Fascículo 8

Cursos y seminarios de matemática

Serie B

ISSN 1851-149X

Javier Etcheverry

Ignacio Ojea

(Editores)

TAMI 2012

Taller de Matemática Industrial

Departamento de Matemática

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

2014

Cursos y Seminarios de Matemática – Serie B

Fascículo 8

ISSN 1851-149X (Versión Electrónica)
ISSN 1851-1481 (Versión Impresa)

Universidad de Buenos Aires.
Departamento de Matemática
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria – Pabellón I
(1428) Ciudad de Buenos Aires
Argentina.
http://www.dm.uba.ar
e-mail. secre@dm.uba.ar
tel/fax: (+54-11)-4576-3335

Optimización de métodos de detección y diagnóstico de fallas en la industria petroquímica.

YPF

Responsable: Gabriel Ignacio Horowitz, YPF

Participantes: Manuel Benjamin[†], Jorge Gotay[‡], Tatiana Hartinger[†], Mauricio Maestri[†], Federico Navarro[‡], Lucio Pantazis[†], Lucas Sánchez[‡], Maximiliano Valle[†], Adrián Will[‡] y Sergio Andrés Yuhjtman[†]

[†]FCEyN - UBA
[‡]FACET - UNT

1 Descripción del problema

Las plantas petroquímicas poseen una cierta cantidad de instrumentos que registran periódicamente los valores de diferentes variables físicas (usualmente Presión, Temperatura, Caudal, etc). Además en esta industria, debido al interés económico y a la peligrosidad de los procesos y materiales involucrados (combustibles y/o explosivos en su gran mayoría), los procesos están sumamente controlados y monitoreados con precisión, además de contar con numerosos procesos automáticos de control y estaciones de monitoreo. En algunos casos, a veces debido a falla de instrumentos, error humano, problemas en algún proceso, roturas de equipo, o accidentes, uno o más procesos salen de su estado normal de trabajo. En esas condiciones, se dispone de un tiempo máximo de respuesta que usualmente ronda alrededor de 2 horas, para retornar el proceso a su estado original antes que el mal funcionamiento derive en una falla catastrófica provocando desde demoras serias en el proceso, pérdidas económicas, hasta accidentes de gravedad, incendios y explosiones. Es un problema severo que, sólo en la industria petroquímica en los Estados Unidos, causa pérdidas superiores a los USD 20.000.000.000 al año.

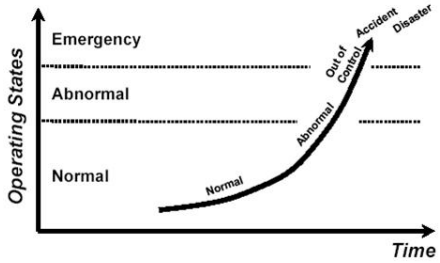

Figura 1: Eventos Anómalos en la Industria Petroquímica

La falla de algún instrumento o sensor de medición es una de las causas de este tipo de problemas. Se distinguen: el caso de la falla aislada, donde un sólo instrumento falla y la medición de los otros parámetros y mediciones del proceso no se ve alterada, y el caso de falla múltiple. Para el caso de falla de una sola variable existen procesos eficientes (Fisher Rosemount, Figura 2), pero que no funcionan correctamente en el caso de falla múltiple. En ocasiones debido a la

presencia de sistemas automáticos de control y a veces por causa de operadores humanos, la falla en un instrumento ocasiona cambios en otras variables de procesos conectados con él. En ese caso, el sistema detecta múltiples variables en diferentes procesos, lo que hace que resulte sumamente difícil para un operador humano distinguir el origen de la falla (puede suceder, incluso, que se trate de una falsa alarma provocada por el mal funcionamiento de un sensor). Mas aún, debido a la presencia de sistemas automáticos que actúan bajo la hipótesis de que el instrumento está funcionando de manera errónea, se pueden ocasionar accidentes reales y de gran peligrosidad (por ejemplo cuando el sistema aumenta la presión o temperatura en un proceso más allá de los límites permitidos).

