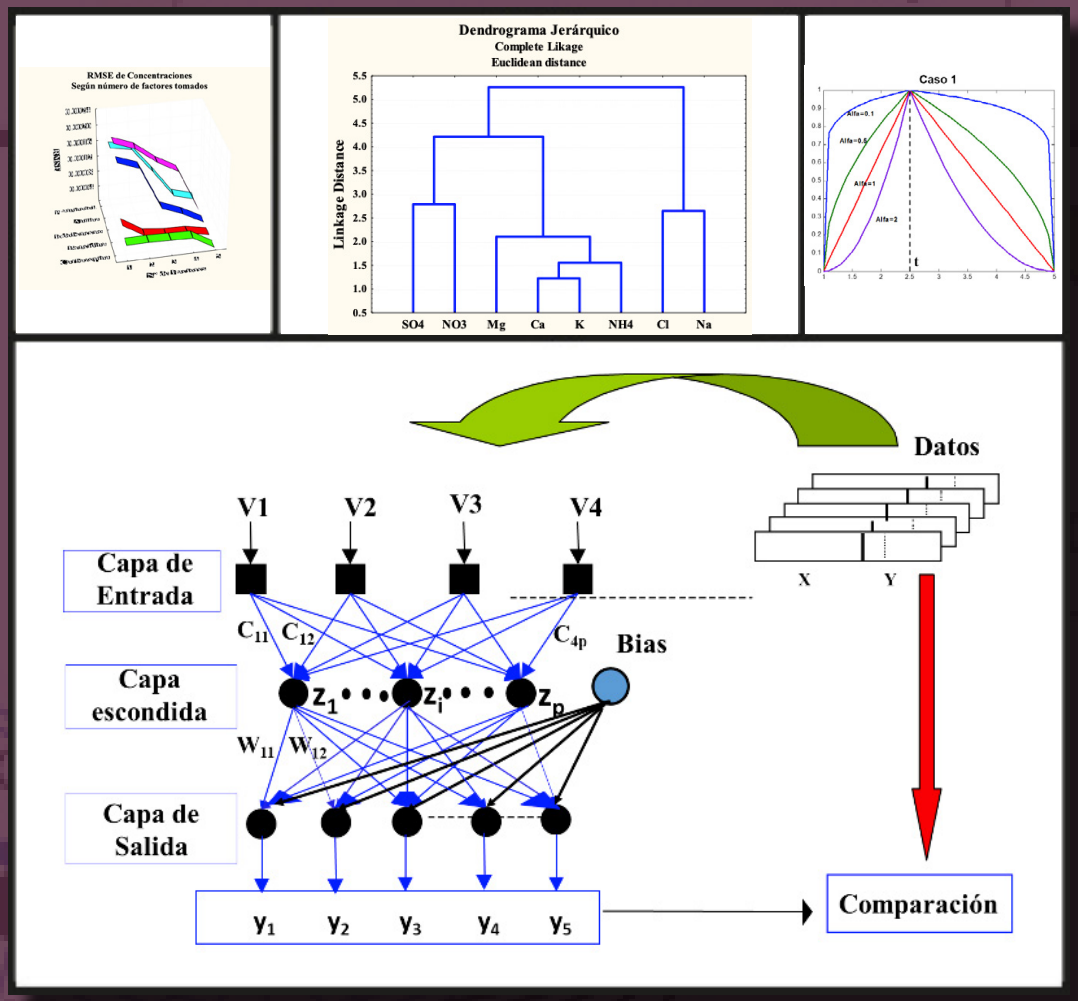
Introducción a la Quimiometría (o Infometría)

Para Científicos e Ingenieros



Jorge Federico Magallanes

Jorge Federico Magallanes

Doctor en Química (UNLP). Desarrolló su carrera en el área de Química Analítica de la Comisión Nacional de Energía Atómica (CNEA) donde actualmente es Investigador Consulto. Desde 1996 fue orientando progresivamente su especialidad hacia la quimiometría.

Fue Profesor adjunto del Departamento de Química Inorgánica, Analítica y Química Física de la Universidad de Buenos Aires entre 1989-2005 y Profesor contratado para el Curso de posgrado en Quimiometría en la Universidad Nacional de San Martín entre 2005 y 2019.

Fue Investigador responsable de la contraparte Argentina del Convenio Bilateral Argentino-Esloveno SLO/08/12 entre la Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT) y MHEST de Eslovenia, y experto en Chile por el Organismo Internacional de Energía Atómica, entre otras actividades internacionales.

Fue vicepresidente (1999-2001) y presidente (2001-2003) de la Asociación Argentina de Químicos Analíticos, asociación que le otorgó en 2019 el Premio a la Trayectoria en Química Analítica.

Magallanes, Jorge Federico
Introducción a la Quimiometría : o Infometría : para científicos e ingenieros / Jorge Federico Magallanes. - 1a ed. - Ciudad Autónoma de Buenos Aires : Asociación Argentina para el Progreso de las Ciencias, 2023.
Libro digital, PDF
Archivo Digital: descarga y online
ISBN 978-987-48617-2-6
1. An·lisis de la Información. 2. Interpretación. 3. Ingeniería. I. Título.
CDD 620.002

Ficha de catalogación

Introducción a la Quimiometría (o Infometría).

Apuntes Autobiográficos de un Científico Argentino.

Jorge Federico Magallanes

Diseño: AAPC

Maquetador: Gabriel Gil

Editado en 2023 por

AAPC
ASOCIACION ARGENTINA PARA EL PROGRESO DE LAS CIENCIAS

ISBN Nº 978-987-48617-2-6

Queda hecho el depósito que marca la Ley 11.723

http://www.aargentinapciencias.org/

A mis padres, por su amor y sacrificios:

María Catalina Rodríguez y Federico Gabino Magallanes

Agradecimientos

A la Comisión Nacional de Energía Atómica:

Por su apoyo en todos los emprendimientos profesionales de mi carrera.

A la Universidad Nacional de San Martín:

Donde he dictado cursos de posgrado en quimiometría desde el año 2005.

A la Asociación Argentina de Químicos Analíticos:

Por ser el principal foro de discusión del desarrollo de la quimiometría en el país, además de los temas concernientes a la química analítica.

PRÓLOGO

Este libro surge como resultado de muchos años de docencia en el dictado de un curso de posgrado sobre el tema. La necesidad de estos conocimientos hizo que el curso se dictara anualmente en forma regular y presencial, pero también se dictó en forma *full time* en universidades del interior del país e incluso en el exterior del país. También hubo una ocasión de dictado a distancia para ocho países sudamericanos simultáneamente. Durante toda esta experiencia he tenido alumnos, la mayoría de ellos profesionales, o próximos a recibirse, en carreras tan diversas como geología, farmacia, matemática, medicina, medioambiente, alimentación, física y biología entre otras; donde rara vez los químicos fueron mayoría.

Esto es debido a que la quimiometría tiene cierta universalidad y técnicas iguales o similares se aplican, por ejemplo, en psicometría, econometría y biometría entre otras. De allí la extensión del título a *Infometría*, que suena más apropiado para resolver problemas de interpretación y análisis de la información en muchas áreas de la ciencia e ingeniería.

Es conveniente aclarar que esta especialidad es muy joven, ya que ha surgido como consecuencia de la evolución de la electrónica y la instrumentación. Es así que tuvo un exponencial crecimiento cuando las computadoras comenzaron a formar parte de los laboratorios y al mismo tiempo los instrumentos comenzaron a brindar cantidades de información mucho más grandes que en el pasado. Entonces comenzó a hacerse complicada la utilización numérica ventajosa de toda esa información y al mismo tiempo se pudieron utilizar técnicas matemáticas conocidas desde tiempo atrás, pero imposibles de aplicar manualmente a grandes cantidades de datos; el ejemplo más típico es el álgebra de matrices.

Todo esto comenzó en la década de los años 80.

Varias veces en el pasado hemos coincidido con mis colegas en que un curso de quimiometría o infometría debería dictarse como curso de grado en los primeros años de las carreras de ciencias. Esto permitiría a los alumnos aplicar estos conocimientos a todo lo que aprendiesen en el resto de sus carreras. Como esto aún no ha sucedido y anualmente recibo consultas, en su gran mayoría de profesionales, interesados acerca de cuándo se dictará el próximo curso, es que he decidido encarar la tarea de reunir los conocimientos más esenciales en un libro. De él espero que sea lo suficientemente claro

para todos quienes deseen iniciarse en el tema y al mismo tiempo lo suficientemente profundo como para que les permita su aplicación a los problemas de sus especialidades.

Para esto último es imprescindible que se ejerciten con las prácticas asociadas a cada capítulo de este libro. Este libro puede usarse también en un segundo curso de estadística enfocado en el diseño estadístico de experimentos para estudiantes de licenciatura, ingeniería, física, ciencias físicas, químicas, matemáticas y otros campos de las ciencias.

También es mi deseo, que les sirva de guía a aquellos quienes quieran dedicarse a profundizar en esta actividad; para ellos es la introducción de referencias bibliográficas en cada capítulo.

Conocimientos previos convenientes para facilitar el aprendizaje

En principio cualquier estudiante de grado de las carreras de ciencias o ingeniería podría comenzar a interiorizarse en este libro si tiene al menos conocimientos de estadística básica y práctica con programas tales como Excell® y Word®. Sin embargo, en lo que sigue, se necesitará introducir cierto bagaje de nuevos conocimientos para aquellos que no los han tenido en su formación previa. De todos modos, el inicio con la falta total de estos nuevos conocimientos nunca fueron motivo de abandono o desánimo para alumnos que estaban, debido a su particular formación, muy alejados de ellos, sobre todo ante la carencia en matemática y programación, éstos los fueron adquiriendo durante el progreso del curso.

A medida que se avance en los capítulos, los nuevos conceptos resultarán más sencillos para aquellos que tengan conocimientos de álgebra lineal, al menos en operaciones con vectores y matrices.

También, debido a la escasez de literatura en español y a la poca adaptación de términos a nuestro idioma, el conocimiento de inglés facilitará la comprensión y memorización de conceptos nuevos. En muchos temas se introduce la nomenclatura en inglés ante la falta de divulgación en términos aceptados en español.

Otro aspecto que es necesario mencionar es el cálculo matemático mediante programas específicos. He implementado los ejemplos y las prácticas bajo un supuesto conocimiento elemental del programa Matlab®. Quien no posea este conocimiento puede adquirirlo fácilmente hoy en día mediante la consulta de cualquier manual, libro o curso en páginas web, introductorios en el tema. A continuación, se muestran algunas citas para tratar estos temas:

https://la.mathworks.com/help/matlab/getting-started-with-matlab.html

http://www.mat.ucm.es/~jair/matlab/notas.htm

Aprenda Matlab 7 como si estuviera en primero. Javier García de Jalón, José Ignacio Rodríguez, Jesús Vidal. Adobe Acrobat Document®. Pdf

Más avanzado:

Matlab para ingenieros. Holly Moore. Pearson-Prentice Hall®.

Declaración de responsabilidad sobre el uso de los programas de cálculo desarrollados para este libro

Este libro consta de una parte teórica y otra de prácticas. En varios de sus capítulos se utilizan programas de cálculo computacional, en adelante el "*software*", para resolver los ejercicios de las prácticas o comprobar los desarrollos teóricos. El software incluido es un complemento destinado al uso exclusivo y específico para este libro o ejercitaciones que deseen practicar los lectores a problemas similares a los aquí tratados. En ningún caso está permitida la modificación, parcial o total del software para otras aplicaciones. Tampoco está permitido el uso comercial del software. Cualquier aplicación no autorizada del mismo fuera del propósito de este libro es responsabilidad exclusiva de quién lo ejecute. No se admitirán reclamos por el uso y resultados obtenidos, cualesquiera fueran, de quién no esté autorizado a hacerlo.

Material adicional

Los programas que se usan en este texto se pueden encontrar en el siguiente link:

https://aargentinapciencias.org/introduccion-a-la-quimiometria-o-infometria/

Hay diferentes tipos de software que debe poseer el lector para utilizar en este libro, a saber: Word, Excell y Matlab®.

Los programas especiales, no los comandos, en Matlab han sido desarrollados por el autor y pueden usarse tanto para resolver las prácticas como para chequear los resultados expuestos en la teoría.

Si bien las prácticas tienen cuada una, un archivo “Práctica resuelta”, se sugiere al lector calcular los ejercicios propuestos y luego chequearlos con el archivo mencionado, especialmente si se pretende aplicar los conocimientos adquiridos a problemas propios.

ÍNDICE

…Había aprendido sin esfuerzo el inglés, el francés, el portugués, el latín. Sospecho, sin embargo, que no era muy capaz de pensar. Pensar es olvidar diferencias, es generalizar, abstraer. En el abarrotado mundo de Funes no había sino detalles, casi inmediatos.

De “Funes el memorioso”

Jorge Luís Borges

PRIMERA PARTE

Técnicas de Cálculo Lineal

CAPÍTULO 1

Introducción a conceptos básicos elementales

Definición

Es mi parecer, que todas las definiciones de la quimiometría que conozco son incompletas. Sin embargo, expondré una que me parece la mejor; ignoro quién la mencionó originalmente en inglés y la reproduzco traducida.

La quimiometría es la ciencia de extraer información de sistemas químicos a través de la interpretación de los datos. Es una disciplina altamente interfacial, que utiliza métodos frecuentemente empleados en disciplinas de análisis de datos tales como estadística multivariada, matemática aplicada y ciencias de la computación con el fin de resolver problemas en química, bioquímica, medicina, biología e ingeniería química.

Técnicas Quimiométricas

En muchas áreas de la Ciencia y tecnología, como por ejemplo en química, biología, medioambiente y todas las áreas en las cuales se trabaja con muchas variables, los datos obtenidos experimentalmente son generados con el objeto de "***caracterizar"*** un fenómeno.

Esta tarea es sencilla cuando se manejan solo 2 variables (incluso 3, en una computadora) ya que podemos representar gráficamente estos datos y analizar sus interrelaciones. Las técnicas instrumentales modernas asistidas por computadoras pueden proporcionar mucho más que 3 variables por muestra u objeto*. En el caso de la química, por ejemplo, existen hoy en día técnicas multi-elementales, como son fluorescencia de rayos X (FRX) o plasma inductivamente acoplado combinado con espectrometría de masa (ICP-MS), entre otras, que proporcionan simultáneamente el espectro o las masas y la composición cuantitativa de varios (o todos) los componentes de una muestra.

Cabe preguntarse ahora si toda esta recopilación de datos es absolutamente imprescindible a los fines de nuestro análisis y además, ¿como vamos a manejarla?, ya que no podemos representar más de 3 variables a la vez. Preguntas de este tipo podemos hacerlas en diferentes campos de estudio, por ejemplo:

a) Un análisis de sangre proporciona unas 20 variables, como caracterizar las posibles enfermedades según la combinación de estas variables?

b) ¿Cuantas variables necesito realmente para caracterizar un tipo agua o producto?

c) ¿Cómo construyo un modelo que me relacione las variables con el comportamiento del sistema en estudio?

Entonces…

El problema central y general del cual nos ocuparemos, es ¿Cómo extraer información sustantiva desde una gran cantidad de datos?

El problema grave de no analizar la información adecuadamente puede resultar no solo en una pérdida de interpretación del sistema en estudio, sino también en una interpretación errónea de los resultados, que conducirán a conclusiones falsas.

**Llamaremos objeto a una serie de parámetros que constituyen un vector asociado a un evento de medición: por ejemplo, se extrae una muestra de agua y de ella se toman en cuenta su cantidad de sólidos en suspensión, sólidos disueltos, 7 especies químicas, fecha y lugar de muestreo. Todas estas variables constituyen un vector que será un objeto para el análisis de un conjunto de ellos.*

Un ejemplo tipo - Supongamos un simple sistema con sólo 2 variables: pesos y alturas promedio de la población de adultos mayores, cuya tabla ha sido elaborada con datos de la Organización mundial de la salud OMS (Fig. 1.1, Ref. 1). Las líneas llenas que cruzan el centro de la figura marcan los valores medios de ambas variables, las líneas punteadas paralelas a las anteriores marcan una desviación estándar (DS) desde la media, los puntos en el cuadro son los pesos y alturas registrados.

Veamos cómo puede interpretarse esta información con el objeto de mantener una buena salud:

a- Analizando independientemente una variable a la vez se puede pensar que, si estamos dentro de una DS con el peso, o sea entre 55.4 y 104.6 Kg, nuestro peso es normal. Y Lo mismo puede pensarse respecto de la altura, entre 1.60 y 1.95 metros. Con lo que todos los individuos normales y supuestamente saludables caen dentro del rectángulo de líneas punteadas.

b- Analizando las variables en forma sucesiva, una persona que supere por poco el límite de peso y en igual forma el de altura podría considerarse saludable (lo mismo regiría para los límites inferiores).

c- **Analizando ambas variables a la vez:** si se comparan los datos de la tabla con el Índice de Masa Corporal (IMC), que incluye a ambas variables, los datos quedan divididos en zonas, de las cuales la de individuos normales es la que representa un índice entre 18.5 y 25 metros/Kg^2 (triángulos celestes). Vemos ahora que muchos individuos saludables están fuera del rectángulo y que otros que antes estaban dentro y parecían normales y saludables en realidad no lo eran.

Más adelante volveremos varias veces sobre este tema, con fundamentos matemáticos. La importancia de este ejemplo con solo dos variables es mucho mayor aún cuanto mayor sea el número de variables que incluye el problema. Se destaca este principio en la siguiente regla general básica:

No es lo mismo analizar un problema en forma univariante (Una variable a la vez, *one variable at time, 'OVAT' en inglés*) que en forma multivariante (todas en conjunto) debido a las interrelaciones entre ellas.

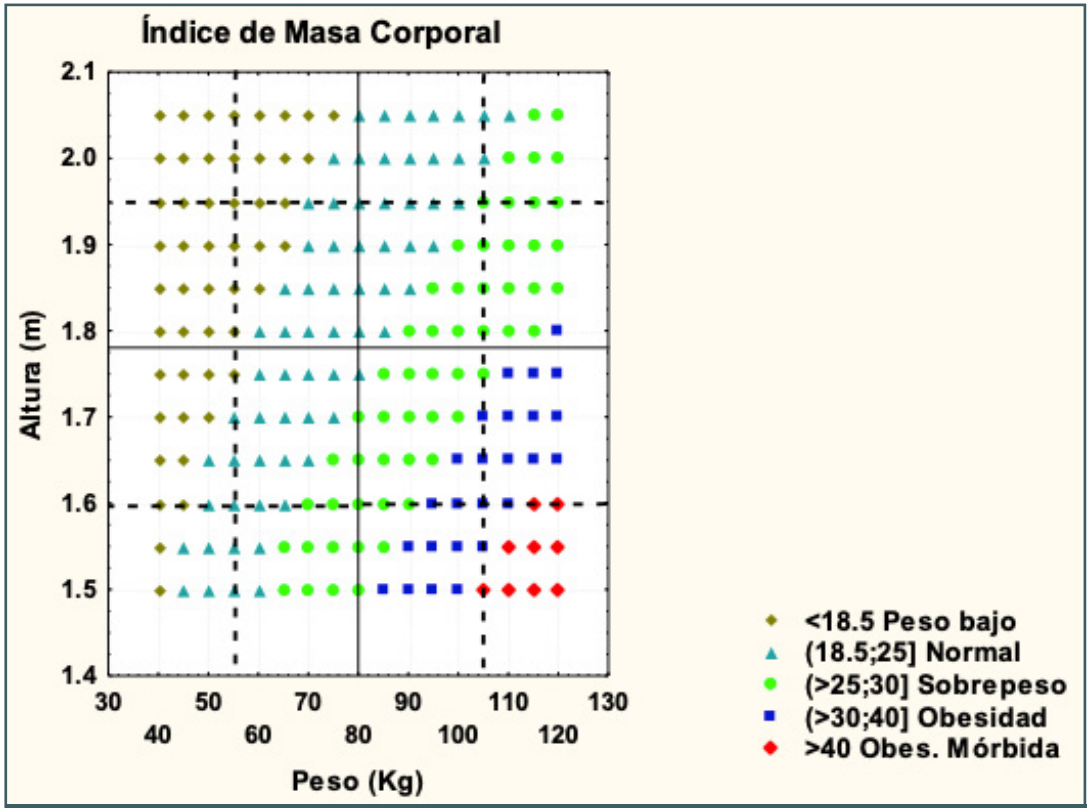


Figura 1.1

La ignorancia de esta regla puede conducir a grandes errores de interpretación. Por lo tanto, de aquí en más se desalienta el uso del sistema OVAT.

La Información Visual

Aunque no parezca evidente al sentido común, el modo habitual (y casi exclusivo) de incorporar conocimiento, es en forma visual (gráfica). Alguien podría argumentar que si expreso la ecuación de una recta cualquier profesional de ciencias comprenderá la relación entre variables. Pero en realidad esto ocurre porque éstos conocen de antemano la relación entre la ecuación y la imagen de la recta. Alguien no avezado a las matemáticas no comprendería la ecuación. **¿Podría** acaso el lector comprender la relación entre x e y en la siguiente ecuación, sin representarla?

$$y = \frac{x}{ax^2 + bx + c}$$

Aceptando este concepto podemos analizar qué ocurre cuando agregamos más dimensiones espaciales a nuestros datos. Supongamos que tenemos un lote de objetos y queremos saber si hay grupos diferentes o no entre ellos midiendo solo una propiedad (una variable). Representando los resultados en un gráfico univariante como el de la figura 1.2 (a), podríamos tener la duda de si se trata de dos grupos (distribuciones en negro) o de uno solo (distribución azul). Podríamos recurrir al análisis estadístico y definir esta situación. Pero en lugar de ello mediremos otra propiedad de los objetos (agregamos una variable) y volvemos a representar los puntos. Como resultado obtendríamos un gráfico en **dos dimensiones** y en él observamos claramente que hay dos grupos bien diferenciados sin necesidad de interpretar a través de un cálculo.

Podríamos entonces, por curiosidad, agregar una tercera variable para ver que obtenemos y representar los resultados **tridimensionalmente**.

Ahora aparecen diferenciados tres grupos de objetos, uno de los cuales podría estar subdividido.

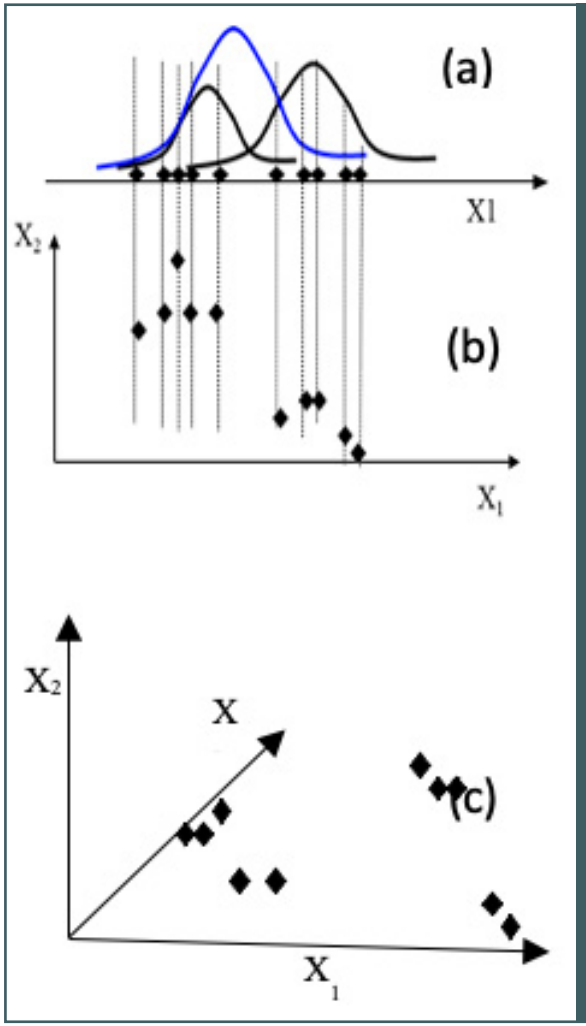


Figura 1.2

"Es obvio que cuando agregamos más variables de análisis tenemos mejor información visual".

Observar que el número de variables es igual al número de dimensiones del problema. ¿Cómo beneficiarnos agregando más variables y mantener una representación comprensible?

Para hacer uso del reconocimiento visual habrá que representar los datos en 2 o 3 dimensiones. Lo que debemos resolver entonces es: **como condensar** información m-dimensional en 2 o 3 dimensiones.

Dendrogramas

Como su nombre lo indica, esta es una representación en forma de ramas de un árbol. Está inspirado en las clasificaciones de especies biológicas como plantas, animales, bacterias, etc. Como ejemplo se muestra un lote de muestras de aire donde se han determinado 7 componentes. Queremos saber cuáles son más similares entre sí respecto a las concentraciones de sus componentes y si podemos, además, ver agrupaciones entre los componentes. Tengamos en cuenta que estamos manejando 7 variables (o sea 7 **dimensiones**) y representaremos este sistema en sólo dos dimensiones. Nos referiremos ahora a la figura 1.3.

En el eje horizontal se identifica cada una de las muestras (objetos) a través de una línea vertical. Estas líneas se unen a cierta distancia del eje vertical. En éste, se indican distancias entre las muestras (aunque todavía no las hemos definido), cuanto menor es la distancia a la cual se unen las muestras, más similares son entre sí, e inversamente cuanto más alto es el nivel de unión de dos líneas, mayor es su diferenciación. Obsérvese que dos muestras que se unen, se conectan a su vez con otra muestra o grupo de muestras a una altura mayor, pudiéndose así reunirlas en **grupos** de diferente similitud. Un grupo 'de baja altura' indica muestras muy similares entre sí, como el grupo de la derecha entre las muestras 1 y 33. Más adelante volveremos detalladamente sobre este tema para poder calcular un dendrograma desde un lote de muestras. También veremos otros modos de representaciones m-dimensionales

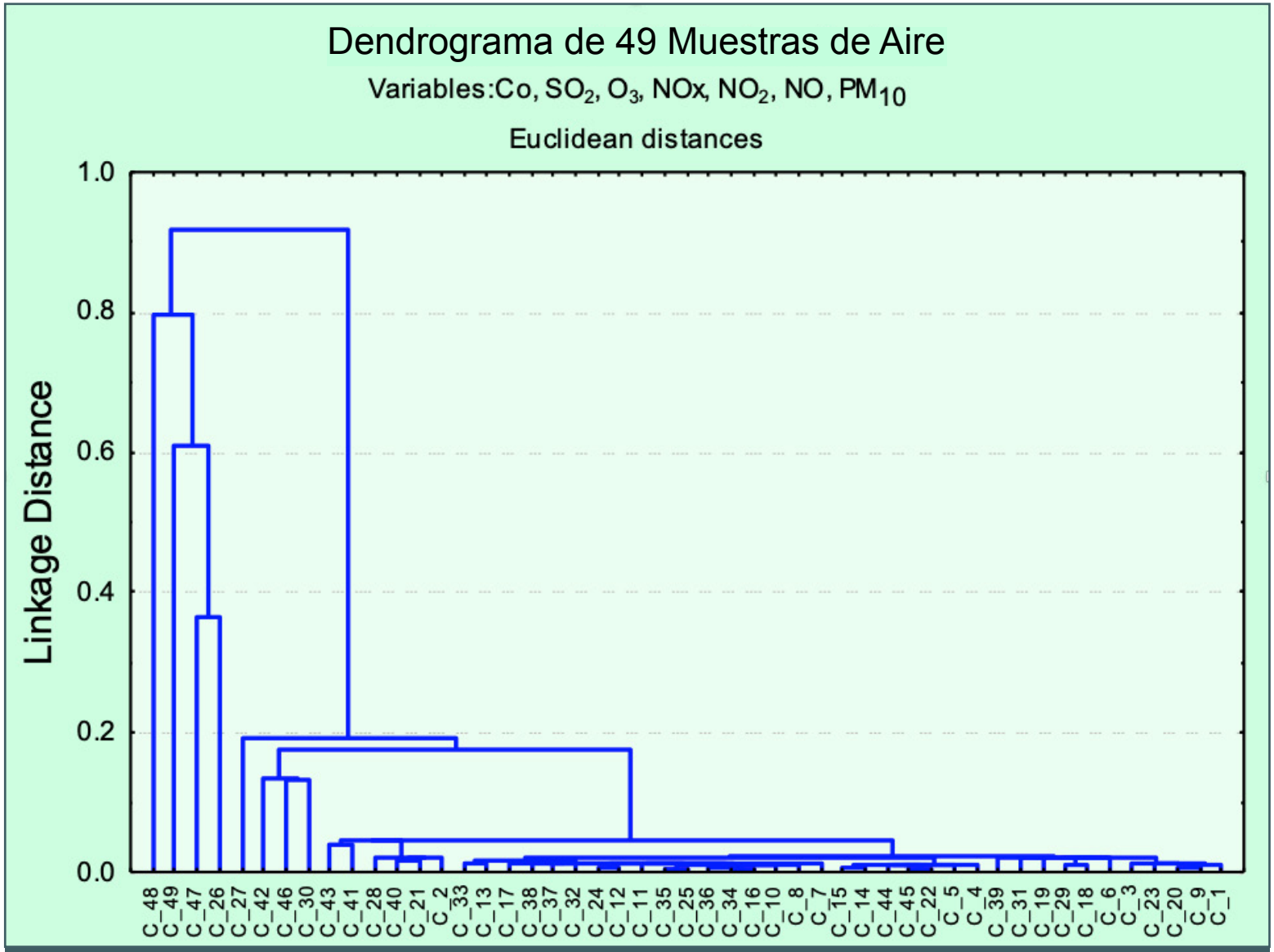
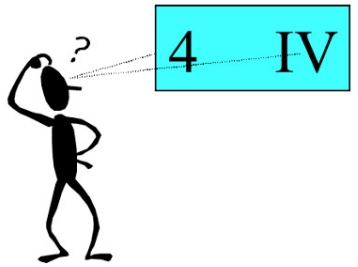

Dendrograma de 49 Muestras de Aire
Variables:Co, SO_2, O_3, NOx, NO_2, NO, PM_{10}
Euclidean distances
Linkage Distance
1.0
0.8
0.6
0.4
0.2
0.0
C_48 C_49 C_47 C_26 C_27 C_42 C_46 C_30 C_43 C_41 C_28 C_40 C_21 C_2 C_33 C_13 C_17 C_38 C_37 C_32 C_24 C_12 C_11 C_35 C_25 C_36 C_34 C_16 C_10 C_8 C_7 C_15 C_14 C_44 C_45 C_22 C_5 C_4 C_39 C_31 C_19 C_29 C_18 C_6 C_3 C_23 C_20 C_9 C_1

Figura 1.3

Modelos



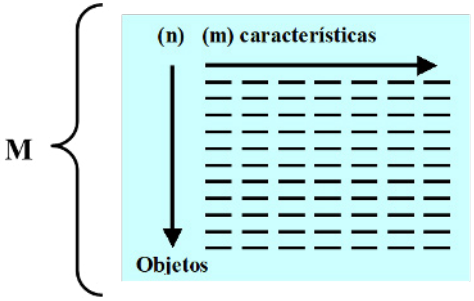
Hemos dicho que habitualmente la información que incorporamos es de tipo visual. Muchas veces, cuando se quiere representar la relación entre un grupo o subgrupos de objetos, se utilizan modelos, ya sea gráficos o algorítmicos. Por ejemplo, todos sabemos que los números no son una entidad física, pero los representamos con distintos modelos, como en la figura, de modo de ponernos de acuerdo a qué nos referimos.

Objetos caracterizados por sus variables pueden ser clasificados y relacionados a través de un modelo que discrimine alguna/s cualidad/es de interés. La obtención de modelos es **el fin último** del análisis de datos, que permite comprender, optimizar y dominar un sistema o proceso.

Un modelo reúne y relaciona una gran cantidad de información que habitualmente se representa en un plano. Suele ser la mejor síntesis del sistema que queremos interpretar y dominar. Un modelo no es siempre alcanzable, su concreción depende de la complejidad del sistema en estudio y de la cantidad y calidad de información reunida para su desarrollo.

Representación Algebraica de un Sistema Multivariable

Cuando un **objeto** está caracterizado por más de un parámetro, llamémoslos m>1, será conveniente representarlo por un ***vector*** m-dimensional. Así, los 'm' parámetros definen un ***espacio*** m-dimensional. El vector estará compuesto por 'm' valores escalares. Cada vector (o muestra) compondrá una fila dentro de una *matriz* **M** de 'n' muestras (u objetos).



M es una matriz de n filas por m columnas y la representaremos como $\mathbf{M}_{n,m}$ o $\mathbf{M}_{nxm}$ (observar el orden de los subíndices fila, columna). En adelante, tanto vectores como matrices son representadas en **negrita** para diferenciarlas de los valores escalares.

Si fuera posible observar la distribución de objetos en este espacio m dimensional, se vería que éstos tienden a formar grupos o "*clusters"* que nos permitirían comenzar a interpretar resultados. Como esto no es directamente posible, habrá que recurrir a diferentes técnicas de cálculo para conseguir el objetivo. Comenzaremos por el más sencillo de los métodos que usaremos posteriormente en técnicas más complejas.

Concepto de Covarianza y Correlación

Introduciremos ahora nuevos conceptos sobre covarianza y correlación para poder manejarlos en forma multivariable.

Supongamos que estamos estudiando el sistema $\mathbf{X}_{\mathbf{n,m}}$, el cual contiene m variables y del cual se han hecho n mediciones. Ignoramos aún, si existe o no alguna relación entre estas variables. Un primer paso para estudiarlas es tomar dos columnas, por ejemplo, **k** y **l**) de ellas ($\mathbf{x}_\mathbf{k}$ y $\mathbf{x}_\mathbf{l}$ en la matriz de datos) y ver en qué grado estas varían conjuntamente (en alguna dirección). Esa estimación se hace mediante la covarianza cuya definición viene dada por la ecuación [1].

$$\mathbf{X} = \begin{vmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{vmatrix}$$

$$\mathbf{cov}(\mathbf{x_k}, \mathbf{x_l}) = \frac{1}{\mathbf{n}-1} \sum_{\mathbf{h}=1}^{\mathbf{n}} (\mathbf{x_{h,k}} - \overline{\mathbf{x_k}})(\mathbf{x_{h,l}} - \overline{\mathbf{x_l}}) \qquad [1]$$

Donde $\bar{X}_k$ y $\bar{X}_l$ son los valores medios de las columnas k y l respectivamente (ver en el Anexo el significado de la media de un vector o *centrado*. La covarianza puede tomar valores entre +∞ y -∞. El problema principal de la covarianza es que su valor depende de las unidades de medida.

Este problema se soluciona si dividimos la covarianza por las respectivas desviaciones estándar, S_k, S_l, de las variables. Entonces obtenemos el Coeficiente de Correlación de Pearson, $r(x_k,x_l)$, que es adimensional (no depende de las unidades).

La correlación de Pearson

De acuerdo a la definición anterior, la correlación se expresa según la ecuación [2].

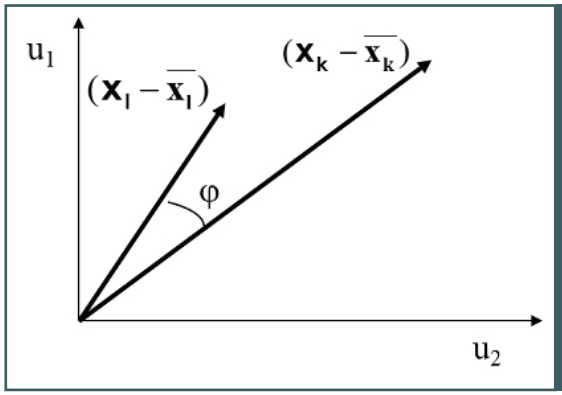
$$r(x_k,x_l) = cov(x_k,x_l)/(s_k.s_l) \qquad [2]$$

Esta relación la podemos expresar en forma vectorial, lo que nos permite definiciones más claras, tanto algebraica como geométricamente. $\mathbf{X_k} - \overline{X_k}$ y $\mathbf{X_l} - \overline{X_l}$ son vectores centrados de las columnas k y l, con lo que la ecuación queda:

$$r(x_k, x_l) = \frac{(\mathbf{x_k} - \overline{x_k})^T.(\mathbf{x_l} - \overline{x_l})}{\left\|\mathbf{x_k} - \overline{x_k}\right\| \left\|\mathbf{x_l} - \overline{x_l}\right\|} = \cos\varphi$$

Ver Anexo sobre este punto. [3]

El superíndice ^{T} indica la operación de *transposición* del vector. Algebraicamente, ahora podemos decir que la correlación es el **producto escalar de los vectores medios normalizados** de las variables. Por ser un producto escalar, este es igual al coseno del ángulo entre las variables. Por lo tanto, r puede variar sólo entre –1 y +1.



Un r<<|1| indica que las variables no están correlacionadas. Esto no implica que no haya relaciones entre las variables, sino solamente que no hay una relación ***lineal*** entre ellas.

Cuando se estudian muchas variables a la vez, contenidas en una matriz como $\mathbf{X}_{n,m}$, no es necesario calcular cada par de variables individualmente. En operación matricial podemos calcular la ***matriz de correlaciones***, donde aparecen las correlaciones entre todas las variables entre sí. Sea $\mathbf{X}_{nxm}$ una matriz de n observaciones por m variables. Calculamos la matriz autoescalada $\mathbf{X}_{n\bullet a}$, o sea *centrada en columnas* y dividida por la desviación estándar en cada columna. Entonces **cada elemento de** $\mathbf{X}_{nxm}$ debe ser transformado en:

$$\mathbf{x}_{i,j} = \left(\mathbf{x}_{i,j} - \overline{\mathbf{x}_j}\right) / \mathbf{s}_j$$

Finalmente, la matriz de correlación $\mathbf{R}_n$ es:

$$\mathbf{R}_n = 1/(n-1).\mathbf{X}_{n\bullet a}{}^{T}.\mathbf{X}_{n\bullet a}\,.$$

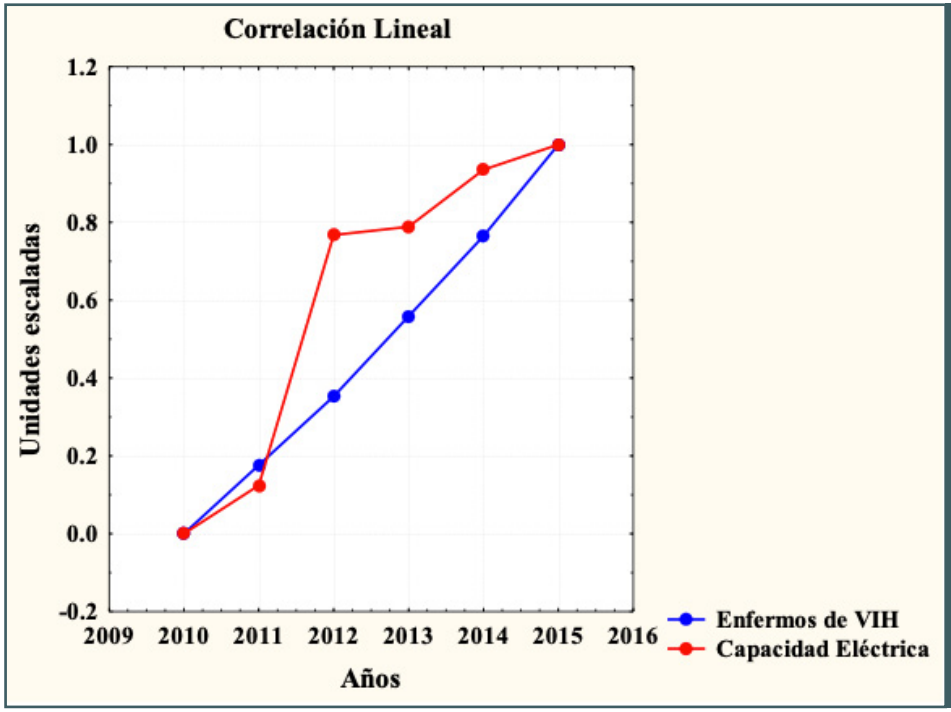


Figura 1.4

Debe tenerse cuidado con una interpretación apresurada acerca del significado de una correlación. Cuando una correlación está establecida entre 2 variables se la puede interpretar a través de su coeficiente de correlación usualmente incorporado a una expresión matemática como es por ejemplo la ecuación de una recta. Pero muchas veces se interpretan comportamientos entre correlaciones como si fueran similares y siguieran las mismas reglas.

Sin embargo, aunque tengan una misma expresión matemática, esto no significa **causalidad**. Es necesario tener en cuenta esto porque se suele tener tendencia a encontrar **una causa** cuando se observan estas relaciones, cuando en realidad puede tratarse sólo de casualidad. Las figuras 1.4 y 1.5 (Ref. 2-5) muestran lo aquí expresado.

Observemos ahora los valores algebraicos y gráficos de otra matriz de correlación entre 6 variables (X e Y1... Y5) obtenida desde 100 datos, o sea desde una matriz $\mathbf{M}_{100,6}$.

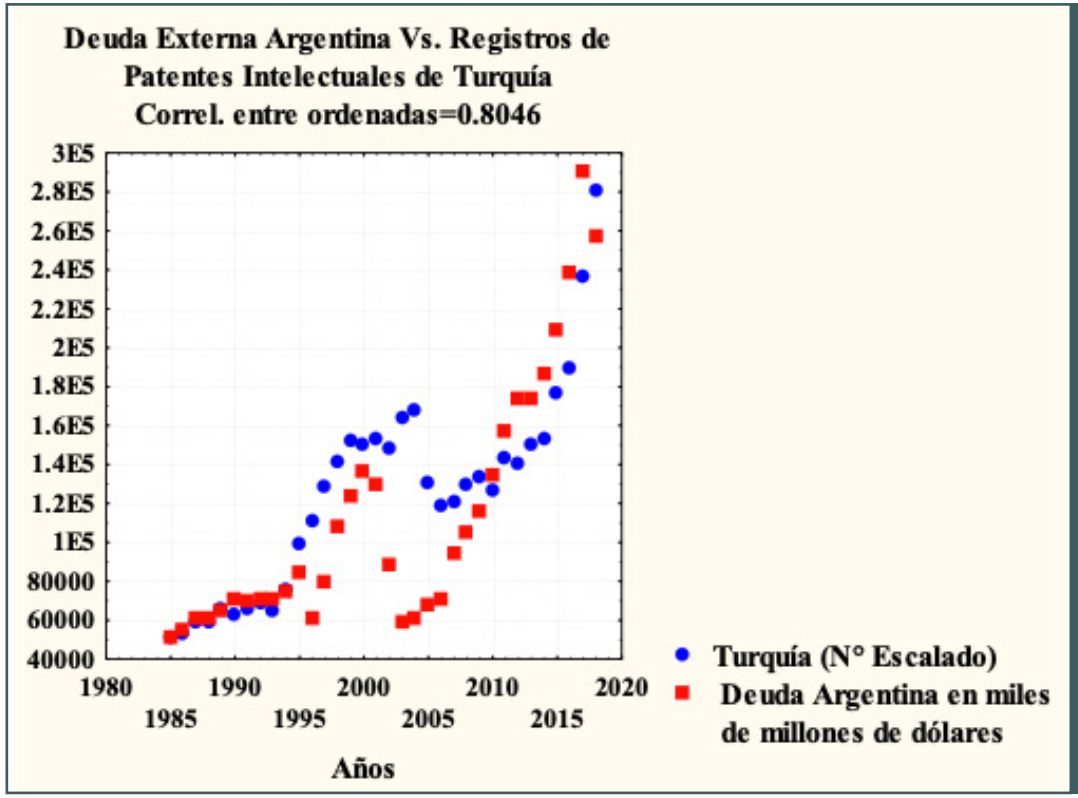


Figura 1.5

Matriz de correlaciones					
X	Y1	Y2	Y3	Y4	Y5
1.00	1.00	0.99	-0.99	-0.00	0.70
1.00	1.00	0.99	-0.99	-0.00	0.70
0.99	0.99	1.00	-0.98	-0.04	0.71
-0.99	-0.99	-0.98	1.00	0.03	-0.70
-0.00	-0.00	-0.04	0.03	1.00	-0.08
0.70	0.70	0.71	-0.70	-0.08	1.00

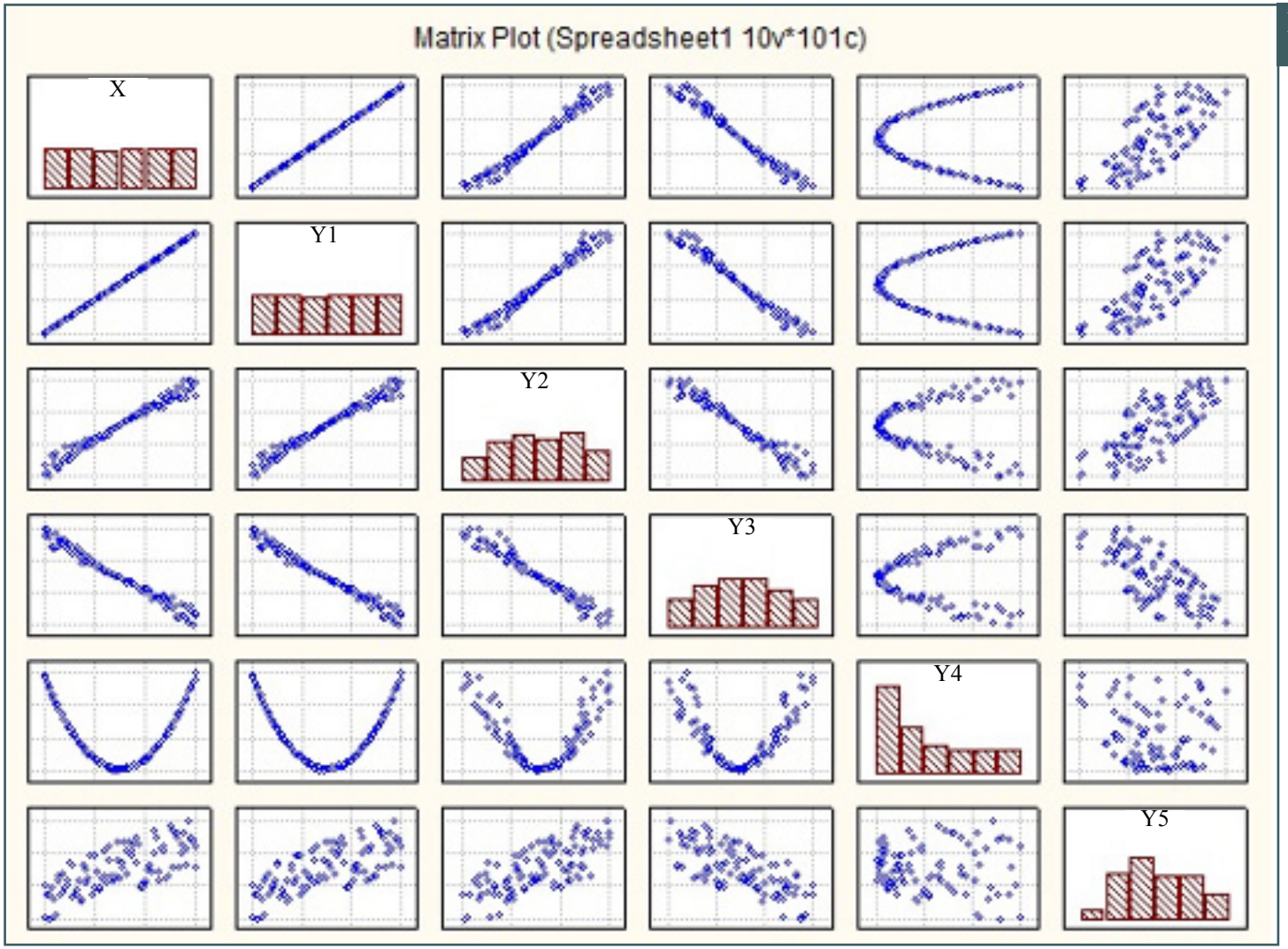


Figura 1.6

En primer lugar, vemos que la matriz de correlación es de 6x6 y simétrica respecto de la diagonal principal, de modo que la parte inferior izquierda o superior derecha guardan los valores de correlación entre todas las variables. La diagonal principal está compuesta por 'unos' debido a que el ángulo entre el vector medio de una variable con sí mismo es 0 y su coseno vale 1.

