos faltantes (Nomikos y MacGregor 1995).

4. Control estadístico *mega*variante mediante técnicas de proyección sobre estructuras latentes

Las técnicas clásicas de SPC son extremadamente ineficientes en los contextos *mega*variantes de los procesos del siglo XXI dada la elevada dimensionalidad de la matriz de covarianzas y su alto grado de colinealidad debido a la estructura de correlaciones existente dentro y entre las variables de proceso y las de calidad, así como la probable presencia de datos faltantes y el bajo ratio señal-ruido. En estas situaciones analizar cada variable por separado, como si se tratara de variables independientes, hace que la interpretación y el diagnóstico de problemas con el enfoque univariante sean realmente difíciles. Por otra parte, el mal condicionamiento de la matriz de covarianzas provoca problemas en su inversión, inestabilizando el estadístico T^2-Hotelling utilizado en el enfoque multivariante clásico (ecuación (1)). En el caso de que haya más variables que individuos, la matriz de covarianzas es singular, lo que impide el cálculo de este estadístico. Por otra parte, la T^2-Hotelling también se ve afectada por los datos faltantes, pues en el caso de que estos existan imposibilitan la utilización de la información registrada en el resto de variables para el cálculo de este estadístico en la monitorización de nuevos individuos.

En este contexto *mega*variante la utilización de técnicas estadísticas multivariantes de proyección sobre estructuras latentes tales como el Análisis de Componentes Principales (*Principal Component Analysis*, PCA) (Jackson 1991) y la Regresión en Mínimos Cuadrados Parciales (*Partial Least Squares*, PLS) (Geladi y Kowalski 1986,

Wold *et al.* 1987), relativamente robustas a la presencia de datos faltantes, y que manejan bien grandes matrices de datos mal condicionadas (incluso en el caso de existir más variables que individuos), aparece como la alternativa para adaptar el SPC clásico a estos contextos, y permite desarrollar lo que se podría denominar como Control Estadístico *Mega*variante de Procesos.

Las técnicas multivariantes mencionadas, PCA y PLS, comprimen la información multidimensional en unas pocas variables latentes que explican una gran parte de la variabilidad de las variables medidas, así como de sus relaciones. Es en este nuevo subespacio, de dimensión mucho más reducida que el espacio original de las variables de proceso y de calidad, donde las técnicas clásicas de SPC pueden utilizarse sin problemas, permitiendo a los operarios de los procesos controlar indirectamente la multitud de variables del proceso mediante la monitorización de unas pocas variables latentes, así como predecir los valores de las características de calidad a partir de la información registrada del proceso mediante la construcción de modelos inferenciales, también llamados *soft sensors*.

Considérese la situación en la que se dispone para una serie de I individuos (instantes de tiempo o lotes) de los registros de J_X variables de proceso (o trayectorias de variables de proceso) y J_Y variables de calidad registradas en condiciones operativas normales de funcionamiento (bajo control estadístico). Esta información puede organizarse en dos matrices de datos: una de variables de proceso $\boldsymbol{X}$, de dimensión IxJ_X, y otra de variables de calidad $\boldsymbol{Y}$, de dimensión IxJ_Y, que generalmente se suelen centrar y tipificar a varianza unitaria.

PCA puede utilizarse para descomponer tanto la matriz de variables de proceso ($\boldsymbol{Z}$=$\boldsymbol{X}$) como la de variables de calidad ($\boldsymbol{Z}$=$\boldsymbol{Y}$) en un conjunto de A matrices de rango 1

$$\boldsymbol{Z} = \sum_{a=1}^{A} \boldsymbol{t}_a \boldsymbol{p}_a^T + \boldsymbol{E} = \hat{\boldsymbol{Z}} + \boldsymbol{E} \tag{2}$$

donde los vectores de cargas $\boldsymbol{p}_a$ son las direcciones principales de máxima varianza en el espacio $\boldsymbol{Z}$ y definen el subespacio latente de dimensión A ($A \leq$rango $\boldsymbol{Z}$); los vectores de *scores*