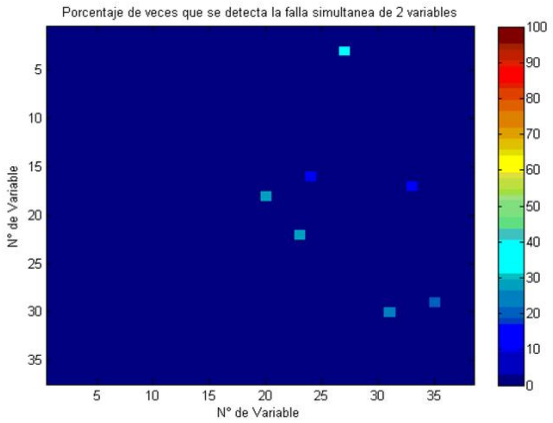

Figura 2: Fisher Rosemount - Detección de menos del 1% de los casos

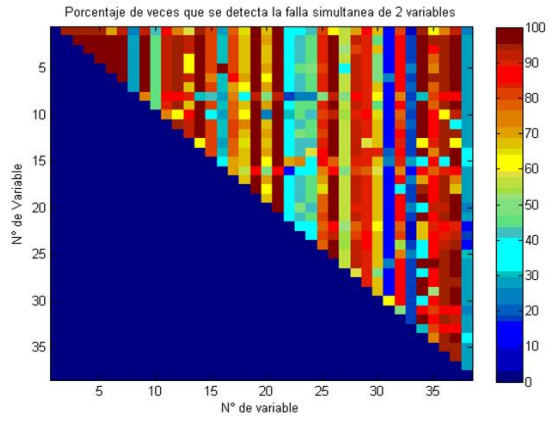

Figura 3: Sistema de YPF caso 2 fallas simultáneas

En este sentido, YPF cuenta con un proceso de monitoreo de este tipo de fallas, del tipo de Asistencia a la Toma de Decisiones. El proceso funciona eficientemente incluso en el caso de fallas múltiples (Figura 3. El sistema está basado en el hecho que, a pesar de que en teoría existen relaciones no lineales en las variables involucradas, en la práctica la mayoría de las variables en las plantas químicas observadas, presentan entre sí relaciones lineales. Esto permite resolver el

problema por lo menos en algunas de sus variantes, detectando cuando las relaciones lineales históricamente presentes entre las variables, cambian.

El proceso descripto consiste entonces en una primera parte en detectar, a partir de datos históricos, relaciones entre las variables utilizando en este caso la matriz de correlación lineal. Es un proceso combinatorio donde se listan todos los posibles subconjuntos de k variables, con k desde 2 en adelante, extrayendo la submatriz correspondiente de la matriz de correlación lineal, y buscando los casos en los que se puede despejar, aproximadamente, una variable en función de las otras, y sólo una. O sea, cualquier subconjunto de $k-1$ variables no presenta correlaciones lineales notables entre ellas. La segunda parte del proceso, el proceso de diagnóstico, corre aproximadamente cada hora y detecta cuantas variables no respetan su correlación histórica. La o las variables que presenten más alteraciones en sus correlaciones históricas con las demás variables, será la candidata a haber sufrido un mal funcionamiento. Este proceso se denomina "Angel Guardián" (Figura 4)

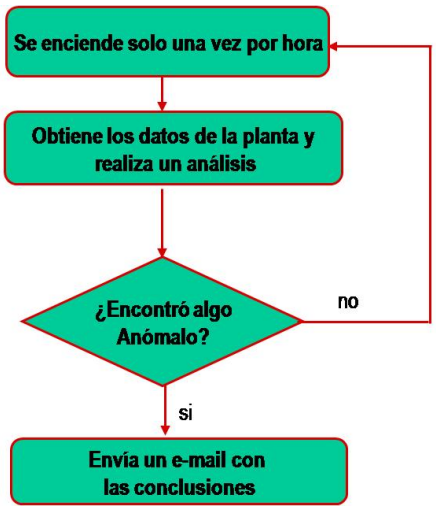

Figura 4: Sistema de YPF "Angel Guardián"

El problema planteado en TAMI 2012 consiste en reducir el tiempo necesario para llevar a cabo la primera parte del proceso, haciendo que éste resulte eficiente y se lo pueda correr en menos tiempo y/o para valores mayores de k.

El sistema en su diseño original, toma la matriz de datos A, se normaliza restando la media aritmética (con lo que los datos quedan centrados en cero), y se divide cada coordenada de los datos, por la desviación standard de la variable correspondiente. Este es un procedimiento que evita numerosos problemas numéricos y es usual en Análisis Numérico y Estadística (ver [1]). A continuación se arma la matriz de correlación A^tA, y buscando submatrices de esta matriz de correlación, recorre la lista de todas las combinaciones posibles de variables, desde 2 en adelante.