Veamos ahora el significado gráfico de las correlaciones. La correlación entre X e Y1 vale 1.00. En el segundo cuadro desde la izquierda, que es igual al segundo hacia abajo y que relacionan en este caso a X con Y1, se observa que los 100 datos guardan la relación de una perfecta línea recta entre estas 2 variables. La correlación entre X e Y2 vale 0.99, la relación entre ellas es claramente lineal, pero los datos están algo dispersos alrededor de la recta. Si tomamos X e Y3 (r = -0.99) vemos que la relación es también lineal pero su pendiente es negativa, como lo indica el signo de r. Para X e Y4 r tiene un valor muy bajo, redondeado a 0, sin embargo **esto no significa que no haya una relación entre las variables**, como se ve en la figura, sino solamente que la relación existente **no es lineal**.

Finalmente vemos que otros valores de r como 0.7, entre X e Y5, o X(1 a 3) e Y5, o -0.08 entre X5 e Y5 no muestran una clara relación entre las variables.

Así como hemos obtenido la matriz de correlación R_n, también puede obtenerse en forma similar la matriz de varianza-covarianza que se expresa como:

$$\mathbf{R_v} = 1/(n-1).\mathbf{X_c}^{\mathbf{T}}.\mathbf{X_c},$$

Donde $\mathbf{X_c}$ es la matriz centrada en columnas de X (pero sin dividir por s). La diagonal principal de $\mathbf{R_v}$ contiene la varianza de cada variable y los valores fuera de la diagonal son las covarianzas entre las 2 variables correspondientes. Esta matriz es también de mucha importancia porque es parte de cálculos quimiométricos más elevados que veremos más adelante.

Algunas Matrices Características

Las matrices de correlación (Corr) y covariancia (Cov) de una matriz X son del tipo $M=X^TX$, muchas veces trabajaremos con este tipo de matriz que tiene la particularidad de ser simétrica respecto de la diagonal principal. Veamos los ejemplos:

$$X=\begin{vmatrix} 0.8318 & 0.19343 & 0.37837 \\ 0.50281 & 0.68222 & 0.86001 \\ 0.70947 & 0.30276 & 0.85366 \\ 0.42889 & 0.54167 & 0.59356 \\ 0.30462 & 0.15087 & 0.49655 \\ 0.18965 & 0.6979 & 0.89977 \end{vmatrix}$$

$$Y=\begin{vmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0.5 & 0.866 & 0 \\ 0.5 & 0.289 & 0.816 \\ -1 & 0 & 0 \\ -0.5 & -0.866 & 0 \\ -0.5 & 0.866 & 0 \\ 0.5 & -0.866 & 0 \\ -0.5 & -0.289 & -0.816 \\ -0.5 & 0.289 & 0.816 \\ 0.5 & -0.289 & -0.816 \\ 0 & -0.577 & 0.816 \\ 0 & 0.577 & -0.816 \end{vmatrix}$$

$$Cov(x)=\begin{vmatrix} 0.0587 & -0.0282 & -0.0179 & 0.0523 \\ -0.0282 & 0.0596 & 0.0390 & -0.0147 \\ -0.0179 & 0.0390 & 0.0486 & -0.0094 \\ 0.0523 & -0.0147 & -0.0094 & 0.0530 \end{vmatrix}$$

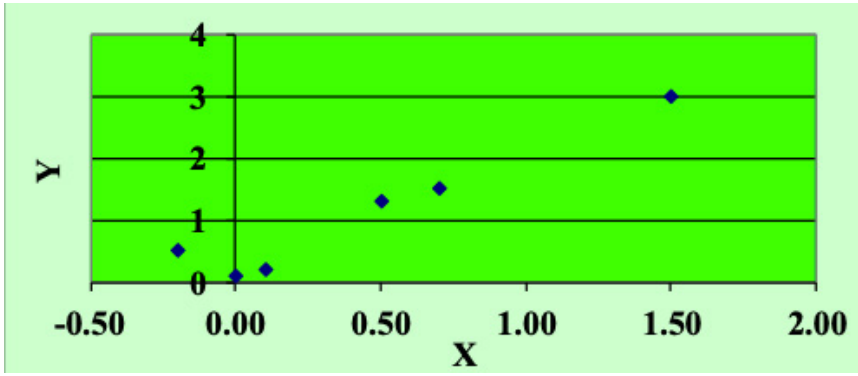
$$Corr(x)=\begin{vmatrix} 1.0000 & -0.4771 & -0.3349 & 0.9383 \\ -0.4771 & 1.0000 & 0.7252 & -0.2607 \\ -0.3349 & 0.7252 & 1.0000 & -0.1851 \\ 0.9383 & -0.2607 & -0.1851 & 1.0000 \end{vmatrix}$$

Matrices Ortogonales

$$Cov(Y)=\begin{vmatrix} 0.3333 & 0 & 0 \\ 0 & 0.3333 & 0 \\ 0 & 0 & 0.3329 \end{vmatrix}$$

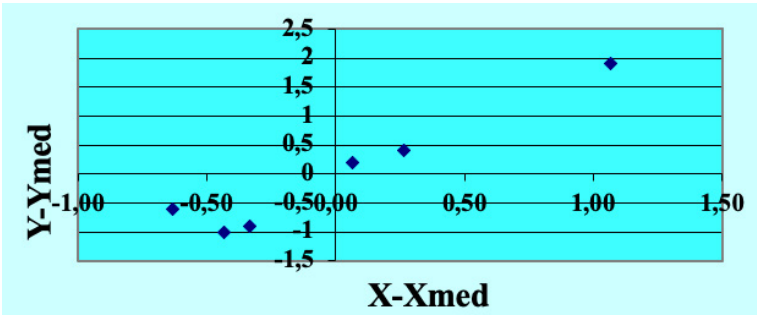
$$Corr(y)=\begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix}$$

Interpretación de la operación de centrado de datos



X	Y	X-Xmed	Y-Ymed
0.50	1.3	0.07	0.2
0.70	1.5	0.27	0.4
-0.20	0.5	-0.63	-0.6
1.50	3	1.07	1.9
0.10	0.2	-0.33	-0.9
0.00	0.1	-0.43	-1
0.43	1.1	<--valores medios	

Supongamos que tenemos la tabla de datos para las variables X e Y. Si las representamos directamente, obtenemos el gráfico superior. Pero si a cada dato de una columna le restamos su promedio y luego representamos estos nuevos datos transformados, obtenemos el gráfico inferior. Esta operación de ***transformación de variables*** se llama ***centrado*** y su efecto es correr el origen de coordenadas al centro (valor medio) de las variables.



Observe que la posición relativa entre los puntos ha quedado exactamente igual.

La operación de centrado es muy común en las diferentes técnicas de cálculo en quimiometría. Se puede aplicar a columnas o filas de datos, de allí los nombres de *centrado en columnas* o *centrado en filas.*

Expresión algebraica de la correlación en el espacio multivariable

Muchos lectores coincidirán inmediatamente en que nuestra definición de correlación no es la que han aprendido en los cursos básicos de álgebra. Justificaremos aquí la generalización de aquella definición a través de dos observaciones:

La primera es que la definición algebraica básica es aplicable a un plano (dos variables) pero no sirve para el espacio multivariable, a menos que calculemos la relación entre cada par de ellas. Tenga en cuenta que para 10 variables se necesitaría calcular 45 pares de éstas.

| La segunda observación es la demostración de que la definición multivariable contiene a la del álgebra básica y es mucho más fácil de recordar. ¿Acaso recuerda de memoria la expresión del álgebra básica?

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_i (x_i - \bar{x})^2][\sum_i (y_i - \bar{y})^2]}}$$

Llamemos ahora $\mathbf{V_x}=(x_i - \bar{x})$ y $\mathbf{V_y}=(y_i - \bar{y})$

…entonces $\sum_i (x_i - \bar{x})(y_i - \bar{y})=\mathbf{V_x}^{\mathrm{T}}.\mathbf{V_y}$ y

el denominador $\sqrt{[\sum_i (x_i - \bar{x})^2][\sum_i (y_i - \bar{y})^2]} = \|\mathbf{V_x}\|.\|\mathbf{V_y}\|$.

Entonces: $r = \frac{\boldsymbol{V_x}.\boldsymbol{V_i}}{\|\boldsymbol{V_x}\|.\|\boldsymbol{V_i}\|} = \cos\varphi$

Que es la ecuación multivariable de r. Observar que **dos vectores cualesquiera** en el espacio multidimensional forman un plano, por eso, la ecuación multidimensional abarca a la del álgebra básica.

Referencias del capítulo

1. https://adultosmayores.info/salud/tablas-imc-por-edad/

2. Instituto Nicaragüense De Energía Ente Regulador Capacidad Instalada Sistema Eléctrico Nacional (Mw) Tipo de Generación 2010-2015. https://www.ine.gob.ni/DGE/estadisticas/serieHistorica/capacidad-instalada-energia-2010-2017-actabril18.pdf

3. Statista (España) Para Gráfico. https://es.statista.com/estadisticas/598982/numero-de-personas-infectadas-por-el-vih-en-todo-el-mundo/

4. WIPO Intellectual Property Statistics Data Center. https://www3.wipo.int/ipstats/index.htm?tab=patent

5. Trading economics. https://tradingeconomics.com/argentina/government-debt-to-gdp

CAPÍTULO 2

Análisis por Componentes Principales y Análisis de Factores

Análisis por Componentes Principales

Introducción: El análisis por componentes principales (PC en inglés) es una técnica muy difundida debido a que se aplica a una gran generalidad de problemas. Cuando se comienza a analizar un sistema multivariable las primeras preguntas a contestarse son: ¿Cuáles son las variables más importantes que describen el sistema?, ¿Guardan éstas variables alguna relación entre sí?, ¿Se pueden descartar variables que no son importantes?, y algunas otras más. Todas estas preguntas pueden ser analizadas mediante PC.

Ejemplo práctico y soporte teórico

Supongamos que tenemos una serie de n observaciones (vectores, objetos) de m dimensiones expresadas como $\mathbf{x}_1, \mathbf{x}_2,...\mathbf{x}_n$. Componentes Principales, CP (ó PC en inglés), es un camino para transformar esta serie de vectores en otra serie $\mathbf{y}_1, \mathbf{y}_2,...$

$\mathbf{y}_n$, también m-dimensional. Los **y**'s tienen la propiedad de que la mayor parte de la información está contenida en unas pocas primeras dimensiones, pudiéndose descartar el resto con muy poca pérdida de información.

La idea entonces, es que CP es un método de reducción dimensional. La transformación en CP da algunas facilidades para analizar la información:

- Obtener información gráfica de los datos en 2-D o 3-D.
- Aplicar métodos computacionales intensivos sobre los datos reducidos.
- Inspeccionar la estructura de los datos, la cuál se presenta de manera más confusa en las p dimensiones originales.
- Obtener información cuantitativa acerca del 'peso' de cada variable original.

Antes de continuar con el ejemplo es imprescindible, para quienes no dominen los principios del álgebra lineal, que vean el anexo de este capítulo para poder continuar.

La idea principal para entender el cálculo de los CP es que la información más importante está asociada con valores de varianza grandes. Es posible demostrar, para quienes quieran profundizar en el álgebra lineal, que la dirección de máxima varianza es paralela al *eigenvector* correspondiente al mayor *eigenvalue* de la **matriz de varianza-covarianza** de los datos. También es posible demostrar que, de todas las direcciones **ortogonales a la dirección de más alta varianza**, la varianza que sigue inmediatamente a la mayor, la segunda mayor, es la dirección paralela al *eigenvector* correspondiente al segundo más alto *eigenvalue* (Ref. 3,4). Estas relaciones se extienden hasta completar las m dimensiones originales (en este ejemplo, 4). Observe que en álgebra lineal (o sea, en el espacio multidimensional) la ortogonalidad entre las direcciones de los vectores existe para dimensiones aún más grandes que 3.

Una interpretación alternativa a ésta y totalmente equivalente es pensar en la línea (en el espacio multidimensional m) que pase más cerca de los puntos determinados por $\mathbf{x}_i$ según el criterio de **distancias cuadráticas mínimas**. Esto es fácil de apreciar en 2 o 3 dimensiones (figura 2.1). En el primer caso se ve la proyección de los puntos sobre una recta y en el segundo sobre un plano. Los puntos amarillos de la figura son los originales en el espacio multidimensional y los negros son sus respectivas proyecciones.

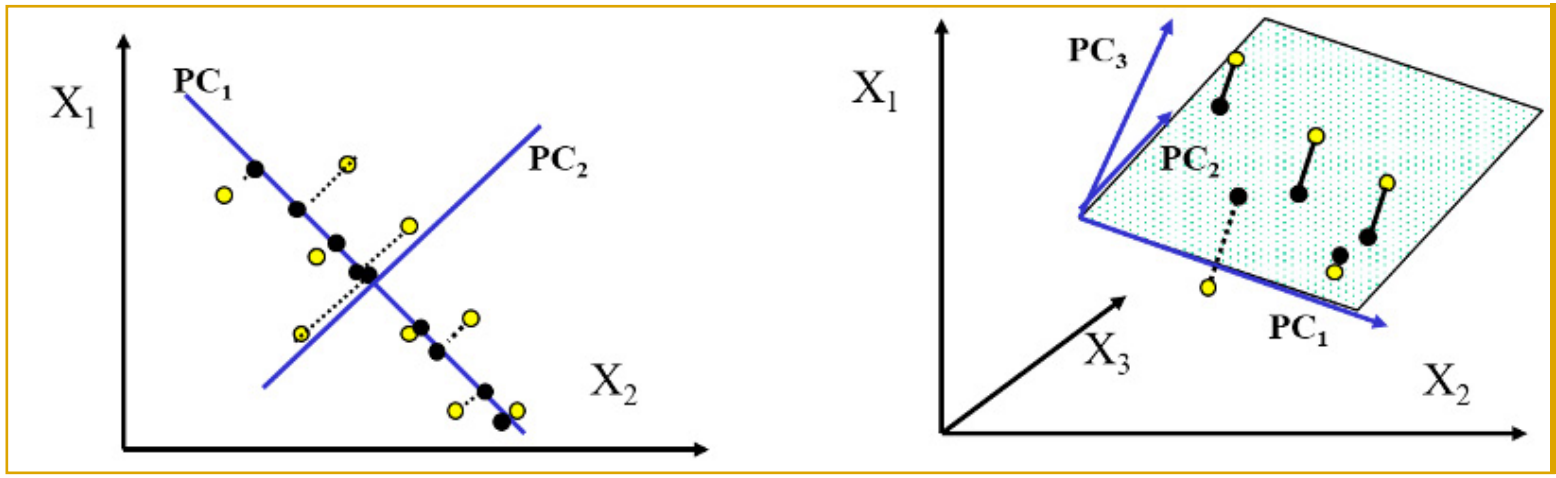


Figura. 2.1

El cálculo de PC se hace sobre la base de una descomposición matricial llamada:

eigenvalue decomposition cuya expresión es: **M=E.Λ.E'**. Donde E' es la matriz inversa de E.

Esta operación puede realizarse sólo sobre matrices cuadradas, M, por eso se utilizan habitualmente las matrices de varianza-covarianza o la de correlación.

La descomposición da como resultado una matriz $\mathbf{E}_{m,m}$ de eigenvectors y un matriz diagonal $\mathbf{\Lambda}$ de m eigenvalues.

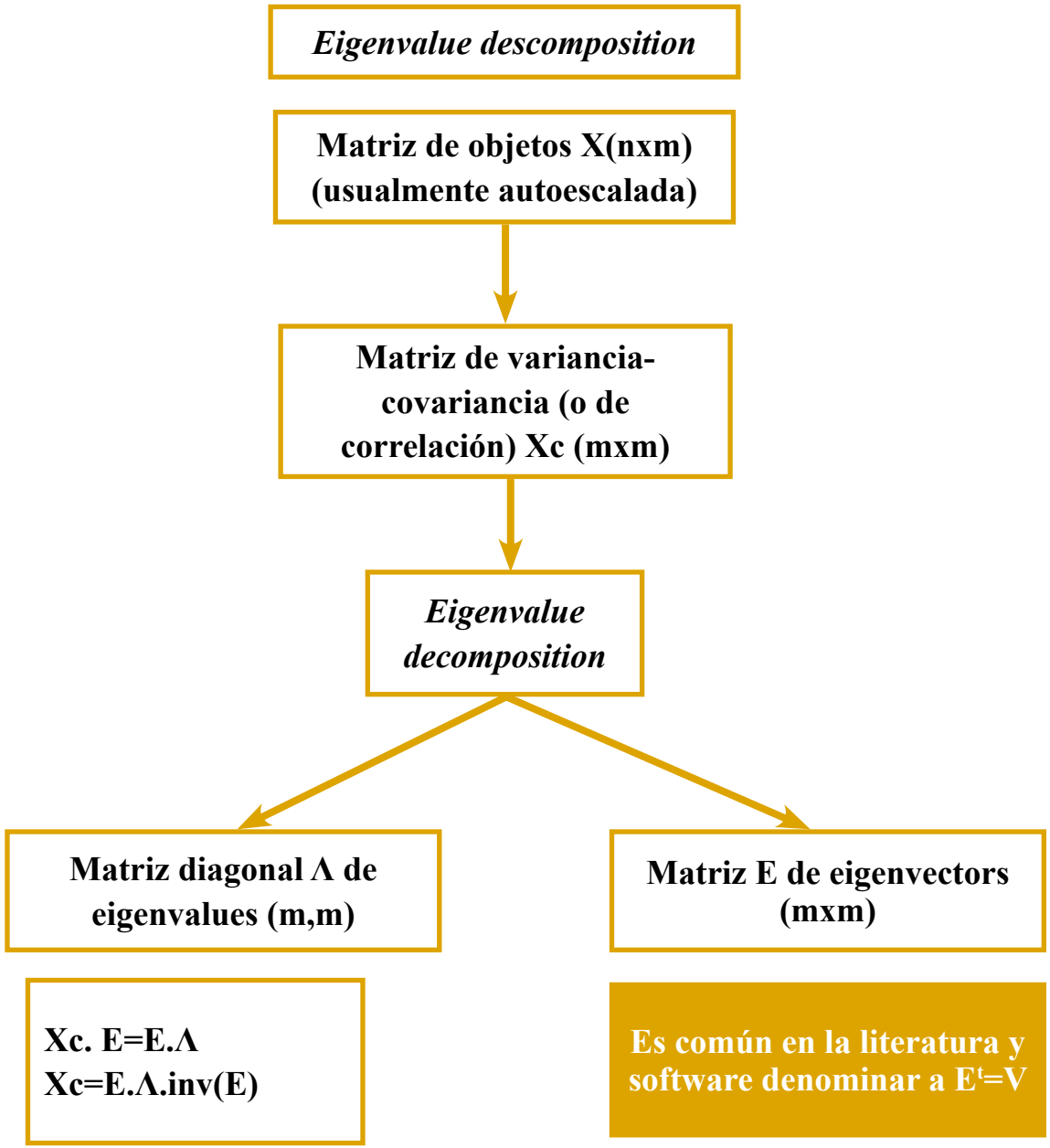
En la práctica calculamos la matriz de varianza-covarianza **S**, de (m x m) como:

$\mathbf{S}_{m,m}=1/(n-1)\ \mathbf{M}^T\mathbf{M}$. Donde **M** es la matriz centrada de los datos originales $\mathbf{X}_{nxm}$.

A los componentes del **vector** de *eigenvalues* (la diagonal principal de **Λ)** de **S** los denominaremos $\lambda_1 \geq \lambda_2 \geq \lambda_3 ..... \geq \lambda_m \geq 0$. Y a los correspondientes *eigenvectors* $e_1, e_2, e_3, ..., e_m$.

Los eigenvectors constituyen vectores columna de la matriz de descomposición (m x m) que llamaremos **E**.

En la figura 2.1 podemos apreciar que los objetos en el espacio multidimensional permanecen incambiables y que los Componentes Principales son **nuevos ejes** que imponemos al conjunto de datos.


Eigenvalue descomposition
Matriz de objetos X(nxm) (usualmente autoescalada)
Matriz de variancia-covariancia (o de correlación) Xc (mxm)
Eigenvalue decomposition
Matriz diagonal Λ de eigenvalues (m,m)
Matriz E de eigenvectors (mxm)
Xc. E=E.Λ
Xc=E.Λ.inv(E)
Es común en la literatura y software denominar a E^t=V

Matrices y definiciones de los Componentes Principales

La matriz $\mathbf{\Lambda}$ es una matriz diagonal de mxm y sus valores son los ***eigenvalues*** de **Xc**. A los ***eigenvalues*** los ordenaremos así: $\lambda_1 \geq \lambda_2 \geq \lambda_3 ..... \geq \lambda_p \geq 0$.

Debido a la ortonormalidad de **E** los nuevos ejes tienen **longitud 1** y son **ortogonales entre sí**, lo que algebraicamente significa imponer las siguientes restricciones:

$$e_i^T e_j = \begin{cases} 1 & si \;\; i = j \\ 0 & si \;\; i \neq j \end{cases} \qquad \begin{bmatrix} \lambda_i & 0 & . & 0 \\ 0 & \lambda_2 & . & 0 \\ . & . & . & 0 \\ 0 & . & . & \lambda r \end{bmatrix} = \mathbf{\Lambda}$$

Observar que los ceros fuera de la diagonal principal están indicando que la covarianza, en el nuevo sistema de coordenadas, es nula. PC, es entonces, un método de obtener una serie de nuevas variables no correlacionadas.

En los nuevos ejes, cada objeto tendrá un nuevo juego de coordenadas que se llaman **score.** Definimos las coordenadas de la matriz de ***scores*** **S** de **las n filas** de **X** como: $\mathbf{S=XE\Lambda^p}$

Similarmente definimos las coordenadas de las m **columnas** de **X** para la matriz de ***loadings*** $\mathbf{L=E\Lambda^p}$**,** p es un factor de escala y puede valer 0, ½ ó 1. A la matriz **L** se la suele llamar también **matriz de coeficientes** debido a que los componentes principales **CP**, de los objetos, son una combinación lineal de las variables de **X**.

Algunas propiedades

Usualmente, los primeros CP contienen la mayor parte de la varianza del sistema, los demás se pueden ignorar y de ese modo obtenemos una importante reducción dimensional.

Típica tabla de resultados de los programas comerciales de CP. Observe con cuidado todas las columnas

Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative%
2.058540	51.46351	2.058540	51.4635
1.022178	25.55446	3.080718	77.0180
0.667820	16.69551	3.748539	93.7135
0.251461	6.28653	4.000000	100.0000

Además, La suma de la varianza muestral para los CP es igual a la suma de la varianza muestral de **X**.

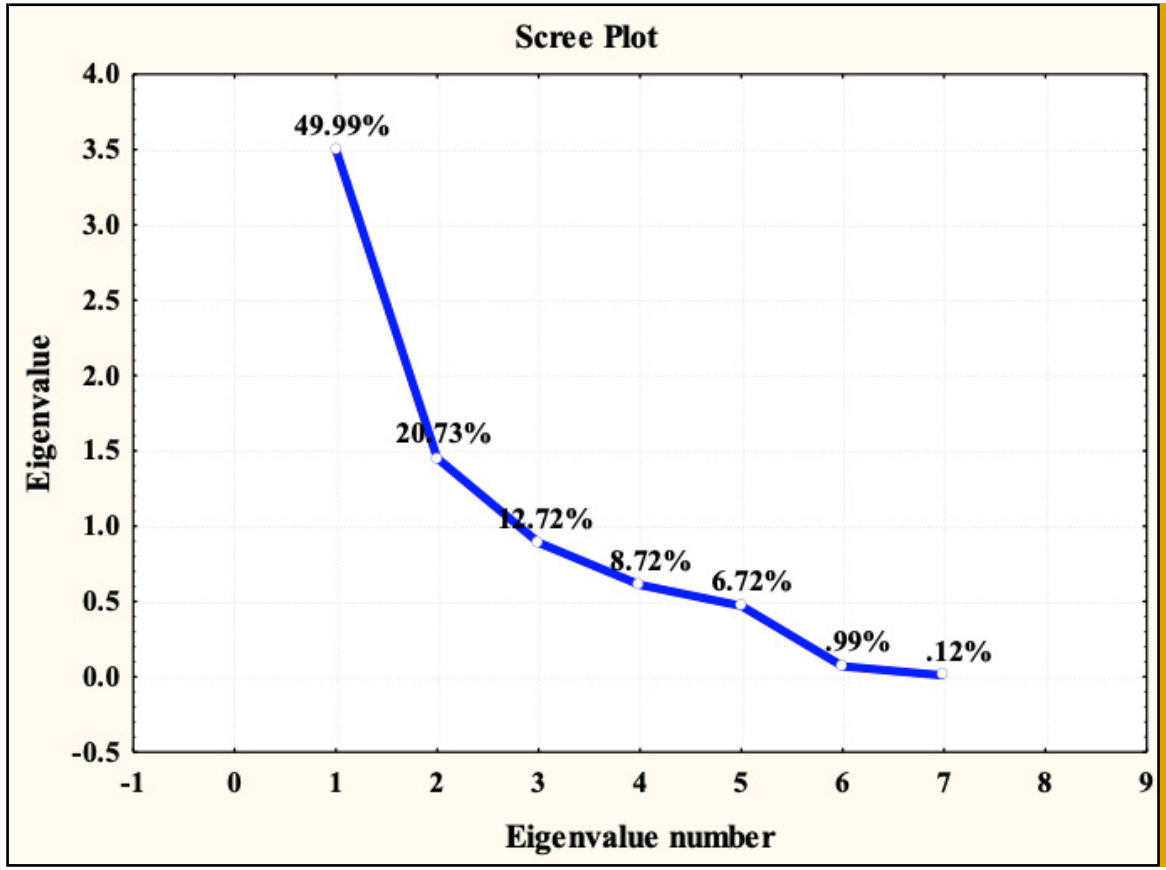
Si la matriz **X** hubiese sido autoescalada para el cálculo de los CP, entonces:

$$\sum_{i=1}^{m} \lambda_i = \sum_{i=1}^{m} s_i^2 \qquad \textbf{[1]}$$

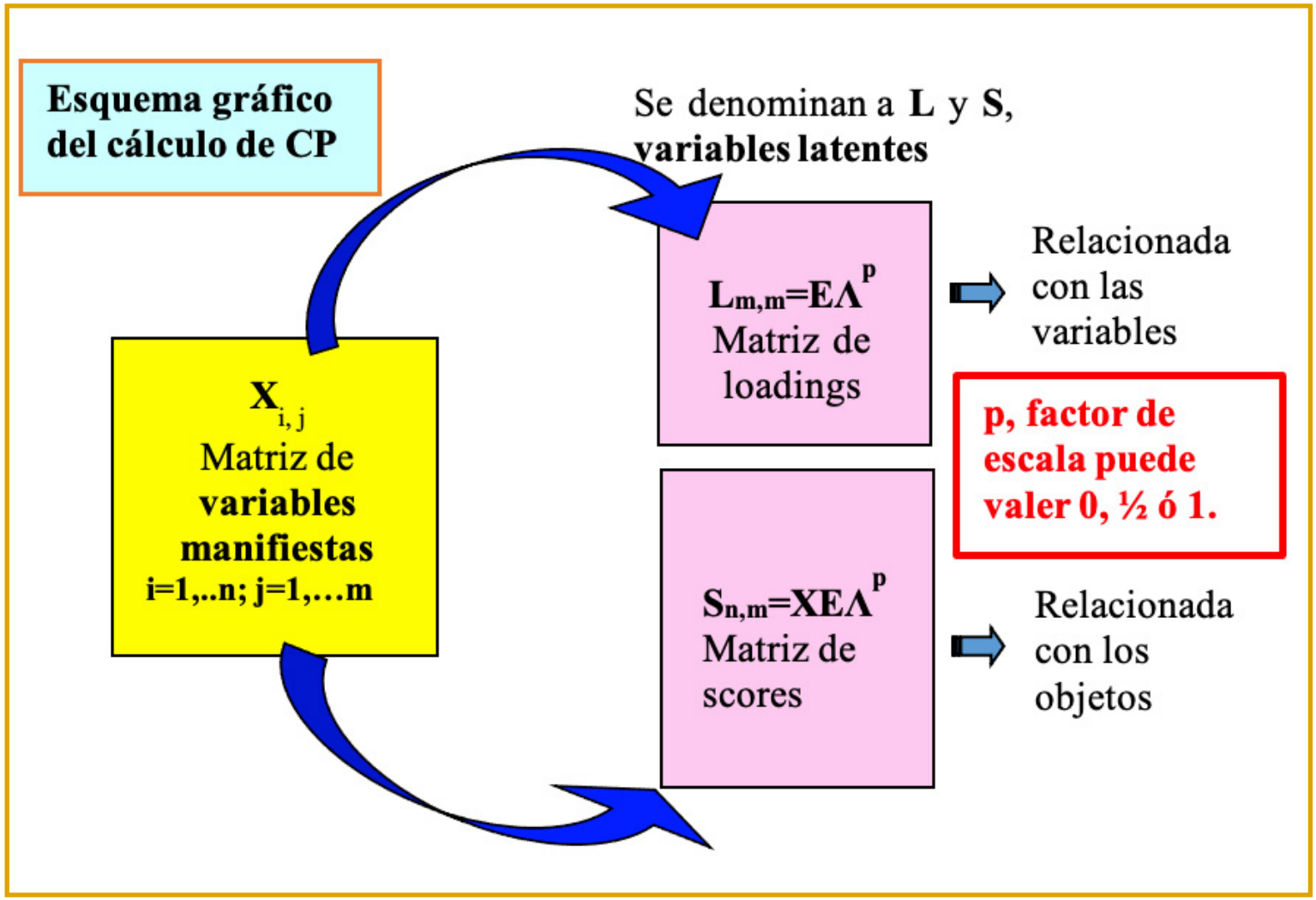
$$\sum_{i=1}^{m} \lambda_i = \mathrm{D}, \text{ la dimensión del problema} \qquad \textbf{[2]}$$

Donde los S_i^2 son los valores de la segunda columna de la tabla anterior. Los nuevos vectores y (y_1, y_2, ...y_m) **no están correlacionados**, esto queda evidenciado en la matriz $\mathbf{\Lambda}$, cuyos elementos fuera de la diagonal son ceros.

De acuerdo a la mencionada propiedad [1], un gráfico de las varianzas (*eigenvalues*) vs. el número m de CP, denominado ***scree plot***, indica cuántos CP's conviene tomar en cuenta para representar una buena proporción de la información; esta es una representación gráfica de una tabla similar a la anterior.



Vemos aquí que el sexto CP ya tiene muy poca variabilidad y podría considerarse que sólo aporta 'ruido' a la información sustantiva. Por lo tanto, los CP a considerar son los 5 primeros como máximo.



Si además estamos ante la condición [2], entonces cada eigenvalue λ valdría 1 en promedio, entonces se desechan aquellos CP menores a 1.

Ejemplo introductorio

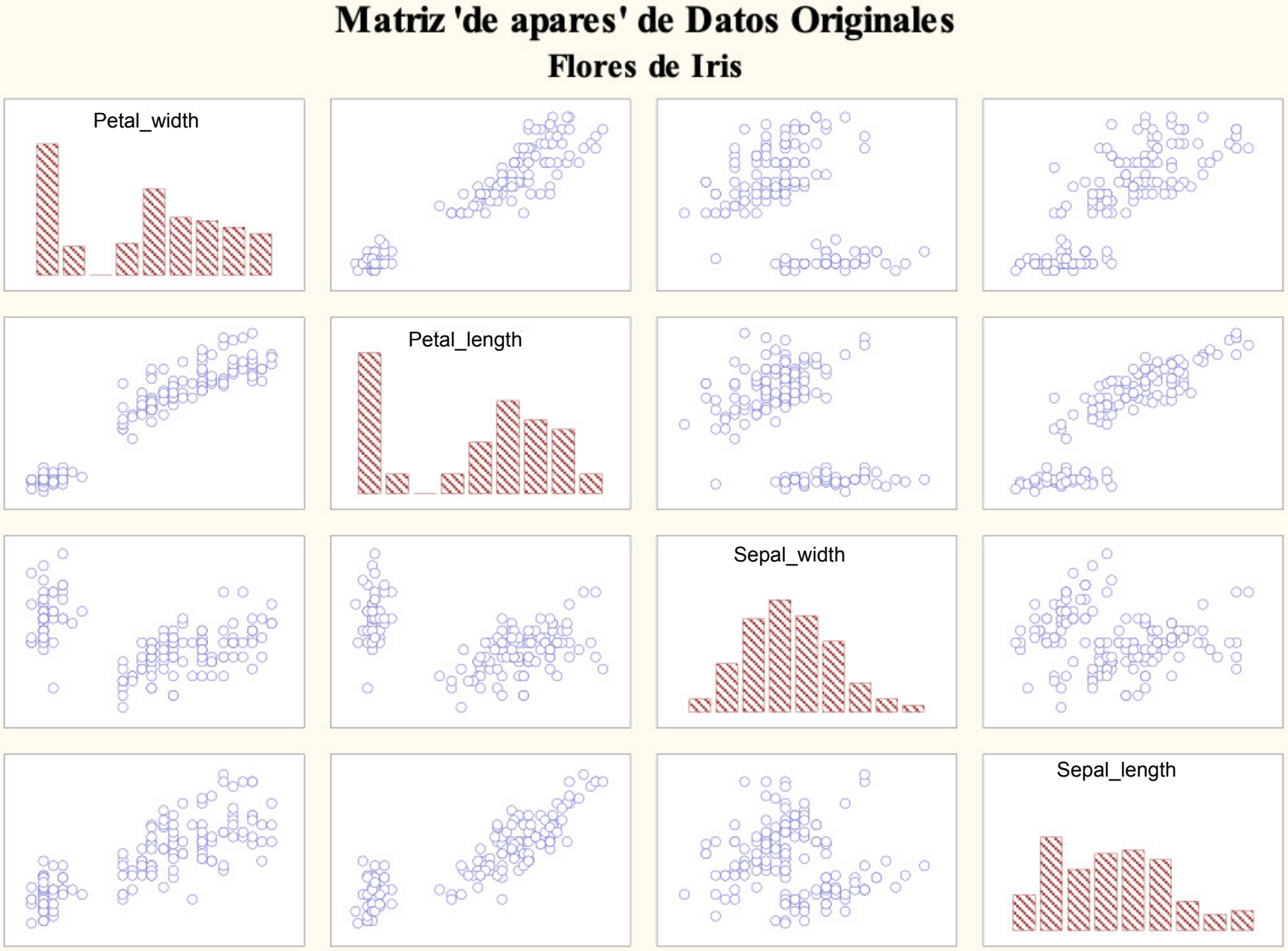
Veamos una aplicación sobre un ejemplo sencillo. Tomaremos un conjunto de 150 observaciones de flores de Iris (Ref.1).

Hay varias *clases* de flores de Iris y pretendemos distinguirlas a partir de un considerable número de muestras sobre las que se hacen una serie de mediciones, éstas son:

- **Ancho de pétalos**
- **Lasgo de pétalos**
- **Ancho de sépalos**
- **Largo de sépalos**

Cada muestra (objeto) contiene 4 variables que constituyen un vector de cuatro dimensiones. Si ignoramos Componentes Principales, un método primario de analizar la información que poseemos es un hacer gráfico 'de a pares' entre las variables, similar a una matriz de correlaciones. Reuniendo todos los gráficos organizadamente obtenemos una matriz gráfica como la de la figura siguiente (2,2). La diagonal principal es la distribución de cada variable.

Figura. 2.2


Matriz 'de apares' de Datos Originales
Flores de Iris
Petal_width
Petal_length
Sepal_width
Sepal_length

Lo que podemos observar es que para cualquier combinación de variables parecería haber 2 grupos de datos. Lo que nos conduce a la idea de que probablemente haya 2 clases de flores de Iris. Más tarde compararemos este gráfico con el que originaremos con los componentes principales.

Los resultados del cálculo de CP para la flor de Iris son los siguientes:

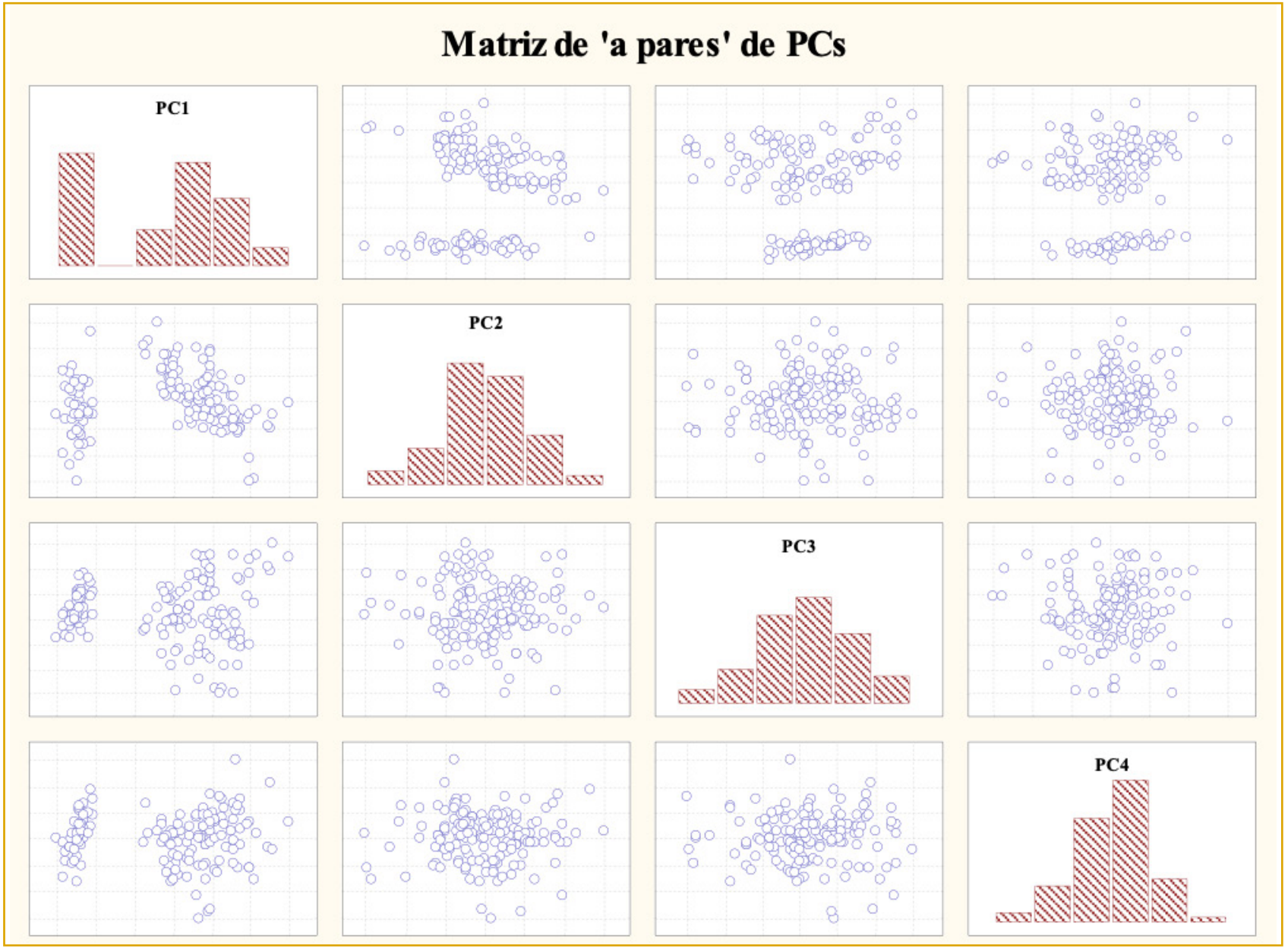
Observemos la matriz de *eigenvectors* **E** y *eigenvalues* λ para los datos autoescalados (Ref. 2). λ= (**2.92, 0.91, 0.15**, 0.02) que representan 73%, 22.8%, 3.75% y 0.50% de porcentage de varianza, respectivamente. La tabla **E** muestra los CP en cada columna y los pesos o 'loadings' en cada fila.

Variable	PC1	PC2	PC3	PC4
Largo de sépalos	0.521	0.377	0.720	0.261
Ancho de sépalos	-0.269	0.923	-0.244	-0.124
Largo de pétalos	0.580	0.024	-0.142	-0.801
Largo de pétalos	0.565	0.067	-0.634	0.524

- Observando los *eigenvalues* vemos que los primeros 3 PC's representan la mayor parte de la variabilidad de los datos, el restante podría despreciarse.

- En el primer *eigenvector* (columna 1 de **E**), se ve que éste toma valores grandes cuando las variables 1, 3 y 4 son grandes y la variable 2 toma valores chicos.

Entonces, $\mathbf{y}_1$ (PC1) toma valores grandes para flores de Iris con largos y finos sépalos y grandes pétalos. El segundo CP mide esencialmente el tamaño de los sépalos (variables 1 y 2) con predominio del ancho de sépalos (diferencia con y1 y no hay contribución del tamaño de los pétalos). En el tercer CP vemos un valor 0.72 para el largo de sépalos.



Comparemos ahora el gráfico ‘de a pares‘ obtenido con las variables originales, que mostramos al principio, con uno similar pero hecho con los *scores* de los PC.

La diferencia es que los 2 grupos de objetos que se observaban para cualquier par de variables manifiestas, aparecen ahora visiblemente sólo para el CP1 (combinado con cualquier otro CP) osea la primera fila del gráfico o la primera columna (porque el gráfico es simétrico respecto a la diagonal principal). Las características de estos dos grupos está ahora determinada por particulares **combinaciones** de las variables originales. Una tenue separación se alcanza a ver en PC3-PC2 que podría estar indicando otro grupo de objetos menos diferenciado. Volveremos sobre estas apreciaciones, más tarde, en la conclusiones.

El ***scree plot*** nos dirá cuántos CP's debemos tomar en cuenta para representar una buena proporción de la información. Esto confirma lo que concluimos de la tabla anterior.

Vemos en la Fig. 2.3 que el tercer CP ya tiene muy poca variabilidad y podría considerarse que sólo aporta 'ruido' a la información sustantiva. Por lo tanto, los CP a considerar son los 2 primeros. Hemos reducido las 4 variables originales a sólo 2, éste es el sentido de la 'reducción dimensional'.

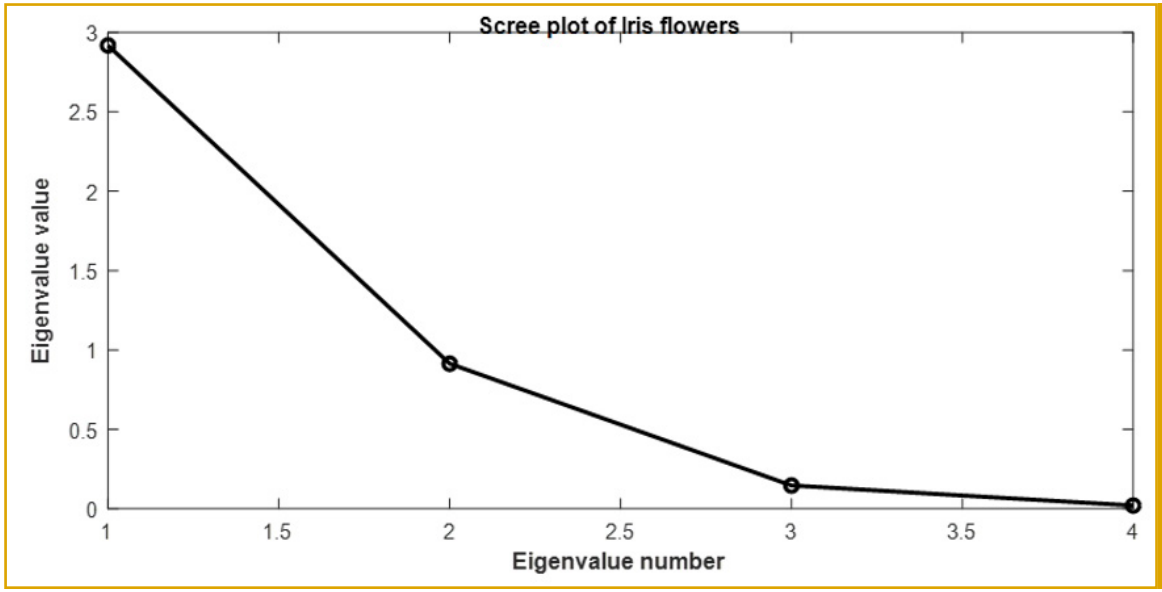


Figura. 2.3

En la figura 2.4, claramente el CP1 diferencia dos grandes grupos, uno ubicado casi totalmente en el plano superior CP>0 y el otro en el semieje negativo.

Usualmente, en los gráficos representamos los **objetos**, con los *scores,* si además marcamos alguna **variable** por nosotros conocida, como 'la clase' de flor de Iris (observe que este dato no ha sido tenido en cuenta en el cálculo), a este tipo de gráfico se los llama ***biplot*** (Fig. 2.4).

Ahora revelamos que hay 3 clases de flores de Iris, denominadas *Iris setosa=1*, *Iris Versicolor* =2 e *Iris virginica=3*.