$$\boldsymbol{t}_a = \boldsymbol{Z}\boldsymbol{p}_a \tag{3}$$

son las nuevas variables latentes, proyección de los I individuos en el subespacio latente A-dimensional definido por las direcciones principales; y $\hat{\boldsymbol{Z}}$ es la predicción de $\boldsymbol{Z}$ a partir de las A componentes principales extraídas. Las variables latentes son ortogonales y pueden ordenarse respecto al porcentaje de varianza explicada. Los vectores $\boldsymbol{p}_a$ y $\boldsymbol{t}_a$ son los a-ésimos vectores propios de las matrices $\boldsymbol{Z}^T\boldsymbol{Z}$ y $\boldsymbol{Z}\boldsymbol{Z}^T$, respectivamente. El número de componentes extraídas suele determinarse por técnicas de validación cruzada (Wold 1978) de forma que la matriz $\boldsymbol{E}$ de residuos del modelo no contenga componentes predictivas significativas. El error cuadrático de predicción de la observación i-ésima viene dado por:

$$ECP_i = \sum_{j=1}^{J} e_{ij}^2 = (z_i - \hat{z}_i)^T (z_i - \hat{z}_i) \tag{4}$$

y representa la distancia euclídea de la observación $\boldsymbol{z}_i$ a su proyección en el subespacio latente A-dimensional, mide por tanto la bondad de ajuste de esa observación al modelo.
El modelo bajo control viene definido por las direcciones $\boldsymbol{p}_a$, el vector de medias $\boldsymbol{\mu}$ de las variables originales, y la matriz de covarianzas de las variables latentes (matriz diagonal que contiene las varianzas s_a^2 de las A variables latentes). Cada vez que llega una nueva observación, $\boldsymbol{z}_{nueva}$, los *scores* t_a y el *ECP* se calculan a partir de (3) y (4), respectivamente, y se comparan con las regiones bajo control de los gráficos de control de estos estadísticos, definidos a continuación.
Los *scores* t_a son combinaciones lineales de las variables del proceso, por lo que, en virtud del Teorema Central del Límite, pueden considerarse distribuidos como una normal multivariante. En la monitorización de los *scores* se utilizan gráficos T^2 construidos con las A componentes principales extraídas:

$$T_A^2 = \sum_{a=1}^{A} \frac{t_{nuevo,a}^2}{s_a^2} = \sum_{a=1}^{A} \left[\frac{\boldsymbol{p}_a^T (\boldsymbol{z}_{nuevo} - \boldsymbol{\mu})}{s_a} \right]^2 \tag{5}$$

donde s_a^2 es la varianza de la variable latente a-ésima. El límite de control superior del gráfico T_A^2 puede calcularse de diversas formas, siendo frecuente utilizar la expresión obtenida por Tracy *et al.* (1992):

$$T_{LCS}^2 = \frac{(I^2-1)A}{I(I-A)} F_{1-\alpha}(A, I-A) \tag{6}$$

donde $F_{1-\alpha}(A, I-A)$ es el percentil (1-α)x100 de la distribución F-Snedecor con (A, I-A) grados de libertad.
El límite de control superior para el gráfico *ECP* puede calcularse a partir de soluciones aproximadas de la distribución de formas cuadráticas o mediante la construcción de una distribución de referencia a partir de datos históricos (Jackson 1991, Nomikos y MacGregor 1995).
El gráfico T_A^2 comprueba si la nueva observación permanece dentro de la región definida por las condiciones operativas normales (bajo control estadístico) en el subespacio proyectado; por su parte, el gráfico *ECP* analiza si la distancia de la nueva observación respecto a su proyección en el subespacio latente es semejante a la de las observaciones registradas bajo control. Variaciones anormales que respetan la estructura de correlación del modelo bajo control se reflejarán en valores anormalmente altos del estadístico T_A^2 para la nueva observación. Por otro lado, si la causa especial provoca una ruptura de la estructura de correlación del modelo bajo control, esto dará lugar a valores elevados del estadístico *ECP*.
Si la nueva observación tiene datos faltantes, varios autores han estudiado diferentes métodos de estimación de los *scores* en esta situación (Arteaga y Ferrer 2002, Nelson *et al* 1996), así como la incertidumbre asociada a los *scores* y *ECP* utilizados en la monitorización, provocada por las variables faltantes (Arteaga y Ferrer 2003, Nelson 2002).