Para determinar si una determinada correlación entre $k < n$ variables resulta útil, se debe cumplir que una de las variables tiene correlación con las demás, y las restantes no presentan correlación entre sí. Para analizar esto, una vez que se extrae la submatriz correspondiente a la combinación de variables analizada, se le realiza un **Análisis de Componentes Principales (*Principal Components Analysis*, PCA)**, un análisis estadístico basado en la **Descomposición en Valores Singulares (*Singular Value Decomposition* SVD)**. Esta descomposición, si bien es costosa desde el punto de vista operacional, presenta la ventaja de ser numéricamente estable.

La Descomposición en Valores Singulares es una descomposición que permite factorizar cualquier matriz real $A_{m\times n}$ como producto $A_{m\times n} = U_{m\times m}S_{m\times n}V_{n\times n}$, donde U y V son matrices ortogonales y S es una matriz diagonal del mismo tamaño que la matriz original, de la forma

$$S_{m\times n} = \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_k \\ & & & \\ & & & \end{pmatrix}$$

Los números $\sigma_1, \ldots, \sigma_k$ se denominan "Valores Singulares" de la matriz A. Son reales positivos y aparecen ordenados de mayor a menor. Los coeficientes restantes de $S_{m\times n}$ son iguales a 0. Desde el punto de vista de Álgebra Lineal Numérica, estos números están relacionados con el núcleo y la condición de la matriz A (ver [1] y [2]). Desde el punto de vista estadístico, las aplicaciones más importantes de esta descomposición, conocida en Estadística como Análisis de Componentes Principales, son entre otras

- Generar nuevas variables no correlacionadas entre sí, que puedan expresar la información contenida en el conjunto original de datos
- Reducir la dimensionalidad del problema en estudio como paso previo a futuros análisis (en el caso en que el problema original contenga variables correlacionadas linealmente)

En el caso que nos ocupa, una combinación de variables que presente las características descriptas arriba será reconocida porque el menor valor singular σ_k será suficientemente pequeño, pero esto no sucede si se realiza la descomposición en valores singulares de cualquier subconjunto propio.

Para el problema presentado, la matriz de datos de entrada contiene 18.000 registros y entre 100 y 200 variables dependiendo de la planta considerada. El problema presenta en este punto una explosión combinatoria, dado que incluso para el caso de 100 variables, tenemos que

$$\begin{aligned}
\begin{pmatrix} 100 \\ 2 \end{pmatrix} &= \begin{pmatrix} 100! \\ 2!98! \end{pmatrix} = 4950 \\
\begin{pmatrix} 100 \\ 3 \end{pmatrix} &= \begin{pmatrix} 100! \\ 3!97! \end{pmatrix} = 323400 \\
\begin{pmatrix} 100 \\ 4 \end{pmatrix} &= \begin{pmatrix} 100! \\ 4!96! \end{pmatrix} = 23527350 \simeq 2.3\ 10^7 \\
\begin{pmatrix} 100 \\ 5 \end{pmatrix} &= \begin{pmatrix} 100! \\ 5!95! \end{pmatrix} = 1806900480 \simeq 1.8\ 10^9 \\
\begin{pmatrix} 100 \\ 6 \end{pmatrix} &= \begin{pmatrix} 100! \\ 6!94! \end{pmatrix} = 143046288000 \simeq 1.4\ 10^{11} \\
&\vdots
\end{aligned}$$

El primero es realizado en unos segundos, el segundo en un par de minutos, pero los dos últimos son claramente irrealizables en la práctica si recordamos que es necesario realizar una Descomposición en Valores Singulares de una matriz de orden k en cada caso. Por lo tanto el problema es

inabordable en la práctica, si como en este caso se pretende recorrer la lista completa, incluso eliminando las correlaciones de menor nivel ya encontradas

El problema consiste entonces en detectar, a partir de la matriz de correlación, todos los posibles subconjuntos de k variables tales que una de ellas esté correlacionada con las demás, pero las otras no estén correlacionadas entre sí (entendiendo que si hay otras variables correlacionadas entre sí, esa relación fue encontrada y extraída al recorrer los subconjuntos de variables de cardinal menor, $1, \ldots, k-1$).

2 Soluciones Presentadas

Se presentaron varias aproximaciones distintas a este problema. Algunas fueron exploradas en profundidad y completadas, mientras que otras fueron sólo esbozadas. Presentamos las principales soluciones planteadas.