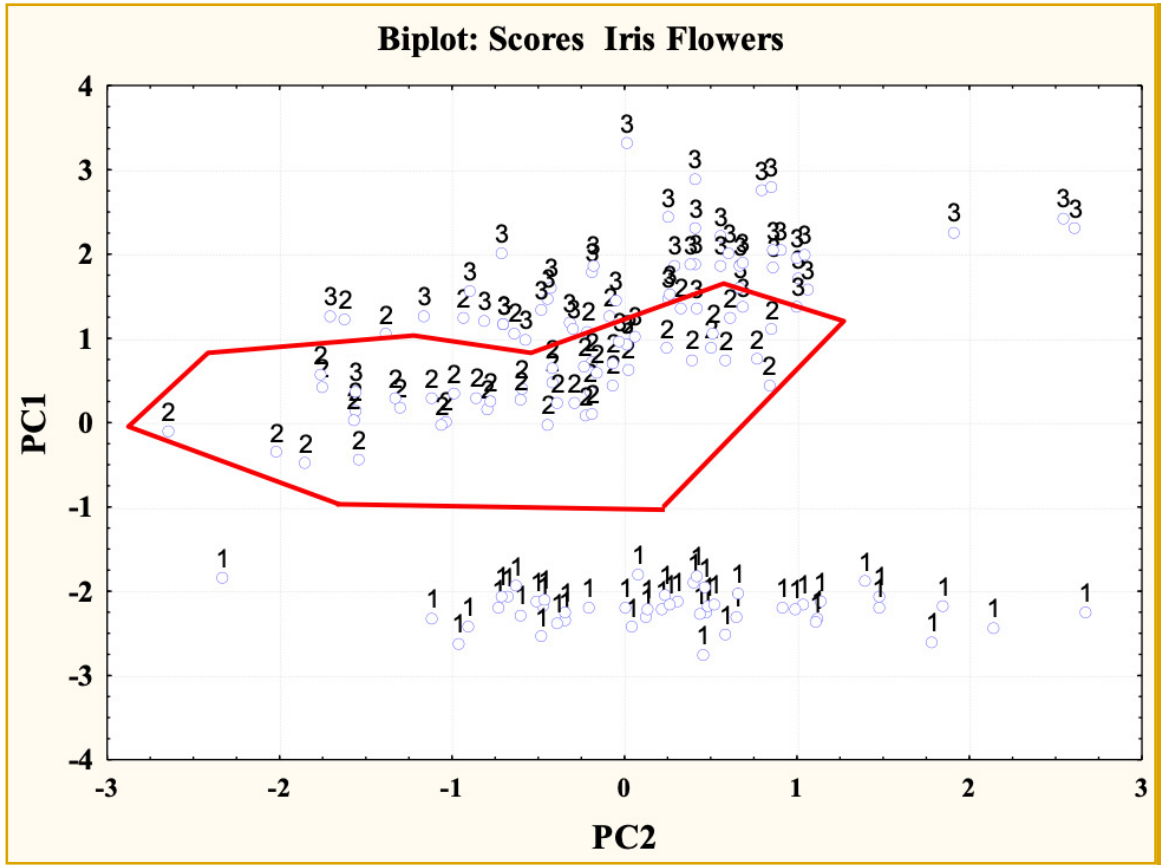
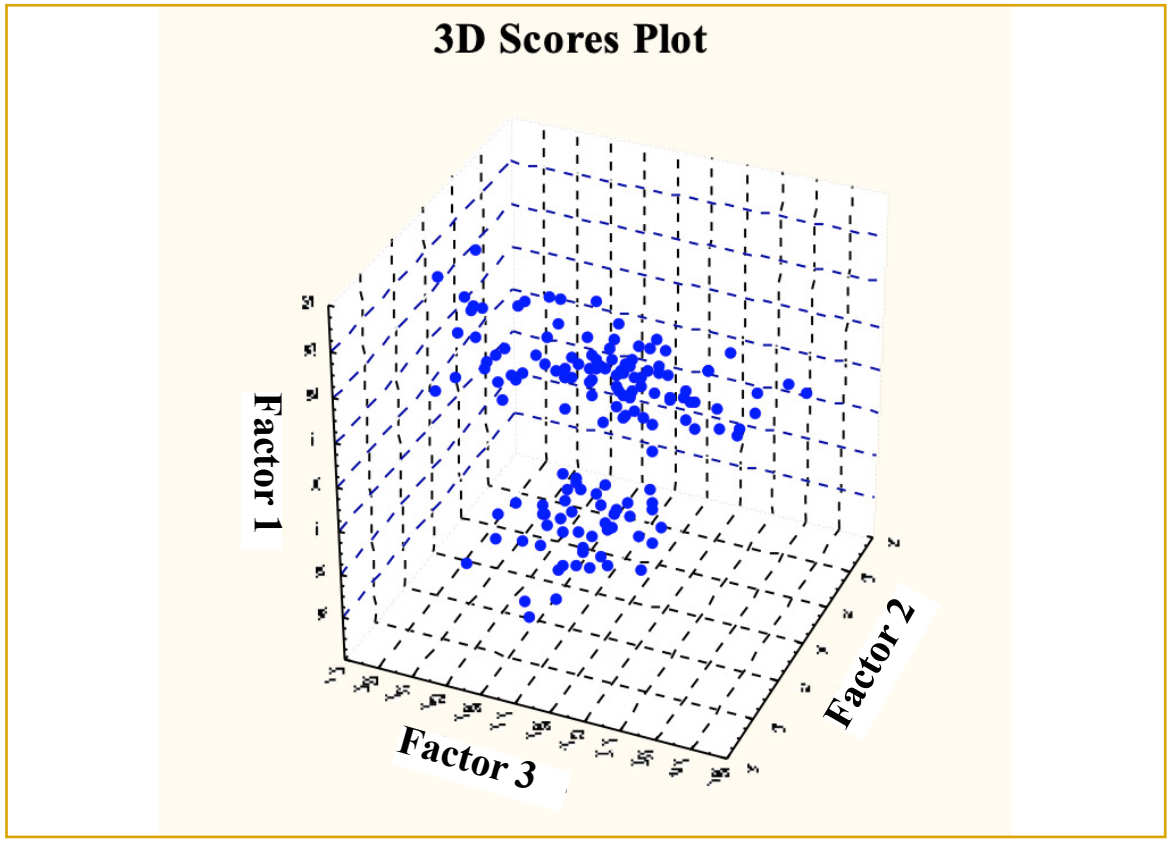


Figura. 2.4

En el gráfico anterior puede verse encerrado por la línea roja, el área que abarca la *Iris Versicolor*, muy cercana a la de *Iris virginica* y que era más difícil de apreciar en el gráfico 'de a pares'. Vemos que hemos *clasificado* las tres clases en tres áreas diferentes. Éste gráfico podría considerarse un modelo, aunque rudimentario, de las especies de Iris.



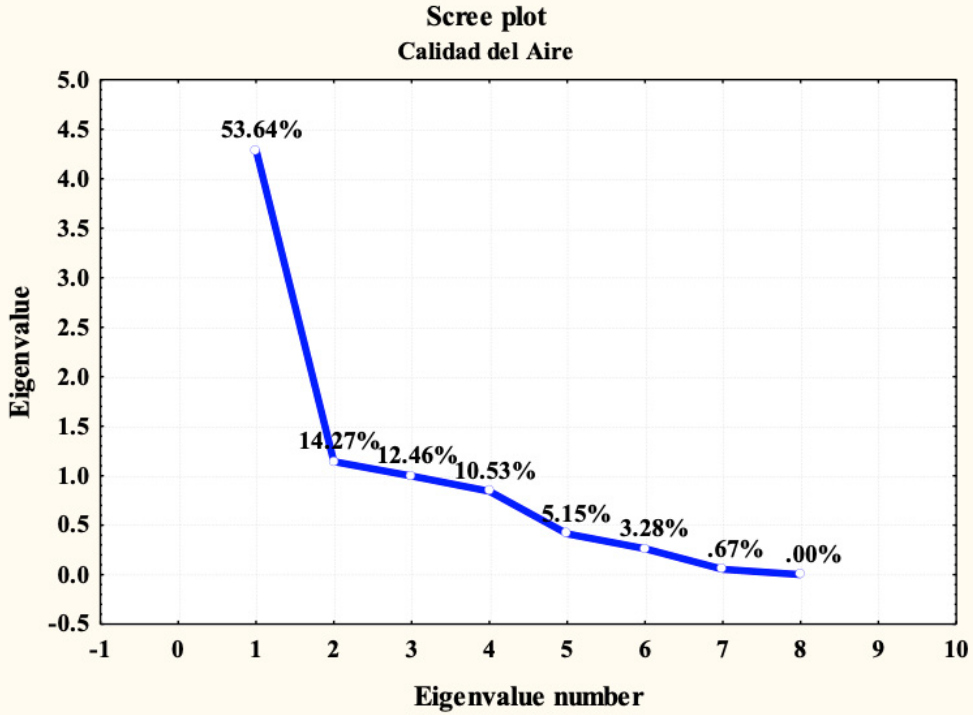
Un gráfico *biplot* en tres dimensiones, agregando el tercer CP, a veces brinda una mejor observación de los *scores*, aunque en el ejemplo presente éste no sea el caso porque el tercer CP ya aporta muy poca información.

Ejemplo Aplicado a la Química en un Problema Multidisciplinario

Problema que se pretende atacar: Aplicación a polución atmosférica

Tomaremos sólo una fracción de una base de datos de mediciones de calidad del aire, en la Capital Federal (Ref. 8), con la intención de hacer este ejemplo suficientemente didáctico para los objetivos actuales. Nuestra base resumida extraída de la referencia 8 consiste de 577 datos y 8 variables. 6 de las variables son gases atmosféricos: CO, NO, NO_2, NO_x, O_3 y SO_2; además, material particulado (PM) y dirección del viento (Dir V) distribuida en 8 sectores (1=N, 2=E, 3=S, 4=O, 5=NE, 6=SE, 7=SO y 8=NO).

El objetivo del ejemplo es averiguar si existen diferencias en la calidad del aire respecto a las condiciones meteorológicas, especialmente debido a la dirección del viento. El cálculo de los CP en base a la matriz de correlación da como resultado el siguiente scree plot:



Variable	Loading factors		
	PC1	PC2	PC3
SO2	0.820847	-0.190817	-0.011594
O3	-0.618254	-0.503694	0.002790
NOx	0.968311	-0.130723	0.008794
NO2	0.735861	0.376706	-0.024752
NO	0.934108	-0.208795	0.014040
PM	0.025350	0.377283	0.884589
CO	0.939035	-0.149560	-0.001055
Dir V	0.029474	0.695977	-0.461654

Observe que los 2 primeros PCs reúnen el 67.91% de la varianza del sistema. El 3er. PC ya aporta sólo 1.81% respecto del 2° y no lo tomaremos en cuenta.

En la tabla de Loading Factors se ve que en PC1, salvo PM y Dir V, los loadings son altos y que el único negativo es O_3. En PC2, Dir V es el único importante, mientras que PM figura alto recién en PC3. Esta descripción se aprecia claramente en un gráfico, como el que se muestra en la Fig. 2.5. El gráfico de loadings muestra además la correlación entre las variables, aquellas que caen en una misma línea están muy correlacionadas, como dirección del viento y PM. Y también lo están NO, NOx, CO y SO_2, aunque algo menos.

Se ve que O3 y NO2 no correlacionan con ninguna otra variable. Observe también que el origen del gráfico está en cero para ambos factores, esto es así cundo los PCs se calculan sobre matrices simétricas, como las de correlación o covarianza.

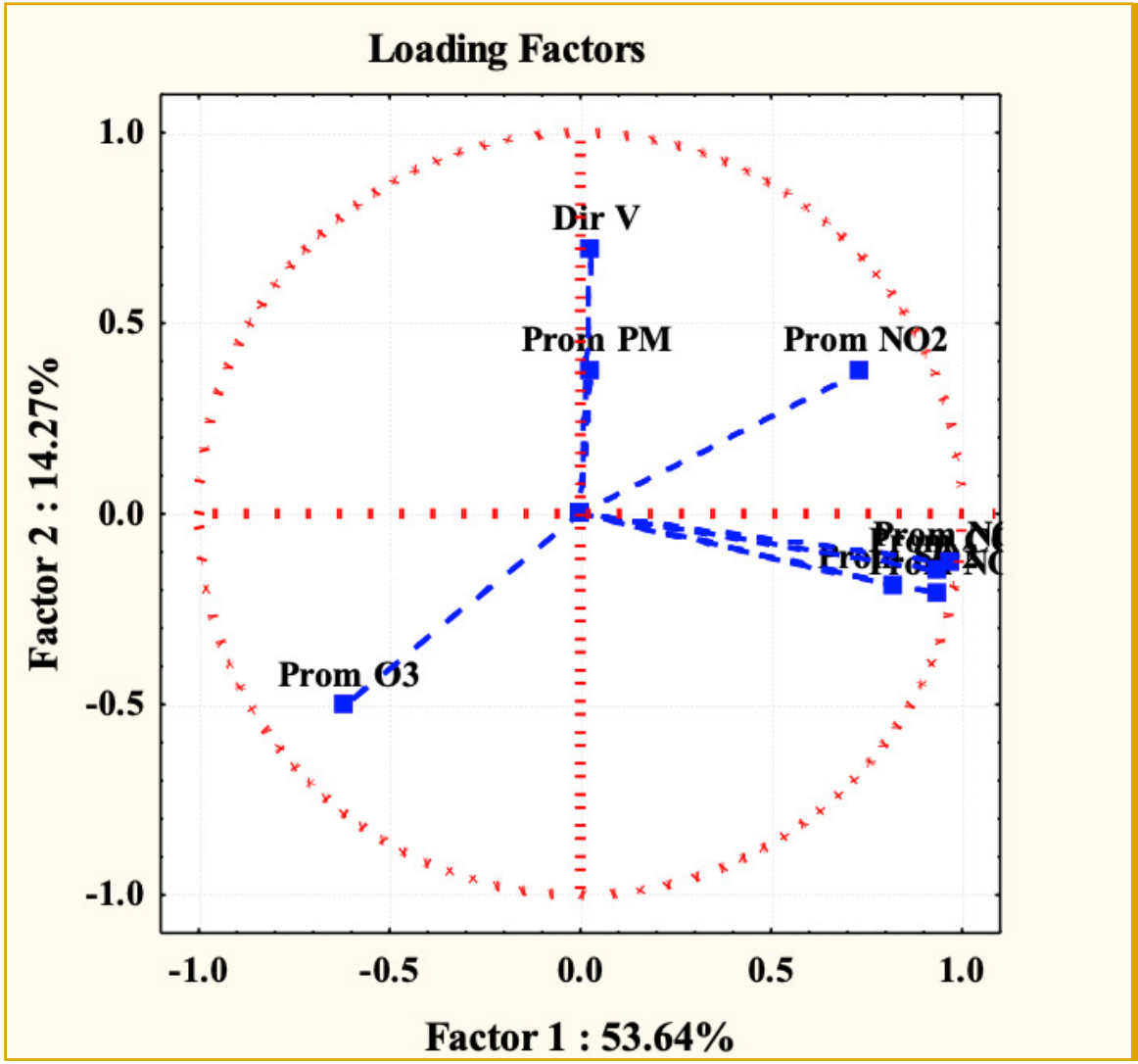


Figura 2.5

Para que esto ocurra en otras matrices los datos deben estar autoescalados, por ejemplo, si se calculan PCs sobre una matriz cuadrada que no sean las mencionadas previamente.

Analicemos ahora el gráfico de los scores. En la Fig. 2.6 se observa claramente una estructura de datos donde aparece un núcleo centrado en 0,0 y una rama hacia valores negativos de PC2 y positivos y altos de PC1. Los números sobre los datos indican el valor de Dir V y predominan el 2, el 3 y en menor grado el 5. Estas son las direcciones Este, Sur y Noreste respectivamente desde donde vendrá la mayor cantidad de PM.

Téngase en cuenta que éstas son conclusiones parciales y sencillas debido a la limitación en el número de datos y variables. Un análisis completo incluye 976 datos y 5 variables meteorológicas más.

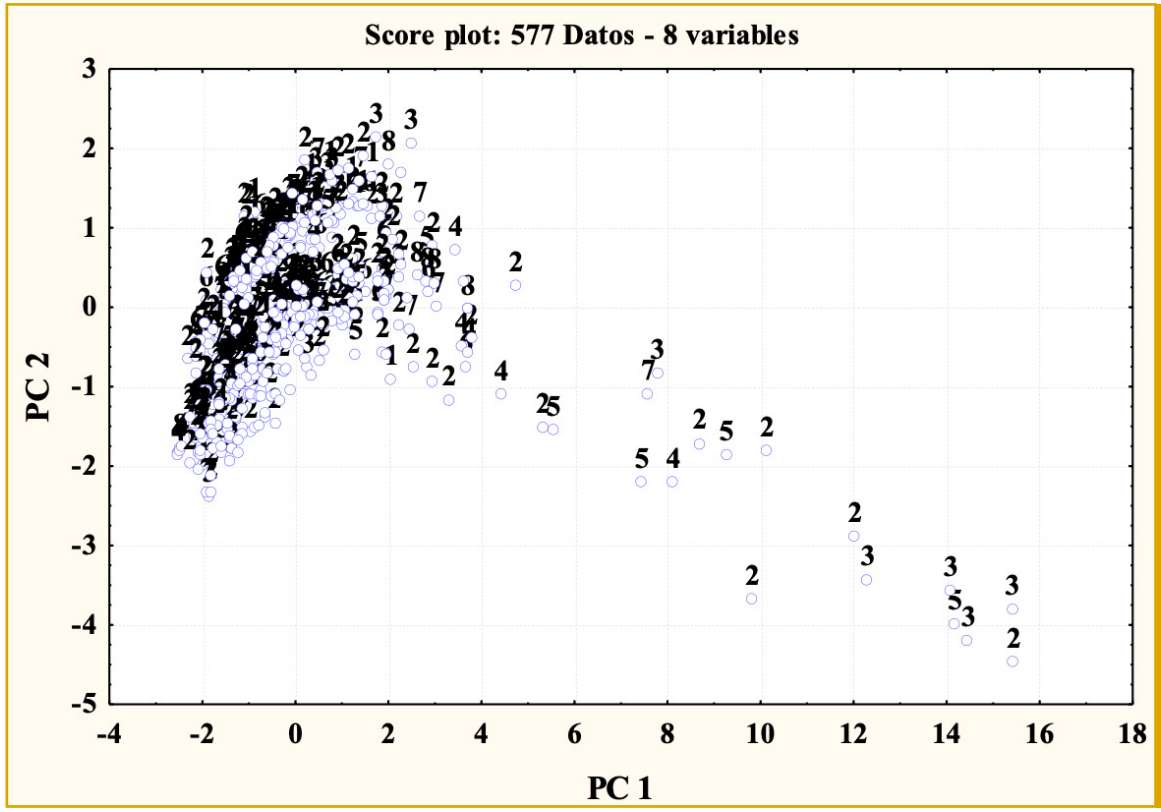


Figura 2.6

PC's y Modelos

Aprendizaje supervisado

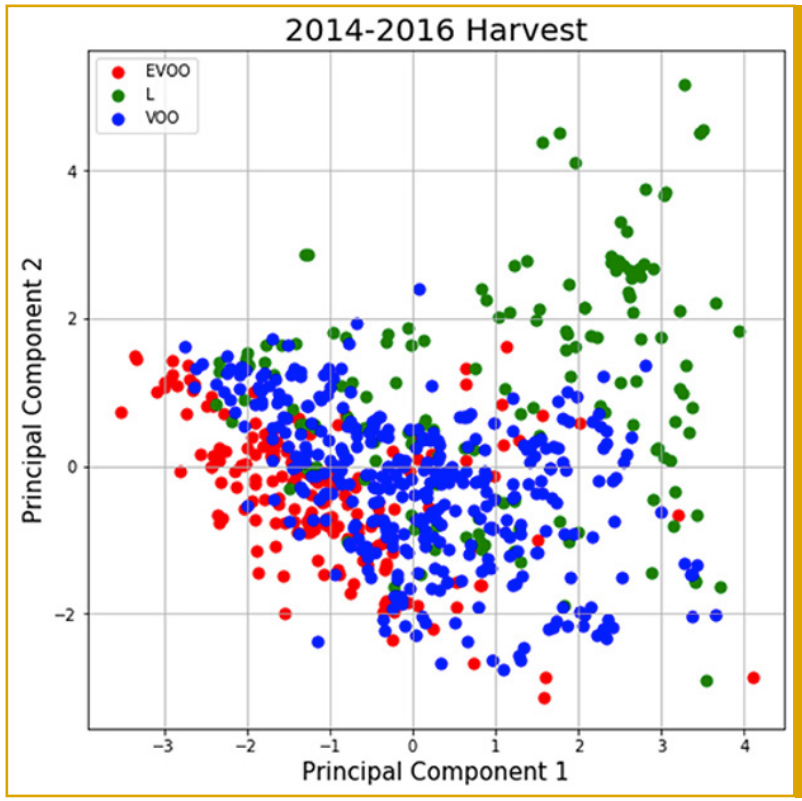
Ejemplo de aplicación a "Autentificación de alimentos"

Problema: Se pretende averiguar si es posible determinar la procedencia geográfica de los aceites de oliva de 9 regiones de Italia (Ref. 3-5).

Datos: Se poseen **500 muestras** de aceite de las **9 regiones** en estudio. A éstas se les ha determinado la concentración de **8 ácidos grasos** por cromatografía gaseosa.

El procedimiento de generar un modelo a partir del conocimiento de **todas las variables** de un lote de muestras se llama ***aprendizaje asistido***. En éste caso, cuando se resalta t**odas las variables**, significa que no solo conocemos las variables de las muestras, sino también **la respuesta correcta de ellas**.

Intentamos usar estos resultados para diseñar un procedimiento capaz de predecir la procedencia geográfica de muestras desconocidas. Para ello debemos elaborar un *modelo* por medio del cual puedan *reconocerse* las nuevas muestras. En terminología de *modelos de reconocimiento* podemos formular el procedimiento de la siguiente manera: Use las muestras ***de origen conocido*** para deducir una *regla de clasificación*. Pruebe la calidad de las *predicciones* sobre **otro lote** de muestras ***de origen conocido***. Luego, si el modelo es suficientemente bueno, podría *clasificar* muestras desconocidas.



Matemáticamente, el problema consiste en representar un espacio 8-dimensional en el cuál se puedan clasificar 9 regiones o clases. Un primer camino para clasificar las muestras es calcular los PC's y representar los scores de las muestras conocidas para ver cómo se ubican en el gráfico. Observe en la figura que puede haber zonas sobrepuestas de poca definición (Ref. 7).

¿Cuáles podrían ser las causas más sospechosas del poco éxito de los CP para resolver este problema? La poca resolución del mapa de scores podría ser causa, entre muchas otras, de una mala selección de variables o de insuficientes variables. Pero queremos referirnos aquí a un problema más profundo: la respuesta, 'y', a un problema multivariable podemos expresarla en forma polinomial dependiente de sus variables x_1, x_2,...x_m. Supongamos por sencillez que tenemos sólo 2 variables, entonces para un modelo ***algebraicamente*** **lineal en sus variables**, aunque incluya variables cuadrtáticas:

$$y=a + b.\ x1+c.x_2+d.x_1.\ x_2+e.x_1^{\ 2}+f.x_2^{\ 2},$$

los coeficientes a, b,…,f, son constantes y en este caso el modelo **es también *estadísticamente* lineal** porque se puede expresar x1.x2=u, $x_1^{\ 2}$=v y $x_2^{\ 2}$ =w y tendríamos el mismo modelo anterior con sus 6 coeficientes y 6 variables en total. Pero

pueden encontrarse algunos tipos de problemas donde, independientemente de si el modelo es *algebraicamente* lineal o no, este es ***estadísticamente* no lineal**. Un ejemplo de este tipo de modelos para 2 variables sería:

$$Y=a+ b.x_1^{(x2)} + c.\log(x_1/x_2)$$

O sea, el modelo no puede expresarse como un polinomio. En este caso, todos los métodos de cálculo lineales, como lo es CP, fallan para describir el modelo. Esto es lo que ocurre con el problema del origen geográfico del aceite de oliva. Más adelante, en la parte de métodos no lineales, veremos otros tipos de técnicas aptas para resolver este tipo de problemas.

Una Mirada a Otros Modelos de Reconocimiento

Sobre la base de la proyección de objetos sobre sus PCs hay una amplia variedad de ***modelos de reconocimiento*** para lograr la ***clasificación*** de objetos (Ref.3).

Para mejorar la clasificación por la simple proyección de objetos, como es el caso con los scores de los PCs, se utilizan métodos de ***Análisis Discriminantes***, de los cuales mencionaremos los más comunes. Estos métodos se pueden aplicar directamente a los datos originales o como complemento de otras técnica, por ejemplo a los scores de un previo análisis por PCs. En general hay dos tipos de estrategia para lograr la discriminación: algunos métodos exploran la máxima diferencia entre clases y otros utilizan la máxima similaridad dentro de las clases. Éste también es un método de aprendizaje asistido. Veremos más detalladamente otros métodos de clasificación en el capítulo de análisis de grupos (*cluster análisis*).

Veamos un gráfico obtenido por *Linear Discriminat Análisis* (LDA). A diferencia de CP, LDA selecciona una recta cuya dirección logre la máxima discriminación (óptimas vecindades) entre las clases dadas. Tomaremos como ejemplo el análisis de las flores de Iris y sobre el gráfico de scores trazaremos las rectas discriminantes Fig.2.7.

Observe que hay zonas en que los objetos de una clase se mezclan con los de otra. Esto se debe a que ésta técnica apunta a definir los límites óptimos entre las clases determinando el ***centroide*** (centro de gravedad geométrico de cada clase, ver definición en el Anexo) y la distancia media entre éstos. Los puntos rojos en la figura marcan el centroide de cada grupo y los puntos negros marcan la distancia media entre los centroides. Pero el límite entre clases no necesariamente estará en la distancia intermedia.

Ese problema puede ser mejorado utilizando mejores métodos para establecer los límites en una clasificación, por ejemplo, utilizando el método potencial. En lugar de utilizar rectas para separar las clases, se utilizan polinomios.

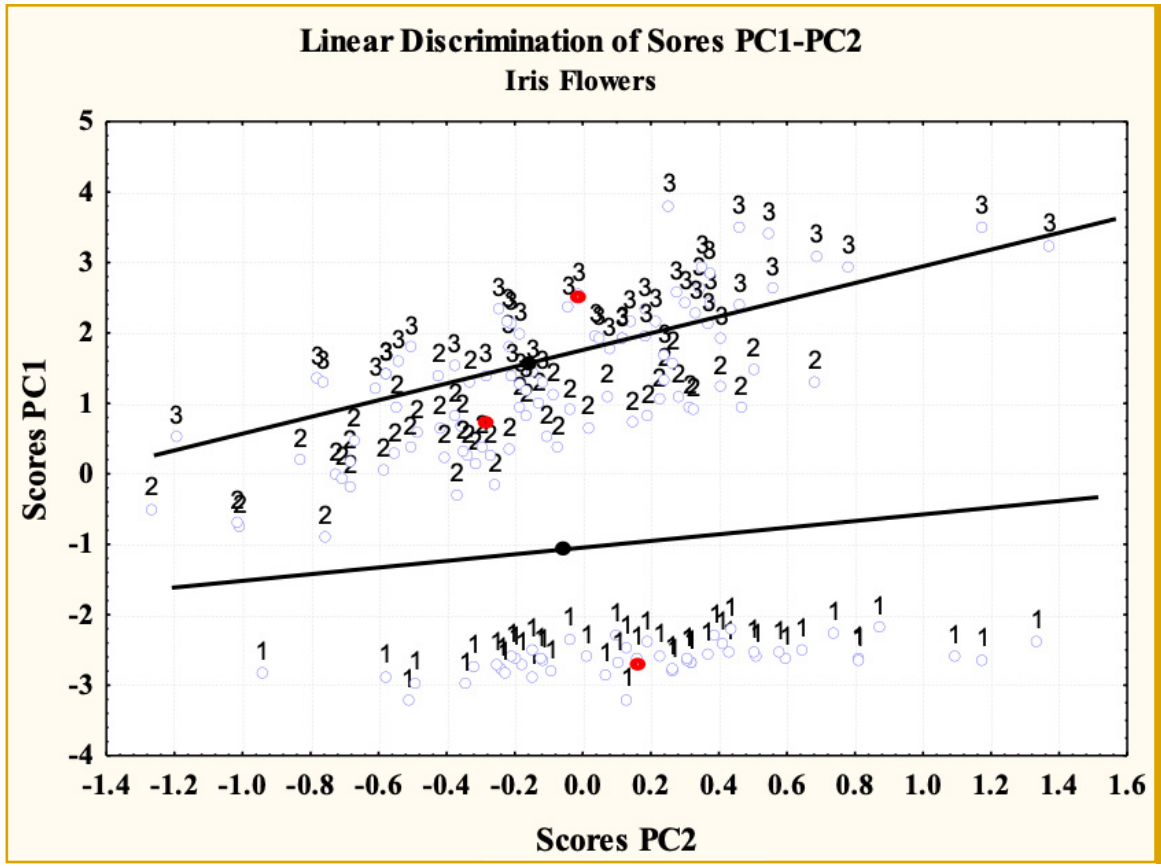


Figura 2.7

Selección Óptima de Variables

Hasta ahora no hemos hecho muchas consideraciones entre la relación que debiera existir entre el número de variables y el número de experiencias para llevar a cabo un análisis por CP. Mencionaremos por el momento algunos aspectos.

Como se ha visto en el cálculo de los *eigenvectors*, la matriz **E** tiene dimensión mxm donde m es el número de variables. Sin embargo, si el número de objetos o ensayos, n, es menor que m, entonces la matriz **E** resultará de una dimensión nxn y no podríamos evaluar el efecto de todas las variables sobre el problema que pretendemos estudiar. Hay dos preguntas que uno debe hacerse en estos casos:

- ¿Son necesarias todas las variables?

- ¿Cuáles me convendría quitar?

Para contestar la primera pregunta cabe indagar si se ha hecho un estudio previo de las variables o no. Si esta etapa fue salteada, éste es el momento de volver atrás y calcular y observar la matriz de correlación. Si el número de variables es el correcto, el mejor camino es agregar nuevas observaciones siempre que sea posible. Cuando lo anterior no es posible, si existen dos o más variables muy correlacionadas, entonces podemos dejar afuera para el cálculo de CP una de ellas (al menos por el momento). Esto está justificado por el hecho de que, si existen correlaciones, entonces una variable es proporcional a la otra, lo que dicho de otra manera significa que, salvo una constante de proporcionalidad, una informa (aproximadamente) lo mismo que la otra. O sea, no se agrega mucha más información incluyendo a las dos. Si no hay otro camino, lo único que queda es partir el número de variables en dos bloques y analizar el problema por partes, pero con la dificultad de perder la relación entre variables de cada bloque.

Algunos inconvenientes de los métodos precedentes

- Algunos de los problemas es clasificar un objeto nuevo, perteneciente a una posterior tanda de datos, que no pertenezca a ninguna clase previa y que, entonces, sería ubicado dentro de alguna clase pre-existente. Si se incorporan objetos de una nueva clase se debe recalcular el problema.

- Otro inconveniente es que estos métodos no cuantifican el **grado de pertenencia** a una clase, de un objeto en relación a otro, ya que los CP son combinaciones de algunas variables y no todas.Ej. En el caso del ejemplo de la calidad del aire, aún en la rama de datos fuera del núcleo, no es posible decidir cuán contaminante es un objeto en relación a otro.

Singular Value Descomposition

(una generalización de PCs)

El teorema Singular ***value decomposition*** (SVD) establece que toda matriz Xnxp, no necesariamente cuadrada, como lo exige componentes principales, puede ser escrita como el producto de tres matrices:

$$\mathbf{X = U\,\Lambda\,V^T} \quad \text{o} \quad \mathbf{\Lambda = U^T X V} \qquad [3]$$

$\mathbf{U_{nxr}}$ **y** $\mathbf{V_{pxr}}$ son matrices ortonormales en columnas y $\mathbf{\Lambda_{rxr}}$ es una matriz diagonal (r=rango de **X**; r ≤ p) similar a la matriz de eigenvalues en PC.

Si la matriz es cuadrada, tal como las de covariancia o correlación, entonces

$$\mathbf{C_n = XX^T = U\Lambda^2 U^T} \quad \text{y} \quad \mathbf{C_P = X^T X = V\,\Lambda^2 V^T} \qquad [4]$$

La ecuación [4] define la ***eigenvalue descomposition*** de una matriz cuadrada simétrica (también llamada **descomposición espectral**). En este caso **U** y **V** son iguales y se la denomina matriz de *eigenvectors* E_{pxp} de $\mathbf{C_p}$. Seguiremos con la escritura de **U** y **V** para conservar las ecuaciones anteriores que son más generales que para las matrices cuadradas.

Debido a la ortonormalidad de **U y V** los nuevos ejes en sus espacios matriciales tienen **longitud 1** y son **ortogonales entre sí**, lo que algebraicamente significa imponer las mismas restricciones algebraicas que para PCs.

En los nuevos ejes, cada objeto tendrá un nuevo juego de coordenadas que se llaman **scores.**

Definimos las coordenadas de la matriz de ***scores*** **S** de **las n filas** de **X** como:

$$\mathbf{S = U\,\Lambda} \qquad \text{y de [3]} \rightarrow \mathbf{S_{n,r} = UU^T XV = X_{n,r} V_{r,r}}$$

$\mathbf{S}$ es ortogonal porque U lo es $\rightarrow \mathbf{S^T S = \Lambda U^T U\,\Lambda = \Lambda^2}$

Cada **elemento** de **S** es un producto escalar:

$\mathbf{s_{ik} = x_i . v_k = \sum_j^p x_{ij} v_{jk}}$ i=1…n k=1…r. Scores de objetos i,…,n y componente principal 1, 2,…, r.

Cada **columna** de **S** (recuerde que **S** tiene r columnas) representa un ***componente principal fila*** de **X** y puede ser interpretado como una combinación lineal de las columnas (variables) de **X** con los elementos de **V** como coeficientes de peso.

$$\mathbf{s_k = X.v_k = } \sum_j^p x_j v_{jk} \qquad \mathbf{k=1\ldots r}$$

Del mismo modo definimos las coordenadas de **las p columnas** de **X** para la matriz de ***loadings*** L.

$$\mathbf{L=V\Lambda=X^TU} \text{ y de } \mathbf{[3]} \rightarrow \mathbf{U^TX= \Lambda V^T \rightarrow X^TU=V\Lambda}$$

También, cada elemento de L puede escribirse como un producto escalar:

$$l_{jk} = \mathbf{x_j^t.u_k} = \sum_i^n x_{ij} u_{ik} \qquad j=1\ldots p \quad k=1\ldots r$$

Cada **columna** de **L** representa un ***componente principal columna*** de **X** y puede ser interpretado como una combinación lineal de filas (objetos) de **X** con los elementos de **U** como coeficientes de peso.

$$\mathbf{l_k = X^T.u_k} = \sum_i^n x_i u_{ik} \qquad k=1\ldots r \quad \text{; y desarrollando…}$$

$$\mathbf{l_1} = x_{11}u_{11}<\Lambda_1^{1/2})> + x_{21}u_{21}<\Lambda_1^{1/2}> + \ldots + x_{n1}u_{n1}<\Lambda_1^{1/2}>$$

$$\mathbf{l_2} = x_{12}u_{12}<\Lambda_2^{1/2}> + x_{22}u_{22}<\Lambda_2^{1/2}> + \ldots + x_{n2}u_{n2}<\Lambda_2^{1/2}>$$

.

.

.

$$\mathbf{lr} = x_{1r}u_{1r}<\Lambda_r^{1/2}> + x_{2r}u_{2r}<\Lambda_r^{1/2}> + \ldots + x_{nr}u_{nr}<\Lambda_r^{1/2}>$$

Este desarrollo de SVD es imprescindible cuando hay que trabajar directamente con matrices de datos que regularmente no son cuadradas, o sea, el número de observaciones (objetos) es mayor que el número de variables. Esto lo veremos en el tema siguiente y lo retomaremos en el capítulo calibración multivariada.

Análisis de Factores

Una notación tradicional en quimiometría equivalente a la definición de SVD es:

$$\mathbf{X=T.P^T} \qquad \textbf{[5]},$$

donde $\mathbf{T=U\,\Lambda}$ (matriz de scores) y $\mathbf{P=V}$ (matriz de loadings), con lo que, como era originalmente $\mathbf{X= U\,\Lambda V^T}$.

Cada fila **xi** de **X** se calcula con la correspondiente fila **ti** de la matriz de scores:

$\mathbf{x_i = t_i\,V^T}$. **X** puede ser reconstruida con los primeros k PCs adicionando una matriz E que toma en cuenta el error cometido al no tomar todos los PCs, $\mathbf{X_{n.r}=T_{n,k}\,P_{k,r}+E_{n,r}}$.

Ahora bien…, muchos problemas en el campo de la ciencia pueden expresarse en forma isomorfa a la ecuación general **[5]**. Cuando un problema se expresa de esta manera (una combinación lineal del producto de 2 matrices) se lo denomina ***análisis de factores***. Observe que, aunque las técnicas de cálculo para evaluar **U** y **V** sean similares al cálculo de CP, el problema está planteado en forma diferente, debido a que nosotros **debemos saber de antemano** que nuestro problema se puede expresar de esa manera. A modo de ejemplos mostraremos dos problemas concernientes a la química del medio ambiente.

Ejemplo 1: En un problema medioambiental llamado el problema fuente–receptor se muestrean aerosoles en los cuales se determinan m componentes químicos en n períodos de tiempo. Tenemos una matriz de datos X de (nxm) y existen p fuentes contaminantes. Con esto, de acuerdo al análisis de factores, podemos escribir:

$$\mathbf{X_{n,m} = C_{n,p} \cdot S_{p,m}}$$

C es la matriz de 'masa total de contaminantes' (para cada período o muestra u objeto) proveniente de cada una de las p fuentes. **S**, matriz llamada *perfil de las fuentes*, es la composición de contaminantes de cada una de las p fuentes.

Sabiendo que nuestra matriz de datos puede descomponerse en este producto de matrices, podríamos evaluar el **número de componentes principales que describen la mayor parte de la variabilidad** del sistema (80-90%) y asignarlo al valor p. Como difícilmente puede conocerse el número de fuentes que afectan a una región, debido a que estas pueden estar ocultas o muy distantes, esta sería una manera de determinar el número de fuentes que afectan al lugar, para el tipo de compuestos medidos.

Aunque la expresión $\mathbf{X=T.P^T}$ proviene del cálculo de los componentes principales, el cálculo de **U** y **V** para PC no representa los valores de $\mathbf{C_{n,p}}$ y $\mathbf{S_{p,m}}$ reales.

Porque, mientras que **U** y **V** han sido calculados bajo ciertas restricciones (ortogonalidad, normalización), $\mathbf{C_{n,p}}$ y $\mathbf{S_{p,m}}$ son matrices con significado físico y pueden tener restricciones muy diferentes a **U** y **V**. **El análisis de factores consiste entonces en transformar las matrices U y V para darles un significado físico acorde al problema tratado.** Existen un variado número de métodos de transformación, muchos de ellos sobre la base de la rotación de los ejes principales.

Como se ha dicho, según el sistema que se esté estudiando, habrá restricciones diferentes (o adicionales) al cálculo de los CP. Por ejemplo, en los cálculos medioambientales no debería haber resultados con concentraciones negativas. Hay una variedad de métodos para 'adaptar' los CP al problema en estudio y así obtener el análisis de factores adecuado. Quien quiera profundizar este tema puede consultar la bibliografía dada, especialmente (Ref. 4), Part B, chapter 34.

Cuando se representan los scores en el plano F1-F2, (en análisis de factores se suele denominar F en lugar de C para los factores) debido a la relación física de las variables, los scores guardan cierta estructura. El gráfico 2.8 muestra los scores del problema siguiente (ejemplo 2) con ejes rotados respecto de la posición original.

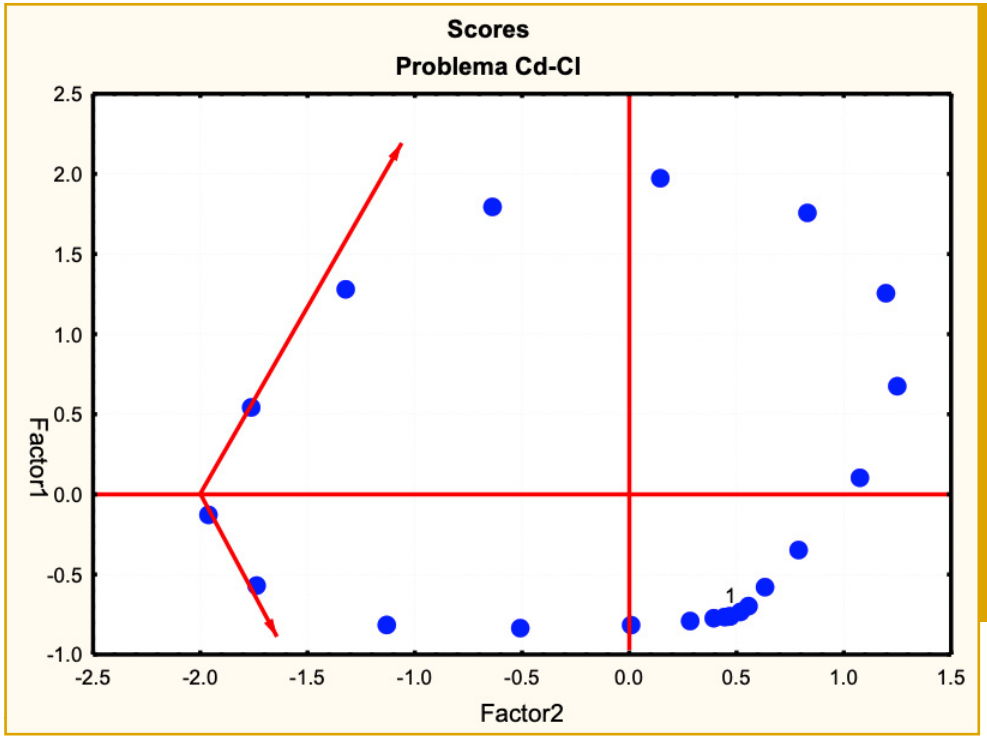


Figura 2.8

Aquí se puede ver como una rotación de los ejes F1-F2 'acomoda' los valores de las matrices a soluciones reales.

Uno de los métodos más conocidos para la rotación de los ejes es el *Varimax*. Hemos visto que en CP los ejes de las variables son rotados; aquí rotaremos los CPs significativos en el subespacio V^t.

F=V^t*R, donde R es una matriz simétrica de ángulo de rotación (Ø), tal como:

$$R = \begin{vmatrix} sen(\emptyset) & \cos(\emptyset) \\ -sen(\emptyset) & \cos(\emptyset) \end{vmatrix}$$

En este método, la rotación de los ejes busca la *máxima simplicidad,* definida como:

$$Simp = \text{var}(f_i^2) = \frac{1}{p}\sum_{i=1}^{p}(f_i^2 - \overline{f_i^2})^2$$

f_i es un vector fila de la matriz **F**, f_i^2 es el cuadrado de los elementos del vector, p es el número de factores tomados en **V**. La simplicidad de una matriz es la suma de simplicidades de sus vectores, o simplemente la variancia de f_i^2.

Carécteristicas de algunos métodos de transformación

Varimax: Máxima simplicidad sobre las columnas de la matriz de loadings.

Quartimax: Máxima simplicidad sobre las filas de la matriz de loadings.

Biquartimax: Máxima simplicidad sobre las columnas y filas de la matriz de loadings.

Equamax: mezcla *pesada* entre varimax y Quartimax.

Ejemplo 2: Aplicación del Análisis de Factores a la determinación del número de especies presentes en sistemas Catión-ligando por métodos polarográficos (Ref. 6).

Cuando un catión metálico está en equilibrio con una solución que contiene un ligando, éste puede estar formando especies con un número variado de ligandos. Por ejemplo, el Cd^{++} en una solución de CN^- puede formar los complejos $CdCN^{+1}$,

$Cd(CN)_2$ y $Cd(CN)_3^{-1}$. Por diversas razones, no siempre se forma toda la serie de complejos posibles, sino que algunos complejos intermedios no se forman en cantidad apreciable. Es por supuesto de interés, conocer cuántas especies se forman, con el fin de comenzar a estudiarlas. Se supone aquí un sistema químico donde las velocidades de reacción son muy rápidas, de modo que éste está siempre en equilibrio.

La polarografía es una técnica electroquímica en la cual se aplica un barrido de potencial entre dos electrodos sumergidos en una solución y se va registrando la corriente.

La expresión de la corriente polarográfica total, It, para un sistema catión–ligandos múltiples en solución es:

$$\mathbf{I_t = \sum_{j=0}^{n} \Phi_j(E, K_e, k_c, k_a, C_x).C_j} \qquad \textbf{[6]}$$

Donde E es el potencial instantáneo de electrodo, $K_{e,}$ es la constante de estabilidad del complejo j, k_c y k_a son las corrientes límite catódica y anódica respectivamente, C_x es la concentración de ligando y C_j es la concentración del complejo j. Para una serie de **s** soluciones con diferentes concentraciones de ligando, obtenemos sus polarogramas y medimos su intensidad a ne potenciales prefijados. Formamos entonces una matriz $\mathbf{Zn_{e,s}}$ cuyos elementos son las intensidades experimentales $I(E,C_x)$. De acuerdo a la sumatoria **[6]** podemos expresar la matriz **Z** como:

$\mathbf{Z=\Phi_{ne,n}.C_{n,s}}$ Donde n es el número de especies en solución. Observe que lo que hemos hecho es expresar el problema en la forma de la ecuación **[5]**.

	C_{L1}	C_{L2}	C_{L3}	.....	C_{Ls}
E1	I_{11}				
E2	I_{12}	I_{22}			
E3					
.	-		**Z** -		
.	-				
Ee	I_{1e}				Ies

	n_1	n_2 ...	n_n
E1			
E2			
E3			
.		**ϕ**	
.			
Ee			

	C_{L1}	C_{L2}	C_{LS}
n_1			
n_2			
n3			
		C	
n_n			

Nuevamente, si obtenemos el número de factores que contienen la mayor parte de la variabilidad del sistema, este número se puede atribuir a n, el número de especies en solución.

Table I. The *k* values estimated by various methods

System		Method					
		RV%	CRV%	IND	S_k *vs. k*	Our method	Literature values
(I)	Cd-Cl	6	9	8	1	3	3
(2)	Cd-Cl (3 species)	2	2	6	2-3	2-3	
(2)	Bi-Cl (6 species)	5	5	11	2	5-7	6
(2)	Cd-SCN (3 species)	2	2	5	2	2 3	2 3
(1)	Cu-morpholine	5	5	5	2	2	

(I) Experimental. (2) Calculated.

En la tabla de la izquierda puede verse el número de especies obtenido por este método (column Our) y otros métodos para algunos sistemas experimentales y calculados.

Otra Técnica Lineal Emparentada con SVD

Análisis de Correlación Canónica (canonical correlation analysis)

El Análisis de Correlación Canónica es una técnica que se utiliza para estudiar la correlación entre **dos series de datos** o matrices de medidas $\mathbf{X}_{n,k}$ e $\mathbf{Y}_{n,m}$ (Ref. 4). Estas matrices pueden tener diferente número de variables, pero deben tener el

mismo número de observaciones. Se trata entonces de encontrar correlaciones **entre grupos** de variables (combinacones lineales) de **X** y de **Y.** Por ejemplo, una alta correlación entre

$$w_1.\mathbf{x_{v1}}+w_3.\mathbf{x_{v3}}+w_5.\mathbf{x_{v5}} \quad y \quad q_3.\mathbf{y_{v3}}+q_4.\mathbf{y_{v4}}+q_6.\mathbf{y_{v6}}+q_8.\mathbf{y_{v8}}.$$

Donde wi y qi son pesos asignados por el método y vi es el número de variable asignado en cada matriz. Un ejemplo real es el siguiente: en un estudio sobre contaminación de aguas en el Río Reconquista, Provincia de Buenos Aires, se quería averiguar si existía cierta correlación entre la contaminación orgánica, Matriz **X** ≡ (DBO, DQO, surfactantes, nitrógeno orgánico y nitrógeno total Kjeldahl) y la de cationes metálicos **Y** ≡ (Cd, Zn, Cu, Cr, Mg, Ni and Pb) (Ref. 9-10). La base de datos contaba con 270 muestras de 26 variables colectadas durante 10 años. La principal relación encontrada fue una correlación ρ=0.83 entre las combinaciones lineales UR y VR.