A diferencia del PCA, PLS utiliza conjuntamente la información de las variables de proceso (***X***) y de calidad (***Y***) registradas en condiciones operativas normales de funcionamiento para definir el modelo bajo control. PLS simultáneamente reduce las dimensiones de ***X*** e ***Y*** para encontrar variables latentes en ***X*** que no sólo expliquen variación asociada a las variables de proceso, sino aquella variación en las variables de ***X*** que predice mejor las variables de calidad, ***Y***. Con este modelo es posible disponer en tiempo real de una predicción de la calidad del producto, anticipándose a los resultados obtenidos en el laboratorio. En el caso de procesos por lotes, además, se puede predecir la calidad del producto final del lote incluso antes de la finalización del mismo. El modelo PLS puede expresarse como:

$$\begin{aligned} \boldsymbol{t}_a &= \boldsymbol{X}_{a-1}\boldsymbol{w}_a;\ \boldsymbol{X}_a = \boldsymbol{X}_{a-1} - \boldsymbol{t}_a\boldsymbol{p}_a^T \\ \boldsymbol{X} &= \sum_{a=1}^{A} \boldsymbol{t}_a\boldsymbol{p}_a^T + \boldsymbol{E} \\ \boldsymbol{Y} &= \sum_{a=1}^{A} \boldsymbol{u}_a\boldsymbol{c}_a^T + \boldsymbol{F} = \sum_{a=1}^{A} \boldsymbol{t}_a\boldsymbol{c}_a^T + \boldsymbol{G} \end{aligned} \qquad (7)$$

donde $\boldsymbol{w}_a$ y $\boldsymbol{c}_a$ son las direcciones en ***X*** e ***Y***, respectivamente, tales que en cada dimensión se maximice la covarianza entre las variables latentes asociadas a los dos subespacios, t_a y u_a. Las direcciones $\boldsymbol{p}_a$ son las que permiten reconstruir mejor (en el sentido de mínimos cuadrados) la matriz ***X***, permitiendo que los vectores de *scores* $\boldsymbol{t}_a$ y de cargas $\boldsymbol{w}_a$ sean ortogonales. En este modelo las nuevas variables latentes en el espacio ***X***

$$\boldsymbol{t}_a = \boldsymbol{X}_{a\text{-}1}\,\boldsymbol{w}_a \qquad (8)$$

representan la proyección de los I individuos en las direcciones de gran varianza en el espacio ***X*** que más correlacionadas están con las variables de interés más importantes en ***Y***. Al igual que en PCA, el número de componentes A a extraer puede determinarse por validación cruzada. Para cualquier observación ($\boldsymbol{x}_i$, $\boldsymbol{y}_i$) pueden calcularse dos errores cuadráticos de predicción, uno en el espacio ***X*** (ECP_X) y otro en el ***Y*** (ECP_Y):

$$ECP_{X_i} = \sum_{j=1}^{J} e_{ij}^2 = (\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i)^T(\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i);\ ECP_{Y_i} = \sum_{j=1}^{J} g_{ij}^2 = (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i) \qquad (9)$$

La monitorización de nuevos individuos se lleva a cabo de forma similar al caso PCA mediante gráficos de control, construidos a partir del modelo bajo control obtenido de los datos registrados en condiciones operativas de funcionamiento normales, en los que se monitorizan los estadísticos T_A^2 y ECP para los nuevos individuos. En este caso T_A^2 monitoriza la variación en las variables de proceso que más influye en las variables de calidad.