2.1 Branch and Bound

Se plantea un método basado en el árbol de decisiones posibles, podando los casos no útiles. Para ello se considera la matriz de datos de entrada, de aproximadamente 18.000 registros, y variables entre cien y doscientas, dependiendo de la planta. De esta forma, la entrada es una matriz de m x n, donde m es la cantidad de mediciones y n es la cantidad de variables. Para decidir si un conjunto de variables está correlacionado, utilizamos el "análisis de componentes principales" (PCA). Un conjunto de variables se considera correlacionado si el autovalor más pequeño de la matriz de covarianza correspondiente a esas variables es menor a 0,05 veces la suma de todos los autovalores, ya que esto garantiza que existe un hiperplano tal que las mediciones son puntos cercanos al mismo. En este caso, diremos que el grado de correlación de esas variables es al menos 95%.

Es posible recorrer todos los subconjuntos de dos, tres y acaso cuatro variables, estableciendo uno por uno cuáles están correlacionados y cuáles no. Sin embargo este método no sirve, por el tiempo que insume, para los conjuntos más grandes. Por lo tanto se requiere un algoritmo más eficiente.

Para plantear la solución al problema, comenzamos señalando los siguientes hechos:

1) Si C es un conjunto de variables y S es un subconjunto de C, el grado de correlación de S es menor o igual al de C.

2) En particular, si un conjunto C de variables está correlacionado (en la medida deseada), todo conjunto que contenga a C también lo estará. Por lo tanto estaremos interesados en hallar los conjuntos correlacionados minimales, es decir aquellos tales que ningún subconjunto posee correlación.

3) En el caso de la planta en estudio, la cantidad total de variables es 116. En base a pruebas preliminares, se determinó que el 82% de los conjuntos de 9 variables están correlacionadas en al menos un 95%.

En base a esto propusimos el siguiente algoritmo para hallar conjuntos de variables correlacionadas minimales. Se parte de una muestra al azar de nueve variables y se calcula su grado de correlación. Si éste da menor al 95%, se descarta la muestra y se elige otra. En caso de dar al menos 95%, quitamos del conjunto una variable y calculamos el grado de correlación del conjunto resultante. De esta forma, se recorre el árbol de subconjuntos del conjunto original, deteniendo la inspección en aquellos subconjuntos con grado de correlación inferior al 95% (ya que por (1) todos sus subconjuntos tampoco tendrán correlación). En otras palabras, sólo se

recorre aquellos nodos del árbol que poseen correlación, encontrándose así todos los conjuntos de variables correlacionadas minimales que son subconjuntos de la muestra de nueve tomada al azar.

Este proceso se repite tantas veces como se quiera. Se observa que cada iteración encuentra en promedio aproximadamente 30 correlaciones minimales demorando en promedio 0,1 segundos, de manera que es posible analizar 1.000.000 de muestras en el lapso de un día, aproximadamente. Para evitar guardar soluciones repetidas, proponemos lo siguiente: la elección al azar de una muestra de nueve variables se hará eligiendo las variables de a una comprobando que ningún subconjunto de las elegidas esté en la lista de correlaciones minimales halladas hasta el momento. Observación importante: es posible hacer esta comprobación sin necesidad de recorrer toda la lista, cuyo tamaño puede llegar al orden de los millones.

Este método permite encontrar una gran cantidad de correlaciones minimales de todos los tamaños menores a 10. Estimaremos a continuación la cantidad de conjuntos de k variables contemplados por el algoritmo, sobre el total.

Supongamos que la cantidad total de variables es N. Dado un conjunto de k variables, con $1 \leq k \leq 9$, calculemos la probabilidad p de que al elegir un conjunto de 9 variables al azar (con distribución uniforme), las k variables originales pertenezcan a la muestra de 9.

La cantidad de conjuntos de 9 variables que contienen a las k dadas es igual a $\binom{N-k}{9-k}$. Dividiendo por la cantidad total de conjuntos de 9 variables:

$$p = \frac{\binom{N-k}{9-k}}{\binom{N}{9}} = \frac{{}^{(N-k)!}/_{(9-k)!(N-9)!}}{{}^{N!}/_{(N-9)!9!}} = \frac{9\ .\ 8.....(9-k+1)}{N.(N-1)....(N-k+1)}$$

Para r muestras independientes la probabilidad de que el conjunto original de k variables no haya sido contemplado en ninguna muestra es $(1-p)^r$.