UR = 1.23% DBO +1.91% DQO + 3.49% SURFACT + 66.53% N_ORG + 26.83%NTK

VR = 3.54% Cd +1.85% Zn +1.30% Cu +0.01% Cr +0.07% Mg +90.14% Ni +3.07% Pb

Volviendo al desarrollo teórico, las combinaciones lineales de **X** y de **Y** que tienen máxima correlación se denominan primeras variables canónicas $\boldsymbol{t_1}=\boldsymbol{X_1}\,\boldsymbol{w_1}$ y $\boldsymbol{u_1}=\boldsymbol{Y_1}\,\boldsymbol{q_1}$. Los vectores de coeficientes $\boldsymbol{w_1}$ y $\boldsymbol{q_1}$ son los pesos canónicos de las variables en **X** e **Y** respectivamente. La correlación entre estas dos primeras variables canónicas es la primera correlación canónica, $\boldsymbol{\rho_1}$. El siguiente par de variables canónicas $\boldsymbol{t_2}$ y $\boldsymbol{u_2}$ también tienen máxima segunda correlación $\boldsymbol{\rho}$, sujeto sin embargo a la condición de que este segundo par no debe estar correlacionado con el primer par, i.e. $\boldsymbol{t_1^T t_2 = u_1^T u_2 = 0}$. El análisis prosigue extrayendo pares adicionales de variables canónicas ortogonales a las previas, hasta que la matriz de datos con el menor número de variables haya sido transformada completamente en el mismo número de variables canónicas ortogonales. Resulta que las correlaciones inter grupo de variables canónicas de diferentes dimensiones son también ortogonales, i.e. $\boldsymbol{t_i^T u_j = 0}$ para $i \neq j$. Bajo el supuesto de multinormalidad de ambas poblaciones se pueden testear las correlaciones canónicas y así lograr una considerable reducción dimensional.

ANEXO

Los siguientes términos son sinónomos:

Eigenvalue = Valor propio = Valor característico = Valores latentes

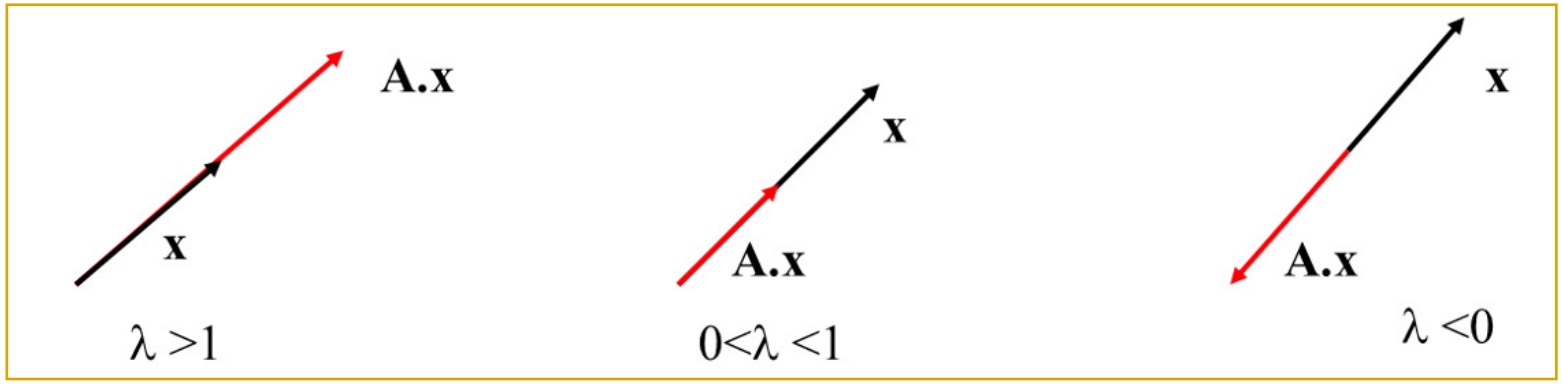
Eigenvector = Vector propio = Vector característico = Vector latente

Sea una **matriz cuadrada A** de nxn. El número real λ es un ***eigenvalue*** de **A**, si existe un vector **x** (***eigenvector***), distinto de cero, tal que:

$$\mathbf{A.x} = \lambda.\mathbf{x} \qquad [a]$$

Todo vector **x** ≠0 que satisfaga [a] es un ***eigenvector*** de A asociado al eigenvalue λ.

El producto de una matriz por un eigenvector da como resultado un vector. Entonces, el vector **A.x** resulta paralelo a **x**.



Definición: El **centroide** de un grupo de objetos es el punto, en el espacio multivariable, cuyas coordenadas son el valor medio de cada variable.

Referencias

1- Fisher R. A. *Annals of Eugenics*, 7:179-188, 1936.

2- Michael Berthold and David J. Hand (Eds.). Intelligent Data Analysis. Springer,Berlin Heidelberg New York 2003. Second edition.

3- Chemometrics: a textbook. D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman. ELSEVIER. The Netherlands, 1988.

4- D.L. Massart; B.G.M. Vandeginste; L.M.C. Buydens; S. De Jong; P.J. Lewi and J. Smeyers-Verbeke. Hanbook of Chemometrics and Qualimetrics. Part A chapter 17 and Part B chapter 31. Elsevier, Amsterdam 1997.

5- Jure Zupan; Johann Gasteiger, Neural Networks in Chemistry and Drug Design. 2nd Edition,. Wiley-VCH, Weinheim, 1999.

6- Application of Factor Analysis to Polarographic Data. Determination of the Number of Species Present in Metal Ion-Ligand Systems. Erwin Baumgartner, Raquel G. Gettar, Francisco D. Mingorance and Jorge F. Magallanes.

7- Deep Learning Techniques to Improve the Performance of Olive Oil Classification. Belén Vega-Márquez, Isabel Nepomuceno-Chamorro, Natividad Jurado-Campos and Cristina Rubio-Escudero. https://doi.org/10.3389/fchem.2019.00929 https://www.google.com/search?sxsrf=ALeKk00SAPYlZF5Y_k_C6l_s9mjBQi_0Mg:1593184598187&source=univ&tbm=isch&q=chemometric+classification+of+olive+oil+data+base&sa=X&ved=2ahUKEwiGgKWQ45 qAhX0HbkGHdMxACoQsAR6BAgHEAE&biw=1360&bih=663.

8-Ananalysisofsecondarypollutantsin BuenosAiresCity.SilviaReich,JorgeMagallanes,LauraDawidowski,DaríoGómez,Neva Grošelj, Jure Zupan. Environmental Monitoring and Assesment (2006) 119:441- 457. DOI: 10.1007/s10661-005-9035-2

9- Multivariate Chemometric Analysis of a Polluted. River of a Megalopolis. Alejandro Gabriel García-Reiriz, Jorge Federico Magallanes, Marjan Vracko, Jure Zupan, Silvia Reich, Daniel Salvador Cicerone. Journal of Environmental Protection, 2011, 2, 903-914. doi:10.4236/jep.2011.27103 Published Online September2011. (http://www.SciRP.org/journal/jep) Copyright © 2011 SciRes. JEP 903

10- The use of an electronic nose to characterize emissions from a highly polluted river". Alberto Lamagna, Silvia Reich, Daniel Rodríguez, Alfredo Boselli, Daniel Cicerone. Sensors&Actuators B. 2008, 131, 121-124

CAPÍTULO 3

Análisis de grupos (Clusters)

Clasificaciones

Clustering (agrupamiento) o ***cluster analysis*** (análisis de grupos) es una técnica utilizada para clasificar objetos caracterizados por una serie de variables, en base a su ***similaridad*** o ***dis-similaridad***.

Existen dos formas básicas de hacer las clasificaciones:

Clasificaciones
- Jerárquicas – (dendrogramas)
- No jerárquicas – (clustering)

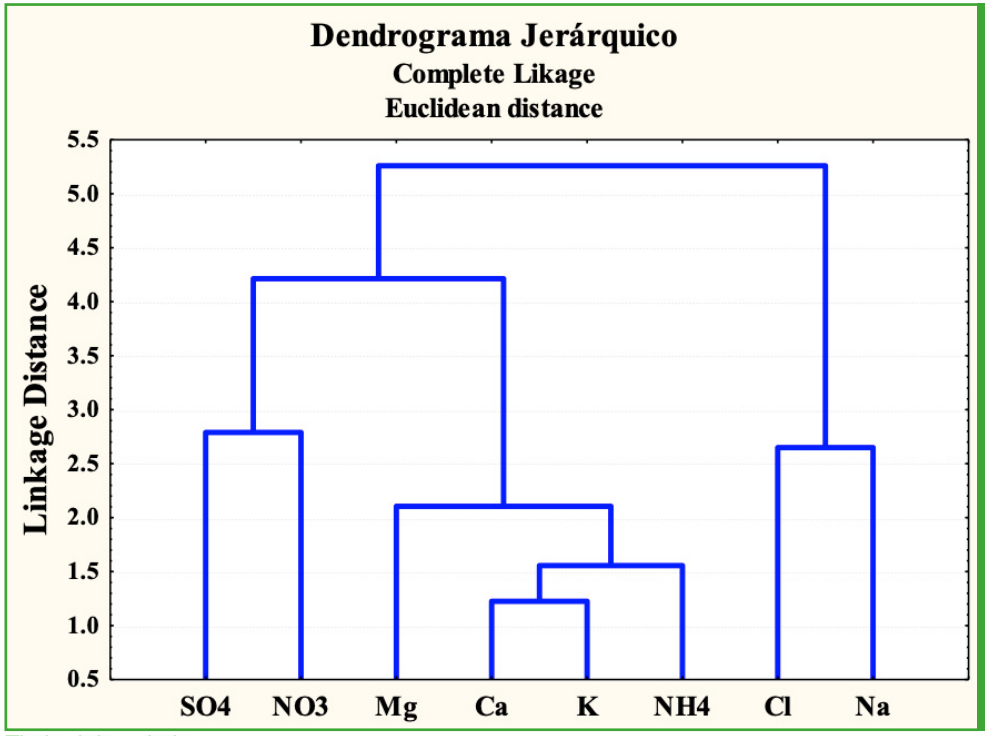


Figura 3.1

Ya hemos visto que los dendrogramas tienen estructura de árbol y que permiten una eficiente reducción dimensional porque pueden representarse en un plano, independientemente del número de variables que intervengan en el problema. Cabe señalar que las variables pueden ser manifiestas o latentes, y que las primeras pueden tener cualquier pretratamiento de datos que las adecuen.

En la figura 3.1 vemos un dendrograma del siguiente sistema: En un área geográfica bastante extendida se han recolectado muestras de agua en 18 puntos y en cada uno se ha medido la concentración de varios iones por cromatografía iónica. Veremos, más adelante, como calcular este dendrograma.

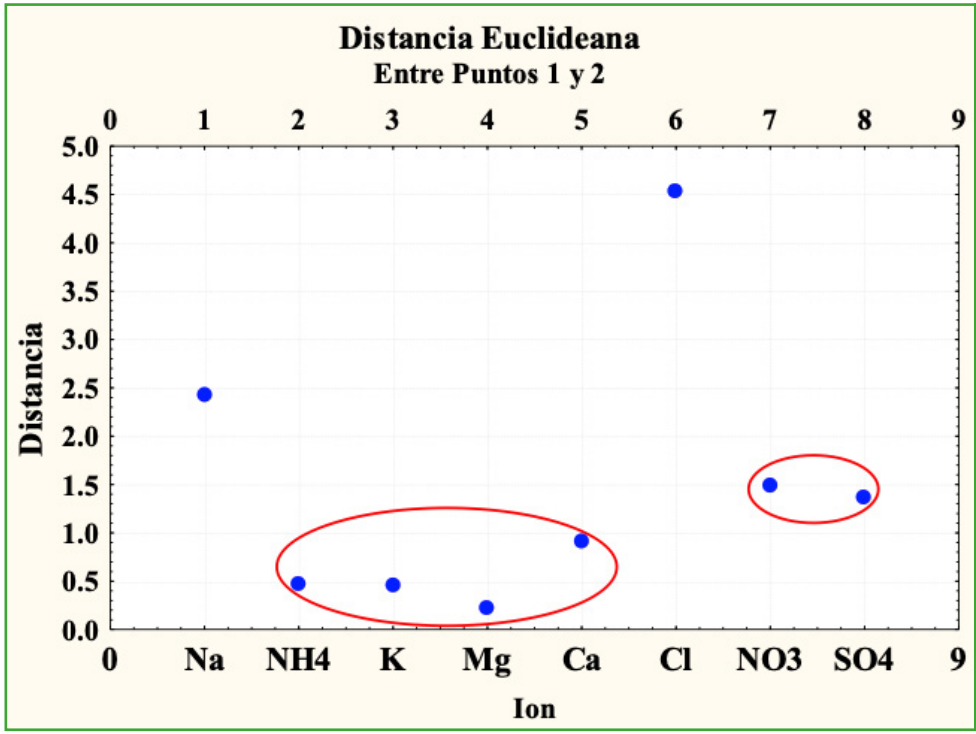


Figura 3.2

Los clusters no jerárquicos no se pueden visualizar a menos que los objetos dependan de sólo 2 o 3 variables, ya que no podemos representar dimensiones mayores. De modo que deberemos analizar resultados numéricos para comprender este tipo de clusters. Una excepción lo constituye el caso en que los objetos son variables latentes y entonces se representen los agrupamientos en el plano de 2 o 3 componentes principales.

La figura 3.2 muestra la distancia para cada ion entre los 2 grupos principales: (aniones) y (cationes) del cluster 3 (ver tabla siguiente), pero no podemos observar los 8 puntos a la vez porque necesitaríamos 8 dimensiones en este tipo de clusters.

N° Clusters	Tabla 1
Inicio	Na NH4 K Mg Na Cl NO3 SO4
2 Clusters	Na-Cl-NO3-SO4 NH4-K-Mg-Ca
3 Clusters	NO3-SO4 Na-Cl NH4-K-Mg-Ca
4 Clusters	NO3-SO4 Mg Na-Cl NH4-K-Ca
5 Clusters	SO4 NO3 Na-Cl NH4-K-Ca Mg
6 Clusters	Na NH4 Cl K-Ca Mg NO3-SO4
7 Clusters	Na NH4 Cl K-Ca Mg SO4 NO3
8 Clusters	Na NH4 K Mg Na Cl NO3 SO4

En la Tabla 1 se muestra la evolución del cálculo con el aumento del número de clusters. A diferencia de lo que ocurre en los clusters jerárquicos, donde una vez ubicado un objeto en un cluster, permanece fijo en el lugar, aquí los objetos cambian de lugar a medida que el número de clusters aumenta.

Podemos concluir entonces en que el número de clusters es subjetivo y así como en la tabla 1 se puede elegir el número correcto de clusters en función de la información adicional que el investigador tiene y analiza, en los clusters jerárquicos el número de clusters correcto depende de la altura en que se corte la distancia en la Fig. 3.1. Por ejemplo, si se corta a la altura de 3 tenemos 3 clusters y a la altura de 4.5 hay solo 2. Esto quiere decir que el número de clusters que se determinen **depende del grado de similitud o dis-smilitud que se considere**.

Similitud - dis-similitud – distancia

Para poder agrupar objetos uno debería medir su ***similitud***. La ***distancia*** entre objetos podría usarse como una de tales medidas, pero muchos tipos de ***coeficientes de similitud*** pueden ser aplicados. Similitud y dis-similitud no tienen definiciones claras, y en cambio la distancia puede definirse mucho mejor:

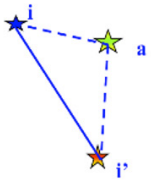
Dis-similitud = Distancia = 1 – similitud

Una dis-similitud entre dos objetos i e i’ es una distancia, D, sí

1- $D_{ii'} \geq 0$; $D_{ii'} = 0$ si $x_i = x_{i'}$ "Las distancias son 0 ó positivas"

2- $D_{ii'} = D_{i'i}$ "Las distancias son simétricas"

3- $D_{ia} + D_{i'a} \geq D_{ii'}$ "Desigualdad métrica: establece que la suma de distancias desde cualquier objeto a los objetos i e i' no puede ser nunca menor a la distancia entre i e i' "



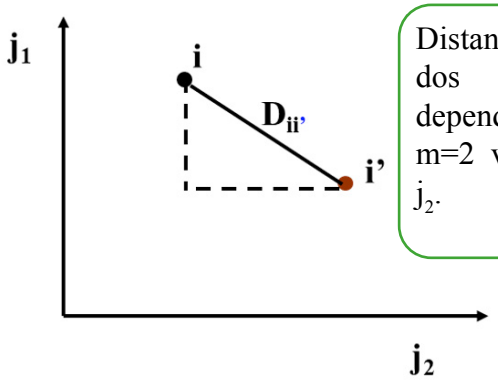
Medidas de dis-similitud para Variables Continuas

La ***distancia Euclidiana*** entre dos objetos i e i', para el conjunto de j=1....m variables se define como:

$$D_{ii'} = \sqrt{\sum_{j=1}^{m} (x_{ij} - x_{i'j})^2}$$

Equivalentemente, $D_{ii'}^2 = (x_{ij} - x_{i'j})^t . (x_{ij} - x_{i'j})$

Observe que esta es una aplicación del teorema de Pitágoras al espacio multidimensional



Distancia entre dos objetos que dependen de solo m=2 variables, j_1 y j_2.

En notación vectorial $D^2_{ii'} = (\mathbf{x}_i - \mathbf{x}_{i'})^T (\mathbf{x}_i - \mathbf{x}_{i'})$ Recuerde que las $\mathbf{x}_i$ son objetos (vectores) que dependen de varias variables. En los casos en que se desea dar un ***peso*** diferente a algunas variables se recurre a la ***distancia Euclidiana pesada***: esto puede ser necesario cuando las variables de una misma unidad tienen muy distinta magnitud. Un ejemplo sería, si en un problema las concentraciones en solución para algunas variables pueden estar en el orden del % y otras en niveles de trazas. Según que se quiera considerar a ambas con igual importancia para el problema, o no, habrá que aplicar o no, respectivamente, pesos a las variables.

$$\mathbf{D_{ii'}} = \sqrt{\sum_{j=1}^{m} \mathbf{w_j}(\mathbf{x_{ij}} - \mathbf{x_{i'j}})^2} \qquad \text{Con } \sum w_j = 1$$

wj es el peso dado a la variable j en los objetos.

Otra variante: la ***distancia Euclidiana estandarizada***: ésta consiste en autoescalar los datos de modo de tener unidades adimensionales y darle similar consideración (importancia) a todas las variables, ya que sus valores oscilarán aproximadamente entre +3 y –3 para variables ***normalmente*** distribuidas (no aplicable si este no es el caso).

$$\mathbf{D_{ii'}} = \sqrt{\sum_{j=1}^{m} [(\mathbf{x_{ij}} - \mathbf{x_{i'j}}) / \mathbf{s_j}]^2}$$

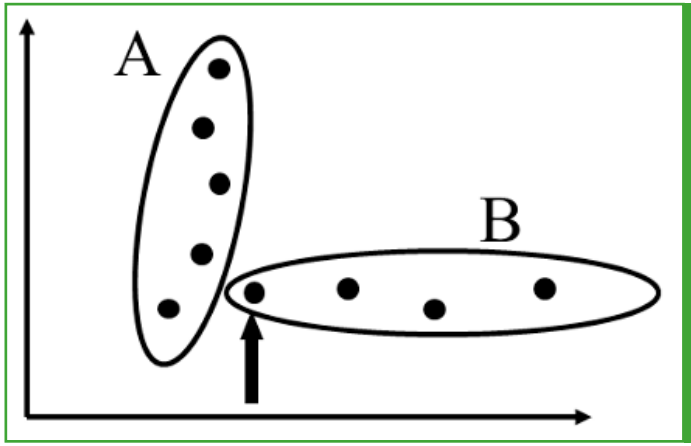
S_j es la desviación estandar de la j-ésima columna $X_{n,m}$

Otra variante muy importante es la ***distancia de Mahalanobis***, dada por:

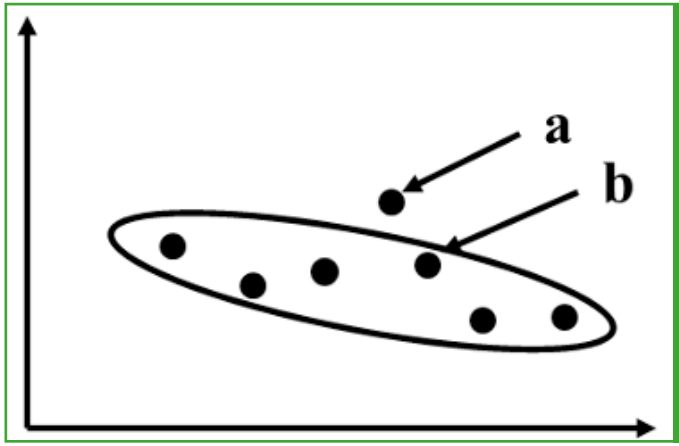
$$D^2_{ii'} = (\mathbf{x}_i - \mathbf{x}_{i'})^T \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{x}_{i'})$$

La distancia de Mahalanobis representa **la distancia entre un grupo de objetos y un objeto simple i'. C** es la matriz de varianza-covarianza de un cluster representado por $\mathbf{x}_i$ que es el *centroide* del cluster.

Ocurre que ciertos clusters tienen una determinada **estructura**, y ésta hace que ciertos objetos sean incorporados a un cluster en función de la estructura de éste y no de la distancia euclidiana. Por ejemplo...



En la figura, el objeto indicado por la flecha, tiene a sus vecinos más cercanos en el cluster A. Sin embargo, considerando la estructura geométrica de ambos clusters, éste pertenece claramente al cluster B.



Del mismo modo, el objeto b está más cerca de a que de sus otros vecinos, sin embargo 'a' no puede considerarse dentro del cluster.

La distancia de Mahalanobis entonces, toma en cuenta la estructura del cluster observando la relación entre sus objetos a través de la matriz de varianza-covarianza y calculando la distancia entre el centroide y un objeto fuera del cluster. Para los casos de las figuras, ésta es la manera de determinar si un objeto pertenece o no a un cluster.

La ***distancia generalizada*** es una expresión general que reúne todas las variantes dadas hasta aquí. Antes habíamos expresado las distancias en forma de cálculo vectorial. Ahora vemos que la ventaja de expresarlas así permite generalizar las distancias anteriores en una sola ecuación.

$$D^2_{ii'} = (\mathbf{x}_i - \mathbf{x}_{i'})^T \mathbf{W} (\mathbf{x}_i - \mathbf{x}_{i'}) \qquad [1]$$

W es una matriz de pesos, cuadrada de m x m.

W = I (matriz unidad) representa la distancia euclidiana ordinaria.

W = diag(**w**) reproduce la distancia euclidiana pesada. w es un vector de pesos.

W = diag($1/d^2$) reproduce la distancia Euclidiana estandarizada. d representa el vector columna de desviaciones estándar.

W = $\mathbf{C}^{-1}$ define la distancia de Mahalanobis. $\mathbf{C}^{-1}$ es la inversa de la matriz de varianza – covarianza del cluster de referencia.

Medidas de similitud utilizando el coeficiente de correlación

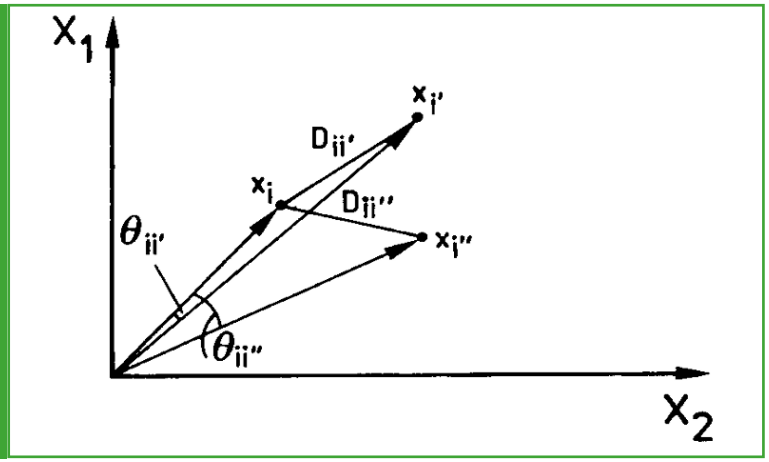
En la introducción ya hemos explicado el *coeficiente de correlación* de Pearson. Rescribamos su expresión:

$$r(y_1,y_2) = \mathrm{cov}(y_1,y_2)/(s_{y1},s_{y2}) = \frac{\frac{\sum (y_1-\overline{y_1})(y_2-\overline{y_2})}{n-1}}{\sqrt{\frac{\sum (y_1-\overline{y_1})^2 (y_2-\overline{y_2})^2}{(n-1)\quad(n-1)}}} = \frac{(\mathbf{y}_1-\overline{\mathbf{y}_1}).(\mathbf{y}_2-\overline{\mathbf{y}_2})}{\|\mathbf{y}_1-\overline{\mathbf{y}_1}\|\,\|\mathbf{y}_2-\overline{\mathbf{y}_2}\|}$$

Haciendo $\mathbf{y}_1' = \mathbf{y}_1 - \overline{\mathbf{y}_1}$ e $\mathbf{y}_2' = \mathbf{y}_2 - \overline{\mathbf{y}_2}$

$$r(\mathbf{y}_1', \mathbf{y}_2') = \frac{y_1' \cdot y_2'}{\|y_1'\| \cdot \|y_2'\|} = \cos(\theta)$$

Observe la diferencia entre *distancia euclidiana y distancia de correlación* para el caso en que x_i, $x_{i'}$ y $x_{i''}$ sean los **vectores medios de las variables**. ¡No confunda con la distancia angular entre objetos!



Ejemplo: La diferencia entre distancia Euclidiana y correlación en un problema químico.

En la Tabla 2 se muestran las concentraciones de un curso de agua, el primero, que recibe los cursos afluentes 2 y 3, se pregunta ¿cuál es el curso que más afecta al curso principal?

Tabla 2

Curso	Sólidos disueltos (g/l)	Sólidos en suspensión (g/l)	Cloruros ppm	Sodio ppm	Potasio ppm	Calcio ppm	Magnesio ppm
1	0.02	0.05	1.5	1.7	0.7	0.5	0.3
2	0.01	0.02	1.7	2	0.5	0.3	0.1
3	0.04	0.1	3	3.4	1.4	1	0.6

A simple vista, parecería que el más similar es el curso 2, porque tenemos tendencia a mirar hacia números más similares. Sin embargo, el curso más similar al primero es el 3. Una observación más cuidadosa muestra que la correlación entre el curso 1 y 3 es 1.0, ahora observamos que el curso 3 tiene concentraciones que duplican exactamente a las del curso 1. O sea que esencialmente el curso 1 es una dilución del curso 3. Tenga en cuenta entonces que **las medidas angulares revelan la proporción** entre los vectores considerados.

La Similitud y las Medidas de Distancia Angular

Hay una gran diferencia conceptual y de cálculo entre *distancias longitudinales o Euclidianas* y *distancias angulares* y también respecto de las de correlación.

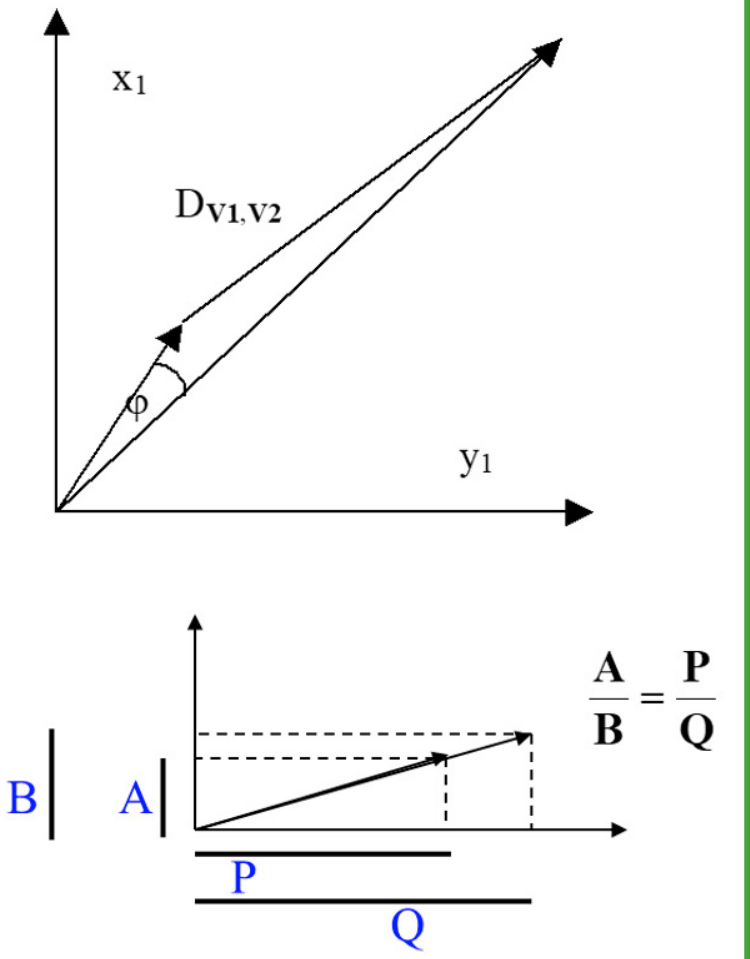
Las *distancias angulares entre objetos* son **él ángulo que existe en el espacio multidimensional entre dos objetos cualesquiera**. Recuerde que la correlación se refiere a ángulos **entre variables centradas**.

Desde el punto de vista del cálculo, las distancias angulares pueden obtenerse a través del producto escalar entre dos vectores. Si tenemos dos objetos representados por V_1 y V_2, tales como:

$\mathbf{V_1}=x_1, x_2,..., x_m$. y $\mathbf{V_2}=y_1,y_2,..., y_m$. Entonces su producto escalar (o producto 'punto') se define como:

$$\mathbf{V_1}.\mathbf{V_2} = x_1.y_1+x_2.y_2+...+x_m.y_m = |\mathbf{V_1}|.|\mathbf{V_2}|.\cos(\mathbf{V_1},\mathbf{V_2})$$

$\cos(\varphi) = \cos(\mathbf{V_1},\mathbf{V_2}) = \mathbf{V_1}.\mathbf{V_2}/(|\mathbf{V_1}|.|\mathbf{V_2}|)$ Observe que este es el ángulo **entre los dos objetos**, $\mathbf{V_1}$ **y** $\mathbf{V_2}$. Sin embargo, observe que **los elementos** de $\mathbf{V_i}$ pueden tener distintas unidades (no así en el caso de la correlación). ¿Cómo procedería para evitar este problema?



La figura muestra, para el caso en que V1 y V2 tuvieran sólo 2 dimensiones, la diferencia geométrica que existe entre distancias longitudinales y angulares. Se ve que aunque la distancia longitudinal puede ser grande, la distancia angular podría llegar a ser muy pequeña, e incluso con ángulo 0, o sea: cos (φ)=1.

Más importante aún que las diferencias de cálculo y geométricas, es la diferencia conceptual. Observe en la figura inferior, que para 2 objetos cuya distancia angular entre ellos es 0, sus magnitudes escalares guardan idéntica **proporción.**

La **proporción** es una cualidad importante en muchos problemas, como en el ejemplo de la Tabla 2, piense por ejemplo que, en química dos soluciones que difieran meramente en su dilución, desde el punto de vista de **la proporción entre sus elementos**, sus distancias angulares valen cero, o sea son iguales. Veamos un ejemplo aplicado a mediciones medioambientales:

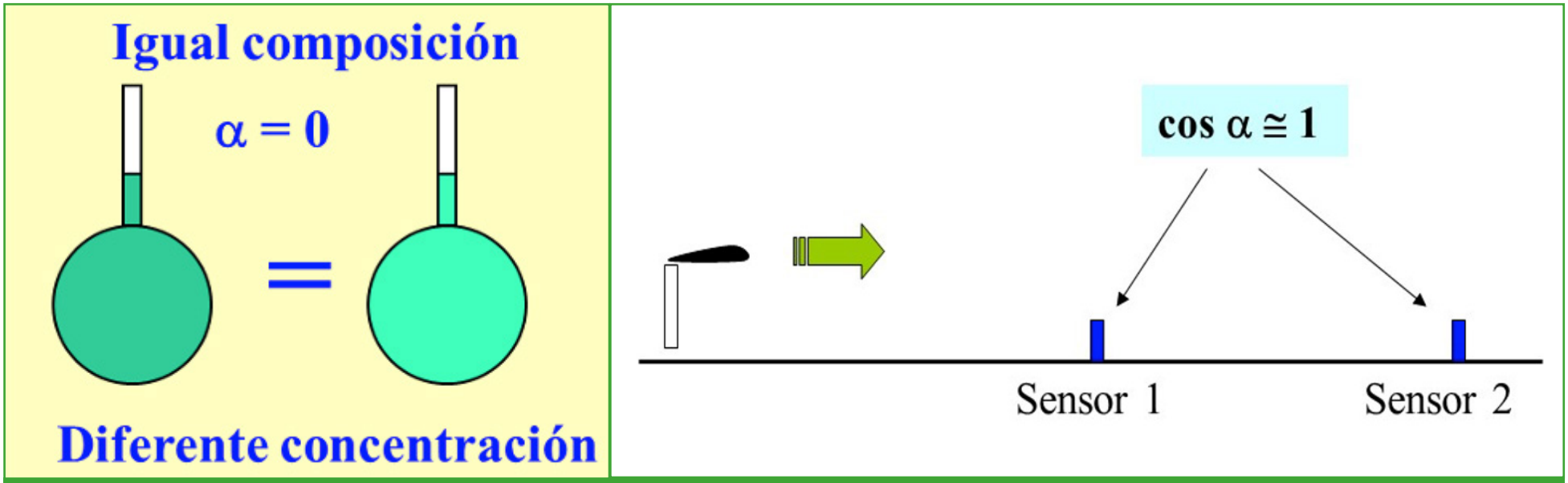


Figura 3.3

La figura 3.3 ilustra la medición de la calidad del aire con el concepto de distancia angular.

Suponga que en cierta área se instalan una serie de sensores distribuidos en un sector. Así como dos soluciones guardan distancia angular cero cuando difieren en su dilución, la disposición de fuentes y sensores, cuando dos de ellos están alineados con una fuente emisora y el viento sopla en la misma dirección fuente → sensores, hace que las mediciones del sensor 2 respecto del 1 también tengan una distancia angular próxima a cero debido a la dispersión de sustancias entre ambos. Por lo tanto, cuando esta condición se detecte, la dirección de los sensores delatará la dirección de una fuente emisora, aunque esta no esté a la vista.

Medida de Dis-similitud para Variables Binarias

Las variables binarias toman usualmente el valor de '1' para un atributo, j, presente y '0' para el atributo ausente.

Para dos objetos i e í' y un atributo cualquiera 'j':

$s_{ii'j} = 1 \quad si \quad x_{ij} = x_{i'j}$

$s_{ii'j} = 0 \quad si \quad x_{ij} \neq x_{i'j}$

El "matching coeficient" (coeficiente de similitud, CS) es la media de S **para los m atributos** entre los objetos i e i':

$$S_{ii'} = 1/m \sum_{j=1}^{m} s_{ii'j}$$

El **coeficiente de similitud de Jaccard** es algo diferente, considera que la presencia simultánea de un atributo significa similitud pero que **la ausencia simultanea no significa nada.**

$$s_{ii'j} = 1 \quad si \; x_{ij} = x_{i'j}$$
$$s_{ii'j} = 0 \quad si \quad x_{ij} \neq x_{i'j}$$
$$\text{se ignora si } x_{ij} = 0 \text{ y } x_{i'j} = 0$$

Sii'se calcula luego, como en el caso anterior. Este CS es llamado a veces **similitud de Tanimoto** y ha sido usado en química combinatoria para describir la similitud de compuestos, por ejemplo, sobre la base de los grupos funcionales que tengan en común.

La similitud de Tanimoto puede expresarse en forma matemáticamente más sencilla como el cociente entre la intersección y la unión de dos objetos A y B:

$$J(A,B) = \frac{A \cap B}{A \cup B}$$

La Distancia= Dis-similaridad de Tanimoto es **1-J(A,B).**

La *Hamming Distance* o Distancia Martillo

Como veremos, existen otros tipos de distancias para variables binarias, las Diferencias entre ellas, más allá de su cálculo algebraico, es el camino que se recorre para medir estas distancias.

La ***Hamming distance*** (distancia martillo) viene dada por:

$$d_{ii'j} = 1 \quad si \quad x_{ij} \neq x_{i'j}$$
$$d_{ii'j} = 0 \quad si \quad x_{ij} = x_{i'j}$$
$$D_{ii'} = \sum d_{ii'j}$$

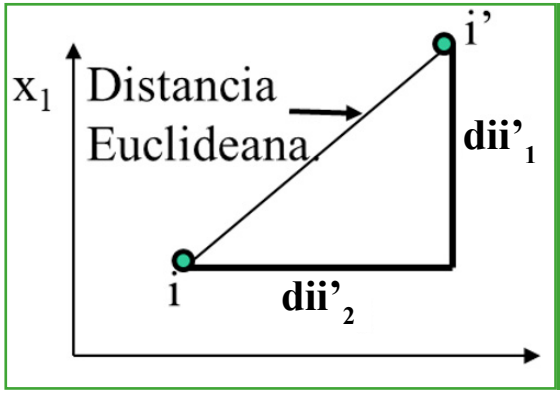


Figura 3.4

A veces se usa la *Hamming distance* como el equivalente de la distancia Euclidiana en el campo binario: $D_{ii'} = \sqrt{\sum d_{ii'j}}$. También existen versiones normalizadas de la *Hamming distance*, a saber:

$$D_{ii'}=1/m\sum d_{ii'j} \text{ y } D_{ii'} =(1/m)^{1/2}. \sum d_{ii'j}$$

Cabe aclarar que dado que Similitud=1-Distancia, la *hamming distance* es igual a los coeficientes de similitud. Algunos autores usan una u otra y debería aclararse su definición para su uso.

Medida de Dis-Similitud para Variables Enteras

En estos casos se utiliza la ***city block distance o Manhatan distance***. También se la conoce como *norma L1*, para una variable j viene expresada por:

$$d_{ii'j} = |x_{ij} - x_{i'j}| \qquad D_{ii'} = \sum d_{ii'}$$

La *Hamming distance* (——— en figura 3.4) es equivalente a la *city Block distance* en versión binaria. La distancia Manhatan es a veces utilizada para variables continuas.

Si bien las ecuaciones para *distancia Euclidiana* y para *Hamming o City block* son muy similares, obsérvese en la figura 3.4 que **los caminos** son muy distintos. Ambas ecuaciones pertenecen a la distancia Minkowsky:

$$D_{ii'} = \left(\sum_{j=1}^{m} \left| x_{i,j} - x_{i',j} \right|^{r} \right)^{1/r}$$

Para r=1 se obtiene la distancia Manhatan y para r=2 la distancia Euclidiana. En este contexto la distancia Euclidiana es referida como la norma L_2.

Medida de Dis-Similitud para Variables Mixtas

En muchos casos los objetos pueden estar descriptos por diferentes clases de variables.

Para eliminar el efecto de los diferentes rangos entre las variables se utiliza el escalado. Las distancias se calculan entonces como:

$$d_{ii'j} = |z_{ij} - z_{i'j}|, \qquad z_{ij} = (x_{ij} - x_{jmín})/r_j, \qquad 0 < z < 1 \quad , \quad r_j : \text{rango de la variable j}$$

$$D_{ii'} = \sqrt{1/m \sum_{j=1}^{m} (d_{ii'j})^2}$$

Hasta ahora, hemos aprendido a medir diferentes clases de distancias, el próximo paso es ver como utilizamos estas distancias para calcular los diferentes tipos de clusters.

Matriz de Similitud

La ***matriz de similitud*** (o dis-similitud) es una matriz cuadrada simétrica que reúne los valores de similitud (o dis-similitud, distancia) entre cada par de objetos. Esta matriz es la base para el cálculo de los dendrogramas jerárquicos. Haremos este procedimiento a través de un ejemplo: En un área bastante extendida se han recolectado muestras en 18 puntos, (m), y en cada uno se ha medido la concentración de varios ,(n), iones por cromatografía iónica, $\mathbf{M}_{\mathbf{n,m}}$. La tabla muestra los datos originales de los gráficos 3,1, 3.2 y Tabla 1. Pretendemos calcular la similitud entre los niveles de concentración de iones. Tenga en cuenta que en este problema cada ion es un vector (objeto) conformado por la concentración en cada punto (las variables).

	P1	P2	P3	P4	P5	P6	P7	P8	P9
Na	0.293	2.414	0.279	0.525	0.123	1.401	0.101	0.849	0.199
NH4	0.435	0.187	0.326	0.306	0.787	1.349	0.394	0.129	0.340
K	0.437	0.144	0.163	0.290	0.083	1.534	0.140	0.075	0.186
Mg	0.049	0.214	0.037	0.103	0.007	0.051	0.017	0.065	0.007
Ca	0.736	0.544	0.273	0.264	0.251	0.985	0.347	0.569	0.337
Cl	0.343	4.522	0.578	0.863	0.216	0.052	0.000	1.138	0.065
NO3	1.384	0.539	1.030	0.784	0.703	0.378	0.201	0.502	0.707
SO4	1.027	0.894	0.786	0.728	2.093	0.792	0.653	0.387	0.637

	P10	P11	P12	P13	P14	P15	P16	P17	P18
Na	0.490	0.077	0.114	0.492	1.346	0.256	0.151	0.240	0.729
NH4	0.621	0.609	0.601	0.821	0.132	0.173	0.294	0.161	0.134
K	0.286	0.133	0.175	0.389	0.113	0.072	1.045	0.477	0.486
Mg	0.036	0.092	0.015	0.052	0.212	0.034	0.466	0.072	0.125
Ca	0.672	0.402	0.604	0.432	0.219	0.251	1.109	0.616	0.839
Cl	0.562	0.253	0.171	0.423	1.930	0.135	0.029	0.225	0.824
NO3	1.772	1.367	0.458	2.464	0.951	0.418	0.386	0.432	0.756
SO4	0.862	0.746	0.885	1.365	1.068	1.038	1.243	1.073	1.750

La ***matriz de similitud*** (o dis-similitud) es una matriz cuadrada simétrica que reúne los valores de similitud (o dis-similitud, distancia) entre cada par de objetos. Esta matriz es la base para el cálculo de los dendrogramas jerárquicos. Haremos este procedimiento a través de un ejemplo: En un área bastante extendida se han recolectado muestras en 18 puntos, (j=1…18), y en cada uno se ha medido la concentración de 8 iones (i=1…8), vectores, por cromatografía iónica. O sea, una matriz M_{18x8}. Pretendemos calcular la similitud entre los niveles de concentración de iones. Tenga en cuenta que en este problema cada ion es un vector (objeto) conformado por la concentración en cada punto j (las variables).

Con la ecuación [1] calculamos la matriz de dis-similitud=distancia, $D^2_{ii'}$

$$D^2_{ii'} = (x_i - x_{i'})^T \; W \; (x_i - x_{i'})$$

$$8x8 \quad 8x18 \; 18x18 \; 18x8$$

	Na	NH4	K	Mg	Ca	Cl	NO3	SO4
Na	0							
NH4	2.9334	0						
K	2.8966	1.468	0					
Mg	3.1413	2.1053	1.8174	0				
Ca	2.6139	1.5547	**1.2248**	2.0795	0			
Cl	2.6484	5.1662	5.2594	4.9843	4.7226	0		
NO3	3.8338	2.9713	3.6979	4.0651	3.0688	5.1503	0	
SO4	3.6972	3.1463	3.5562	4.2136	2.7988	5.0417	2.7888	0

Algoritmos de Clustering

Así como hay varios tipos de distancias entre los objetos, también hay varios **tipos de algoritmos para formar los clusters jerárquicos**. Los más típicos son: el ***average linkage***, el ***single linkage*** y el ***complete linkage***. El mecanismo consiste en determinar primero los dos objetos más similares entre sí, $(D_{ii'})_{mín}$, supongamos que estos sean los objetos p y q (en el ejemplo serían Ca y K).

Ahora, en la matriz de similitud, reemplazamos los dos objetos por uno virtual calculado según alguno de los algoritmos siguientes:

Método	Algoritmo
Average linkage	$(D_{iq} + D_{ip})/2$
Single linkage	$Mín(D_{ip} , D_{iq})$
Complete linkage	$Máx(D_{ip} , D_{iq})$

Observe que para average linkage estamos calculado la distancia promedio entre un objeto i y los otros dos, p y q. En los otros dos casos elegimos simplemente la distancia mínima o máxima

En este ejemplo utilizaremos el *average linkage*, observe que, al eliminar 2 columnas y 2 filas, y reemplazarlas por un objeto virtual nuevo, A* la matriz quedará reducida en una unidad como muestran las tablas siguientes.

	Na	NH4	K	Mg	Ca	Cl	NO3	SO4
Na	0							
NH4	2.9334	0						
K	2.8966	1.468	0					
Mg	3.1413	2.1053	1.8174	0				
Ca	2.6139	1.5547	1.2248	2.0795	0			
Cl	2.6484	5.1662	5.2594	4.9843	4.7226	0		
NO3	3.8338	2.9713	3.6979	4.0651	3.0688	5.1503	0	
SO4	3.6972	3.1463	3.5562	4.2136	2.7988	5.0417	2.7888	0

	Na	NH4	A*	Mg	CL	NO3	SO4
Na	0						
NH4	2.9334	0					
A*	2.75525	1.51135	0				
Mg	3.1413	2.1053	1.94845	0			
CL	2.6484	5.1662	4.991	4.9843	0		
NO3	3.8338	2.9713	3.38335	4.0651	5.1503	0	
SO4	3.6972	3.1463	3.1775	4.2136	5.0417	2.7888	0

La distancia desde A* a los otros objetos se calcula promediando las filas y columnas eliminadas, lo que equivale a calcular las nuevas distancias $(D_{iq} + D_{ip})/2$

Las flechas indican los valores promediados. En el dendrograma los objetos Ca y K se unen al valor **1.2248** (vea el dendrograma final más abajo y los números en color rojo). Ahora el procedimiento se repite con la nueva matriz reducida, y así sucesivamente hasta armar el dendrograma completo. Las tablas siguientes muestran el sucesivo cálculo de las matrices.