Tanto en PCA como en PLS las componentes se suelen calcular mediante el algoritmo iterativo *NIPALS* (Geladi y Kowalski 1986), especialmente útil en los contextos *mega*variantes con grandes matrices de datos mal condicionadas. La utilización de algoritmos no secuenciales, aparte de requerir más tiempo de cálculo (pues extraen todas las componentes posibles), puede ocasionar problemas en la obtención de las

componentes no significativas debido a la alta colinealidad que genera valores propios prácticamente nulos. Otra ventaja adicional del algoritmo *NIPALS* es su buen comportamiento en presencia de datos faltantes (Nelson *et al.* 1996).
El buen funcionamiento de estas técnicas en la monitorización y diagnóstico de fallos en procesos continuos y, sobre todo, su superioridad respecto a los métodos clásicos de SPC univariantes ha sido probada en muchos procesos industriales reales (Dayal *et al* 1994, Kourti y MacGregor 1996, Tano *et al.* 1995), así como en redes de distribución de agua (Palau *et al.* 2003). Sólo los métodos multivariantes, que tratan todos los datos simultáneamente, pueden extraer información sobre la dirección de las variaciones del proceso, esto es, sobre cómo todas las variables se comportan entre sí. Más aún, cuando algo anormal ocurre en el proceso, con el enfoque univariante suele ser difícilmente detectable dado el bajo ratio señal/ruido que se genera en cada variable. Sin embargo, los métodos multivariantes son capaces de amplificar la señal y reducir el ruido, facilitando la detección de estas anomalías.
Nomikos y MacGregor (1995) adaptan los métodos de proyección estadísticos multivariantes PCA y PLS, desarrollados para la monitorización de procesos continuos, al caso de los procesos por lotes basándose en el Análisis de Componentes Principales en tres dimensiones (*Multiway Principal Component Analysis* -MPCA-) y en la Regresión en Mínimos Cuadrados Parciales en tres dimensiones (*Multi-way Partial Least Squares* -MPLS-) (Wold *et al.* 1987, Geladi 1989). La idea básica de estas técnicas consiste en desplegar la matriz tridimensional lotes x variables x tiempo convirtiéndola en la matriz $\boldsymbol{X}$ **(**IxJ_X), donde las columnas representan las trayectorias de las variables del proceso a lo largo del procesado del lote. Aplicaciones de estas técnicas en la monitorización de procesos por lotes pueden encontrarse en Kourti *et al* (1996), y más recientemente en Ferrer (2002) y Zarzo *et al.* (2002a y 2002b).
Los esquemas de monitorización propuestos, basados en las técnicas PCA y PLS, pueden extenderse a situaciones donde los procesos se pueden considerar constituidos por una serie de bloques, dentro de los cuales las variables están muy correlacionadas, siendo menor la correlación entre variables de bloques distintos. Estos bloques pueden corresponder a unidades físicas distintas del proceso, a diferentes subprocesos, a diversas secciones de una misma unidad física, etc. En algunos procesos por lotes puede haber una etapa de pre-procesado antes de que comience el proceso propiamente dicho. A veces también se dispone de información extra relevante al proceso como calidad de las materias primas, condiciones iniciales de temperatura y presión, composición de la carga inicial, cambios de turno de operarios o de proveedores de materias primas, etc. Todas estas fuentes o bloques de información, junto con las variables de calidad del producto final, pueden integrarse en un único esquema SPC mediante el uso de los denominados métodos de proyección multi-bloque (Wold *et al* 1987). Estos métodos permiten el establecimiento de gráficos de monitorización tanto para cada bloque de variables, como para el proceso en su conjunto y, en general, reducen el tiempo de reacción en la detección de anomalías, a la vez que facilitan el diagnóstico de las causas de fallo. Ejemplos de utilización de estos métodos multi-bloque en procesos reales pueden encontrarse en Kourti *et al* (1996), Kourti y MacGregor (1996), MacGregor *et al* (1994), Qin *et al.* (2001), Westerhuis *et al.* (1998a) y Wold *et al.* (1996).