Suponiendo $N = 116$ y $r = 1.000.000$ obtenemos:

$$k = 4 \qquad p = 1,8 \times 10^{-5} \qquad (1-p)^r = 2,2 \times 10^{-8}$$

Es decir que prácticamente la totalidad de los conjuntos de 4 variables ha sido contemplada en alguna muestra.

$$k = 5 \qquad p = 7,9 \times 10^{-7} \qquad (1-p)^r = 0,46$$

De modo que el 54% de los conjuntos de 5 variables ha sido contemplado.

$$k = 6 \qquad p = 2,8 \times 10^{-8} \qquad (1-p)^r = 0,97$$

Sólo el 3% de los conjuntos de 6 variables ha sido contemplado. Dado que la cantidad de correlaciones minimales de tamaño 6 que el algoritmo halla es considerablemente grande, cabe preguntarse si esta muestra del 3% es suficientemente representativa o no para los fines del diagnóstico de fallas.

2.2 Diagonal Dominancia

Se busca un criterio que permita descartar eficientemente conjuntos de variables que no presentan correlación. Dado que el cálculo de los valores singulares insume una cantidad de tiempo apreciable al realizarlo gran cantidad de veces, se busca un método computacionalmente menos costoso.

Una matriz A de orden $k \times k$, se dice *estrictamente diagonal dominante* cuando

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|$$

Es sabido que las matrices estrictamente diagonal dominantes son inversibles. Esto, aplicado a una matriz de correlación, es equivalente a que el menor valor singular sea distinto de 0. De esta forma, es esperable que exista un criterio similar al siguiente: para cierta constante pequeña $C > 0$, las desigualdades

$$|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ij}| + C$$

garantizan que las variables no están correlacionadas en la medida requerida.

Evaluación de la propuesta:

Para realizar una valoración de la propuesta se consideraron 50 variables con 7034 observaciones, para la cual se obtuvo una matriz de correlación de orden 50×50, y se consideraron relaciones entre tres y cuatro variables, de modo que las sub matrices de correlación fueron de orden 3×3 y de orden 4×4. En la tabla 2 se muestran los resultados. La columna que indica el total de sub matrices que son estrictamente diagonal dominantes, el 80,7 % del total matrices de 3×3, y el 47,8% de total de matrices de 4×4, no habría que aplicarles el PCA. La columna encabezada con Fallos, está relacionada con la cantidad de sub matrices que son de diagonal estrictamente dominante, y sin embargo las dos primeras componentes principales explican el 95% o más de la correlación.

Orden de las matrices	Total de combinaciones	Total Diagonal Dominante	Fallos
3×3	14190	11447 (80.7 %)	376 (2,65 %)
4×4	148995	71154 (47.8 %)	2365 (1,6 %)

Table 2: Resultados Obtenidos para $n = 3$ y 4

2.3 Cota del Determinante

Sea B la matriz de correlación de un conjunto de n variables con m observaciones. Al ser B una matriz de correlación, tiene las siguientes propiedades:

- Es Simétrica, semidefinida positiva (o sea, $\langle Bx, x\rangle \geq 0 \ \forall \ x$)
- Tiene unos en la diagonal

Por lo tanto, diagonaliza en base ortonormal (teorema espectral para operadores autoadjuntos reales).

Sean entonces

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$$

los autovalores de B contados con multiplicidad.

Dado que nos interesa discernir si $\lambda_n < cn$, con $c = 0,05$, observamos que en dicho caso el determinante será menor a una constante, como mostramos a continuación. Utilizamos la desigualdad aritmético-geométrica para los restantes $n-1$ autovalores. Notar que si alguno de ellos es 0 la cota vale trivialmente.

$$det(B) = \left(\prod_{i=1}^{n-1} \lambda_i\right) \lambda_n \leq \left(\frac{\sum_{i=1}^{n-1} \lambda_i}{n-1}\right)^{n-1} \lambda_n <$$

$$< \left(\frac{tr(B)}{n-1}\right)^{n-1} cn = \left(\frac{n}{n-1}\right)^{n-1} cn < e.cn$$

Si $\det(B)$ supera esa cota, se puede asegurar que el conjunto de variables considerado no presenta correlaciones útiles y puede ser descartado. Notar que, aplicando la desigualdad aritmético-geométrica para todos los autovalores obtenemos $det(B) \leq 1$ sin importar si existe o no correlación. De modo que la cota pierde efectividad a medida que n crece.