	Na	B*	Mg	CL	NO3	SO4
Na	0					
B*	2.844325	0				
Mg	3.1413	2.026875	0			
CL	2.6484	5.0786	4.9843	0		
NO3	3.8338	3.177325	4.0651	5.1503	0	
SO4	3.6972	3.1619	4.2136	5.0417	2.7888	0

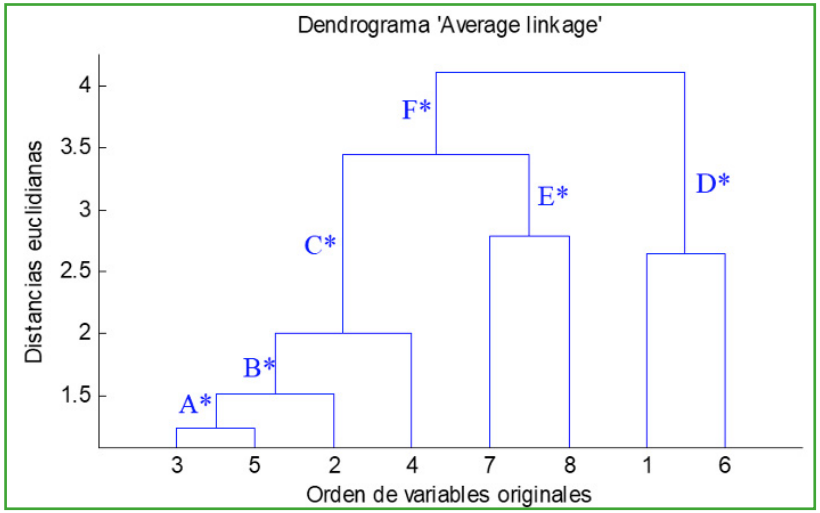
	Na	C*	CL	NO3	SO4
Na	0				
C*	2.9928125	0			
CL	2.6484	5.03145	0		
NO3	3.8338	3.6212125	5.1503	0	
SO4	3.6972	3.68775	5.0417	2.7888	0

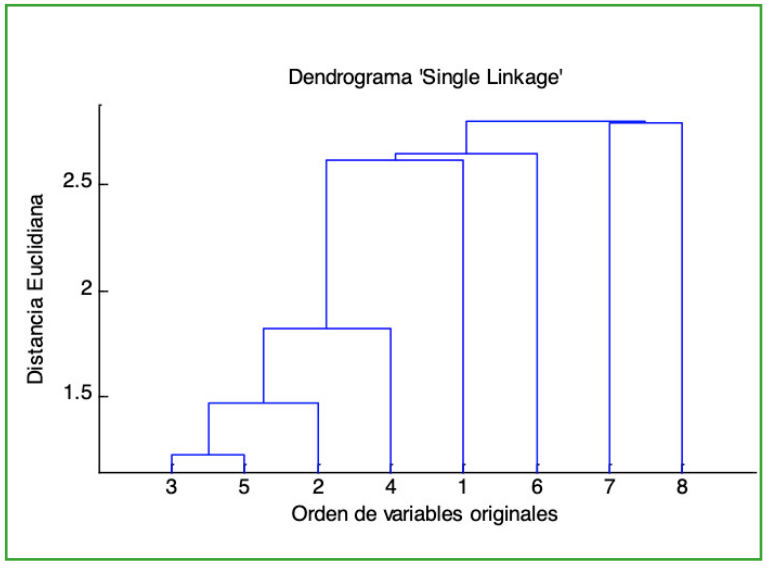
	D*	C*	NO3	SO4
D*	0			
C*	4.01213	0		
NO3	4.20177083	3.6212125	0	
SO4	4.14221667	3.68775	2.7888	0

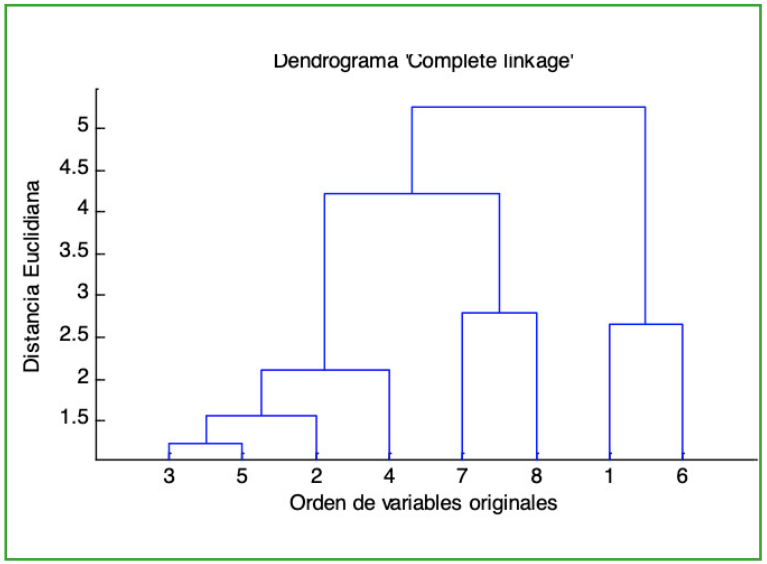
	D*	C*	E*
D*	0		
C*	4.01213	0	
E*	4.17199375	3.4655	0

	D*	F*
D*	0	
F*	4.09206188	0

Nueva variable virtual	Unión de	Orden original
A*	Ca-K	3-5
B*	NH4-A*	2-3*
C*	B*-Mg	2*-4
D*	Na-Cl	1-6
E*	NO3-SO4	7-8
F*	C*-E*	







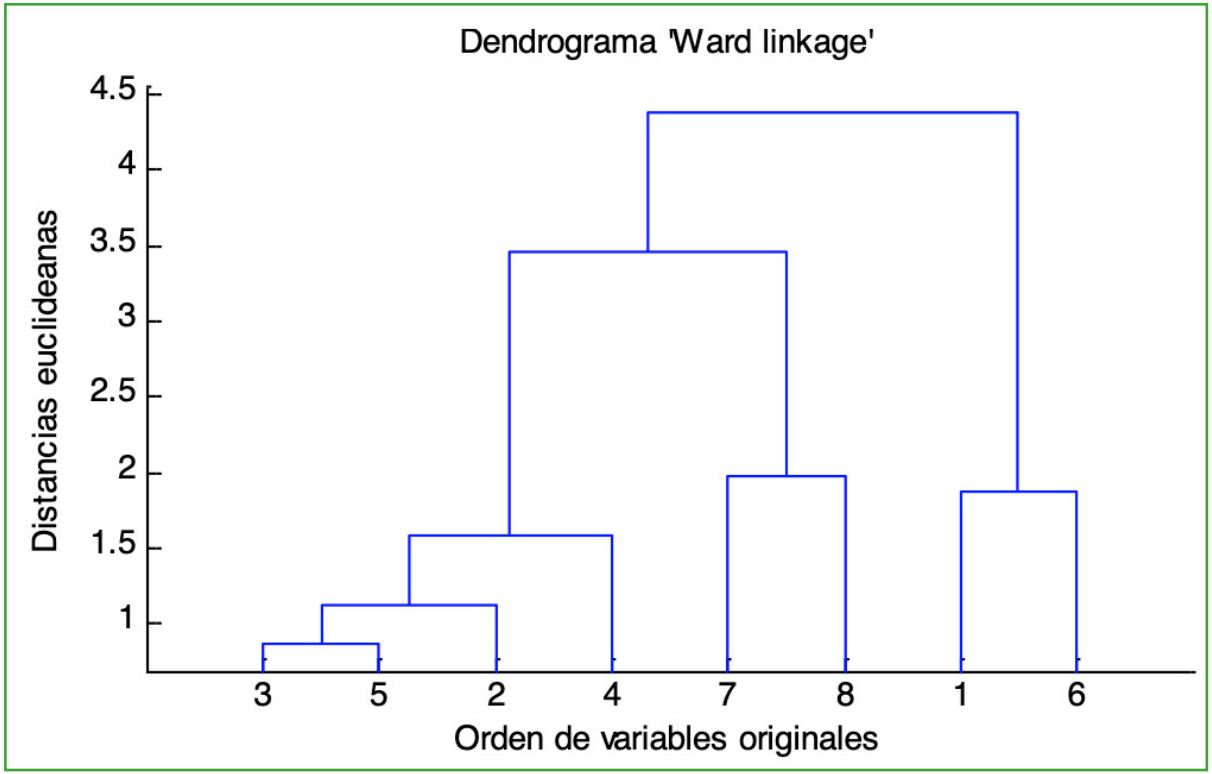
Observe las diferencias en el dendrograma cuando se calculan con single linkage y complete linkage. Los agrupamientos no cambian, pero sí sus distancias.

Conclusiones: hay una similitud muy grande entre las concentraciones de K+ y Ca^{2+} y una similitud algo menor de éstos con los cationes NH^{4+} y Mg^{2+}. El catión Na+ es disímil respecto de este grupo. Por otra parte, Na+ y Cl- son similares entre sí, aunque muy diferentes del grupo anterior (probablemente una fuente de NaCl). Lo mismo ocurre con el grupo constituido por NO3- y SO42- que deberían completar el balance de carga de los cationes, principalmente los que son distintos al Na+.

A primera vista los gráficos no parecen iguales, sin embargo, no se engañe con las uniones de los grupos, en realidad cualesquiera de estos tres algoritmos de cálculo clasifican los mismos clusters. Éstos sólo difieren en la altura (distancia) que alcanza cada grupo.

Cuando la serie de datos es muy grande, las distancias entre clusters pueden ser muy disímiles. En el *average linkege* se pueden utilizar pesos para salvar este inconveniente. Existen entonces dos variantes de este caso, **pesado** y **no pesado**.

Otro método que da buenos resultados es el de Ward. Se basa en un criterio de heterogeneidad, ésta es definida como la suma de distancias cuadráticas entre cada objeto de un cluster y su centroide. Los elementos o clusters son unidos con la condición de que la suma de heterogeneidades de todos los clusters debería incrementarse tan poco como sea posible. Puede verse sin embargo, a pesar de los diferentes criterios, que los dendrogramas de *Ward likage* no difieren mucho, en este caso, de los del *average linkage.*



Métodos no Jerárquicos

A diferencia de formar clusters agrupando objetos en forma jerárquica, ya que no todos los problemas guardan tal relación, ahora nos proponemos formar un cierto número de clusters directamente a través de otra estrategia. Se mostrará el procedimiento con un grupo sencillo de datos en solo 2 dimensiones, para poder visualizarlo. A este método se lo denomina ***MacQueen's k-means method o k means*** (Ref. 2).

Primero elegiremos dos objetos que consideraremos que pertenecen a distintos clusters, Por ejemplo, E y H en la figura siguiente, estos se llaman *seeds* (semillas). Si bien a simple vista se ve que estos objetos estarían mal elegidos, tenga en cuenta que en el espacio multidimensional no se pueden observar, por lo tanto, la elección es intencionalmente mala. Sin embargo, veremos que el cálculo progresa bien, aunque se hayan elegido los peores puntos.

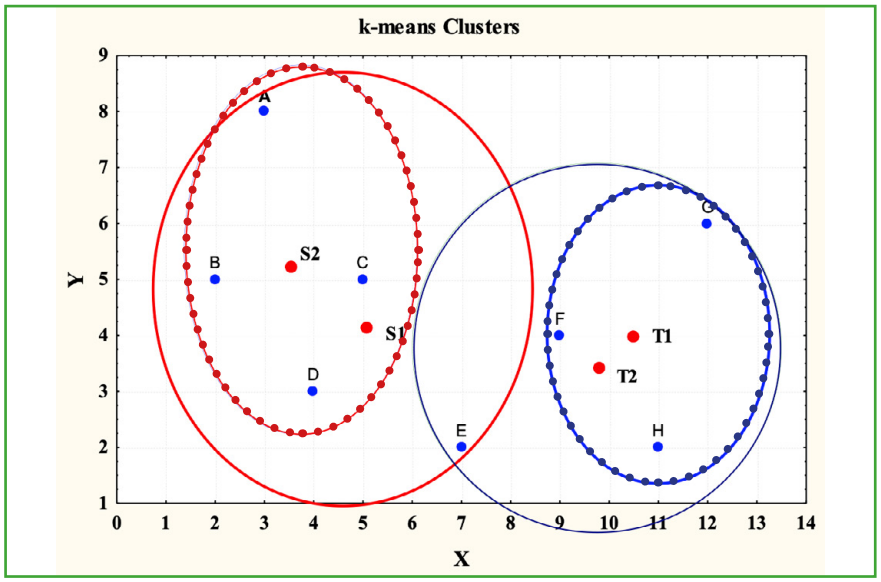


Tabla 3	X	Y
A	3	8
B	2	5
C	5	5
D	4	3
E	7	2
F	9	4
G	12	6
H	11	2

Comenzamos calculando los centroides, S1 y T1, de los clusters formados con los puntos más cercanos respecto a E y H. Los 2 grupos son: A-B-C-D-E y F-G-H.

Ahora volvemos a formar **nuevos clusters** con los objetos más cercanos a estos centroides, S1 y T1 (sin incluirlos), estos son A-B-C-D y E-F-G-H. Y volvemos a calcular los **nuevos centroides** S2, T2.

Si intentamos repetir el procedimiento para ver si los cluster se modifican, descubriremos que éstos quedan igual. Esto significa el final del cálculo.

En lugar de utilizar *seeds* al azar como posiciones iniciales alrededor de los cuales se forman los cluster, se pueden utilizar ***centrotipos***. Estos son objetos elegidos como representativos de los clusters que se quieren formar. El criterio para armar el dendrograma será ahora que la suma de distancias desde los objetos al centrotipo más cercano debe ser mínima.

Los dos métodos descriptos, jerárquicos y no jerárquicos, se llaman “partitioning-optimization” y consisten en partir una serie de objetos en subseries, ya sea alrededor de un objeto de una serie a ser agrupada (centrotipo) o alrededor de un objeto virtual (centroide). En general se **maximiza la distancia entre clusters** y se **minimiza la distancia intra cluster**. Se puede entonces particionar el conjunto de objetos en T = B + W, T es la matriz relacionada con la variancia-covariancia de los objetos, B es la misma matriz para los centroides (relaciona la variación ‘entre grupos’) y W es una matriz similar para los componentes ‘dentro de los clusters’. También es equivalente escribir la ecuación con las trazas ‘tr’ de las matrices: tr(T) = tr(B) + tr(W). Como tr(T) es constante, minimizar tr(W) equivale a maximizar tr(B). tr(B) es la suma de cuadrados de distancias Euclidianas entre centroides.

Antes de iniciar cualquier intento de clasificación se debe razonar cuidadosamente acerca del tipo de relaciones que puede haber entre los objetos. Una clasificación jerárquica puede dar resultados muy diferentes a una no-jerárquica. La misma preocupación debe tenerse para elegir el tipo de distancia más representativa entre los objetos. Y también para adecuar las variables (pretratamiento de datos) según las características del problema que estamos estudiando.

Referencias

1. D.L. Massart; B.G.M. Vandeginste; L.M.C. Buydens; S. De Jong; P.J. Lewi and J. Smeyers-Verbeke. Hanbook of Chemometrics and Qualimetrics. Part B. Elsevier, Amsterdam 1997

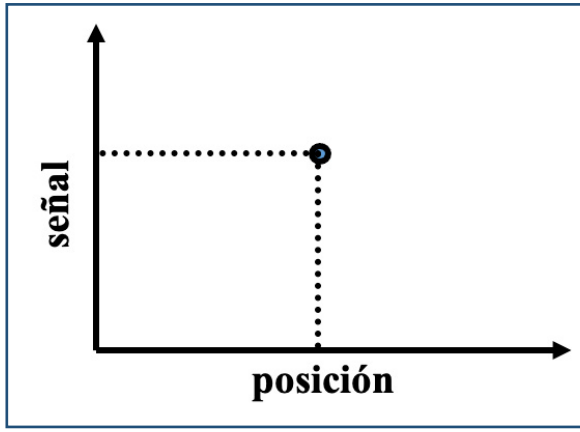
2. J. MacQueen, Some methods for classification and analysis of multivariate observations. In: L. Le Cam and J. Neyman (eds.). Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, CA, 1967, pp. 281-297.

CAPITULO 4

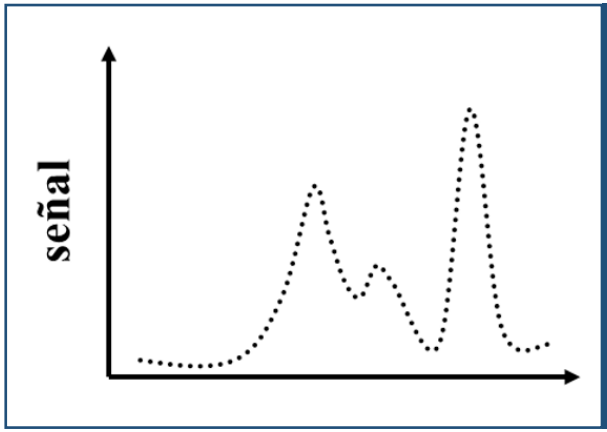
Calibración Multivariada

Generaciones de Instrumentos de Medición

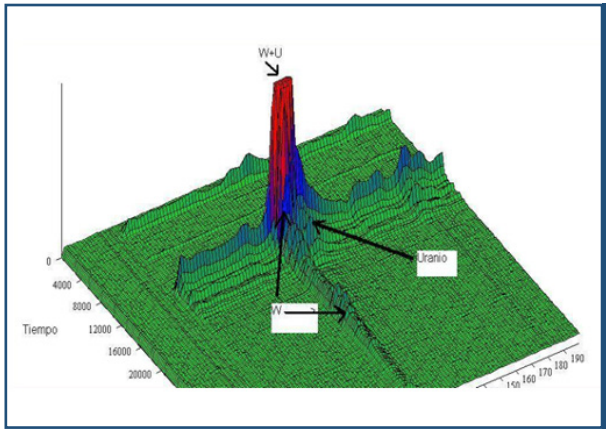
En el curso del tiempo los instrumentos de medición han tenido sucesivas etapas de evolución. Últimamente, el desarrollo de la electrónica ha permitido hacer avances muy importantes. Señalaremos ahora las tres generaciones que aún persisten en cuanto al modo de obtener información.



Instrumentos de orden cero: pHímetros, termómetros, balanzas, densímetros, etc. (un solo dato).



Instrumentos de orden 1: Espectrofotómetros, voltamperímetros, polarógrafos, espectroscopías, etc. (que entregan un vector de datos).



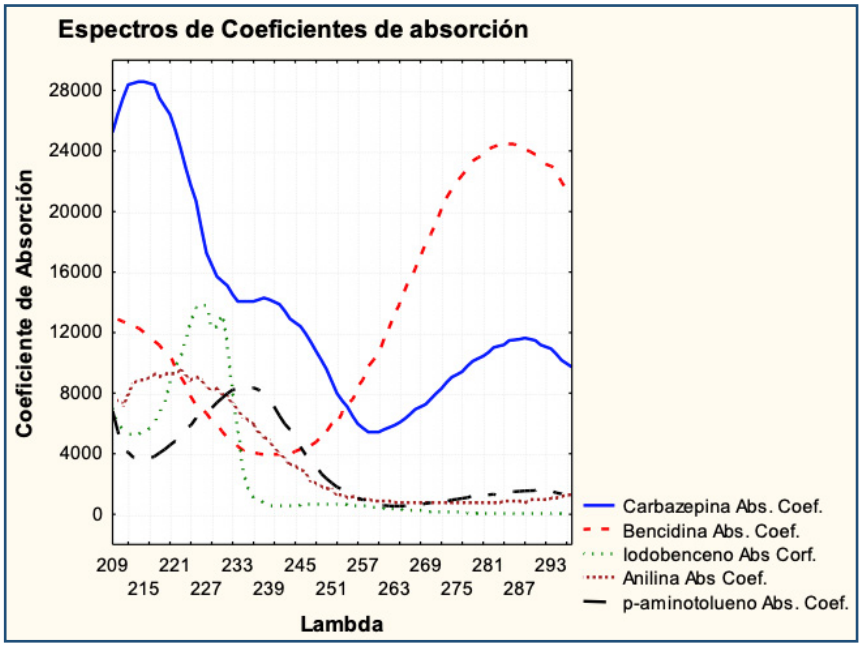
Instrumentos de orden 2: Fluorómetros de excitación-emisión, instrumentos en cascada (ICP-MS, LC-MS, LC-Diode array, etc. Que entregan matrices de datos en 3 dimensiones.

Cuando la información se hacía cada vez más abundante, se necesitaban técnicas matemáticas capaces de analizarlas automáticamente a través de desarrollos de software. Hoy en día existen métodos de análisis para estudiar resultados de mediciones de más de tres dimensiones, más adelante veremos cómo se resuelven estos problemas.

Para analizar estos métodos, comenzaremos por el caso más sencillo que es la calibración univariante. Emplearemos en este capítulo operaciones de álgebra de matrices.

Introducción

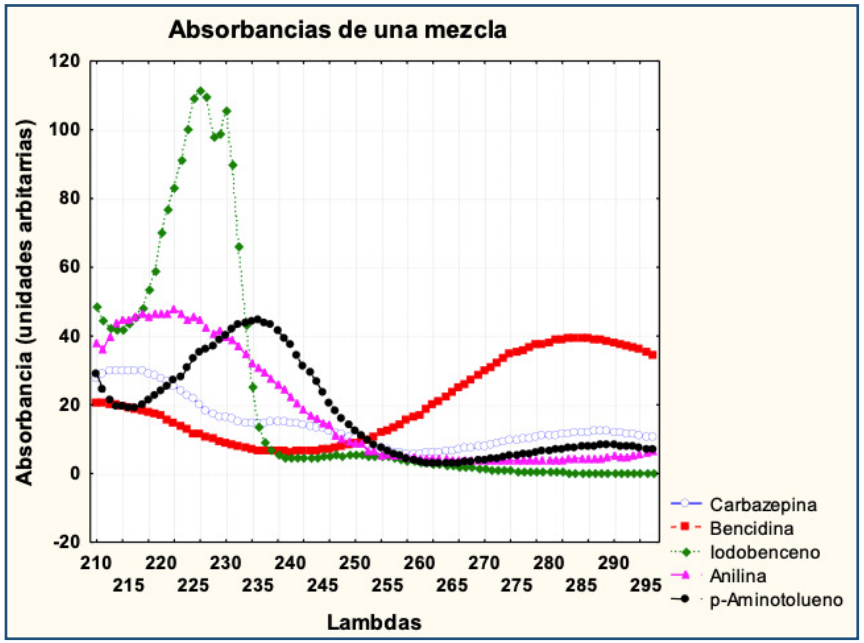
Utilizaremos datos obtenidos de varias fuentes de datos, que, con el objeto de analizarlos en el rango del espectro ultravioleta (UV) y con igual definición, han sido recortados e interpolados o completados a partir de los coeficientes de absorción de cada compuesto.



El gráfico muestra los coeficientes de absorción (ε) en función de la longitud de onda (en adelante Lambda, o λ) en la zona UV del espectro de las 5 especies seleccionadas, a saber: Carbazepina, Bencidina, Iodobenceno, Anilina y p-Aminotolueno. A partir de esta información veremos cómo analizar uno o varios simultáneamente de estos compuestos cuando están presentes en conjunto en solución en un solvente apropiado.

Análisis univariante (a partir de una única Longitud de onda)

En el gráfico vemos que, en una solución de varios compuestos, la absorbancia total será la suma de las absorciones de todos los compuestos. Como se ve, la interferencia entre ellos afectará la posibilidad de determinar su concentración individualmente en alguna única lambda (o sea con una única variable). Pero puede haber alguna zona donde un compuesto de interés tenga poca interferencia del resto. Además, téngase en cuenta que las absorbancias dependen de la concentración de las especies.



Por ejemplo, en estos dos gráficos vemos que, en la zona de 285 nanómetros, si la relación de concentraciones entre bencidina y carbazepina es grande, sería posible determinar el mayor componente en presencia del resto. En la calibración univariante tenemos la opción de dos métodos de cálculo: **El método directo (o clásico)** y **el método inverso**. (Ref. 1).

En ambos métodos necesitamos la misma información: un vector **c** conteniendo las concentraciones de una serie de soluciones, mezcla, de los compuestos y otro vector **x** conteniendo sus respectivas absorbancias **a una simple longitud de onda** (o como se suele decir, **con un único sensor**). Ambos vectores tienen una longitud I que es igual al número de experimentos (muestras), en este caso I=16. En este ejemplo calibraremos la bencidina a λ=283 nanómetros (nm) **en presencia del resto de los componentes**. Los datos completos están en el archivo "Tabla de datos para la teoría (en práctica 4)".

1- El método directo o clásico

c	x
0.456	0.161
0.456	0.176
0.152	0.102
....	....
....	....
....	....
0.76	0.212
0.304	0.142

Matemáticamente, la relación entre x y c puede expresarse en forma vectorial en el método directo como:

$$\mathbf{x} \approx \mathbf{c}.s. \qquad \textbf{[1].}$$

Por qué utilizar ≈ en lugar de = lo explicaremos al tratar los errores de cálculo. El **escalar,** s, está determinado por los experimentos y puede calcularse mediante álgebra lineal:

$$\mathbf{c}^t.\mathbf{x} \approx \mathbf{c}^t.\mathbf{c}.s \rightarrow (\mathbf{c}^t.\mathbf{c})^{-1}.\mathbf{c}^t.\mathbf{x} \approx (\mathbf{c}^t.\mathbf{c})^{-1}.(\mathbf{c}^t.\mathbf{c}).s$$

$$\text{Obviando} \approx \text{es, } s = (\mathbf{c}^t.\mathbf{c})^{-1}.\mathbf{c}^t.\mathbf{x} \qquad \textbf{[2]}$$

Conociendo s podemos estimar **x** como:

$$\hat{x}=c.s \qquad \textbf{[3].}$$

El copete circunflejo indica “valor estimado”. Las muestras desconocidas se estiman como:

$$\hat{c}=s^{-1}.\hat{x} \qquad \textbf{[4]}$$

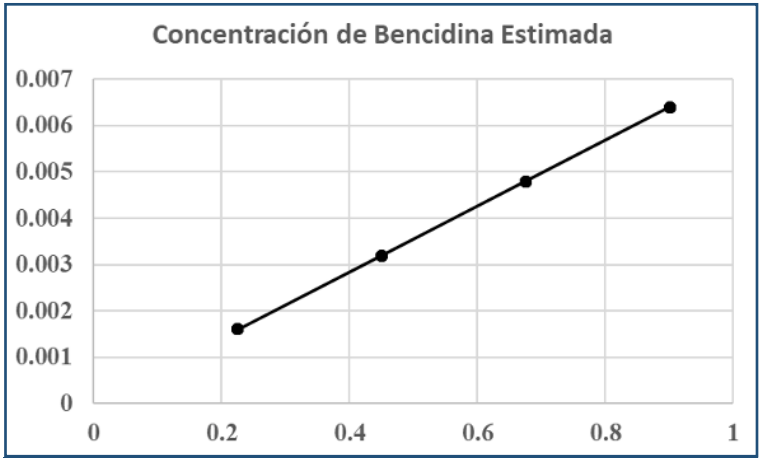
Absorbancia Λ=283	Conc. Bencidina
0.231771	0.0016
0.445997	0.0032
0.660223	0.0048
0.874449	0.0064
0.296143	0.0016
0.488673	0.0032
0.6508605	0.0048
0.8433905	0.0064
0.3008605	0.0016
0.478219	0.0032
0.694628	0.0048
0.8719865	0.0064
0.3121085	0.0016
0.5089925	0.0032
0.697168	0.0048
0.894052	0.0064

Absorbancia Estimada	Conc. Estimada
0.2252	0.0016
0.4505	0.0032
0.6757	0.0048
0.901	0.0064
0.2252	0.0016
0.4505	0.0032
0.6757	0.0048
0.901	0.0064
0.2252	0.0016
0.4505	0.0032
0.6757	0.0048
0.901	0.0064
0.2252	0.0016
0.4505	0.0032
0.6757	0.0048
0.901	0.0064

Ejemplo desarrollado: Las tablas de la izquierda muestran las columnas de la concentración de calibración de bencidina y la absorbancia a Lambda=283 nm. El valor calculado de S es S= **140.775**.

Se observa que la estimación de Lambda tiene cierto error, sin embargo, la estimación de la concentración es perfecta. El cálculo de errores lo consideraremos en general al final del punto sobre calibración univariante

En este caso particular no es necesaria la corrección por ordenada al origen ya que ésta pasa por cero (ver figura).



Absorbancia en abscisas, concentración en ordenadas.

De todos modos, se explica cómo proceder para estimarla. En primer lugar se debe agregar una columna de 1 (unos) a la izquierda de la columna de absorbancias, ahora X será de nx2. El cálculo de s es idéntico al anterior (ecuación 2), pero ahora tendrá 2 términos: el 1° es la ordenada al origen y el 2° es el valor de s. La única diferencia de cálculo es que no podemos estimar la concentración con la ecuación 4, porque al ser s, un vector, no tiene inversa. Recurriendo al álgebra lineal tenemos que despejar C desde la ecuación 3:

$$\hat{x}=c.s\ , \qquad \hat{x}.s^t=c.s.s^t \rightarrow c=\hat{x}.s^t\ (s.s^t)^{-1}$$

2- El método inverso

Para los químicos analíticos es más natural plantear un modelo como

$$\mathbf{c \approx x.s\ ,} \qquad \textbf{[5]}$$

que es el del método inverso en lugar del de la ecuación [1]. Las operaciones matemáticas para su solución son similares a las del método directo, pero hay algunas diferencias a considerar. Desde un principio introduciremos aquí el cálculo con la ordenada al origen:

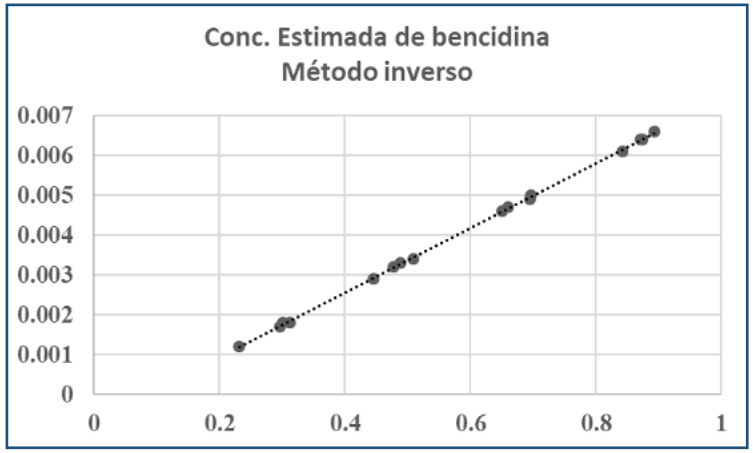
$\mathbf{c \approx b_0 + b_1.x}$ que responde a la ecuación [5] porque ya sabemos que s puede ser un vector.

En este caso **b** es un vector columna conteniendo a b_0 y b_1. **X** es ahora una matriz de dos columnas, la primera es una columna de 1s (unos) para calcular el término independiente y la segunda es la columna de las absorbancias de estándares. Aplicando la aritmética de la calibración inversa, obtenemos:

$$\mathbf{b}=(\mathbf{X}^t\mathbf{X})^{-1}.\mathbf{X}^t.\mathbf{c} \text{ y luego la estimación de c, } \hat{c}=X.b.$$

Recalculemos el ejemplo anterior con este método. b_0= 0.0007 y b_1=0.0081

$\hat{C}$ Bencidina	Continuación
0.0012	0.0018
0.0029	0.0032
0.0047	0.0049
0.0064	0.0064
0.0017	0.0018
0.0033	0.0034
0.0046	0.005
0.0061	0.0066

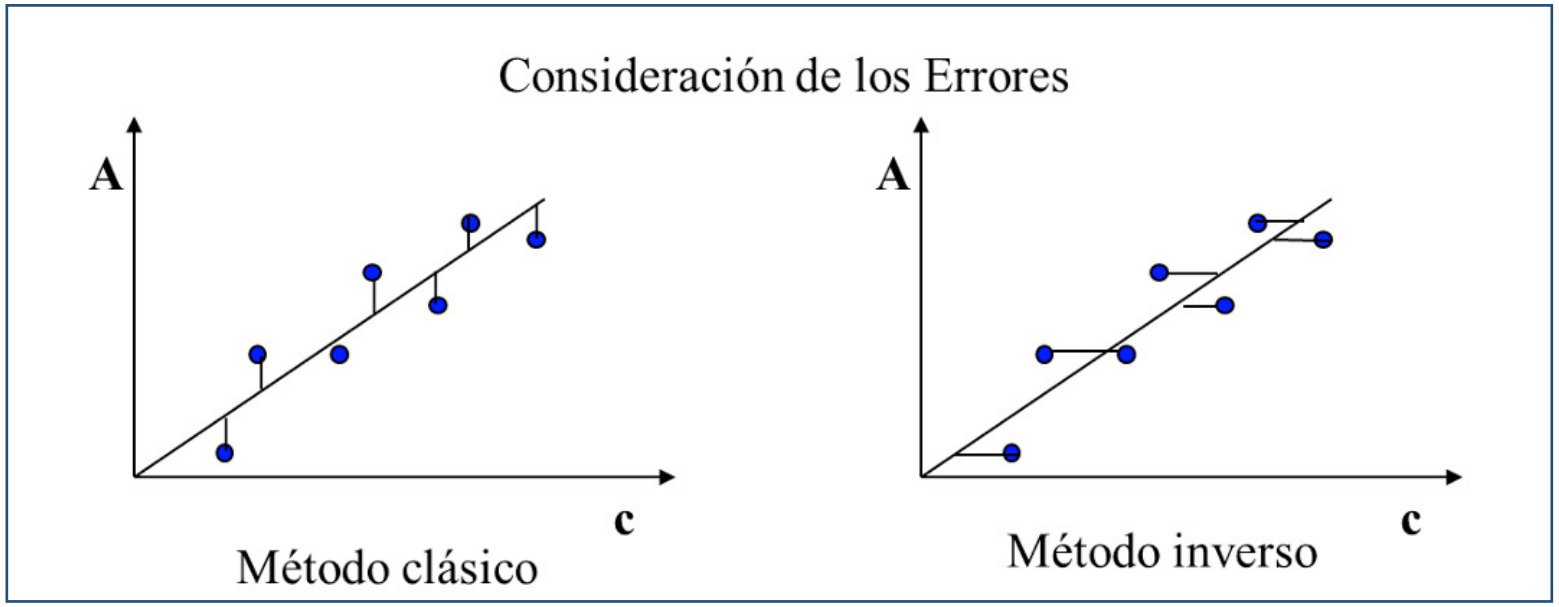


Como se ve, en este caso $\hat{c}$ es menos precisa que en el caso anterior. La diferencia puede deberse a la inversión de $X^t.X$, en la cual X es ahora una matriz más grande.

Consideración de los Errores

Hemos visto que los modelos anteriores se presentaron como **X≈C.s** y **C≈X.s** para los métodos directo (o clásico) e inverso, respectivamente. Esto se debe a que, estrictamente, las estimaciones adolecen de errores y por lo tanto la escritura más correcta es **X = C.s +E** y **C=X.s +E**, donde E es un término de **error de calibración** que completa la igualdad.

Según la definición de cada uno de los modelos, los errores son atribuidos a distintas fuentes. En el método directo, el término, E, complementario a **x**, tiene necesariamente la magnitud de absorbancia, mientras que, en el método inverso, E, es necesariamente un término de concentraciones (Ref.1).



La consideración de estos errores en ambos métodos permite entonces comparar los errores instrumentales (medición de la absorbancia) con los de las operaciones de laboratorio (preparación de estándares).

El cálculo del error absoluto, E, es la raíz cuadrada del error medio cuadrático entre los valores reales y estimados de las variables, o sea:

$$\boldsymbol{E = \sqrt{\sum_{i=1}^{I} \frac{(\hat{v}_i - v_i)^2}{g}}} \qquad \textbf{[6]}$$

I, es el número de muestras (concentraciones o Lambdas).

vi, es la variable considerada (concentración o Lambda).

g, son los grados de libertad.

Los grados de libertad son **g=número de muestras - número de parámetros estimados**. Sin embargo, cabe hacer una aclaración, algunos autores utilizan éste g y otros utilizan g=I en la ecuación del error, dependiendo de si están utilizando la teoría frecuentista (estadística clásica) o *máximum likelihood*, la cual se explicará en otro capítulo.

El error relativo % es: $\boldsymbol{E\% = \frac{E}{\bar{v}}.100 \quad \bar{v} = medias\ de\ C\ o\ X}$ **[7]**

En los ejemplos anteriores, y según la teoría frecuentista, para el método directo E=E%=0 para las concentraciones y E=0.04517 para las absorbancias, cuyo E% es 0.0452/0.5780*100=7.8201. Para las concentraciones, por el método inverso E=0.00020 y E%=5.083.

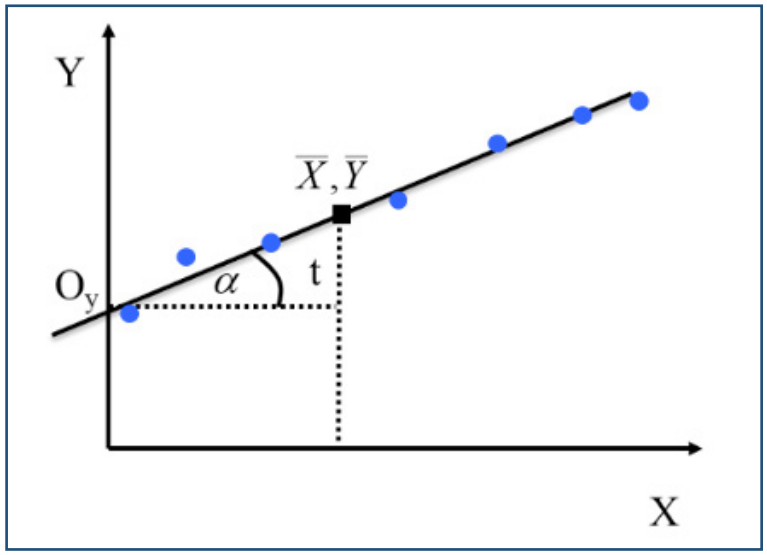
La segunda parte en la consideración de errores es **la etapa del error de predicción.**

En la etapa de calibración las disposiciones de medida y cálculo se hacen con la intención de tener condiciones óptimas para el método. Por ejemplo, las concentraciones se eligen de acuerdo a una distribución y orden determinados e incluso con repeticiones (observar la tabla de valores en el ejemplo del método directo). Pero en una aplicación concreta las concentraciones serían al azar. Para calcular el error de predicción, las concentraciones se eligen dentro del mismo rango de la calibración, pero al azar. El número de ensayos conviene que sea el mismo o uno lo más cercano posible. El error de calibración en los casos reales suele ser menor al error de predicción y este último debe ser tomado como **el error del método**. En nuestro ejemplo se han elegido otros 16 ensayos con las concentraciones determinadas al azar y sus correspondientes absorbancias de las mezclas de los 5 componentes. El error calibración para Bencidina es, $\mathbf{E_c\%=5.08}$, mientras que el error de predicción es, $\mathbf{E_p\%=5.66}$.

Un método más general del cálculo de la ordenada al origen

En los temas siguientes trataremos las técnicas del cálculo multivariante. Entonces, cuando hay que tratar con varias longitudes de onda y concentraciones a la vez, el agregado de un término independiente a cada una de ellas se vuelve menos práctico. Lo que se suele hacer es centrar los datos (ver Capítulo 1, Anexo), tanto para las absorbancias como para las concentraciones. La recta de calibración **calculada sin ordenada al origen** pivoteará entonces sobre el valor medio de las variables. Todo lo que hay que hacer al final del cálculo es sumar el valor medio de las variables para obtener el resultado definitivo. Téngase en cuenta que en calibración multivariada no siempre se suelen representar las rectas de calibración, de modo que hay que acostumbrarse a interpretar numéricamente los resultados.

De todos modos, si se quiere representar la recta de calibración con su ordenada al origen y pendiente para un caso univariante, podemos hacerlo con la figura siguiente:



La ordenada al origen $O_{y,}$ se calcula a partir de los valores medios de X e Y, sabiendo que **Y=X.**s (ecuación [5]), donde s representa ahora a la pendiente. (s=b_0 en el método inverso)

$$\boldsymbol{O_y = \overline{Y} - t = \overline{Y} - \overline{X}.tg(\alpha) = \overline{Y} - \overline{X}.b}$$

Regresión Lineal Múltiple: La Ventaja de la Multidetección

En un espectro digitalizado, cada longitud de onda, λ, puede ser considerada un sensor. ¿Por qué usar entonces un solo sensor, o porqué usar múltiples sensores? Hay varias razones para utilizar varios sensores, una de ellas es que, si la solución contiene **n** componentes, se puede resolver la mezcla contando con al menos **n** sensores (siempre y cuando que las absorbancias de los componentes no estén correlacionadas). Otra razón es que cada sensor contiene algo de información, la información resultante de utilizar **'n'** sensores será el promedio de todos ellos y usualmente es mejor que utilizar uno solo. (Ref.1, 3-5)

Como ejemplo veremos la determinación simultanea de 3 de los 5 componentes indicados en la introducción. Elegiremos Iodobenceno, p-Aminotolueno y Bencidina. Aprovechando el máximo de los 3 componentes, utilizaremos 3 sensores de

esta zona: **λ**=226, λ=235 y λ=290 y la tabla de 16 mezclas ya utilizada para calibración, pero sólo con las concentraciones de los 3 sensores. Tenemos una matriz de $\mathbf{X}_{16,3}$ para las absorbancias y otra $\mathbf{C}_{16,3}$ para las concentraciones.

Planteamos, por ejemplo, el método inverso $\mathbf{C} \approx \mathbf{X.b}$, de donde $\mathbf{b}=(\mathbf{X}^t.\mathbf{X})^{-1}.\mathbf{X}^t.\mathbf{C}$ y obtendremos **b** de 3x3 (3 λ's para cada uno de los 3 componentes).

Valores de b por 1.10^{-4}			
Iodobenceno	p-Aminotolueno	Bencidina	λ
0.7816	-0.1839	0.01	226
-0.8449	0.8268	-0.0994	235
-0.0895	-0.0917	0.4329	290

Errores			
	Iodobenceno	p-amino-tolueno	Bencidina
E	0.000602	0.001285	0.000158
Ec%	12.055	28.559	3.95284708

Este método puede utilizarse suponiendo que **los analitos principales son todos conocidos** y funciona bien si esto es cierto. **Aplicar el método a soluciones donde hay interferentes desconocidos puede conducir a serios errores de estimación.**

Observe que, aunque el error de p-Aminotolueno es alto, el de Iodobenceno puede ser útil en una mezcla semejante y el de Bencidina es bueno. Esto se ajusta a lo señalado en el párrafo anterior ya que anilina y carbazepina están interfiriendo fuertemente el espectro y no están incluidas en el cálculo.

Solución más avanzada utilizando todos los sensores

Solución por el método directo o clásico

Podemos extender el método anterior para determinar los cinco compuestos de la figura 1 utilizando las 87 lambdas o sea, sensores. No hay impedimento para ello siempre y cuando se tengan en cuenta ciertas consideraciones. Por ejemplo, a menos que existan correlaciones, el **número de componentes de la solución debe ser menor al número de experiencias o espectros y también menor al número de sensores.**

La matriz de datos X de los 5 componentes a 87 λ's está relacionada con la matriz de concentraciones C y la matriz de espectros S de la siguiente manera:

$$\mathbf{X}_{(16x87)}=\mathbf{C}_{(16x5)}.\mathbf{S}_{(5x87),} \qquad [8]$$

de donde: $$S=(C^t.C)^{-1}.C^t\,X \quad y \quad C=X.S^t.(S.S^t)^{-1}$$

Téngase en cuenta que en este procedimiento las matrices cuadratizadas **($C^t.C$)** y **($\hat{S}.\hat{S}^t$)** de tamaño 5x5 tienen inversas, previendo que los experimentos hayan tenido un diseño adecuado y que las absorbancias de los compuestos no estén correlacionas

Tabla de resultados

Compuesto	Ec%	Ep%
Carbazepina	10.05	4.88
Bencidina	0	1.42
Iodobenceno	0	1.03
Anilina	0	0.70
p-aminotolueno	1.34	1.12

Observe que los errores no son iguales para todos los compuestos, esto depende de cuán grande sea el solapamiento espectral de un compuesto particular respecto de todo el resto. De aquí se desprende que en lugar de tomar los espectros con λ's igualmente espaciadas, se pueden practicar algunas estrategias para mejorar este aspecto. El error de predicción es mayor que el de calibración y es el verdadero error del método.

Es necesario remarcar que la predicción será aceptable **si todos los compuestos significativos han sido incluidos en la calibración**. Si en lugar de los 5 compuestos de esta mezcla, hubiéramos tomado sólo 3 o 4, las predicciones serían mucho más pobres.

Con la ecuación 8 se pueden reconstruir las absorbancias, éstas suelen tener errores más bajos que las concentraciones.

Solución por el método inverso

En este caso, la relación básica es $\mathbf{C}_{16x15}=\mathbf{X}_{16x87}.\mathbf{B}_{87x5}$

Y el cálculo de **B** es: $\mathbf{B}=(\mathbf{X}^t.\mathbf{X})^{-1}.\mathbf{X}^t.\mathbf{C}$

El cálculo de $(\mathbf{X}^t.\mathbf{X})^{-1}$ introduce varios problemas: uno es el de la dimensión de esta matriz 87x87 en lugar de 5x5 para $(\hat{\mathbf{S}}.\hat{\mathbf{S}}^t)$. Como $(\mathbf{X}^t.\mathbf{X})$ tiene sólo 5 grados de libertad, ésta puede no tener inversa debido a las muy probables correlaciones en los espectros. El método es practicable si:

1) el número de experimentos y sensores es al menos igual al número de componentes de la solución (lo que no es un problema)
2) el número de experimentos es al menos igual al número de sensores (lo que es un serio problema).

Independientemente del método de solución, la desventaja de los métodos basados en regresión lineal múltiple es que **todos los compuestos significativos deben ser conocidos** porque de lo contrario la estimación será muy poco precisa.

Análisis por Componentes Principales (PCA)

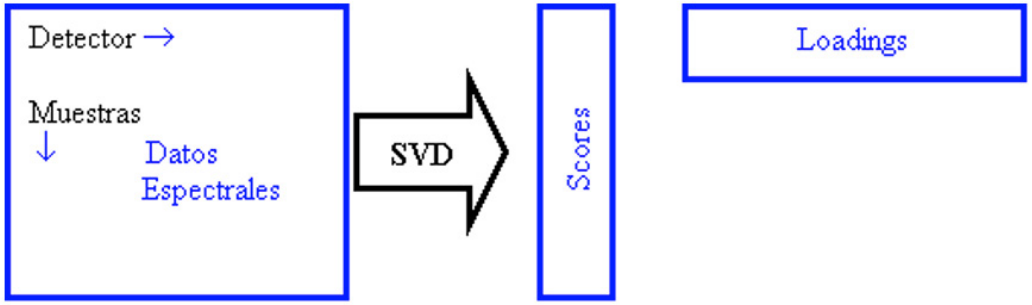
Un teorema del álgebra lineal demuestra que cualquier **matriz rectangular** (pero no necesariamente cuadrada como exige CP), $\mathbf{X}_{nxm}$ puede descomponerse como un producto de 3 matrices a través de *singular value decomposition* (**SVD**) (Ref. 2):

$$\mathbf{X}_{nxm}=\mathbf{U}_{nxn}.\mathbf{\Sigma}_{nxm}.\mathbf{V}^t_{mxm} \qquad [9]$$

$\mathbf{\Sigma}$ es una matriz diagonal relacionada con los *eigenvalues* de la matriz **X,** para una matriz cuadrada $\mathbf{\Sigma}=(\Lambda)^{1/2,}$ y en ese caso la SVD coincide con la *eigenvalue decomposition* (**EVD**). Las matrices **U** y **V** son ortonormales.