Desde el punto de vista de la aplicabilidad y eficacia de estos métodos de proyección multivariante es importante tener en cuenta que, como todo enfoque inferencial, para que funcionen es necesario que se cumplan dos supuestos básicos: que los procesos sean "comparables" y que los sucesos de interés sean "observables". El primer supuesto establece que el método es válido en tanto en cuanto la base de datos histórica de referencia sea representativa del funcionamiento del proceso. Si algo cambia en el proceso (v.g. se cambia el catalizador, o las condiciones de agitación, etc.), entonces es necesario volver a disponer de una nueva base de datos recogida en las nuevas condiciones, para después aplicar el método. El segundo supuesto expresa el requisito de que los sucesos que uno desea detectar deben ser "observables" a partir de las medidas que están siendo recogidas en el proceso.
En los últimos años han aparecido diversos trabajos relacionados con la integración del Control Estadístico de Procesos (SPC) y del Control Automático de Procesos (APC) (Box y Kramer 1992, Capilla *et al.* 1999, Vander Wiel *et al.* 1992). En el caso en que el proceso funcione regulado con algún algoritmo de control *feedback* o *feedforward* (Box *et al*. 1994) es muy importante destacar que los gráficos SPC multivariantes propuestos deben ser construidos a partir de datos recogidos mientras el proceso funciona en ciclo cerrado, pues son éstas las condiciones "normales" de funcionamiento. Si aparece una perturbación importante, los algoritmos de control pueden llegar a compensarla mediante importantes ajustes en las variables de control. Sin embargo, hasta que lo consiguen, la calidad o la seguridad de los procesos puede verse comprometida. Los métodos de proyección propuestos pueden llegar a detectar estas anomalías debido al comportamiento inusual de las acciones de control, permitiendo la implantación de medidas que eviten su reaparición en el futuro. Este es el tipo de mejora continua a largo plazo que puede conseguirse mediante la incorporación de estrategias de monitorización en procesos regulados.

5. Líneas de investigación abiertas

La aplicación de las técnicas de proyección multivariante sobre estructuras latentes (PCA y PLS) al SPC es relativamente novedosa, pues los primeros artículos publicados datan de la década de 1990. Ello indica que es un campo todavía por explorar y en el que existen muchos temas por resolver, algunos de los cuáles se exponen a continuación:

a) Diagnóstico de fallos.

Como ocurre con muchos procedimientos "no direccionales" del SPC (no diseñados para detectar un tipo particular de anomalía, sino cualquier desviación del proceso respecto a su comportamiento normal), estos no identifican la causa de la anomalía que puede llegar a afectar al proceso, sino que únicamente señalan gráficamente que es muy probable que el comportamiento del proceso no sea consistente con el modelo de referencia. En el SPC tradicional suelen ser los técnicos y operarios del proceso los que, basándose en su conocimiento del mismo, proporcionan un diagnóstico de posibles causas y toman medidas correctoras. Este proceso de diagnóstico suele ser mucho más complicado en el SPC *mega*variante.

Diversos autores han propuesto diferentes técnicas para identificar la/s variable/s responsable/s del fallo del proceso una vez detectada la anomalía en los gráficos de control multivariante. De entre todos los procedimientos destaca el de los gráficos de contribución (MacGregor *et al* 1994, Miller *et al.* 1993, Kourti y MacGregor 1996). Recientemente se han propuesto otro tipo de técnicas basadas en la reconstrucción de los fallos (Qin, 2002).

b) Grados de libertad

Los límites de control de los gráficos T^2 y *ECP* exigen el cálculo de los grados de libertad asociados a las distribuciones de probabilidad teóricas utilizadas en la monitorización. La suposición tradicional de que si se dispone de una matriz $\boldsymbol{X}$ con I observaciones y J variables, y se estiman P parámetros, quedan IJ-P grados de libertad residual, no es razonable en los métodos multivariantes de proyección que utilizan matrices megavariantes, dada la alta colinealidad existente en $\boldsymbol{X}$. Van der Voet (1999) presenta una primera aproximación a este complejo problema.

c) Control estadístico megavariante mediante análisis de imágenes

El abaratamiento de la tecnología digital permite que cada vez sea más frecuente disponer de cámaras digitales que proporcionan información visual sobre características de los procesos. Esta información es también de naturaleza *mega*variante, lo que hace imprescindible la utilización de técnicas multivariantes de proyección sobre estructuras latentes para extraer la información latente contenida. Algunos trabajos de aplicación del análisis de imágenes al SPC son: Bharati y MacGregor (2000) y MacGregor *et al.* (1998).

d) Técnicas de alineación de procesos por lotes.