3 Otros métodos propuestos

Entre los otros métodos propuestos para continuar con el problema, se presentan los más importantes

3.1 Métodos para reducir matrices relacionadas a grafos a forma de bloques

Existen algoritmos destinados a reducir matrices relacionadas a un grafo a una matriz en banda, u otras matrices similares con mejores propiedades algebraicas. Se sugirió que un cambio en el orden de las variables (intercambio de dos filas seguido por el intercambio de las dos columnas correspondientes), podría permitir mejorar el caso en que el problema resulta reducible y se puede llevar a forma de bloques. En este sentido, e inspirados en la idea de la diagonal dominancia, se llegó a probar el reordenamiento de filas y columnas de la matriz, de acuerdo a la norma 1 de la fila o columna correspondiente (de mayor a menor, y respetando el mismo cambio en filas y columnas para que represente sólo una permutación de las variables).

Los resultados muestran que si bien no se consigue una forma de bloques, la matriz presenta un mejor aspecto, con los elementos de mayor magnitud cerca de la diagonal, lo que hace que parezca valer la pena un análisis posterior más profundo.

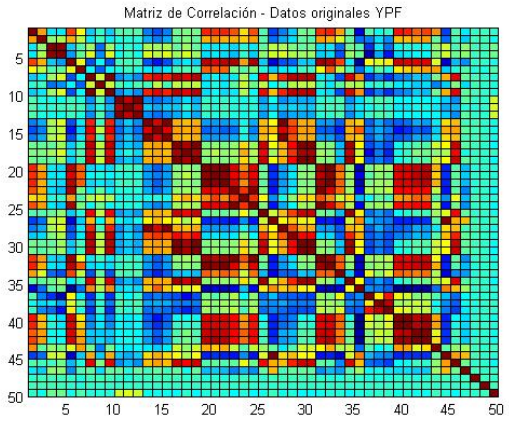

Figura 5: Matriz Original de Correlación

3.2 Eliminación de variables excesivamente correlacionadas

Se observa que existe en la matriz de correlación algunas variables con correlación superior al 99%, lo que hace que el comportamiento de una de estas variables con respecto al resto sea exactamente

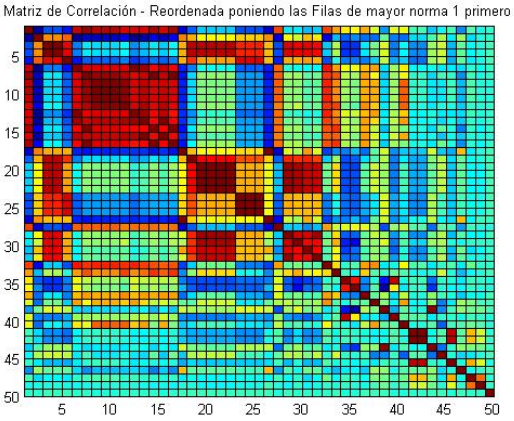

Figura 6: Matriz reordenada según norma 1 de las filas, en orden descendente

el mismo, resultando itercambiables entre sí. Si bien el grupo es pequeño (aproximadamente 10 variables), en una combinatoria tan grande la mejora puede resultar significativa

3.3 Otros métodos algebraicos

Se plantea que se pueden encontrar otros métodos algebraicos similares a Diagonal Dominancia, que permitan mediante ecuaciones simples verificar directamente en la matriz de correlación, que el correspondiente subgrupo de variables va a ser útil o no. Esto requiere mayores investigaciones en las características de las matrices involucradas.

3.4 Cálculo del autovalor más chico de la matriz de correlación en lugar de PCA

El comando PCA de matlab calcula todos los autovalores de la matriz de correlación. Es esperable que exista una forma más eficiente de obtener (o incluso estimar convenientemente) solamente el autovalor más chico, que es el unico que se requiere. Por ejemplo explotando el hecho de que el autovalor mínimo de una matriz real A simétrica y semidefinida positiva es igual a $||A|| - ||A - ||A||.Id||$.

Referencias

[1] Varmuza, K., Filzmoser, P., *Introduction to multivariate statistical analysis in chemometrics*, CRC Press, 2009.

[2] Elden, L. *Matrix Methods in Data Mining and Pattern Recognition*, SIAM, Philadelphia, USA, 2007.