Usualmente se denomina a $\mathbf{T}_{nxm}=\mathbf{U}_{nxn}.\mathbf{\Sigma}_{nxm}$ *scores matrix* y a $\mathbf{P}_{mxm}=\mathbf{V}^t$ *loadings matrix*. De modo que: $\mathbf{X}=\mathbf{T}.\mathbf{P}.$ [10]

La figura muestra esquemáticamente esta descomposición.



Propiedades de **T**:

1- El número de filas es usualmente igual al número de muestras de calibración.
2- El número de columnas es adaptable al número de factores significativos, f, de los datos. **El valor de f se estima desde la matriz Σ.** Idealmente, este número iguala al número de compuestos en **X**, pero puede ser menor debido a información no imprescindible o disminuirse hasta 1.

Propiedades de **P:**

1- El número de columnas es igual al número de detectores, λ's.
2- El número de filas es adaptable al número de factores significativos, f, de los datos.

Regresión por Componentes Principales (PCR)

Cierto tipo de problemas analíticos implican determinar uno o más compuestos en una matriz que contiene otros compuestos desconocidos. En las técnicas de regresión vistas en los puntos anteriores, era necesario conocer **todas** las concentraciones de la muestra para resolver el problema. **PCR** (Principal Components Regresion) tiene la gran ventaja de que **sólo es necesario conocer las concentraciones de los componentes significativos de la muestra** y no la de aquellos que no son de interés (aunque cuantos más componentes conozcamos, mejor será la calibración).

La clave del cálculo está en la matriz **Σ,** esta matriz diagonal está ordenada de mayor a menor y de ella se seleccionará un número de elementos (filas y columnas sucesivas de la diagonal) suficientemente representativo de todo el conjunto de datos. El % de varianza captada es entonces: (Suma de elementos diagonales seleccionados) x100 /traza(**Σ**). Usualmente, unas pocas columnas de **Σ** serán suficientes de modo que la dimensión original de **Σ** se reducirá a fxf. Por lo tanto, reduciremos la matriz T a la dimensión nxf y la matriz P a fxm. En esta técnica, el número de detectores debe ser mayor que n, el número de muestras.

Si tenemos n muestras de calibración de las cuales conocemos en todas ellas la concentración de al menos uno de los componentes (supongamos que sea el único significativo=el calibrante), tendremos un vector **Cnx1** de concentraciones. Los scores pueden ser entonces relacionados así:

$$\mathbf{C}_{nx1} \approx \mathbf{T}_{nxf} . \mathbf{r}_{fx1} \qquad [11]$$

r es un vector de regresión que nos permitirá estimar la concentración en las muestras, idealmente, su longitud debería ser igual al número de componentes de la muestra, pero puede ser menor.

Calculamos entonces **U**, **Σ y V mediante SVD** de la matriz de absorbancias $\mathbf{X}_{nxm}$**.** Luego de reducir **Σ** a un valor apropiado de fxf calculamos los scores y loadings:

$\mathbf{T}_{nxf}=\mathbf{U}_{nxf} . \mathbf{\Sigma}_{fxf}$. Reducir **V** (desde Matlab) a $\mathbf{V}_{m,f}$**;** $\mathbf{P}_{fxm}=\mathbf{V}^{t}_{m,f}$. Téngase en cuenta que con esta reducción **X** debe ser expresada como **X=T.P+E** ya que debido a la reducción de **Σ** es **X≈T.P,** siendo **E** una matriz de errores. Y lo mismo para **C** en la ecuación **[11]**.

Ahora podemos calcular **r** teniendo en cuenta la ecuación **[11]** y la pseudoinversa de **T**:

$$\mathbf{r}=(\mathbf{T}^{t}.\mathbf{T})^{-1}.\mathbf{T}^{t}.\mathbf{C}_{n} \qquad [12]$$

Podemos entonces calcular las concentraciones **estimadas**, $\hat{\boldsymbol{C}}_{n}$ para el lote de calibración:

$\hat{\boldsymbol{C}}_{n}= \mathbf{T}_{nxf} . \mathbf{r}_{fx1}$. Como veremos más adelante, con $\mathbf{C}_{n}$ y $\hat{\boldsymbol{C}}_{n}$ podemos calcular el error del método.

Si en lugar de sólo un componente conociéramos k componentes de la matriz de calibración, en lugar de $\mathbf{C}_{n,1}$ tendríamos una matriz $\mathbf{C}_{nxk}$. Podemos generalizar el cálculo como $\hat{\mathbf{C}} \approx \mathbf{T}_{nxf} . \mathbf{R}_{fxk}$ y $\mathbf{R}=(\mathbf{T}^{t}.\mathbf{T})^{-1}.\mathbf{T}^{t}.\mathbf{C}$**.**

Una particular ventaja de esta técnica es que, si el número de elementos (factores) seleccionados es igual al número de **compuestos de interés**, **T** y **C** tienen las mismas, n k, dimensiones (ver ecuación [12]) y **R** resulta una matriz cuadrada. Entonces, reordenando la ecuación **[10]**:

$$\mathbf{X=T.P=T.R.R^{-1}.P=\hat{C}.\hat{S}}$$

Calculando $\mathbf{\hat{S}=R^{-1}.P}$, se pueden estimar los espectros **de cada componente individual** (sin haber tenido esa información de antemano). Muy conveniente cuando existen componentes que no es posible aislar químicamente. Finalmente, **T.R** estima las concentraciones.

Cálculo de los errores

Para determinar el error del cálculo de concentraciones, el procedimiento más sencillo es obtener la suma de cuadrados de los residuos entre los valores reales y los predichos. Para una matriz donde se calibra con n espectros, el error sobre el componente i es:

$$E = \sqrt{\sum_{i=1}^{n}\left(\frac{(\widehat{c_i}-c_i)^2}{g}\right)} \quad \text{donde g=n, si n>>f,} \qquad E\% = \frac{100.E}{c_i}$$

Recalcularemos nuestro caso ejemplo con la técnica PCR utilizando los 5 factores.

Tabla de resultados

Compuesto	Ec%	Ep%
Carbazepina	0.0002472	0.4853
Bencidina	1.5681E-05	0.1164
Iodobenceno	2.7194E-05	0.0697
Anilina	7.7447E-05	0.0613
p-aminotolueno	2.4637E-05	0.1019

Nuevamente se observa que el error de predicción es mayor al de calibración. Pero además podemos comparar estos resultados con los de calibración multivariada utilizando todos los sensores. Vemos que los errores de predicción para PCR son significativamente menores

Si aún hubiéramos hecho el ejercicio por los dos métodos con sólo alguno de los componentes, en lugar de todos, la diferencia sería aún mayor, debido a que para calibración multivariada la falta de componentes en la calibración empeora mucho los errores.

Se puede considerar otro tipo de errores relacionados con **el ajuste de los espectros**, en lugar de las concentraciones. Recordemos que m es el número total de λ's en $X_{n,m}$.

$$S_x = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \hat{x}_{ij})^2$$

Donde para f factores.

$$\hat{x}_{ij} = \sum_{a=1}^{f} t_{ia}.p_{aj}$$

Finalmente: $E=(\mathbf{S}_x/(n.m))^{1/2}$ y E%=100.E/(media de x) para datos no centrados.

Cuadrados Mínimos Parciales – Partial Least Squares (PLS)

PLS es una muy importante técnica de calibración multivariada. En muchos casos **donde el problema no es linealmente aditivo**, como en QSAR, biométrica o psicométrica, la técnica ha sido aplicada inapropiadamente. Pero en química analítica conocemos que una gran cantidad de problemas son realmente linealmente aditivos y por lo tanto, esta técnica, es de gran utilidad.

Sin embargo, hay que tomar buen recaudo de que en la aplicación del método, las matrices químicas tengan la mismas características espectrales que los datos de calibración y predicción. Por ejemplo, si se trabaja sobre un lote de muestras para determinar componentes en leche de vaca de Argentina, no será apropiado el mismo **método** para leches de otra región o incluso de la misma región y otra raza de vacas, a menos que haya un pretratamiento de muestra para reproducir las matrices, cosa no considerada en nuestros ejemplos. Con estas salvedades, el método es muy robusto.

Existe más de un método para utilizar esta técnica.

Método PLS1

Una de las técnicas de cálculo más difundidas es PLS1. Es necesario advertir que esta técnica tiene muchas variantes de algoritmos de cálculo, de modo que es altamente recomendable que el analista tenga muy en cuenta el tipo de programa que está utilizando.

La característica fundamental de PLS1 es que, en lugar de modelar el sistema con una única ecuación, como vimos hasta ahora, utiliza 2, una que modela los espectros y otra que modela las concentraciones, a saber:

$$\mathbf{X=T.P+E} \quad [13] \quad \text{y} \quad \mathbf{C=T.q+f} \quad [14]$$

Obsérvese que **T**, la matriz de escores que calculamos mediante SVD, es común a ambas ecuaciones, la que relaciona espectros y la que relaciona concentraciones. Si bien la suma de cuadrados (SDC) de los escores (**T**) de cada componente es a veces llamado *eigenvalue*, estos no tienen nada que ver con los de PCA.

El vector **q** es análogo a un *loading* vector, pero no está normalizado. **E** es una matriz de errores y **f** es un vector del mismo tipo. **El hecho de que c sea un vector es debido a que aquí los componentes se calculan uno a uno en forma sucesiva**. Cada componente **genera una matriz de scores distinta** a diferencia de lo que ocurre en PCR donde **T** es una matriz **única**. **P** es una matriz análoga a la de PCA y la SDC de cada fila vale 1. En algunas variantes de cálculo la segunda ecuación, [14], se desarrolla como un producto de tres factores: el primero proporcional a **T**, el segundo una matriz diagonal (*scalig factors*) y el tercero un vector normalizado proporcional a **q**. En muchos programas de cálculo de PLS se suelen centrar los datos de **X** y **c**, sin embargo no hay obligación de hacer esto y cálculos sobre datos no centrados son perfectamente aceptables. Como en PCR, el error asociado a **X** puede ser calculado por una variedad de caminos y debe tenerse cuidado con el número de grados de libertad con que se calculan estos errores.

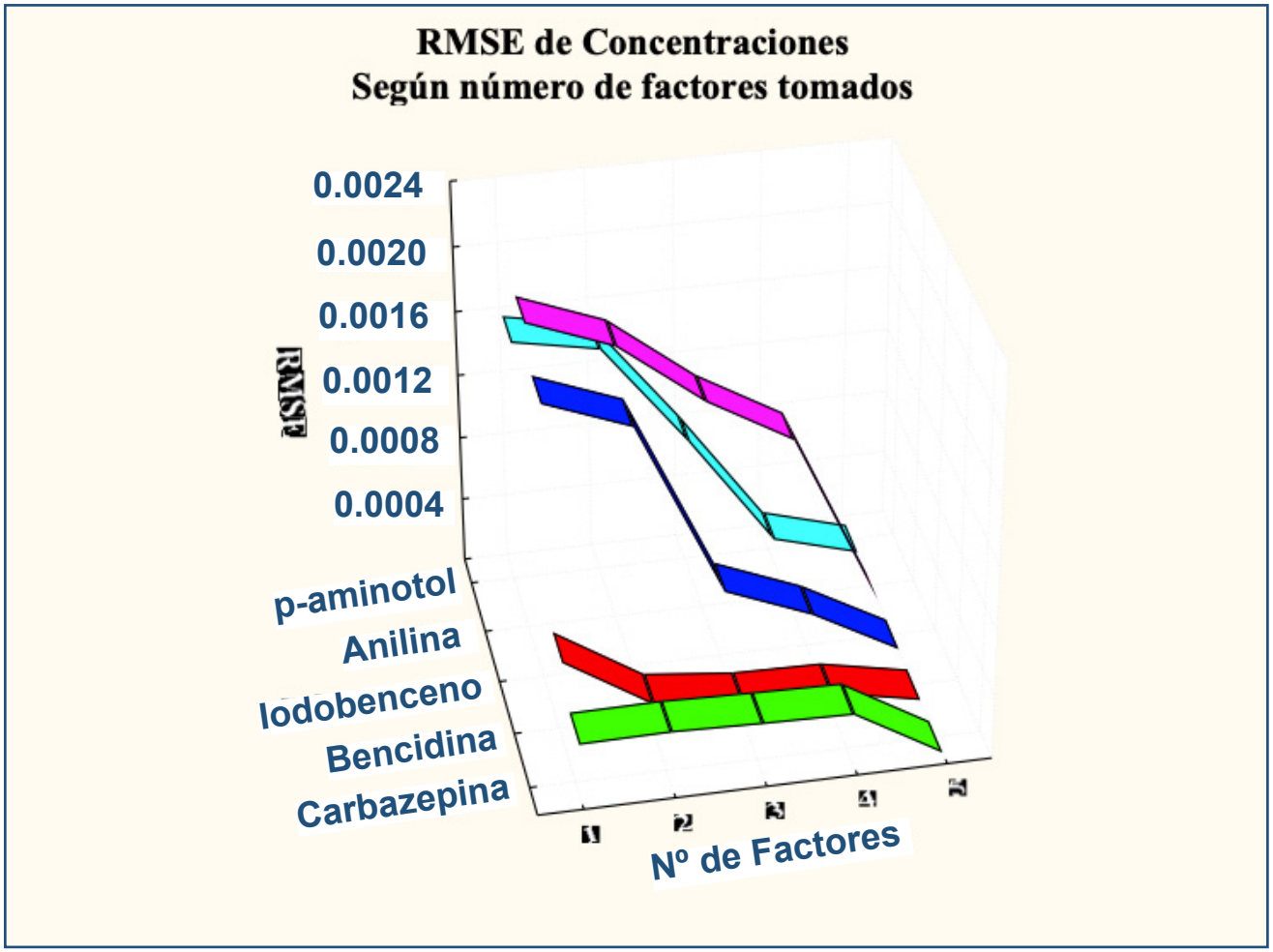
Uno de los algoritmos de cálculo más difundidos para PLS es el **NIPALS** (Non-linear Iterative Partial Least Squares). Los pasos esenciales del cálculo se dan en el apéndice de la Ref. 1, para no intercalar aquí una excesiva cantidad de álgebra lineal. Diremos por ahora que el método es iterativo y que por cada iteración se calcula un factor que se agrega al anterior para mejorar la aproximación. Los sucesivos factores son ortogonales entre sí. La concentracion en el i-**ésimo** espectro para el n-**ésimo** componente calculada con A factores se predice (si fueron centrados) con:

$$\hat{c}_{in} = \sum_{a=1}^{A} t_{ian}.q_{an} + \bar{c}_i$$

Expresado en forma matricial es:

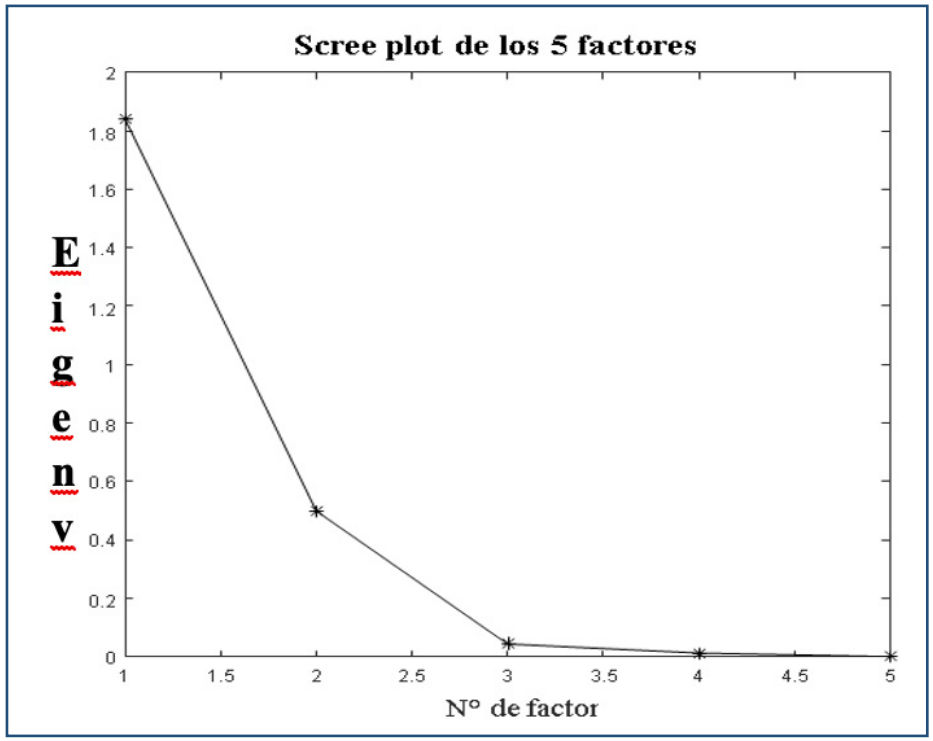
$$c_n = T_n . q_n + c_n$$

En la figura se puede apreciar, para el mismo ejemplo de siempre, la disminución del error a medida que se utilizan más factores y la diferencia para cada componente.



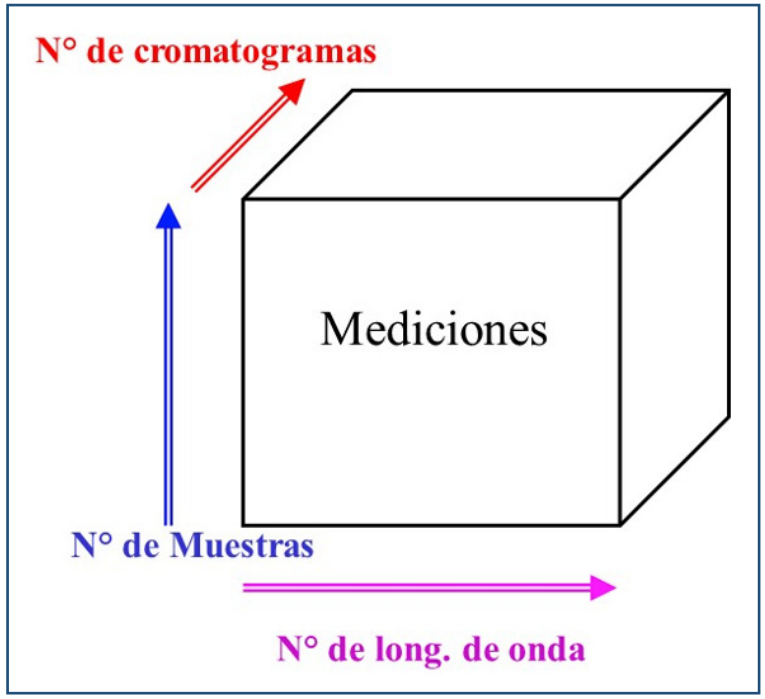
Err. Calibr.	Carbazepina	Bencidina	Iodobenceno	Anilina	p-aminotol.
N°factores	SQRSSY				
1	3.39E-04	5.18E-04	0.0018	0.002	0.0016
2	3.45E-04	1.81E-04	0.0016	0.0018	0.0015
3	3.27E-04	1.20E-04	5.07E-04	0.0014	8.54E-04
4	3.14E-04	1.10E-04	2.92E-04	0.0011	1.51E-04
5	1.78E-09	7.44E-10	1.50E-09	5.57E-09	1.31E-09

Se puede ver, tanto en el gráfico, como en la tabla de errores o el Scree plot a continuación, que después del tercer factor, hay una mínima y poco significativa recuperación de información. En problemas con muchos compuestos, el Scree plot suele utilizarse como guía previa para saber cuántos factores tomar en PLS1.



PLS1 trilineal

Este tipo de cálculo se aplica a instrumentos de segundo orden, que son aquellos que reúnen información bajo la forma de una matriz de tres dimensiones (o sea un cubo de datos). Este tipo de instrumentos es aún muy limitado y se reducen, en química analítica, a la cromatografía acoplada a algún sistema espectrométrico, a la espectroscopía de fluorescencia de excitación-emisión o a un diseño especial del experimentador.



La figura muestra un esquema de la estructura de los datos en una experiencia de cromatografía acoplada a detección espectrofotométrica. Por ejemplo, para un lote de I muestras de calibración, con J períodos de tiempo del cromatograma, se hacen para cada uno de ellos, barridos del espectro en todo el rango con K λ´s. Observe que podríamos considerar a **cada** muestra como una matriz de segundo orden entre un instante del cromatograma y el barrido de su espectro.

La descripción aquí será limitada a una matriz X trilineal (o cúbica) y un solo calibrante c. Si los calibrantes conocidos fueran varios, la solución más simple es aplicar PLS trilineal individualmente a cada variable. Las ecuaciones básicas detalladas para un programa de cálculo se dan en la bibliografía (Ref 1).

En principio no hay ninguna necesidad fundamental para centrar los datos, pero puede hacerse si se considera útil.

Consideremos el caso de la cromatografía acoplada y supongamos que tenemos un lote de datos de I muestras, cada una de las cuales ha sido medida a J tiempos de elución levantando un espectro de K longitudes de onda en cada cromatograma.

Como en PLS1, los factores son calculados en forma sucesiva. Para cada factor tendremos: un vector de *scores*, t, de dimensión I; un *weight* vector análogo a un *loading* vector $^{j}\mathbf{p}$ de longitud J y otro semejante $^{k}\mathbf{p}$ de longitud K. La suma de cuadrados de estos dos últimos vectores es igual a 1. El lenguaje utilizado aquí es similar al de PLS pero *scores* y *loadings* son totalmente diferentes; en PLS trilineal, por ejemplo, estos vectores no son ortogonales.

Como antes, la relación entre $\hat{\mathbf{c}}$ y los *scores* es $\hat{\mathbf{c}}=\mathbf{T.q}$. Sin embargo, **q** debe ser recalculado **después de cada factor** que se agrega, mediante la operación inversa:

$$\mathbf{q}=(\mathbf{T}^{t}.\mathbf{T})^{-1}.\mathbf{T}^{t}.\mathbf{c}$$

T es la matriz de scores cuyas columnas consisten de vectores *score* individuales para cada factor y tiene dimensión IxA donde A es el número de factores calculados. El número de factores, como antes, determina el error del cálculo.Los residuos de X, después del cálculo de cada factor, vienen dados por:

$^{r,a\mathit{new}}\mathbf{X}_{ijk}=^{r,a\mathit{old}}\mathbf{X}_{ijk}-\mathbf{t}_{i}.^{j}\mathbf{p}_{j}.^{k}\mathbf{p}_{k}$, donde r significa residuo. Esto conduce a

$$\hat{x}_{ijk}=\sum_{a=1}^{A} t_{i}.^{j}\mathbf{p}_{j.}\,^{k}\mathbf{p}_{k}$$

Dado que los *scores* y *loadings* de los sucesivos factores no son ortogonales, el método de determinar residuos es simplemente una aproximación. Los residuos del bloque **X** no tienen un significado físico directo, sin embargo, las concentraciones (bloque **c**) son bien modeladas.

Trilinealidad: un llamado de atención

Como hemos visto desde un principio, esta parte del libro refiere a métodos lineales, o sea todos los métodos en los cuales la respuesta de un sistema sea lineal respecto a las variables. Matemáticamente, en forma general esto significa que para un sistema X con variables a,b,…k debe cumplirse que para cada medición *i*, es:

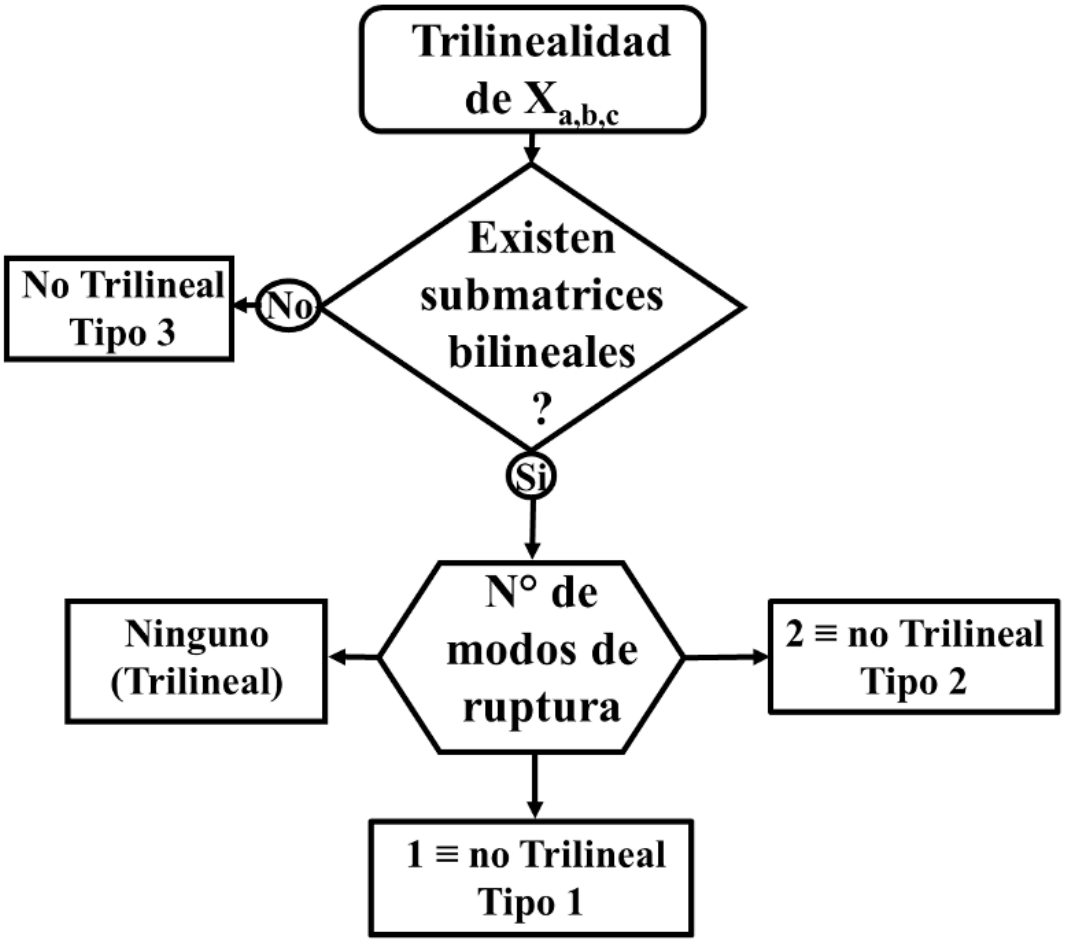
$$X_{i,a,b,\ldots k} = a_i . b_i \ldots k_i \qquad [13]$$

En la sección de regresión lineal múltiple hemos visto, como ejemplo, esta técnica de cálculo para espectroscopía de absorción aplicada a soluciones de mezclas de compuestos. En principio, en este caso no quedan dudas de la linealidad debido a la ley de Lambert Beer, excepto si hubiese desviaciones debido a errores de medición. Hemos visto cómo una matriz rectangular de datos puede ser matemáticamente descompuesta en un producto de 2 o tres matrices (ecuaciones 9 y 10); sin embargo, hay algo más que agregar: los vectores de las matrices deben ser independientes, o sea, no deben existir correlaciones entre ellos. Químicamente hablando, en el caso mencionado de la espectroscopía, por ejemplo, no debería haber espectros iguales entre las mesclas de componentes de las muestras. Matemáticamente esto es fácil de comprobar calculando el rango de la matriz que nos dará el número de vectores no colineales.

Pero en el caso de PLS1 trilineal (o multilineal) interviene más de una técnica y por lo tanto la linealidad debe mantenerse para todas ellas. Existen casos en que esto es difícil de cumplir, como en el ejemplo dado, donde interviene la cromatografía líquida. En ésta, la baja reproducibilidad de los perfiles de elución de cada muestra, para cada constituyente de éstas, introduce desviaciones (Ref. 4,6).

Más allá de las técnicas específicas de preproesamiento para corrección de la linealidad desarrolladas, se debe considerar si el grado de la desviación es mínimo o importante, existiendo una escala de grises entre ambos extremos (Ref. 4, 6).

Para el tratamiento de la trilinealidad sean clasificado varios tipos de ésta en un esquema, tomando en cuenta la causa que las produce (Ref. 4), que es el siguiente:



Los métodos de cálculo más evolucionados que se comentarán a continuación varían, entre otras consideraciones, en su habilidad de resolver el impacto negativo de la multi-linealidad.

Estos **métodos**, que siguen a continuación, son corrientemente utilizados por aquellos que tienen alguna experiencia en calibración multivariada. Existen libros especialmente dedicados y decenas de artículos con aplicaciones de estos méto-

dos, como algunos de los que se citan en la bibliografía del capítulo. Los lectores interesados pueden recurrir a ellos para adquirir el conocimiento pleno y detallado en teoría y práctica.

Está más allá del alcance de este libro, debido a la extensión del tema, hacer un tratado completo en este punto. Aquí se presenta sólo una introducción, a fin de que el lector conozca las posibles aplicaciones con estos métodos. Detalles sobre los programas de cálculo se dan al final del capítulo.

Parallel Factor Analysis (*Parafac*)

Parallel Factor Analysis es uno de los algoritmos reconocidos para analizar sistemas de datos trilineales. Retomando la ecuación [13] para el tratamiento de una serie de N constituyentes de una solución, la expresión matemática que representa al conjunto de datos es

$$\boldsymbol{X_{ijk} = \sum_{n=1}^{N} a_{in} . b_{jn} . c_{kn} + E_{ijk}} \qquad \textbf{[14]}$$

Aquí, *ain* representa el valor del tipo de medición, *a,* del enésimo componente en la muestra i. Análogamente, b_{jn} *y* c_{kn} representan las mediciones del tipo *b* y *c.* Los rangos de I, J y K no son necesariamente iguales.

Observe que la sumatoria expresa X como la adición de las señales individuales de los N componentes más un término de error.

Parallel Factor Analysis (PARAFAC) es uno de los algoritmos de cálculo más aceptado para resolver el sistema. Consiste en la descomposición trilineal de la matriz en tres matrices A, B y C. El algoritmo comienza por estimaciones iniciales de a_{in}, b_{jn} *y* c_{kn} que se van ajustando hasta converger en un valor mínimo para la suma cuadrática de errores (sum of the squared errors, SSE). En cada ciclo del algoritmo, sólo un parámetro es ajustado, por ejemplo, a_{in}, mientras los otros se mantienen constantes, de la misma manera luego sigue el turno para b_{jn} *y* c_{kn}. Esta forma de trabajo se llama cuadrados mínimos alternativos (alternanting least-squared, ALS) y lo veremos aplicado también a otras técnicas.

Una importante característica de esta técnica es la ***unicidad*** de la respuesta, esto es, existe una única solución para todos los parámetros: a_{in} b_{jn} *y* c_{kn}. Una ventaja adicional es que es posible cuantificar los analitos de interés aún en muestras conteniendo especies extrañas no calibradas.

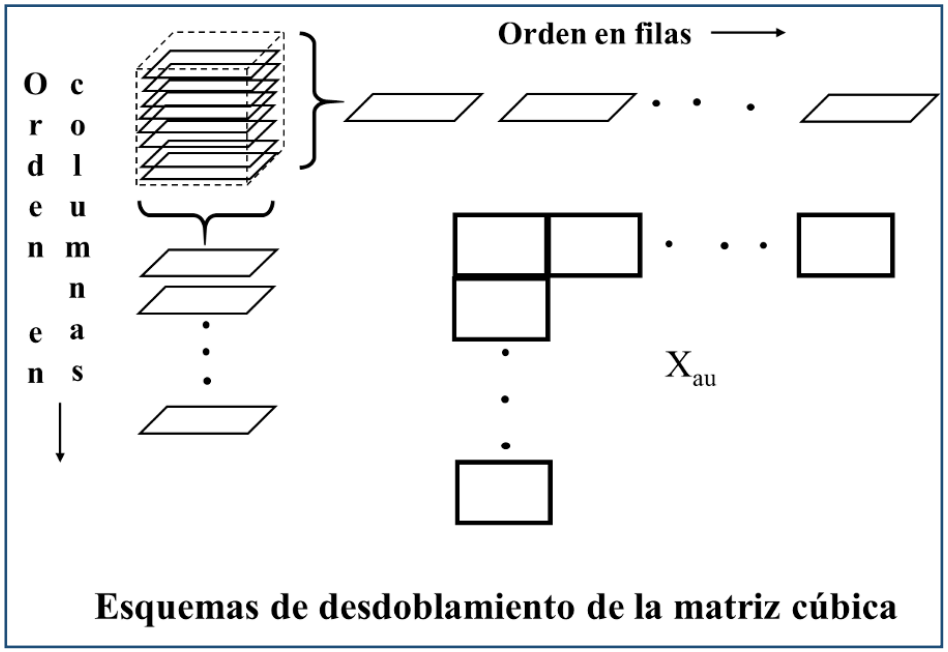
Multivariate Curve Resolution (MCR)

Lo habitual es la utilización de esta técnica en combinación con ALS de modo que su denominación es MCR-ALS (Ref. 7).

Probablemente, la mayor diferencia que tiene esta técnica respecto de las demás, es su aplicación en áreas de conocimiento que van más allá de la química analítica, como son, el medioambiente, cinética y equilibrio químico, microscopía Raman, NIR, procesos químicos de producción, electroquímica, metabolómica y otras áreas aún más alejadas de la química tales como imágenes hiperespectrales.

Este modo de cálculo es aplicable a no trilinealidad del tipo 1.

La estrategia básica de este método es el desdoblamiento de la matriz cúbica (en el caso trilineal) en una serie de matrices bidimensionales. Este desdoblamiento puede hacerse en dos sentidos diferentes: El de las columnas o el de las filas, como muestra la figura siguiente. Al conjunto de matrices bidimensionales desdobladas se lo llama, a los fines del cálculo, *matriz aumentada* $\mathbf{X}_{au}$.



Esquemas de desdoblamiento de la matriz cúbica

La expresión matemática de $\mathbf{X}_{au}$ es $\mathbf{X}_{au} \approx \mathbf{B}_{au}.\mathbf{C}^t$ **[15]**

Las dimensiones de estas matrices son $\mathbf{X}_{au(I,J,K)}$; $\mathbf{B}_{au(I,J,N)}$ y $\mathbf{C}_{(K,N)}$. Aquí, I, J y K tienen el mismo significado que los modos en PLS1 trilineal, la coma separa las dimensiones de las matrices en los sentidos filas, columnas.

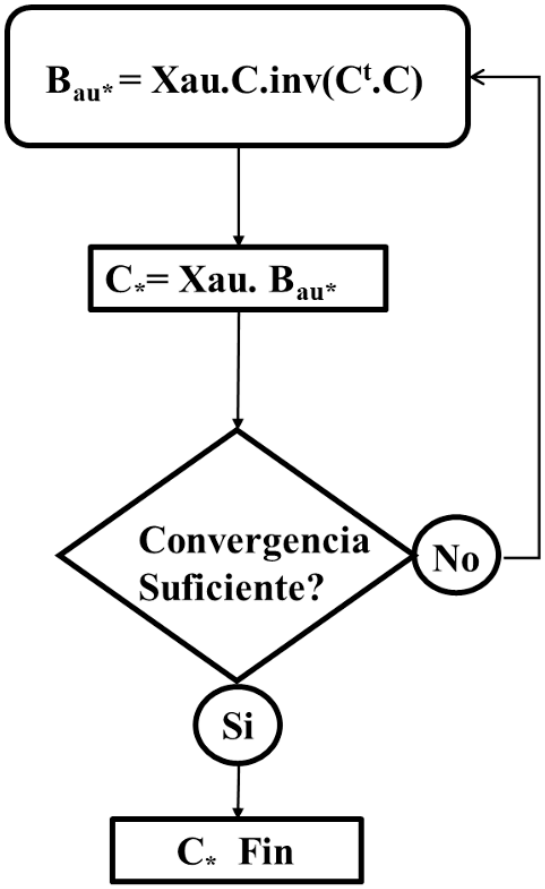
Si se elige, como en PLS1 trilineal, un ordenamiento en columnas para representar los perfiles de elución cromatográficos en las distintas muestras, la representación de la ecuación [15] en forma completa es:

$$\mathbf{X}_{au} = \mathbf{B}_{au}.\mathbf{C}^t + \mathbf{E}_{au} \qquad \textbf{[16]}$$

El término $\mathbf{E}_{au}$ es el error asociado al cálculo. La solución entonces de la ecuación 16 viene dada por el método de minimización de la suma de cuadrados de los elementos de $\mathbf{E}_{au}$**.** El problema aquí es que la descomposici**ón de la ecuación** 16 no tiene solución única como en *parafac*. Una causa de este problema es el **ajuste de las condiciones iniciales**, pero además **deben aplicarse restricciones** al cálculo para obtener soluciones eficientes. Estas restricciones deben corresponder a condiciones fisicoquímicas razonables, tal como la no negatividad de las concentraciones, ph, factor UV, etc.

Otra consideración es la estimación del número de componentes en el cálculo. Entre otros, aquí mencionaremos solamente nuestro conocido *scree-plot* que utilizamos en PC o su correspondiente suma de varianza capturada (ver capítulo 2).

Respecto del crítico ajuste de las condiciones iniciales, estas se pueden hacer disponiendo de una primera estimación de C o de o B_{au}. Por ejemplo, si se dispone de una estimción de C desde los datos experimentales, se puede continuar con una estimación de B, B_{au*}, extraída de la ecuación [15] ($_*$ ≡ 'Estimado') y seguir el siguiente ciclo:



Durante este ciclo es cuando se aplican las **restricciones** para que las soluciones matemáticas adquieran sentido fisico-químico.

UPLS-RBL

Hasta ahora, hemos dado solución al tipo de problemas con trilinealidad del tipo 1. UPLS-RBL tiene la propiedad de resolver también los casos de trilinealidad de tipo 2 y 3.

Tomamos otra vez la designación I, J, K para las dimensiones de la matriz de calibración matriz de segundo orden, $\mathbf{X}_{I,J,K}$. El primer paso de la operación matemática es desdoblarla (unfold) en vectores $\mathbf{x}_{vec}$.

$$\mathbf{x}_{vec,i}=\text{vec}(\mathbf{X}_i) \qquad [17]$$

Designamos vec a la operación de vectorización e "i" a la iésima muestra de calibración. Con todos los vectores "i" armamos una nueva matriz para analizarla por PLS:

$$\mathbf{X}_{PLS}=| \mathbf{x}_{vec,1}\ \mathbf{x}_{vec,2}\ \mathbf{x}_{vec,3}\cdots\ \mathbf{x}_{vec,i}| \qquad [18]$$

Observe que los vectores tienen dimensión JxK y entonces las dimensiones de $\mathbf{X}_{PLS}$ son (JK),I . Como hemos visto en la ecuación [8] del capítulo 2 sobre PLS, la resolución de esta matriz es:

$$\mathbf{X}_{PLS}=\mathbf{P}\mathbf{T}^t + \mathbf{E}_{pls}, \qquad [19]$$

donde **P**, son los loadings; **T**, son los scores y $\mathbf{E}_{pls}$ el error asociado a $\mathbf{X}_{PLS}$. Como sabemos $\mathbf{X}_{PLS}$ se resolverá para capturar el máximo adecuado % de varianza en relación al ajuste del **número de variables latentes,** *a* (ver capítulo 2). Sin embargo, hay que tener en cuenta aquí, que el valor de *a* está, en principio, relacionado con el % de varianza **total** de $\mathbf{X}_{PLS}$, pero aquí nos interesa la concordancia con los valores de las concentraciones de analitos en las muestras de calibración. Para lograr este objetivo la calibración de PLS se efectúa para un analito a la vez, con su particular valor de *a*. Por lo tanto, los valores de *a*, pueden variar de un analito a otro. Para tomar el mejor valor de *a* de cada analito, a_c, se deben tomar en consideración todas las correcciones instrumentales y analíticas de la medición y no sólo el valor matemático. Finalmente, el conjunto de valores **óptimos** a_c de cada analito constituirá la matriz A y retomando la ecuación [19], las dimensones de **P** y **T** son JK,A e I,A respectivamente.

Una técnica muy utilizada para calcular el número de variables latentes, A, es el de la suma de cuadrados del error de predicción, **PRESS,** en la literatura (Predicction of the Error Squared Sum). Nos referiremos a ella en el próximo punto, "validación de los modelos".

No olvidemos que la ecuación [19] ahora reconstituida y optimizada está expresada en variables *latentes* y no en variables *manifiestas*. Para estimar las concentraciones de analitos hay que recurrir al vector de regresión de PLS (ecuación [11], [12] y siguiente), $\mathbf{r}=(\mathbf{T}^t.\mathbf{T})^{-1}.\mathbf{T}^t.\mathbf{C}_I$, y con este estimamos C_I.

$$\hat{C}_I = \mathbf{T}_{I,A}.\mathbf{r}_{A,1} \qquad [20]$$

Si en la matriz de calibración, $\mathbf{X}_{I,J,K}$, hubieran existido componentes inesperados, entonces la ecuación [20] no es adecuada para predecir los analitos. La comprobación de este hecho puede hacerse calculando los residuos de UPLS: $s_{res} = ||vec(\mathbf{X}_{I,J,K} - \mathbf{P}.\mathbf{r}||/(JK-A)^{1/2}$. Ante la presencia de componentes inesperados s_{res} será anormalmente grande en comparación con el nivel de ruido instrumental.

Para corregir este problema se recurre entonces a la "Bilinealización residual" (de allí el nombre UPLS-RBL). La estrategia de RBL es modelar los residuos asumiendo que éstos se pueden ordenar en una matriz bilineal (de allí el nombre bilinealización). Entonces modela la matriz de datos, $\mathbf{X}_{I,J,K}$, como compuesta de dos contribuciones, la parte de ella que se explica por la calibración de los loadings de PLS y la contribución de potenciales interferentes, modelados también, por PLS.

$$\mathbf{X}_{I,J,K} = refold(\mathbf{P}.\mathbf{r}_{RBL}) + \mathbf{B}_{RBL}.\mathbf{T}^t_{RBL} + \mathbf{E}_{RBL}, \qquad [21]$$

que es major ordenarla como $\mathbf{X}_{I,J,K} - refold(\mathbf{P}.\mathbf{r}_{RBL}) = \mathbf{B}_{RBL}.\mathbf{T}^t_{RBL} + \mathbf{E}_{RBL}$

El operador *refold* simboliza la operación inversa a la de vectorización. El lado izquierdo de esta última igualdad es **la matriz residual**, el primer término del lado derecho es el PCA de la matriz residual y el último término es el error residual que debe ser reducido al mínimo. Esta ecuación requiere un método de minimización no lineal para $\mathbf{E}_{RBL}$; debe hacerse con Algoritmos tales como Gauss-Newton o Levenberg-Marquardt que es un poco más robusto. El resultado final será un nuevo valor del vector de regresión $\mathbf{r}_{RBL}$ que será **utilizado en la ecuación** [20] para obtener las concentraciones de la muestra respecto de los analitos calibrantes.

Otros modelos de cálculo

No hemos agotado para nada todas las variantes de cálculo que tiene ésta área de la quimiometría, sólo hemos descripto las más comunes. Pero existe un número importante de variantes debido a que hay actualmente un continuo desarrollo sobre este tema. Por ejemplo, aquí hemos tratado problemas de química analítica de hasta segundo orden. Pero existen

desarrollos para resolver problemas de tercer y más órdenes aplicados también, como ya se ha dicho, a otras áreas de la ciencia. El lector que se haya familiarizado con calibración multivariada y quiera profundizar en este tema puede recurrir a la bibliografía dada en este capítulo y la que se puede rastrear fácilmente con los medios de información computacionales.

Validación de los modelos

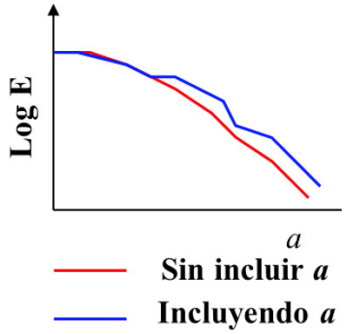
Para validar el modelo calculado se aplican principios similares a todos los métodos multivariantes vistos hasta aquí. La validación pretende responder a las siguientes preguntas:

a- Cuántos factores son necesarios para describir la serie de datos?
b- Cuán precisa es la predicción de una muestra desconocida?
c- Cuán representativos son los datos seleccionados para producir un modelo?

Autopredicción

En los métotos bilineales hemos visto que el error de calibración sobre el lote de muestras utilizado para tal fin viene dado por la expresión de E. Ahora nos referiremos exclusivamente al error en el cálculo de concentraciones con conceptos más generales.

El valor de E en función del número de factores '*a*' empleados en los métodos de **cálculo**, por ejemplo en **UPLS-RBL**, puede ser monótonamente decreciente si no se incluye *a* en el denominador, o ligeramente oscilante o creciente después de un fuerte descenso, si se la incluye. En el último caso esto es debido probablemente a problemas en los datos. El comportamiento más común está esquematizado en la figura siguiente (Ref. 1).



Este tipo de gráfico es muy útil para decidir cuántos factores conviene usar en el modelo, y se puede hacer con diferentes conceptos, a saber:

A- Establecer un error pre-establecido por la experiencia, por ejemplo 2%, debajo del cual no se incluye ningún factor más.
B- Obtener algunas medidas independientes del nivel de ruido e imponer ese nivel como límite para incluir factores.
C- Si los errores alcanzan un *plateau* (en un gráfico no logarítmico), utilizar ese nivel como límite.

Es necesario tomar muy en cuenta aquí que los errores de autopredicción son siempre menores a los errores de predicción y son por lo tanto engañosos para establecer un error del método. Esto es válido para cualquier tipo de modelo que se calcule.

Errores de predicción del modelo

El método más correcto para determinar el error de predicción de un modelo es utilizar un lote de *L* muestras, independiente del lote de calibración, para calcularlo. En este caso la expresión que reemplaza a E es:

$$PRESS = \sum_{l=1}^{L} (c_l - \hat{c}_l)^2 \qquad E_{test} = \sqrt{\frac{\sum_{l=1}^{L} (c_l - \hat{c}_l)^2}{L}} = \sqrt{\frac{PRESS}{L}} = RMSPE$$

PRESS es la sinonimia en inglés de *suma de cuadrados del error de predicción y* RSMPE es la *raíz cuadrada del error medio de predicción*. El número de muestras *L* en el lote de predicción no tiene que ser necesariamente igual al del lote de calibración del modelo, pero preferentemente debería ser del mismo orden. Sin embargo, algo a tener muy en cuenta es que el error de predicción no será igual para cualquier lote de datos y si no se toman recaudos las diferencias entre distintos lotes puede ser muy grande. Para evitar este problema se deben organizar los lotes de cálculo y predicción bajo reglas de *diseño de experimentos* (tema desarrollado en la parte 3 de este libro).