Uno de los temas más problemáticos en la actualidad es cómo implantar en la práctica métodos de monitorización en tiempo real en procesos por lotes de duración variable en los que las diversas operaciones que deben realizarse a lo largo de las etapas del proceso no se ejecutan en los mismos instantes de tiempo entre lotes distintos. En estos casos es necesario recurrir a técnicas de sincronización de trayectorias. Algunas propuestas se basan en el uso de variables indicadoras (Nomikos y Macgregor 1995, Neogi y Schlags 1998), o en la utilización de técnicas desarrolladas inicialmente para el reconocimiento del habla (*Dynamic Time Warping*) (Kassidas *et al* 1998).

e) Técnicas de análisis de matrices de datos tridimensionales (3-way data).

Como ya se ha comentado, la información recogida por los diferentes sensores de las variables de proceso a lo largo del tiempo para cada lote estudiado puede estructurarse en una matriz de tres dimensiones (lotes x variables x tiempo). El análisis conjunto de la información contenida en esta matriz puede realizarse, además de con las técnicas MPCA y MPLS ya comentadas, mediante distintos métodos de análisis de matrices de datos tridimensionales (3-way data), entre los que destacan los siguientes:

- ✓ *Parallel Factor Analysis* (PARAFAC) (Bro 1997, Smilde 1992, Smilde y Doornbos 1991)
- ✓ Modelos de Tucker-3 (Geladi 1989, Smilde 1992, Tucker 1966)
- ✓ *Multilinear* PLS (N-PLS) (Bro 1996)
- ✓ *Multiway Covariates Regression Models* (Smilde y Kiers 1998).

Las diferencias entre estos métodos radica en la forma de analizar la matriz tridimensional obtenida de los procesos por lotes lo que supone monitorizar diferentes

fuentes de información del proceso. Una de las desventajas de los métodos *Multiway* PCA y PLS es el elevado número de parámetros que contienen los modelos derivados. Sin embargo, los modelos obtenidos mediante PARAFAC y Tucker-3 son mucho más parsimoniosos, lo que proporciona más estabilidad y constituye una ventaja a priori en el aspecto de la monitorización en línea. Algunas referencias recientes de la aplicación de estas técnicas en la monitorización de procesos por lotes se pueden encontrar en Boqué y Smilde (1999), Dahl *et al* (1999), y Louwerse y Smilde (2000). Westerhuis *et al.* (1998b) y Gurden *et al.* (2000) proporcionan un estudio comparativo de algunas de estas técnicas.

6. Agradecimientos

El autor quiere agradecer especialmente al profesor John F. MacGregor, quien sin duda pasará a la historia como uno de los investigadores que más ha contribuido en el siglo XX al desarrollo del Control Multivariante de Procesos, todo el apoyo recibido, especialmente, durante la estancia como profesor visitante en *McMaster University*. Gracias, también, a los profesores Dora Kourti y Paul Taylor (*McMaster University*), y Johan Westerhuis (*University of Amsterdam*) por transmitir su entusiasmo y apoyo en esta línea de investigación.

Referencias

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

Arteaga, F.; Ferrer, A. (2003): Monitorización de procesos multivariantes con datos faltantes mediante análisis de componentes principales. 27 Congreso Nacional de Estadística e Investigación Operativa. Lleida.

Bharati. M. H.; MacGregor J.F. (2000): Texture analysis of images using Principal Component Analysis. SPIE/Photonics Conference on Process Imaging for Automatic Control. Boston.

Boqué, R.; Smilde, A.K. (1999): Monitoring and diagnosing batch processes with multiway regression models. *AIChE Journal* 45, (7), 1504-1520.

Box, G.E.P. (1976): Science and Statistics. *Journal of the American Statistical Association* 71, 791-799.

Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C. (1994): Time Series Analysis. Forecasting and Control. Prentice Hall.

Box, G.E.P.; Kramer, T. (1992): Statistical Process Monitoring and Feedback Adjustment – A Discussion. *Technometrics* 34, 251-267.

Bro, R. (1996): Multiway calibration. Multilinear PLS. *Journal of Chemometrics* 10, 47-61.

Bro, R. (1997): PARAFAC. Tutorial and applications. *Chemometrics Intell. Lab. Syst.* 38, 149-171.

Capilla, C.; Ferrer, A.; Romero, R.; Hualda, A. (1999): Integration of Statistical and Engineering Process Control in a Continuous Petrochemical Polymerization Process. *Technometrics* 41, (1), 14-28.

Dahl, K.S.; Piovoso, M.J.; Kosanovich, K.A. (1999): Translating third-order data analysis methods to chemical batch processes. *Chemometrics Intell. Lab. Syst.* 46, 161-180.

Dayal, B.; MacGregor, J.F.; Taylor, P.A.; Kildaw, R.; Marcikic, S. (1994): Application of feedforward neural networks and partial least squares regression for modelling Kappa number in a continuous Kamyr digester. *Pulp Paper Canada* 95, (1), T7-T13.

Ferrer, A. (2002): Detección, diagnóstico de fallos y mejora de procesos "batch" mediante técnicas de control estadístico multivariante. *Automática e Instrumentación* 326, 62-72.

Geladi, P. (1989): Analysis of Multi-way (Multi-mode) Data. *Chemometrics and Intelligent Laboratory Systems* 7, 11-30.

Geladi, P.; Kowalski, B.R. (1986): Partial Least-Squares Regression: A Tutorial. *Analytica Chimica Acta* 185, 1-17.

Gurden, S.P.; Westerhuis, J.A.; Bro, R.; Smilde, A.K. (2000): A comparison of multiway regression and scaling methods. University of Amsterdam. Department of Chemical Engineering. Process Analysis & Chemometrics Group. Internal Report #50.

Jackson, J. E. (1991): A User´s Guide to Principal Components. John Wiley & Sons. New York.

Kassidas, A.; MacGregor, J.F.; Taylor, P.A. (1998): Syncronization of batch trajectories using dynamic time warping. *AIChE J.* 44, 864-875.

Kourti, T.; Lee, J.; MacGregor, J.F. (1996): Experiences with Industrial Applications of Projections Methods for Multivariate Statistical Process Control. *Computers in Chemical Engineering* 20 Suppl., 745-750.

Kourti, T.; MacGregor, J.F. (1996): Multivariate SPC Methods for Process and Product Monitoring. *Journal of Quality Technology* 28, (4), 409-428.

Kruskal, W.H.; Tanur, J.M. (eds.) (1978): International Encyclopedia of Statistics (2 vol.). Collier-MacMillan. New York.

Louwerse, D.J.; Smilde, A.K. (2000): Multivariate statistical process control of batch processes based on three-way models. *Chemical Engineering Science* 55, 1225-1235.

MacGregor, J.F. (1996): Using On-Line Process Data to Improve Quality. Is there a Role for Statisticians?. Are They Up for the Challenge?. *ASQC Statistics Division Newsletter,* 16, 2, 6-13.

MacGregor, J. F.; Bharati, M.H.; Yu, H (1998): Multivariate Image Analysis for process monitoring and control. *Ind. Eng. Chem. Res.* 37, (12), 4715-4724.

MacGregor, J.F.; Jaeckle, C.; Kiparissides, C.; Koutoudi, M. (1994): Process Monitoring and Diagnosis by Multiblock PLS Methods. *AIChE Journal* 40, (5), 826-838.

Medawar, P.B. (1979): Advice to a Young Scientist. Harper and Row.

Miller, P.; Swanson, R. E.; Heckler, C. F. (1993): Contribution Plots: The Missing Link in Multivariate Quality Control. Presented at Annual Fall Technical Conference of the American Society for Quality Control (Milwaukee, WI) and the American Statistical Association" (Alexandria, VA).