Validación cruzada (*cross-validation*)

Si bien el método recién mencionado es el más correcto, obviamente, medir todo un lote de muestras puede ser muy tedioso o costoso, cuando no imposible. Por lo tanto, se pueden practicar otras técnicas más económicas en mediciones. La más conocida de ellas es la 'deje uno afuera' o *leave-one-out* en inglés. Suponiendo que no disponemos de un lote de muestras para predicción, y por lo tanto disponemos solamente de un lote de muestras para cálculo del modelo, se procede de la siguiente manera: Se calcula el modelo dejando fuera del lote una muestra cualquiera. Sobre ésta se calcula el error de predicción c_1-$\hat{c}_1$, luego se vuelve a calcular el modelo incluyendo la muestra que se ha dejado afuera y dejando afuera otra, para obtener un segundo error de predicción c_2-$\hat{c}_2$. El procedimiento se repite hasta que cada una de las muestras del lote hayan quedado una vez afuera, entonces habremos obtenido una colección de errores de predicción cuya suma de cuadrados nos permite calcular PRESS y también RMSPE. El error de predicción calculado por validación cruzada suele

ser menor que el calculado con un lote independiente de datos y por lo tanto no puede tener la misma confiabilidad si debemos precisar muy bien los errores. Si se ha utilizado un diseño de experimentos para elegir el/los lotes, la diferencia entre los distintos tipos de errores, incluido el de autopredicción, será mínima, lo que no quiere decir que sea despreciable.

Número de términos en un modelo

Existen algunas novedades relativamente recientes en este punto, en particular referidas a métodos no algorítmicos. Este tema en particular, aplicable a los modelos en general, será tratado en el capítulo 11, en la sección "Optimización de modelos multivariables".

Programas de cálculo en calibración multivariada

Como hemos visto al principio del capítulo, los cálculos aplicados a las calibraciones univariantes e incluso los de calibraciones multivariantes mediante regresión lineal **múltiple**, pueden ejecutarse simplemente con los comandos Matlab® apropiados y un mínimo de álgebra lineal. Sin embargo, de allí en adelante, desde el análisis por componentes principales, se necesitan métodos computacionales para aplicarlos con eficacia. Existe una amplia variedad de programas de cálculo (*software*) en esta área. Daremos aquí algunas direcciones útiles y mínimas para utilizar programas amigables a quienes comienzan a utilizar estas técnicas de cálculo.

Tanto para PCA como para PCR, PLS y PLS trilineal, pueden utilizarse los programas que se mencionan en la parte práctica de este libro.

Existen paquetes más avanzados de programas aplicados a calibración multivariada, algunos gratuitos, generalmente desarrollados en universidades y otros que son comerciales. Entre los primeros deseo destacar MCV2 y MCV3 (Ref. 4,8), que contienen interfaces gráficas amigables para los que se inician y cuyo soporte de cálculo es con MTLAB®. Entre los programas comerciales cabe destacar PLS_Toolbox Advanced Chemometrics Software for use with MATLAB® (Ref.9). Este tiene dos ventajas, la primera es que puede bajarse una versi**ón de prueba en forma gratuita por algunos días y luego adquirirlo si resulta práctico,** la segunda es que es una herramienta para ser utilizada en MATLAB®, sin necesidad de aprender ningún lenguaje adicional.

Por supuesto, otros paquetes de programas pueden buscarse en Internet.

Bibliografía del capítulo

1. Richard G.Brereton. *Analyst*,2000,125,2125-2154.

2. D.L. Massart; B.G.M. Vandeginste; L.M.C. Buydens; S. De Jong; P.J. Lewi and J. Smeyers-Verbeke. Hanbook of Chemometrics and Qualimetrics. Part B Capítulo 36. Elsevier, Amsterdam 1997.

3. Alejandro C. Olivieri. Calibración Multivariada. Ediciones Científicas Argentinas. Buenos Aires 2001.

4. Alejandro C. Oliviri and Graciela M. Escandar. Practical Three-Way Calibration. Elsevier MA 02451, USA. 2014. ISBN: 978-0-12-410408-2.

5. Barry M Wise and Neal B. Gallagher. PLS_Toolbox *for use with* MATLAB™. Eigenvector Technologies. http://www.eigenvector.com/PLS_Toolbox.html.

6. Silvana Azcarate, Adriano de Araújo Gomes, Arsenio Muñoz de la Peña and Héctor C. Goicoechea. Modeling second-order data for classification issues: data characteristics, algorithms, processing procedures and applications. TrAC Trends in Analytical Chemistry · August 2018. DOI: 10.1016/j.trac.2018.07.022.

7. S.C. Rutan, Anna De Juan, Roma Tauler. Introduction to Multivariate Curve Resolution. Comprehensive Chemometric Chapter: Introduction to Multivariate Curve Resolution. Publisher: Elsevier Editors: S. Brown, R. Tauler, B. Walczak

8. Olivieri AC, WU HL, Yu RQ. MCV3 a MATLAB graphical interface toolbox for third-order multivariate calibration. Chemom Intell Lab Sys 2012;**116**:9-16.

9. PLS_Toolbox. Advanced Chemometrics Software for use with MATLAB®. 2021 Eigenvector Research, Inc. 196 Hyacinth Road, Manson, WA 98831.https://eigenvector.com/software/pls-toolbox/.

SEGUNDA PARTE

Métodos No Algorítmicos

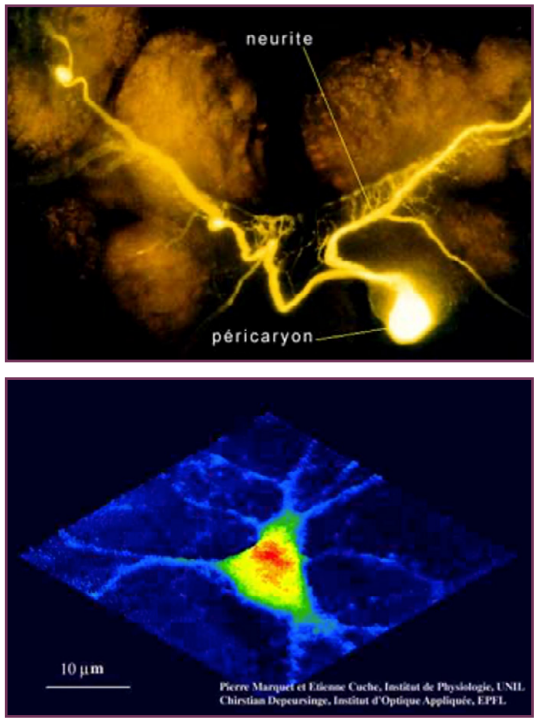
CAPITULO 5

Introducción y Redes de Una Capa

Existen hoy en día muchos métodos no lineales de análisis de la información, algunos diseñados sobre la base de imitación de fenómenos naturales, tales como: las redes neuronales artificiales (Ref. 1), los algoritmos genéticos(Ref. 2), optimización por enjambre de partículas (*particle swarm optimization*) (Ref. 3,4) *ant colony optimization* (Ref. 5,6), entre otros. Todas ellas han sido desarrolladas a partir de estrategias heurísticas.

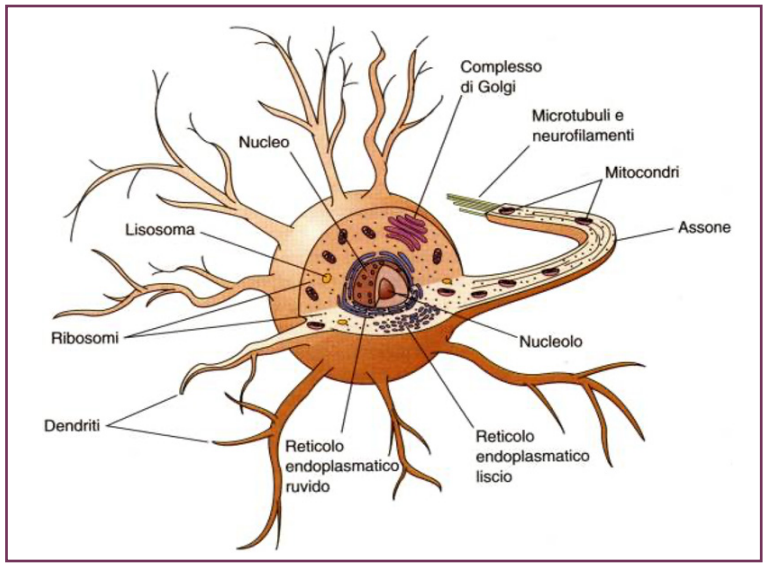
Las dos primeras técnicas tienen aplicaciones más amplias y más difundidas hasta ahora en quimiometría. A ellas nos referiremos especialmente.

Redes Neuronales Artificiales



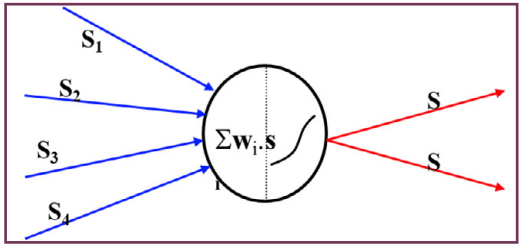
Las redes neuronales artificiales surgieron a partir del avance en el conocimiento de la anatomía y la función de las neuronas en el campo de la biología. Las figuras muestran dos neuronas diferentes: la superior es una neurona de caracol y la inferior una neurona de Purkinje. Las diferencias que se observan entre ellas son debidas, principalmente, a los distintos métodos de observación. Sin embargo, a pesar de la gran distancia en la escala zoológica que existe entre ambas, sus elementos esenciales son prácticamente idénticos: Una gran cantidad de filamentos por donde entra la información (dendritas), el núcleo de la célula y un filamento más robusto por donde sale la información (axón).

Esta compleja estructura real es usualmente esquematizada para su mejor comprensión, tal como se muestra en la figura siguiente.



Lo más interesante de las neuronas no es su particular anatomía, sino preguntarse ¿Qué ocurre con la información durante la entrada-salida? Observemos la inmensa cantidad de señales de entrada (10^3 a 10^4 por neurona), éstas entran en forma **paralela** al núcleo de la neurona y no en serie, lo que las hace más eficiente que las computadoras actuales. La información se reúne en el núcleo y tras un cierto (o más bien incierto) procesamiento, sale una única señal por el axón que en cierta forma **reúne las características de todas las señales de entrada.**

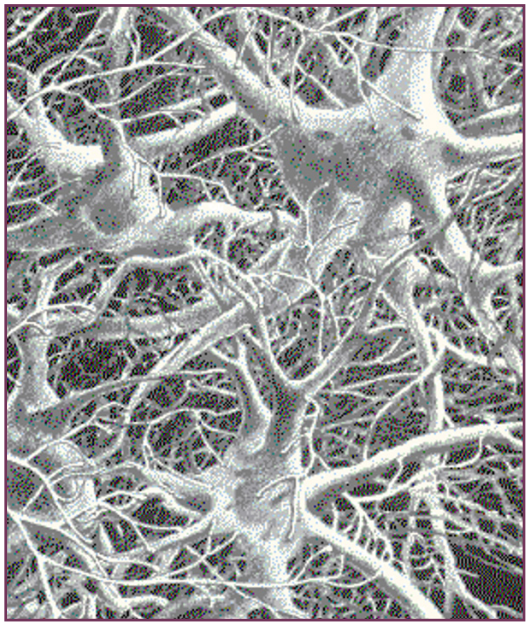
Poco se sabe todavía acerca del procesamiento de la información en el núcleo, pero, aun así, muchos se vieron tentados a tratar de simular el proceso de entrada-salida de las neuronas.



Esquema de neurona artificial para simulación del procesamiento de la información.

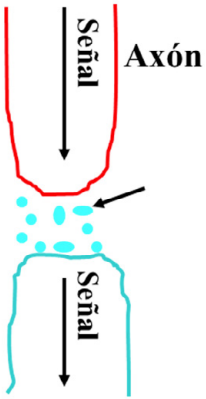
Para ello, se utiliza un esquema que facilite la mecánica de cálculo, tal como el de la figura anterior. Las entradas s1 a s4 pueden ser muchas más, lo mismo que las señales de salida. La función del núcleo la veremos más adelante.

Pero la simulación del funcionamiento de una única neurona no es aún suficiente, nuestro objetivo es armar una **red** de neuronas para simular el proceso de la información en ella. Una vista parcial de una red real puede verse en la figura.

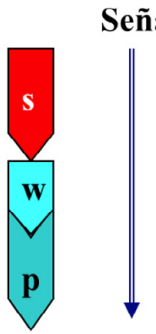


Para tener una idea de su complejidad basta decir que el número de neuronas del cerebro humano se estima que es del orden de 10^{11}!. Por supuesto estamos muy lejos de armar semejantes redes, pero sí es posible armar redes más pequeñas que son muy útiles para muchas aplicaciones en los campos de la ciencia y la tecnología y en esto radica nuestro interés actual. Muchos de los equipos que utilizamos a diario funcionan hoy en día mediante redes neuronales: sintetizadores de voz, mejoramiento de imágenes e instrumentos robóticos entre otros. Nuestros objetivos aquí serán las funciones de clasificación, reducción dimensional y modelado.

Procesamiento de las señales de entrada

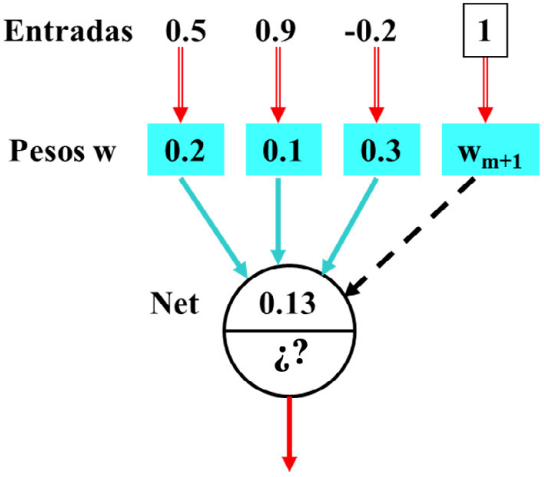


La figura de la izquierda esquematiza el proceso de entrada de la información en una neurona real. El axón y la dendrita no están en contacto directo, sino que se contactan a través de los neurotransmisores. Cuando existe conducción se dice que están en ***sinapsis***. La función del neurotransmisor es ***modular*** la intensidad de señal, o sea modificar su intensidad. Para nuestra simulación usaremos un esquema similar: la señal de entrada 's' pasa por un modulador 'w' que modifica la señal, de modo que la señal resultante que se transmite por esa dendrita se expresa como p=w.s.



A w se lo llama usualmente ***peso***, por su similitud con un peso estadístico. Como se ha visto, nuestra neurona tiene varias dendritas, o sea entradas, de modo que la señal total que llega al núcleo de la neurona es la suma de las intensidades de señal de cada dendrita individual. Podemos expresar entonces la señal de entrada como.

$$\mathbf{Net}=w_1.s_1+w_2.s_2+....+w_m.s_m=\sum w_i.s_i$$



Si expresamos el conjunto de w's como un vector $\mathbf{W}(w_1,w_2,...w_m)$ y hacemos lo mismo con el conjunto $\mathbf{S}(s_1,s_2,...s_m)$, podemos expresar Net como el producto escalar **Net=W.S**. En la figura de la izquierda se ejemplifica este cálculo con una neurona sencilla. Por razones de cálculo es conveniente introducir en **Net** un término independiente que nos dé un grado más de libertad, entonces:

$$Net=\sum_{i=1}^{m+1} w_i.s_i = W.S$$

Las funciones de transferencia

En la figura anterior, el núcleo de la neurona fue dividida en dos partes. En la parte superior hemos colocado el valor de la señal entrante y en la inferior hemos dejado un interrogante. Este lugar lo ocupa el cálculo de la transformación que sufre la señal antes de abandonar la neurona. La transformación se lleva a cabo mediante alguna función matemática, que denominaremos ***función de transferencia*** (FT).

Las FT son de crucial importancia por varias razones. En primer lugar, las FT distinguen a las neuronas dándole distintas propiedades de respuesta. Usualmente en una red neuronal hay un único tipo de neuronas, o sea todas tienen la misma FT, pero puede no ser siempre así. Algunas redes muy primitivas y elementales no tienen funciones de transferencia, pero la introducción de este concepto potenció enormemente la utilidad de las redes. Esto se debe a que la FT puede ser una función no lineal y esto hace que las redes puedan entonces resolver problemas no lineales, difíciles de resolver por otras vías.

Analicemos la FT de la figura 5.1, denominada *hard-limiter* en inglés y que representa una función escalón. En abscisas tenemos el valor de Net y en ordenadas el valor de la salida, I. Esta FT hará actuar a la neurona como si fuera una llave interruptora.

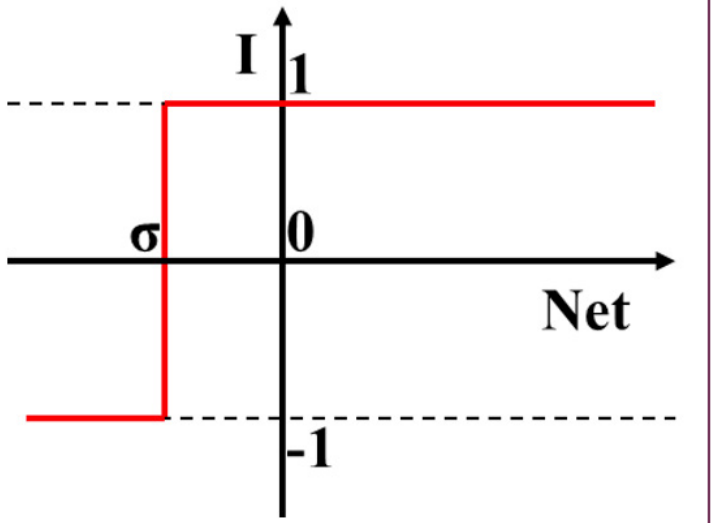
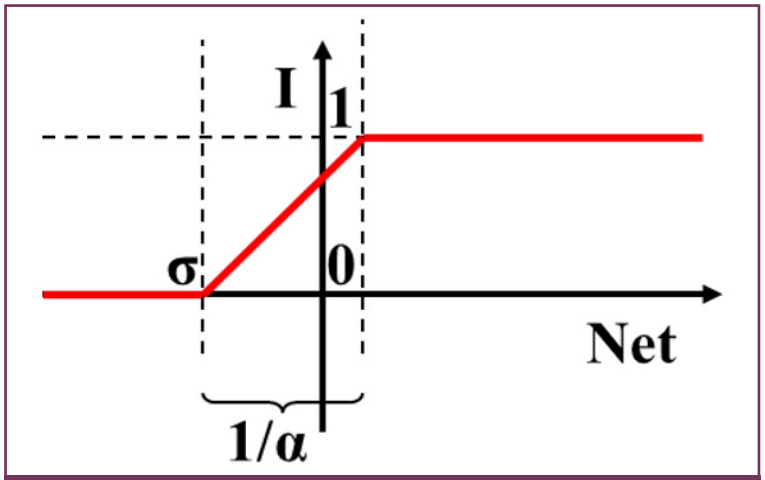


Figura 5.1

Si Net supera a σ la llave estará conectada y la salida es 1, de lo contrario no habrá conducción y la salida es –1. A σ se lo denomina valor ***umbral***.

Es común tratar de que los algoritmos que representan una FT sean sencillos para tener una buena velocidad de cálculo. Por ejemplo, el algoritmo que representa a la FT *hard-limiter* es: I=signo(Net- σ).1 ó también:

$$I = \begin{cases} 1 & \text{si Net} \geq \sigma \\ -1 & \text{si Net} < \sigma \end{cases}$$

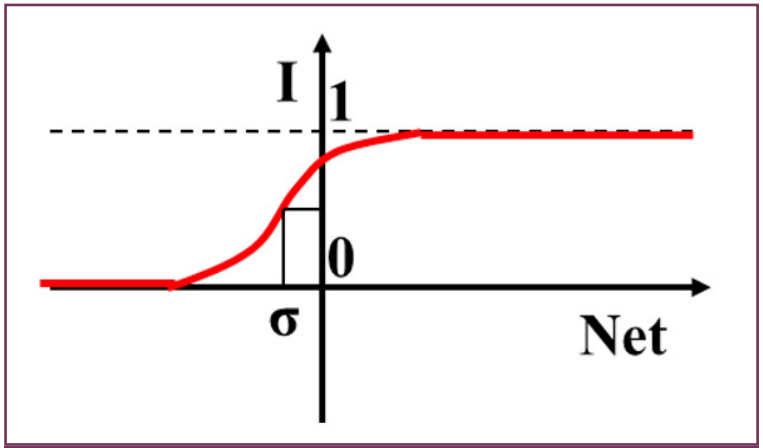


Threshold Logic

Otra FT similar a la anterior que, pero que posee un tramo $1/\alpha$ en el cual se comporta linealmente es la *Threshold logic* que se muestra en la figura. El algoritmo de cálculo para esta función es:

I=max(0, min(1,Net))

Si conoce algo de programación para interpretar las funciones max y min ¿Se anima a comprobar el algoritmo?



Sigmoidea

Una FT que merece especial atención es la función *sigmoidea* que se muestra en la figura. Su característica es que posee zonas de linealidad (próximas al punto de inflexión σ) y zonas de no linealidad, pero el pasaje de una zona a otra es progresivo, a diferencia de lo que hace la *Threshold logic*.

El algoritmo de cálculo de la función *sigmoidea* es*:*

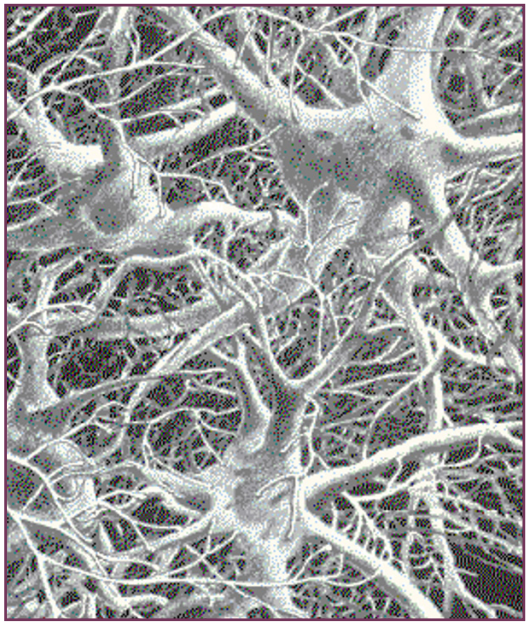
$$I(Net, \alpha, \sigma) = 1/\{1+\exp[-\alpha(Net + \sigma)]\}$$

Esta FT es algo más complicada de definir, sin embargo, tiene una importante propiedad matemática, es que su derivada puede expresarse simplemente como **I'=I(1-I).** Debido a que la derivada de la FT interviene en el cálculo de una de las redes más importantes es que esta función es de suma utilidad.

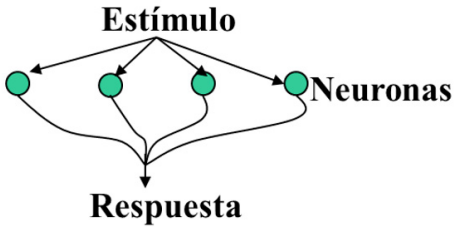
La arquitectura de las redes

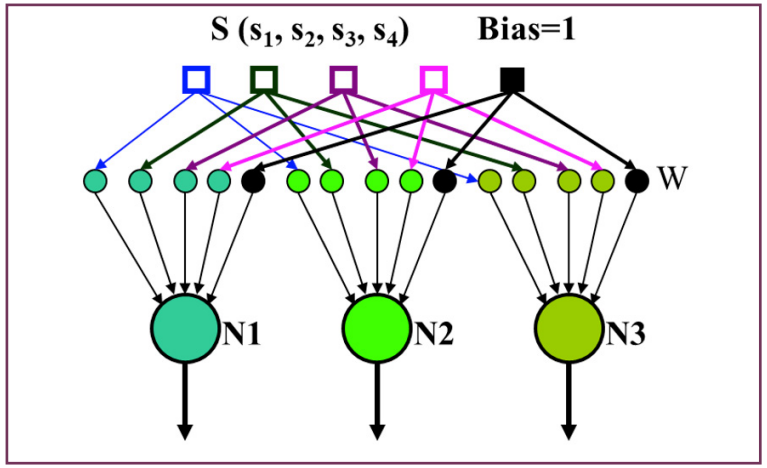
Volvamos a la imagen de una red neuronal real ya presentada.

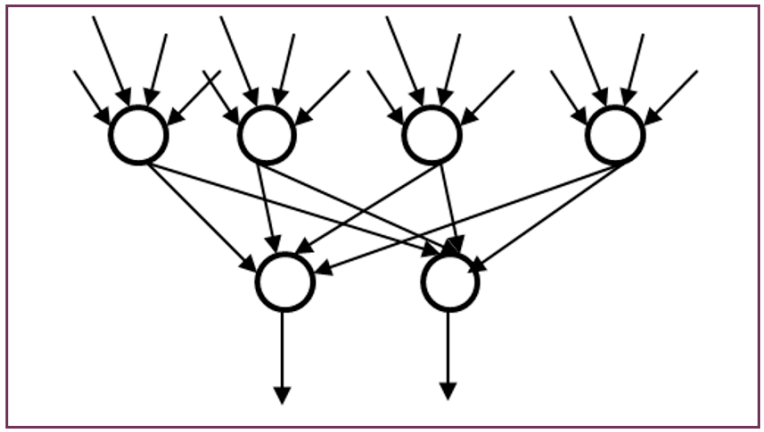
Cuando una parte de un organismo vivo es estimulada por algún fenómeno físico (calor, sonido, presión, etc.), no sólo una, sino muchas neuronas de esa área son estimuladas **al mismo tiempo**.



Las neuronas afectadas emiten entonces diferentes señales en forma conjunta, y esta señal conjunta es la que llega al cerebro indicándole lo que está pasando. Podríamos representar este mecanismo de la siguiente manera:







El mecanismo descripto puede asemejarse combinando las neuronas individuales. Tenemos una señal de entrada **S**, que es un vector conteniendo todas las señales de entrada individuales s_1, s_2, ...s_4. La primera fila (los cuadrados) es una capa de entrada de señales, lo único que hace esta capa es distribuir las señales **a todas las neuronas por igual.** Observe que esta capa incluye un *bias* y que este también está presente en la entrada de cada neurona (entradas en color negro). Las señales son afectadas por los *pesos* de cada dendrita (círculos pequeños) y luego entra a las neuronas N1, N2 y N3. Cada una de ellas produce una salida **diferente** debido a que, si bien han recibido las mismas señales, sus pesos son generalmente diferentes. Las salidas de las neuronas pueden ser la respuesta final o conectarse a una nueva capa como muestra la figura inferior.

Vemos entonces que existen redes de una sola capa de neuronas y otras con capas múltiples. También debe observase que las señales de salida de las neuronas de una capa se transmiten **siempre hacia la capa inferior** y nunca hacia una neurona de la misma capa.

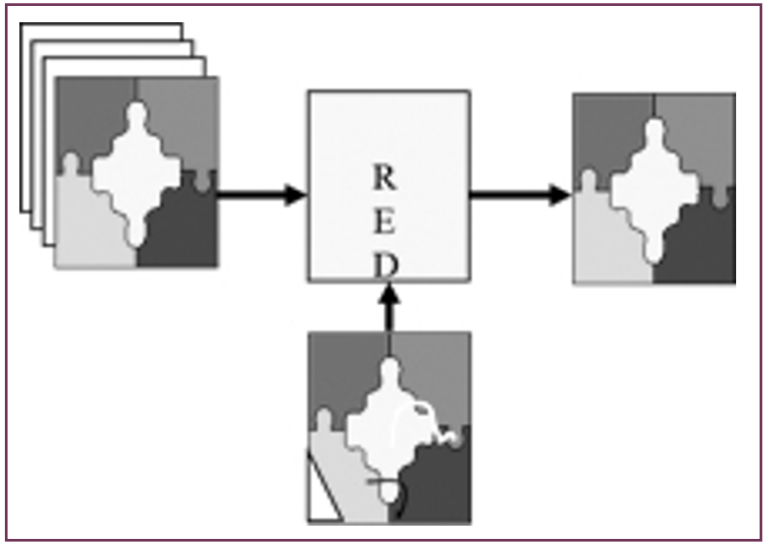
Hasta ahora hemos visto los aspectos básicos de las redes. Con estos elementos pueden armarse, en principio, redes de cualquier tamaño y cualquier complejidad. Pero estamos hablando solamente de la ***arquitectura*** de la red y no hemos dicho que función queremos que la red ejecute. Existen redes con arquitectura bien definida capaces de resolver problemas bien específicos, los que nos lleva a decir que existen distintas ***clases*** de redes. Para analizar en profundidad las redes que nos interesan es conveniente primero describir otras redes más sencillas que nos ayudarán a comprender después cuestiones más complejas.

Parte I: Redes de una sola capa

Existen redes con arquitectura bien definida capaces de resolver problemas muy específicos, los que nos lleva a decir que existen distintas *clases* de redes. Para analizar en profundidad las redes que nos interesan es conveniente primero describir otras redes más sencillas que nos ayudarán a comprender después los mecanismos de cuestiones más complejas.

La red Hopfield

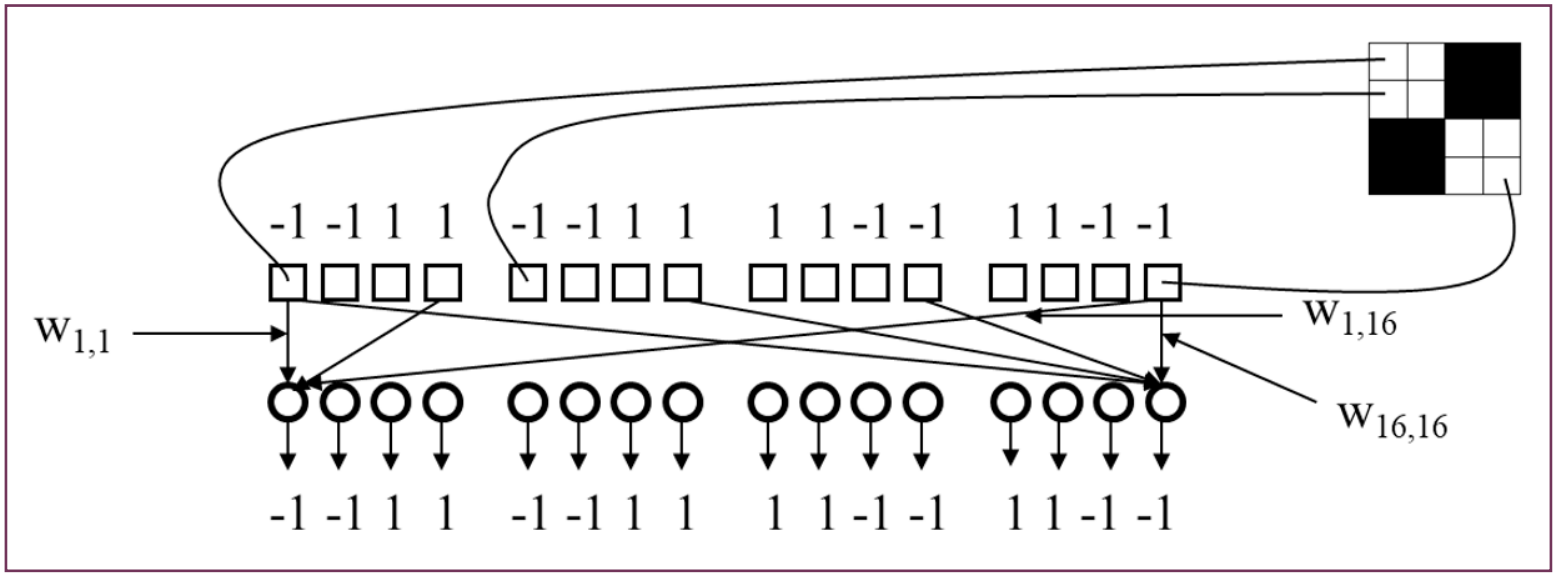
La red Hopfield consta de una única capa. Fue diseñada para manejar imágenes y específicamente es muy capaz de ejecutar tareas de ***autoasociación***. Esta función es la que le da la capacidad al cerebro de reconocer imágenes, aún, aunque estén deterioradas o se representen en forma parcial. Imagine esta situación: a usted le muestran una foto de un ex compañero de la facultad cuando tenía 15 **años y lo desafían**, ahora que tiene 35, a que acierte de quién se trataba; teniendo en cuenta que usted lo conoció años después. Seguramente él había cambiado mucho, probablemente había cambiado su peso y estatura, su peinado puede haber sido muy diferente cuando estaba en la escuela secundaria, su vestimenta tiene diferente estilo. ¿Cómo puede usted haberse dado cuenta de quién se trataba si la imagen no es igual a la que usted tiene ahora de él? La respuesta es: debido a la auto asociación! Con los elementos similares que quedaron de su ex compañero, su cerebro reconstruyó la imagen original para que pudiera reconocerla. Esquematizando, esto es lo que puede hacer la red Hopfield.



A la izquierda de la figura tenemos un juego de imágenes diferentes. En el centro tenemos la red que representamos como una caja, sin importarnos por ahora lo que hay dentro. Procedemos a 'mostrarle' las imágenes de la izquierda a la red hasta que ésta aprenda a reconocerlas. A esta etapa, común a todas las redes se la denomina entrenamiento de la red. Una vez entrenada, la red no sólo reconocerá cada imagen, sino que, además, si se le muestra una imagen deteriorada, como la de la figura inferior, la red reconocerá la imagen original. Para describir las imágenes se dividen en sectores o *pixels* que nos servirán para codificarlas. Por ejemplo, la pantalla de su monitor o televisor está dividida en miles o millones de pixeles y cada uno aporta un punto de la imagen.

La codificación en este ejemplo sencillo se hace asignando un '1' a los pixeles negros y '-1' a los blancos. Para simplificar la explicación, supongamos que se representa una muy pequeña parte de una de las imágenes, como la que se representa en la figura siguiente: la codificación se describe como: X1=(-1,-1,1,1,-1,-1,1,1,1,1,-1,-1,1,1,-1,-1).

Como vemos, la red, para este pequeño sector, consta de una sola capa de neuronas, la capa superior es la de entrada, que como dijimos lo único que hace es distribuir las señales. Por simplicidad se han dibujado solo algunas de las conexiones entre las capas, pero en realidad están todas conectadas entre sí, lo que se denomina ***full connexion***. En la capa de entrada están presentes los números descriptores de la figura, observe que cada bloque de 4 neuronas representa una fila de cuadros de la figura. En la capa de salida se ven los mismos números, lo que indica que a la salida la red informa que se ha detectado la figura 1. También se han indicado las posiciones de algunos de los *pesos* de las neuronas. **Los pesos de la red deben ser calculados previamente a su *entrenamiento***. No entraremos aquí en el álgebra del cálculo, quién esté interesado puede consultar la referencia 1.

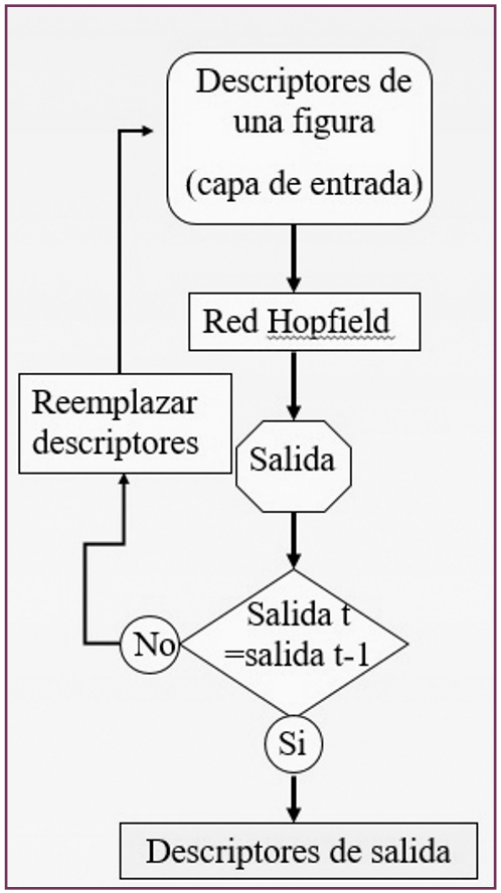


Si llamamos P al número total de figuras cada peso $w_{i,j}$ es:

$$w_{i,j} = \sum_{s=1}^{P} x_{s,j} . x_{s,i} \quad , \quad \text{Para } j \neq i$$

$$w_{i,j} = 0 \quad , \quad \text{Para } j = i$$

'i' y 'j' identifican los pixeles de las figuras. Por ejemplo el peso w1,2 se calcula multi-plicando el primer cuadro ='i' de cada figura por el correspondiente segundo cuadro ='j'.



La *función de transferencia* de las neuronas es la *hard limiter*, pero modificada para trabajar en forma bipolar o sea con valores de salida (-1, +1) en lugar de (0, +1).

Mecánica del cálculo

El diagrama de flujo muestra la mecánica del cálculo. Los descriptores de cualquier imagen codificada se colocan en la capa de entrada y la red producirá un código de salida. Si el código es igual al de entrada, entonces esa es la respuesta (esto ocurrirá si los códigos introducidos coinciden con los de alguna de las figuras originales). Si el código de salida difiere, entonces se introduce esta salida en la entrada, y se obtendrá una nueva salida.

El proceso continúa hasta que dos salidas sucesivas se repiten, entonces esa es la salida definitiva.

Los siguientes tipos de salida pueden ocurrir:

1. **La salida es idéntica a la entrada**

2. **La salida es idéntica a algún otro modelo almacenado**

3. **La salida no es igual a ninguno de los modelos**

4. **Cualquiera de los casos anteriores con los signos invertidos (una imagen en negativo).**

Cuando el número de errores no es aceptable, para reducirlos, hay que aumentar el número de pixels. Por ejemplo, la figura que representa una muy pequeña parte de la imagen, que posee 16 cuadrados o pixels, podría dividirse en 32 cuadros. Esto mejorará el número de errores, pero al mismo tiempo requerirá matrices mucho más grandes y se necesitará mayor potencia de cálculo.

Finalmente, interesa resaltar aquí una particularidad, es la necesidad de que los pesos de las neuronas deben calcularse de antemano, antes de comenzar la mecánica del c***álculo y éstos*** **permanecen fijos**, sin cambiar.

La red ABBAM (Adaptive Bidireccional Associative Memory)

Cuando una red neuronal aprehende a través de la etapa de training, el aprendizaje puede ser ***supervisado*** **ó** ***no supervisado***. Hasta ahora hemos visto sólo el primer tipo, que a su vez, puede dividirse en dos clases:

Aprendizaje supervisado

- Relación intrínseca: Los objetos describen por si mismos un código de identidad.
- Relación arbitraria: A los objetos se les asigna un código de identidad arbitrario.

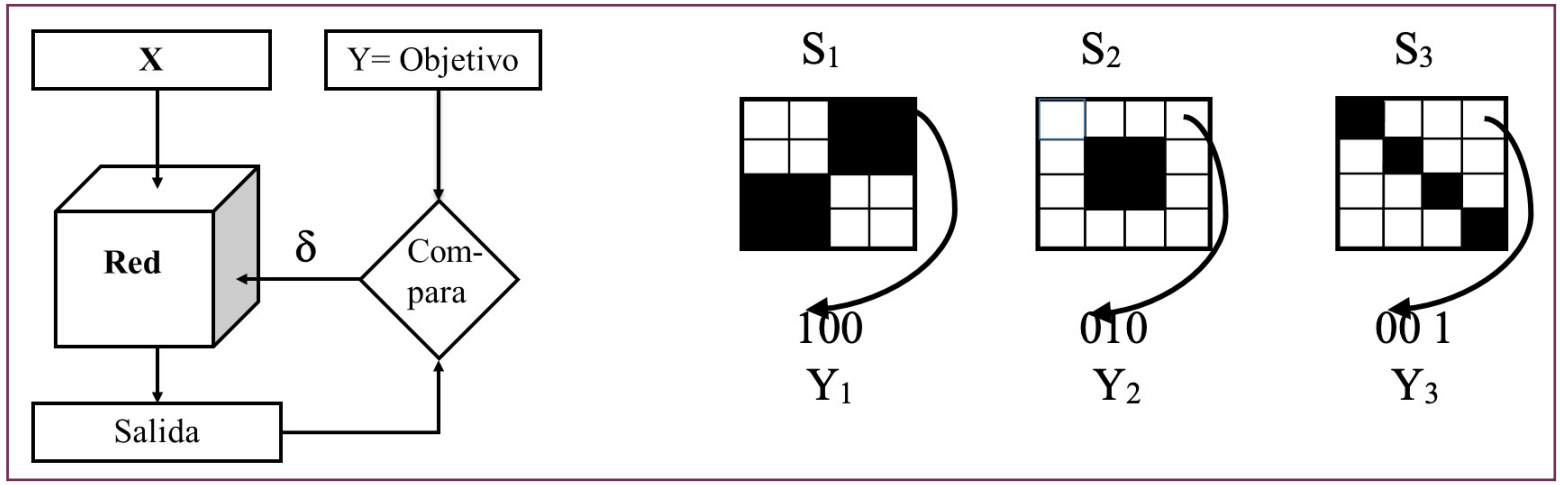
La red Hopfield, como hemos visto, pertenece a la clase de aprendizaje supervisado intrínseco. La red de tipo ABBAM, que veremos ahora, pertenece a la clase de aprendizaje supervisado arbitrario. Este es el tipo más común de aprendizaje, en matemática por ejemplo, donde nosotros asignamos símbolos arbitrarios a los números, las operaciones, las relaciones, etc.

La red ABBAM y otras más avanzadas, trabajan con un mecanismo de cálculo que nos recuerda el de la realimentación de un circuito electrónico para corregir su salida. Como veremos en la figura siguiente, un vector X describe un objeto que debe identificarse con un código que llamaremos 'el objetivo' y designaremos como Y.

Veremos que la operación básica para calcular este sistema es:

$$\mathbf{Y=X.W}$$

donde W es una matriz de pesos. Veremos ahora como se calculan Y, W y X.



Por ejemplo, la figura s_1 tendrá la identificación Y_1=100 y a la s_2 corresponderá Y_2=010. Obsérvese que las W3 figuras quedan siempre identificadas con 2 ceros y un 1, la posición del 1 identifica la figura. Esta forma de identificación se llama ***distribuída***. Por facilidad de cálculo trabajaremos en el modo bipolar. Para ello asignamos el valor –1 a los ceros que identifican a $\mathbf{Y}_s$ mientras que los 'unos' quedan como están. Asimismo, asignamos el valor –1 a los cuadros blancos de

la figura y 1 a los negros. Ejemplificaremos el cálculo de los pesos para componer la matriz **W**. Esta matriz tiene ahora una dimensión de 16x3=nX_s, ¡mucho menor que la matriz de la red Hopfield! Que necesitaría una matriz de 21*21=441. En los pesos se combinan la descripción intrínseca de las, p, figuras $\mathbf{s_1,s_2,s_3}$ con el código $\mathbf{Y_{s,j}}$ asignado mediante la fórmula general:

s≡figura i≡descriptor

$$W^0_{n,i} = \sum_{s=1}^{p} X_{s,i}.Y_{s,j}$$

Por ejemplo, el peso $W^0_{4,1}=X_{1,4}.Y_{1,1}+X_{2,4}.Y_{2,1}+X_{3,4}.Y_{3,1}=1.1+(-1).(-1)+(-1).(-1)=3$

La posición $W^0_{4,1}$ corresponde entonces al cuadro de la primera fila y **cuarta columna de cada figura** y al **primer dígito de cada $\mathbf{Y_s}$** como muestran las flechas de la figura.

La arquitectura de la red ABBAM es muy similar a la de la red Hopfield, excepto en que el número de neuronas de salida es ahora solamente 3 (el número de dígitos de **Ys**). En nuestro caso habrá 16 neuronas de entrada (4x4 pixeles) para introducir los vectores **X**.

La función de transferencia de las neuronas es $hl(Net_j)=sign(Net_j)$ y la señal de entrada, para cada figura, es:

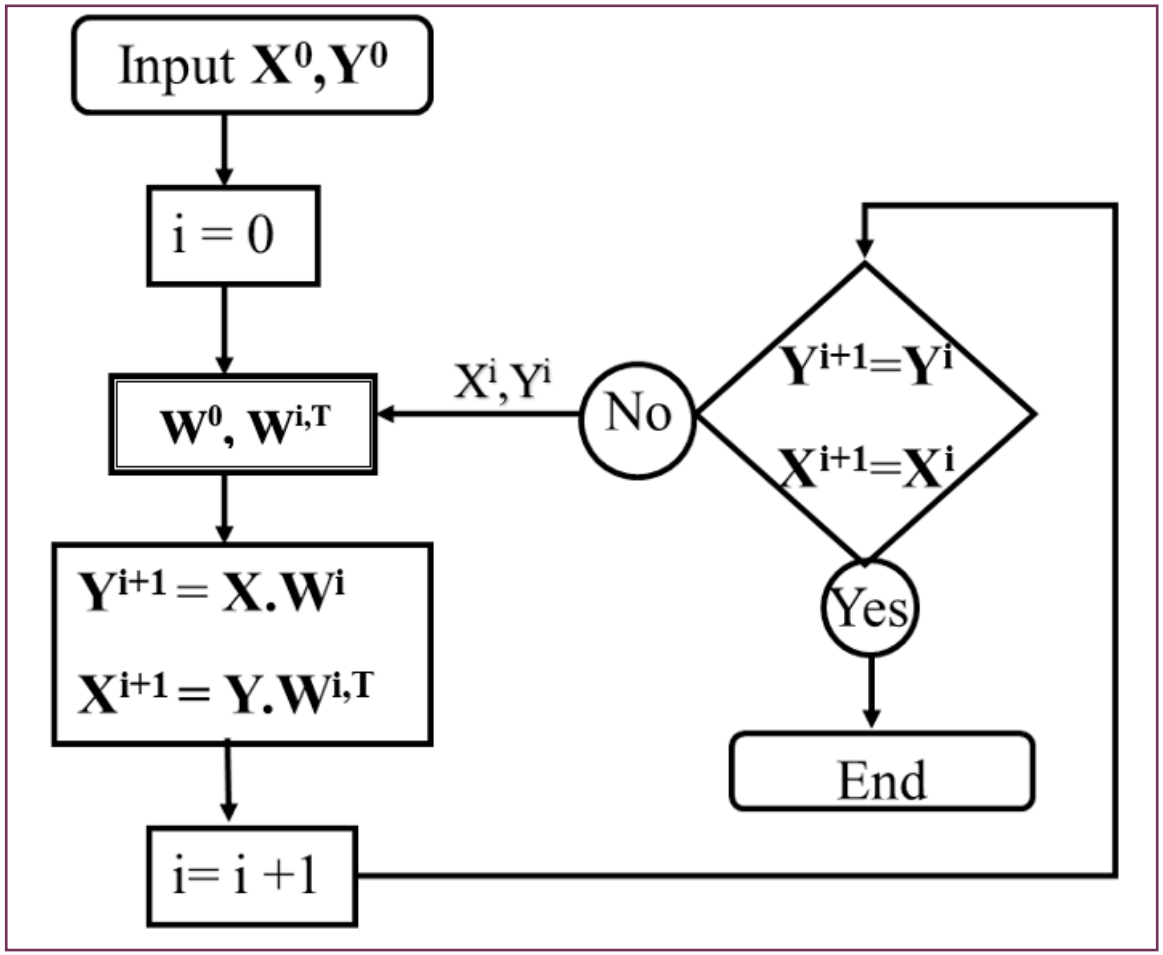
$$\mathbf{Net_j} = \sum_{\mathbf{i=1}}^{\mathbf{m}} \mathbf{w_{j,i}.x_i}$$

Donde m es la dimensión de las figuras, en este caso 16 pixels.