Nelson, P.P.C. (2002): Treatment of missing measurements in PCA and PLS models, M. Eng. Thesis. Department of Chemical Engineering, McMaster University. Hamilton, Ontario, Canada.

Nelson, P.P.C.; Taylor, P.A.; MacGregor, J.F. (1996): Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics Intell. Lab. Sist*. 35, 45-65.

Neogi, D.; Schlags, C.E. (1998): Multivariate Statistical Analysis of an Emulsion Batch Process. *Ind. Engng. Chem. Res.* 37, 3971-3979.

Nomikos, P.; MacGregor, J.F. (1995): Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics* 37, (1), 41-59.

Palau, V.; Arregui de la Cruz, F.; Ferrer, A. (2003): Statistical analysis of the inflows of a network sector applied to the problem detection on the daily consumes. Second International Conference on Efficient Use and Management of Water in Urban Areas. Tenerife.

Qin, S.J.; Valle, S.; Piovoso, M.J. (2001): On unifying multiblock analysis with application to decentralized process monitoring. *Journal of Chemometrics* 15, 715-742.

Qin, S.J. (2002): Process chemometrics: basics and beyond. En revisión.

Shewhart, W.A. (1931): Economic Control Quality of Manufactured Product. D. Van Nostrand Co., Inc. New York.

Smilde, A.K. (1992): Three-way analysis. Problems and prospects. *Chemometrics Intell. Lab. Syst*. 15, 143-157.

Smilde, A.K.; Doornbos, D.A. (1991): Three-way methods for the calibration of chromatographic systems: comparing PARAFAC and three-way PLS. *Journal of Chemometrics* 5, 345-360.

Smilde, A.K.; Kiers, H.A.L. (1998): Multiway covariates regression models. *Journal of Chemometrics* 13, 31-48.

Tano, K.; Samskog, P.O.; Andreasson, B. (1995): Mathematical Modelling in Mining Industry Increases Both Quality and Quantity!.- Multivariate Modelling and On-Line Data Presentation for Process Optimization at LKAB. Presented at the *International Federation of Automatic Control Symposium on Automation in Mining Mineral and Metal Processing.* Sun City. South Africa.

Tracy, N.D.; Young, J.C.; Mason, R.L. (1992): Multivariate Control Charts for Individual Observations. *Journal of Quality Technology* 24, 88-95.

Tucker, L.R. (1966): Some mathematical notes on three-mode factor analysis. *Psychometrika* 31, 279-311.

Van der Voet, H. (1999): Pseudo-degrees of freedom for complex predictive models: the example of partial least squares. *Journal of Chemometrics* 13, 195-208.

Vander Wiel, S.A.; Tucker, W.T.; Faltin, F.W.; Doganaksoy, N. (1992): Algorithmic Statistical Process Control: Concepts and an Application. *Technometrics* 34, 286-297.

Westerhuis, J.A.; Kourti, T.; MacGregor, J.F. (1998a): Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics* 12, 301-321.

Westerhuis, J.A.; Kourti, T.; MacGregor, J.F. (1998b): Comparing alternative approaches for multivariate statistical analysis of batch processes. *Journal of Chemometrics* 13, 397-413.

Wold, S. (1978): Cross-Validatory Estimation of the Number of Components in Factor and Principal Component Models. *Technometrics* 20, 397-405.

Wold, S.; Geladi, P.; Esbensen, K.; Öhman, J. (1987): Multi-Way Principal Components-and PLS-Analysis. *Journal of Chemometrics* 1, 41-56.

Wold, S.; Kettaneh. N.; Tjessem, K. (1996): Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of Chemometrics* 10, 463-482.

Zarzo, M.; Ferrer, A.; Romero, R. (2002a): Fault detection by PLS to improve the quality of batch PPOX production. 3rd International Chemometrics Research Meeting ICRM2002. Veldhoven (Holanda).

Zarzo, M.; Ferrer, A.; Romero, R. (2002b): Multivariate process control to improve the quality of batch PPOX production. 2nd Annual Conference on Business and Industrial Statistics. Rimini (Italia).