Las ventajas de esta red son dos: por un lado, trabaja con matrices mucho más chicas (mayor potencia de cálculo) y por otro, el **ajuste automático y evolutivo de los pesos** significan un paso importante en la evolución de las redes.

La mecánica de cálculo se ve en el diagrama de flujo siguiente.

¡Note que ahora la matriz de pesos W cambia (se ajusta) en cada ciclo!



1. Un par de datos se presenta a la red. Ciclo i=0.

2.Entran los pesos iniciales a la red W^0.

3. Se calcula el valor de X e Y en el ciclo siguiente (i+1). La T en Wi,T, significa matriz transpuesta y expresa la generación de un nuevo set de pesos, en reemplazo del anterior.

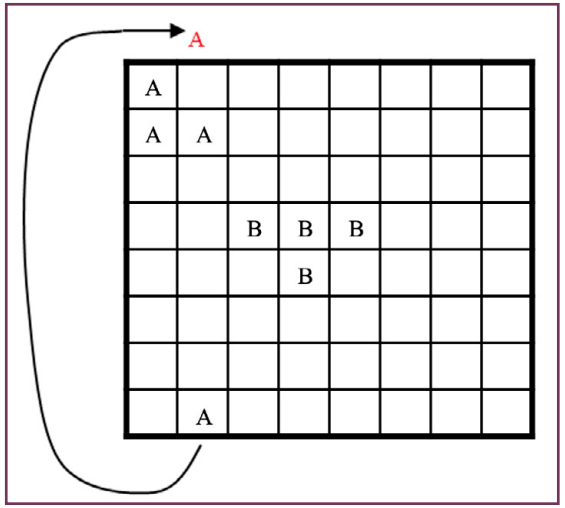
4. Se compara la nueva salida con la anterior. Si es igual termina la identificación de la imagen. Si no es igual, con los nuevos valores del par (Y^{i+1},X^{i+1}) se calculan nuevos pesos y se repite el ciclo hasta que haya coincidencia.

La capacidad de esta red es superior a la Hopfield porque no solo puede detectar autoasociación sino también **heteroasociación**. Esto significa que la red puede identificar que la imagen que se le presenta es combinación de 2 de ellas.

Como antes, los posibles errores de cálculo pueden disminuirse aumentando el número de pixeles en las figuras. Los detalles del cálculo y del ejemplo de heteroasociación pueden consultarse en la referencia 1.

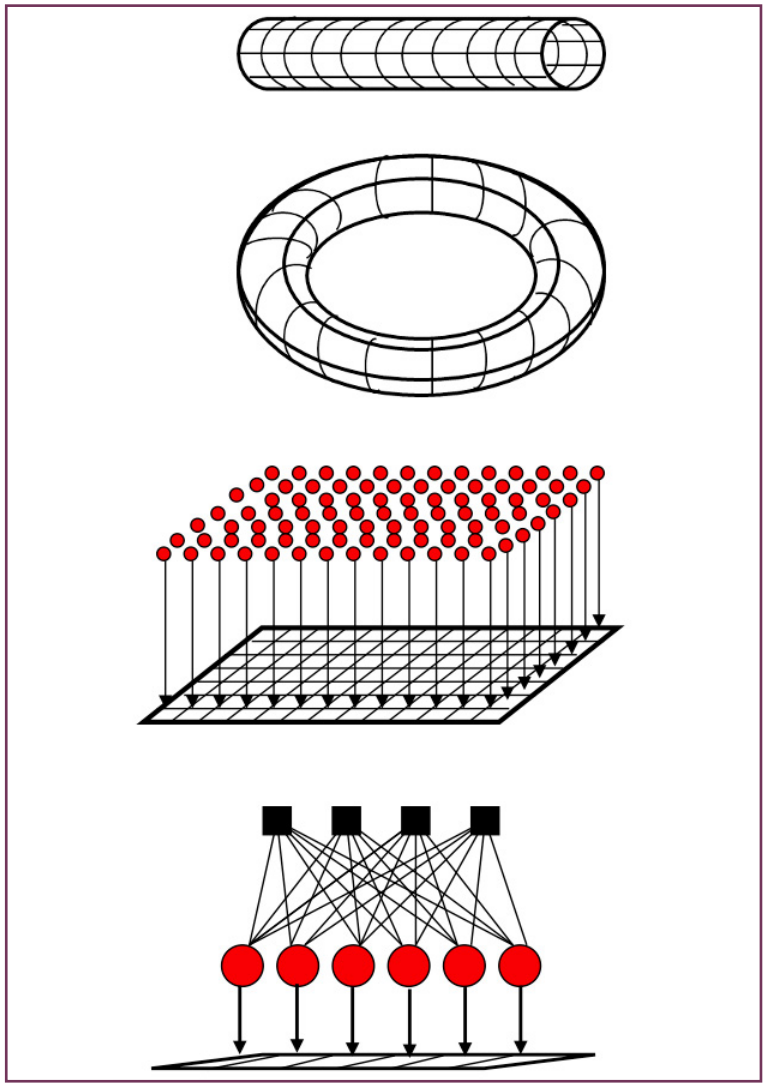
Los Mapas Auto-organizados (self-organizing maps 'SOM') y la Red Kohonen

Este tipo de red es de particular interés porque tiene muchas aplicaciones en el campo de la química y las ciencias en general (Ref 1). Es también una red de una sola capa de neuronas y más avanzada que las anteriores. Su principal característica es que el aprendizaje durante la etapa de entrenamiento es del tipo ***no supervisado***. Esto significa que no le indicaremos a la red la relación que existe entre cada objeto y la respuesta correcta (sea ésta conocida o no) sino que la red misma encontrará las asociaciones entre los objetos. Si volvemos a los ejemplos anteriores para identificar figuras, ahora no describiríamos esas figuras mediante un algoritmo o asignándole un código, sino que la red se encargaría de encontrar las semejanzas y diferencias entre las figuras. Sin embargo, nos alejaremos de las figuras para resolver problemas muchos más cercanos a nosotros. Por ejemplo: imaginemos que tenemos un conjunto grande de objetos que están descriptos por varias de sus propiedades; como podrían ser un conjunto de muestras de aguas descriptas por sus parámetros fisicoquímicos, o un lote de objetos de un producto industrial descrito por sus propiedades adecuadas y parámetros de control (densidad, resistencia a la luz UV, tiempo de degradación, etc., etc.). En ambos casos podríamos desear saber cuáles objetos son muy semejantes entre sí y cuántos grupos diferentes hay. En el primer caso nos puede interesar saber si las aguas son de río, de mar, potables, de deshielo, de lluvia, etc. En el segundo caso puede interesarnos clasificar las muestras como provenientes de productos de primera clase, segunda, importados de A, importados de B, etc. Observe que en los dos ejemplos estamos haciendo una **clasificación** de objetos descriptos por muchas variables, o sea una *clasificación de objetos multivariables*.



Los SOM producen una importante reducción dimensional para la interpretación de los problemas. Su 'forma' de clasificar es ubicar los objetos sobre una superficie cuadriculada, tal como muestra la figura, cuanto más semejantes son dos objetos entre sí, más cercanamente se ubicarán, incluso dentro del mismo casillero. Se suele decir entonces que los SOM establecen una relación topológica entre los objetos.

En la figura los objetos ‘A’ son más similares entre sí que respecto de cualquier objeto ‘B’ y lo mismo puede decirse de los ‘B’. Sin embargo, por ser nuestra ‘grilla’ una figura limitada, un objeto como el A caería fuera del plano aunque tuviese características similares. Para evitar este problema unimos los bordes del plano para formar un cilindro, el cual podemos recorrer sin encontrar límites. Entonces, un objeto ‘A’ como el de la última línea quedaría unido correctamente al grupo.



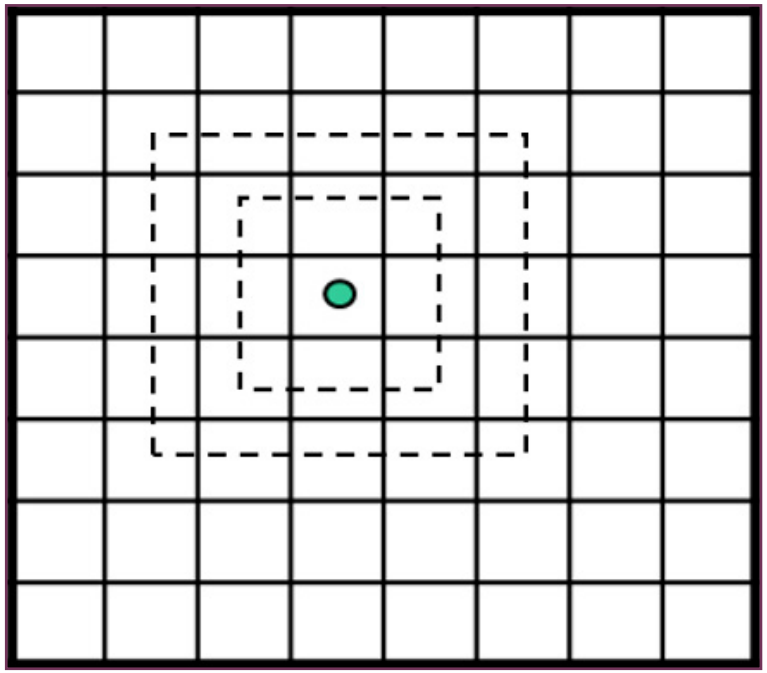
Por supuesto, nos quedan aún las bases del cilindro sin unir, si hacemos esto obtenemos una superficie toroidal que ahora sí podemos recorrer en cualquier dirección sin límite alguno.

Veremos ahora la arquitectura de la red y los aspectos del cálculo para poder estar en condiciones de aplicarla a nuestros problemas. En la figura de la izquierda, cada cuadro de la grilla representa una única salida de una neurona, por lo tanto, tendremos en nuestra red tantas neuronas como cuadros tenga la grilla. En ésta se han representado las neuronas como puntos rojos y por simplicidad se han dibujado sólo las conexiones de dos de los lados de la grilla. Recordemos que a su vez cada neurona está conectada a todas las de la capa de entrada. Un objeto que se presenta a la capa de entrada está descrito por los parámetros que lo caracterizan, cada neurona tiene entonces un número de entradas igual al número de parámetros (o variables) que describen a los objetos. La figura inferior muestra a 6 neuronas recibiendo información de un sistema (o problema) descrito por 4 variables.

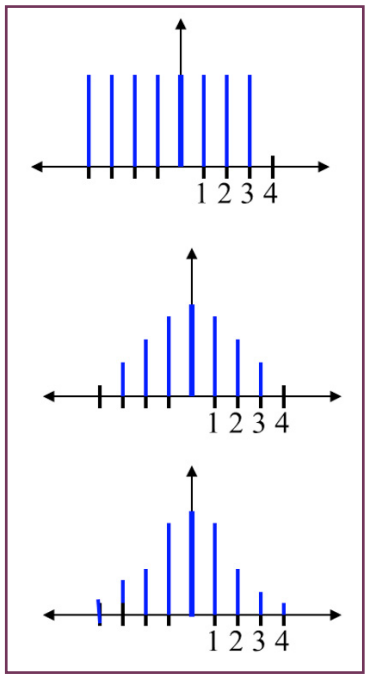
Cuando un objeto es presentado a la red habrá una única neurona que responda y las demás darán una salida nula. La salida de la neurona que ha respondido indica la posición del objeto (sus coordenadas) en la red. La neurona que responde con una señal no nula se selecciona mediante ***aprendizaje competitivo***. El modo más sencillo de hacer esto sería por ejemplo seleccionar la neurona cuyo *Net*=W.S sea el mayor de todos (caso 1). Todas las salidas restantes se anulan; a este procedimiento se lo llama 'la ganadora se queda con todo'. Otro modo muy útil de elegir la ganadora es buscar aquella neurona cuyos pesos sean los más parecidos a los parámetros de entrada de un objeto **s** (caso 2). Si designamos con j a las neuronas y con i a las señales de entrada (o parámetros del objeto), el algoritmo que describe a la neurona ganadora C para el caso 2 es:

$$C_{Ganadora} = mín\left[\sum_{i=1}^{m}\left(x_{s,i} - w_{j,i}\right)^2\right]$$

La suma cuadrática se calcula para cada una de las, m neuronas de la grilla y la de menor valor es la que emitirá la señal de salida.



Volvamos ahora a la grilla original, que por simplicidad está representada en forma rectangular, pero que ya sabemos que puede ser parte de un toroide. Como se ha dicho, la posición relativa de los objetos en la grilla indica su similitud. Una casilla cualquiera como la de la figura tiene 'vecindades', marcadas con líneas punteadas, que son casillas que guardan la misma distancia respecto de la central. Tenemos entonces 'primera vecindad' para la menor distancia (casilla contigua), 2a vecindad para la siguiente, 3a, etc. Una vez que se seleccionó la neurona ganadora C, sus pesos se corrigen en forma proporcional al valor de la señal de entrada. Pero no sólo se corrigen los pesos de C, sino también los de las neuronas vecinas, en proporción a sus distancias respecto de C con una función **Dc,v**, donde v es la vecindad. La intención de esta corrección es que toda la zona cercana a C tenga neuronas con características similares. Las formas de esta corrección pueden ser distintas según la conveniencia y se muestran en la figura siguiente:



Los números en cada figura indican las vecindades respecto de la neurona C. La primera figura muestra que los pesos de todas las neuronas, hasta la vecindad 3, se corregirán en una misma extensión. Esto no significa que los pesos de todas ellas serán iguales porque sus valores iniciales son diferentes. La segunda figura muestra que la neurona C es la que se corrige más fuertemente y luego la corrección disminuye en intensidad hasta la vecindad 3. La figura 4 muestra una corrección de tipo Gausiana.

Las correcciones son diferentes según que se haya elegido el modo ganador caso 1 ó caso 2.

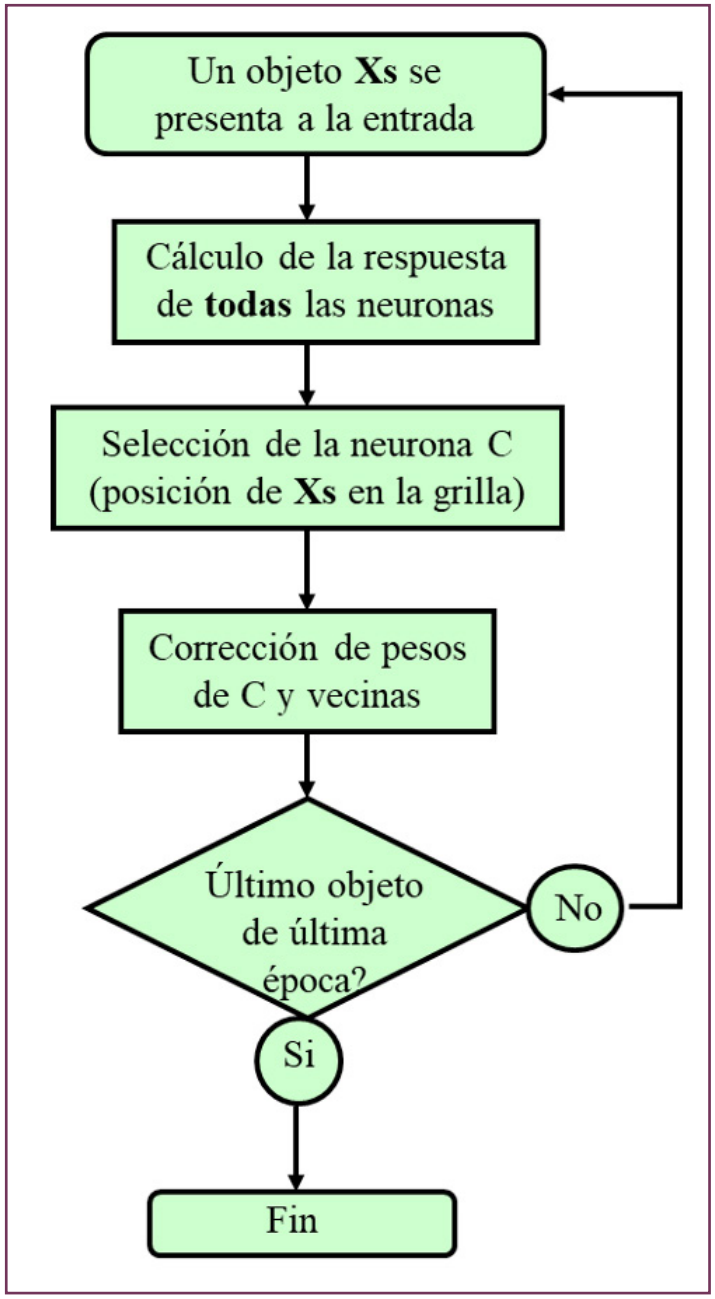
Caso 1

$$\mathbf{w}_{j,i}^{t+1} = \mathbf{w}_{j,i}^{t} + \eta(t).\mathbf{D}_{c,j}.\left(1 - \mathbf{x}_i \mathbf{w}_{j,i}^{t}\right)$$

Caso 2

$$\mathbf{w}_{j,i}^{t+1} = \mathbf{w}_{j,i}^{t} + \eta(t).\mathbf{D}_{c,j}.\left(\mathbf{x}_i - \mathbf{w}_{j,i}^{t}\right)$$

El subíndice j identifica la neurona y el i la señal de entrada(dendrita). El supraíndice t es el ciclo de tiempo. Vemos que los nuevos pesos a t+1 son una función de los anteriores al tiempo t. Los pesos iniciales a t=0 se establecen al azar con valores inferiores a 1. La función η(t) es un coeficiente que va decreciendo la intensidad de la corrección a medida que el cálculo evoluciona. Esta función comienza con un valor máximo no mayor a 1 y decrece linealmente hasta un valor mínimo. Ambos valores se estiman por prueba y error para cada problema particular. La mecánica del cálculo puede verse en el diagrama siguiente:



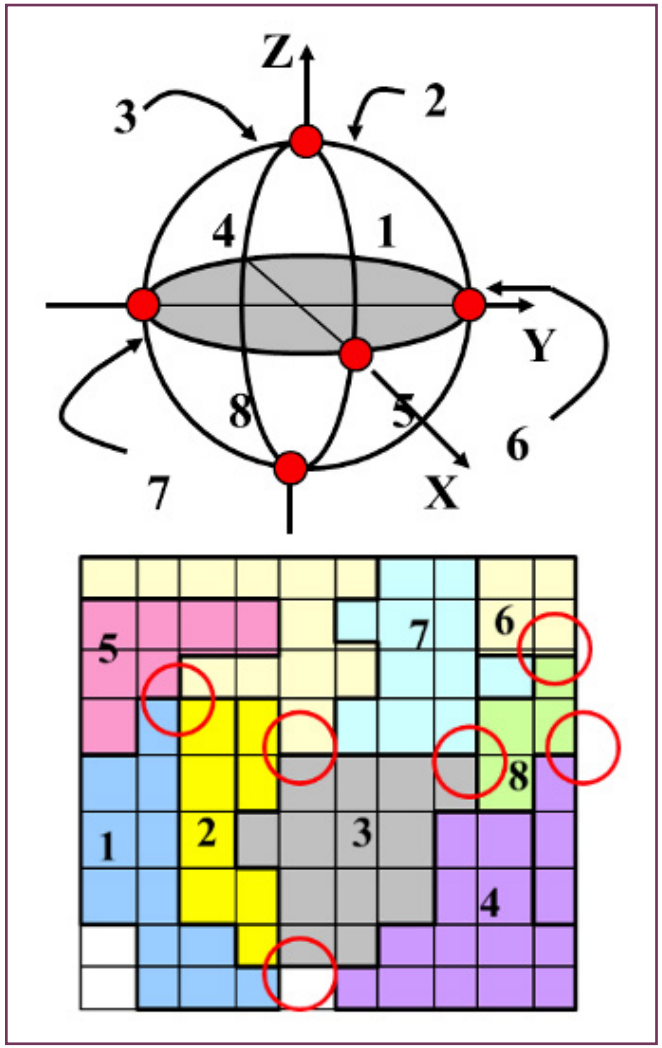
Observe que la táctica del método consiste en a) introducir un objeto y ubicarlo, según el criterio elegido, por la neurona ganadora. b) Corregir los pesos 'de la región' para que todos los objetos similares caigan en la misma área. c) Introducir el próximo objeto y repetir el procedimiento.

Una vez que el lote de datos ha sido pasado por la red, se dice que se ha completado una 'época'. Para que la red converja hasta lograr una clasificación aceptable es necesario procesar varias épocas, el número de éstas depende de la complejidad del problema.

Si comparamos este método de clasificación con el de clusters o componentes principales, la gran diferencia con los SOM es que estos pueden tratar problemas lineales o no lineales, haciendo posibles clasificaciones que no pueden obtenerse por las otras dos vías.

Ejemplos de aplicaciones

1-Ejemplos simulados



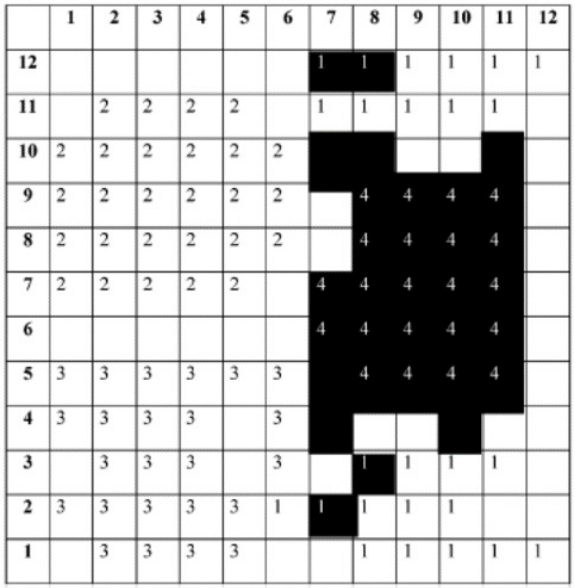
1.1-Descripción de una esfera en un plano

Supongamos que tenemos una esfera dividida en 8 casquetes esféricos iguales. Ponemos sobre la superficie de la esfera puntos al azar, más de 1000 por ejemplo. Observe que, si la esfera es de radio 1, todos los puntos de su superficie serán combinaciones de valores entre –1 y +1 de los ejes X, Y y Z según el radio. Por ejemplo, para el casquete 1 todos los valores serán positivos, pero para el 2, Y y Z son positivos, pero X es negativa. De modo que sabemos de antemano a que casquete pertenece cada punto. Ahora clasificamos los puntos sin darle a la red la información sobre el casquete al que ellos pertenecen. Obtendremos una figura como la inferior, donde todos los puntos caen en sus respectivas áreas (casquetes) y además cada área se vincula con otras correspondientes a sus vecinas y formando los polos. presentes en la esfera original.

Los círculos rojos marcan el punto de unión de 4 casquetes, por ejemplo, el círculo inferior indica la unión de los casquetes 1, 2, 3 y 4, en la figura tridimensional.

Tabla 1. Distancia entre Centroides

	C1	C2	C3	C4
C1	0			
C2	1.828	0		
C3	1.652	1.749	0	
C4	1.417	1.706	2.238	0



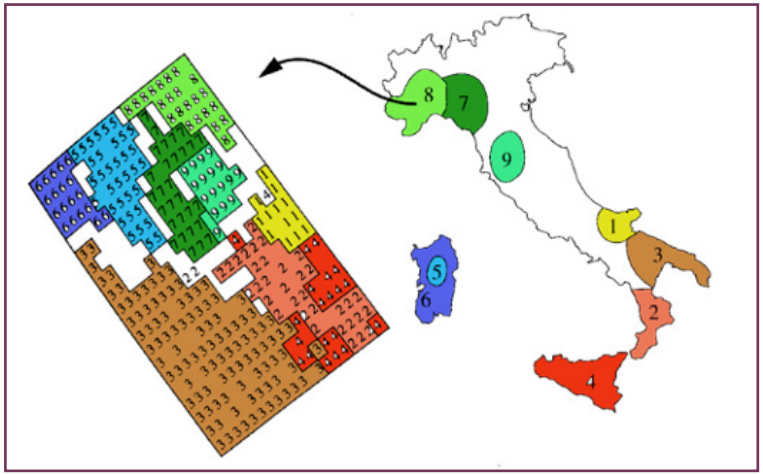
1.2- Clusterización

En este ejemplo (Ref. 7) se han generado 4 clusters con 400 objetos que dependen de 5 variables. En un caso real, estos objetos podrían ser muestras de distintas clases de aguas caracterizadas por 5 variables.

2. La distancia entre clusters puede verse en la tabla 1. Cuando estos objetos son clasificados con la red aparecen perfectamente discriminados en 4 areas, tal como muestra la figura. Observe que los objetos de la clase 1 están unidos por los bordes superior e inferior porque se ha trabajado con una red toroidal.

Ejemplos reales

2.1 Autentificación de Alimentos



En este caso se tiene un lote de 572 muestras de aceite de oliva de 9 regiones de Italia (Ref. 8), las cuales han sido analizadas en el contenido de 8 ácidos grasos (ver regiones y ácidos en las tablas). 250 muestras se utilizaron para la etapa de entrenamiento y 322 para la de predicción. Se utilizó una red Kohonen de 20x20 celdas y los resultados se muestran en la figura. Observe que las regiones adyacentes en el mapa de Italia, también lo son en el mapa de Kohonen. ¿Cómo se explica esto?

Los 8 ácidos grasos:	
Palmítico	Eicosenoico
Palmitoleico	Oleico
Esteárico	Linoleico
Araquídico	Linolenico

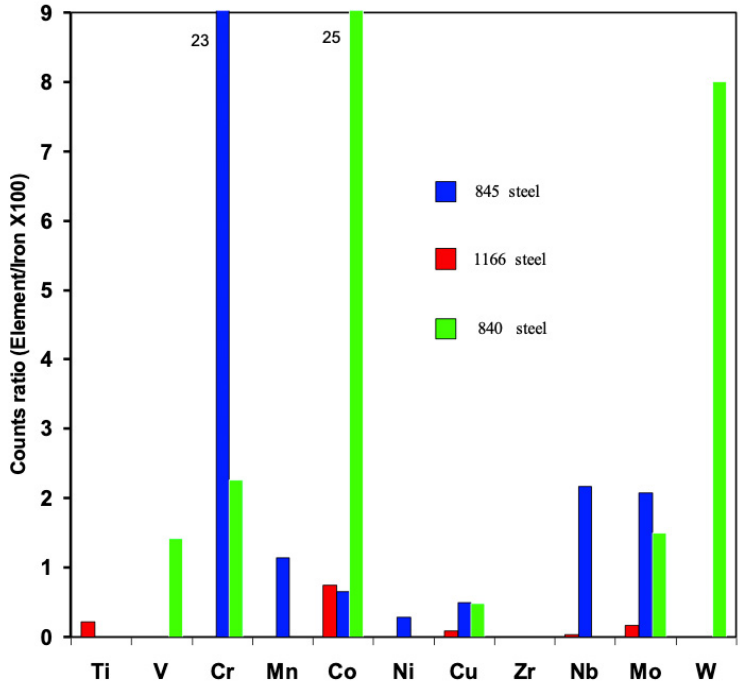
Regiones de Italia y Número de muestras		
1	Apulia Norte	25
2	Calabria	56
3	Apulia Sur	206
4	Sicilia	36
5	Sardenia Interior	65
6	Sardenia Costera	33
7	Liguria Este	50
8	Ligurioa Oeste	50
9	Umbria	51

Lo que ocurre se explica por lo que se llama **'las variables subyacentes del sistema'**. Sencillamente, el parecido climático, el del suelo o de los productos utilizados en la producción, la cultura técnica de los productores, etc., hace que las zonas vecinas tengan un contenido de ácidos grasos más similar entre sí que entre las zonas alejadas.

2.2- Clasificación automática de aceros desde espectros FRX-DE

Características del problema (Ref. 9):

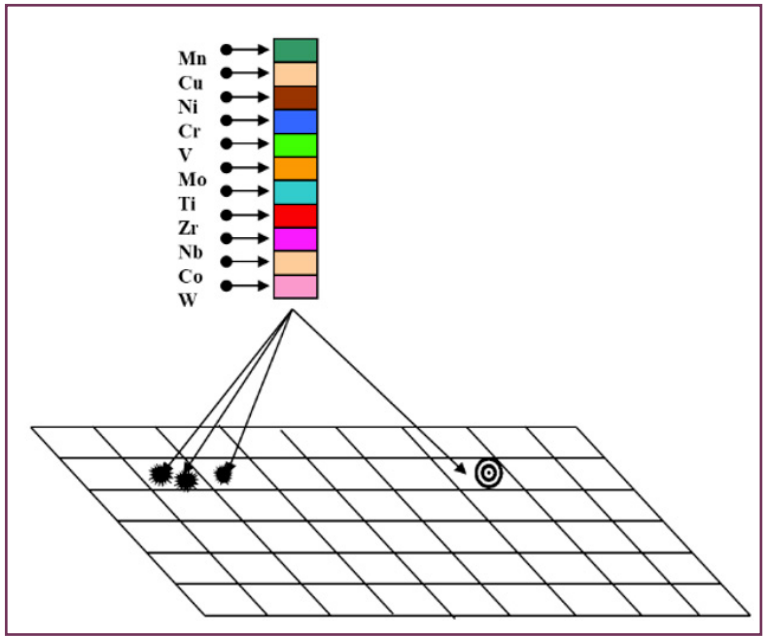
Un método clásico de analizar aceros para conocer su tipo e identificación era hacer un análisis químico de sus componentes, luego buscar en tablas de composición de aceros para identificarlo. Esto llevaba considerable tiempo, de más de un día de trabajo.



3 ejemplos de espectros, uno de cada familia.

La fluorescencia de rayos x (FRX) dispersiva en energía, permite identificar un elemento químico a través de la irradiación de rayos x sobre una muestra, que puede ser sólida o líquida. La muestra irradiada desplaza electrones internos de los átomos y cuando el elemento repone estos electrones con otros de sus capas superiores emite radiación que es típica de **cada elemento**, lo que permite su identificación.

Para probar la capacidad de esta técnica se irradiaron muestras sólidas de 19 aceros estándar pertenecientes a 3 familias diferentes de aceros: dulces, rápidos e inoxidables. Se obtuvieron 190 espectros que contenían 11 intensidades relativas de elementos componentes de los aceros.

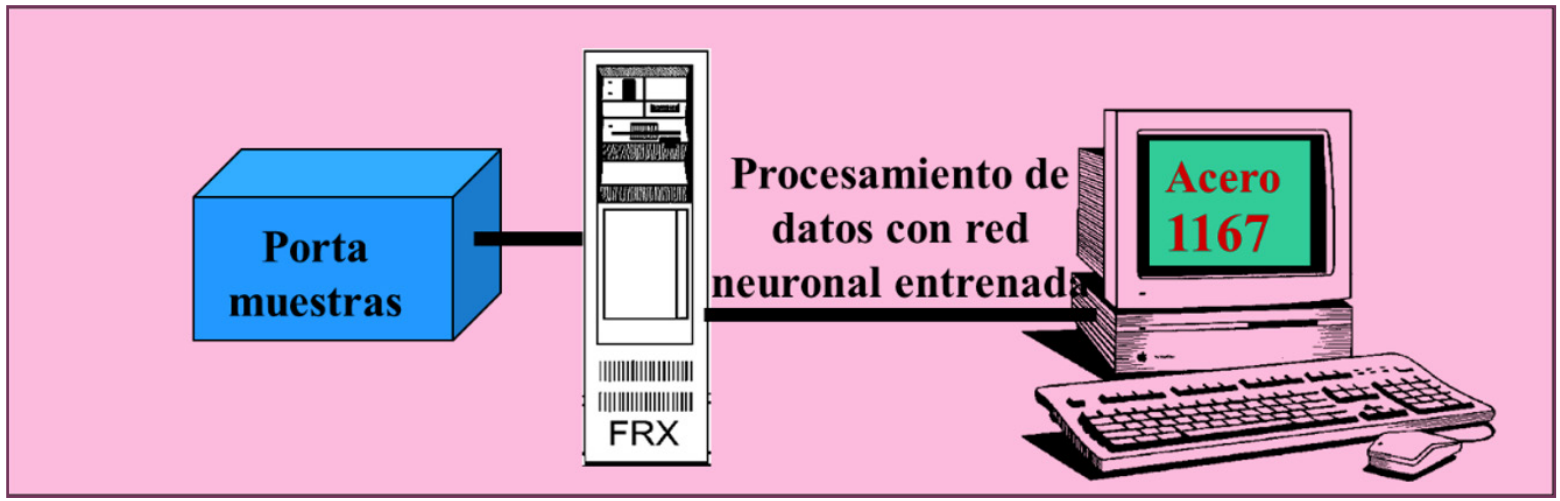


Se entrenó una red Kohonen para clasificar los 19 aceros dentro de sus respectivas familias. La clasificación fue verificada con la etapa de predicción y lotes de prueba. Finalmente, la etapa de predicción, cuyo procesamiento es mucho más rápido que la etapa de entrenamiento, fue incorporada a una computadora.

	838		840				
		836				847	
	839		841			846	
					849	848	
837		845					1161
				1165	1164	1163	
850			1166	1165		1167	
					1162		

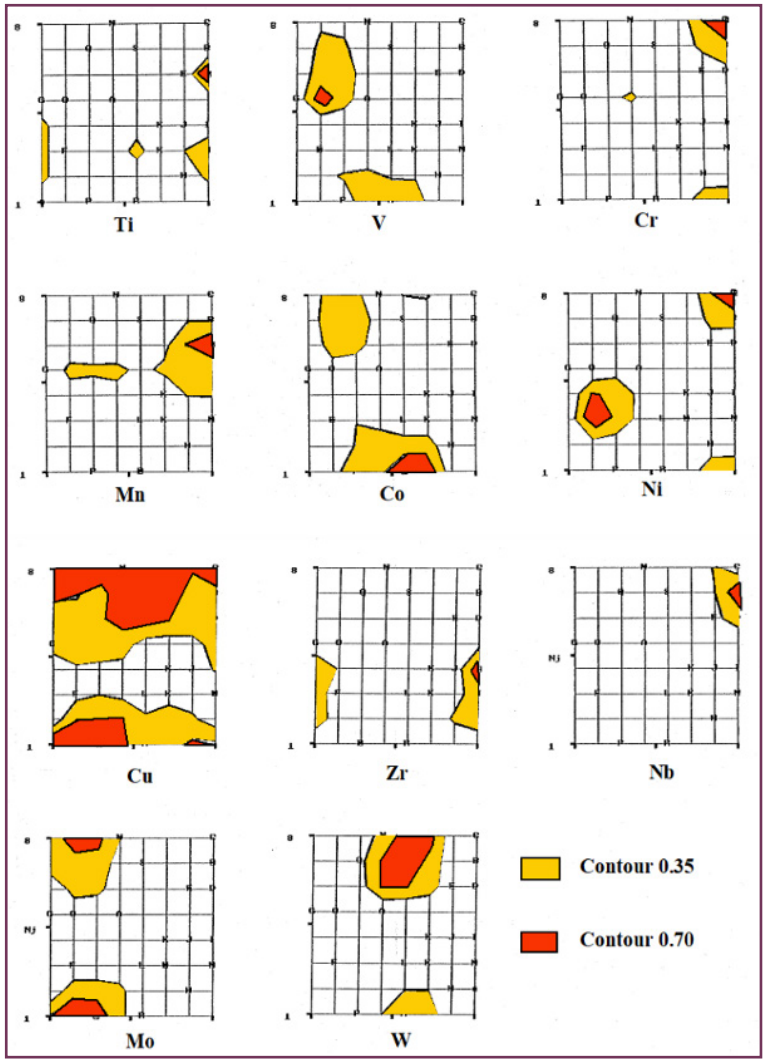
Si se posee un equipo de FRX que pueda transferir los espectros a una computadora se obtiene un sistema que automáticamente podrá entregar resultados en minutos como muestra la figura inferior. La etapa más lenta del análisis será la del corte y pulido de la pieza de muestra.

Red Kohonen de 8x8. Cada color identifica una familia de aceros y dentro de ellas están los números que identifican a cada uno de ellos.



Otro aspecto que debe tomarse en consideración es que no es necesario ningún cálculo de concentraciones de los elementos de la muestra ya que el análisis se torna absolutamente cualitativo, ya que lo que se quiere saber es cuál es el acero de la muestra.

Hay un aspecto más en el cuál la red Kohonen tiene ventajas particulares, se trata de la posibilidad de analizar la influencia de cada variable en el proceso de clasificación una vez optimizados las etapas de entrenamiento y predicción. Recordemos que los pesos en la red, para cada variable, se ajustan automáticamente durante la etapa de entrenamiento.



Si representamos los pesos de cada variable sobre el SOM obtenemos los gráficos de contorno para cada variable. El mapa único y total, llamado TOP MAP es la superposición de todos estos pesos, que ubica a cada objeto en su lugar.

Estos contornos parciales sirven para analizar la eficiencia de cada variable en el proceso de clasificación. Una variable como la del Cu no es muy útil porque su contorno abarca gran parte del mapa. En cambio, el resto de las variables ocupan áreas restringidas que además no se superponen mucho entre sí. Probablemente el Cu podría eliminarse de entre las variables sin afectar demasiado la eficiencia de la clasificación.

2.3 Aplicación de Redes Kohonen a Matrices Incompletas (*Missing data*)

Muchas veces, cuando se colectan grandes cantidades de datos de muchas variables, algunas mediciones pueden fallar, ya sea por malfuncionamiento de un equipo, mal registro de un dato o degradación de la muestra, entre otros posibles. En el caso de campañas ambientales, por ejemplo, cuando una medición ha fallado, ésta no se puede repetir dado que la medición depende de las particulares condiciones del momento de la medición.

Recordemos que cada muestra objeto es un vector conteniendo una serie de variables. Sería importante que, cuando alguna variable es faltante, no se tenga que remover e ignorar el vector completo. Usualmente, ante esta situación los especialistas en el tema tratan de completar la vacante faltante haciendo estimaciones extraíbles desde el resto de la planilla de datos, lo que conlleva un tiempo y esfuerzo considerable además del error de la estimación.

Una estrategia desarrollada con redes de Kohonen (Ref. 10) soluciona este problema. La figura 5.3 muestra la planilla de una hipotéticamente incompleta base de 44 muestras y 7 variables. Los datos faltantes se representan en casillas grises. Si fuera deseable retener los datos en los cuales todos los vectores están completos, entonces sólo 26 de las 44 muestras serían utilizables, lo que representa una pérdida del 41% de los datos. Si en cambio se desea conservar sólo las variables cuyas columnas están completas, sólo 2 de ellas están en estas condiciones, X3 y X5, lo que representa una pérdida del 71% de la información.

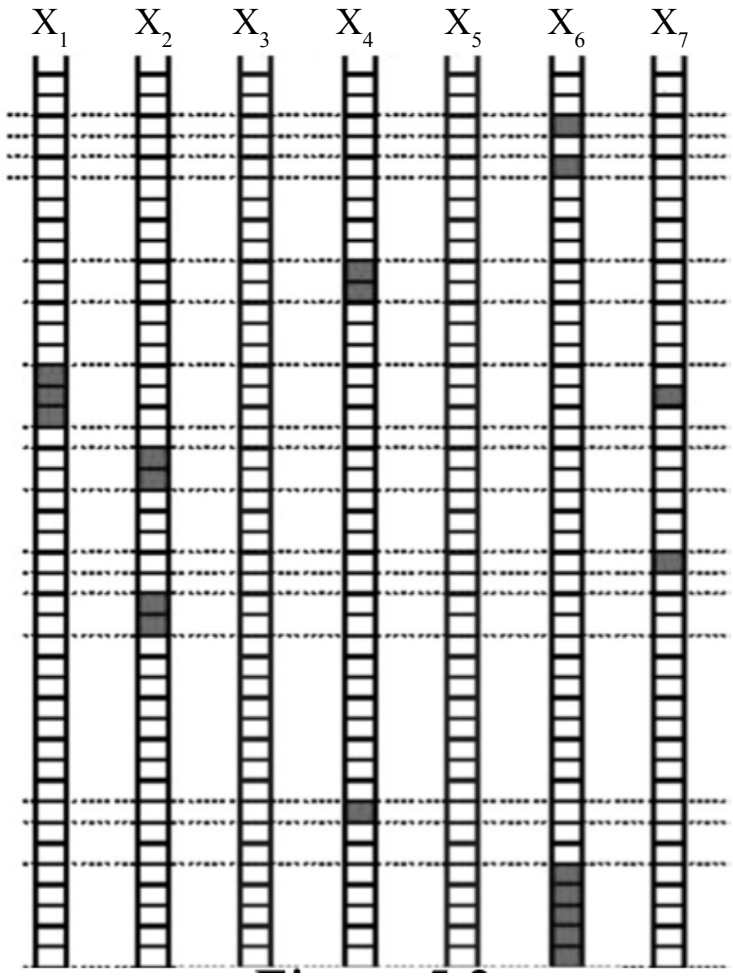


Figura 5.3

Veamos la aplicación a una base de datos real de 267 objetos y 26 variables ambientales. La base fue reducida quitado al azar 50 datos, que consideraremos como los faltantes. Dos especialistas en el tema, trabajando independientemente y sin conocer los 50 datos faltantes estimaron estos valores desde la base de datos reducida. Estos datos se compararon con el procedimiento estimado por la red Kohonen. La red fue modificada ligeramente para tomar en cuenta las distancias entre objetos cuando un vector tiene datos faltantes (ver referencia para los detalles del cálculo).

Tabla2- Datos artificialmente faltantes y evaluados por expertos y red Kohonen								
Dato faltante	Experto N°1	Error	Experto N°2	Error	SOM trainig	Error	SOM predicción	Error
0	0.32	0.291	0.013	0.015	0.056	0.027	0.068	0.046
0.144	0.051	0.093	0.079	0.066	0.069	0.076	0.069	0.038
0.037	0.036	0.001	0.019	0.018	0.048	0.011	0.048	0.008
0.042	0.9	0.859	0.324	0.282	0.58	0.538	0	0.594
0.002	0.005	0.003	0	0.001	0.019	0.017	0.001	0.004
0.144	0.3	0.156	0.033	0.111	0.217	0.073	0.217	0.03
0	0.198	0.198	0.213	0.213	0	0	0.115	0.15
0.08	0.082	0.002	0.059	0.021	0.223	0.142	0.056	0.035
0.489	0.52	0.031	0.28	0.209	0.441	0.048	0.359	0.189
0	0	0	0	0	0.018	0.018	0.001	0.003
0.015	0.042	0.027	0.021	0.006	0.029	0.014	0.019	0.006
0.348	0.36	0.012	0.427	0.079	0.362	0.013	0.362	0.037
0.074	0.135	0.061	0.109	0.035	0.099	0.026	0.032	0.023
0	0.03	0.03	0.014	0.014	0.001	0.001	0	0
0.142	0.12	0.022	0.079	0.063	0.076	0.066	0.076	0.014
0.009	0.009	0	0.015	0.006	0.009	0	0.016	0.004
0.092	0.02	0.072	0.175	0.083	0.07	0.023	0.069	0.016
0.006	0	0.006	0.036	0.03	0.042	0.036	0.008	0.011
0.06	0.018	0.042	0.074	0.014	0.064	0.004	0.06	0.144
0.027	0.023	0.004	0.035	0.008	0.242	0.215	0.021	0.039
0.063	0.063	0	0	0.063	0.063	0	0.063	0.001
0.041	0.004	0.038	0.176	0.135	0.076	0.034	0.003	0.018
0.124	0.078	0.046	0.226	0.101	0.097	0.027	0.097	0.027
0.032	0.024	0.008	0.17	0.138	0.031	0.001	0.136	0.005
0.158	0.148	0.01	0.239	0.08	0.077	0.081	0.176	0.128
0.254	0.1	0.154	0.184	0.07	0.186	0.069	0.088	0.006
0	0.001	0.001	0.002	0.002	0.065	0.065	0.007	0.042
0	0	0	0	0	0.017	0.017	0	0
0.044	0.044	0	0.015	0.029	0.044	0	0.061	0.006
0.086	0.035	0.051	0.024	0.062	0.05	0.036	0.023	0.022
0.129	0.109	0.02	0.165	0.035	0.108	0.021	0.079	0.009
0	0	0	0	0	0	0	0.001	0.004
0.037	0.012	0.025	0.014	0.023	0.042	0.006	0.017	0.008
0.078	0.12	0.042	0.286	0.208	0.101	0.023	0.134	0.009
0.047	0.063	0.016	0.029	0.018	0.04	0.007	0.061	0.036
0.026	0.027	0.001	0.028	0.001	0.064	0.038	0.018	0.008
0	0	0	0	0	0	0	0	0.001
0.029	0.057	0.028	0.052	0.023	0.103	0.074	0.065	0.04
0	0	0	0.157	0.157	0	0	0.031	0.002
0.07	0.08	0.01	0.044	0.026	0.097	0.026	0.081	0.009
0.049	0.06	0.011	0.024	0.025	0.066	0.017	0.01	0.006
0	0	0	0	0	0.023	0.023	0.035	0.009
0.149	0.148	0.001	0.133	0.016	0.116	0.033	0.098	0.029
0.011	0.022	0.011	0.138	0.127	0.251	0.239	0.041	0.037
0.989	0.989	0	0.375	0.614	0.402	0.587	0.388	0.271
0.051	0.078	0.027	0.079	0.028	0.081	0.03	0.068	0.021
0.295	0.262	0.033	0.261	0.033	0.202	0.092	0.218	0.014
0.08	0.125	0.045	0.669	0.589	0.112	0.032	0.058	0.002
0.082	0.055	0.027	0.12	0.038	0.088	0.006	0.008	0.072
0.082	0.12	0.038	0.076	0.006	0.057	0.025	0.134	0.002
	MAE 0	.051 0		.078 0		.059 0		0.045
	RMSE	0.139		0.149		0.129		0.1

La Tabla 2 muestra los resultados de cuatro procedimientos de estimación y sus respectivos errores. Al pie, los dos primeros corresponden a los expertos, el tercero consiste en entrenar la red con la base de datos completa e imputar los datos faltantes con los pesos correspondientes a las neuronas. Por ejemplo, si la tercera variable del objeto 10 en la base de datos es faltante, el peso de la neurona de la variable 3 del objeto 10 en la red entrenada es el imputado al dato faltante.

El cuarto procedimiento consiste en entrenar la red solamente con los vectores completos de la base y luego introducir los vectores incompletos en la etapa de predicción. Los valores imputados deben tomarse igual que en el caso anterior.

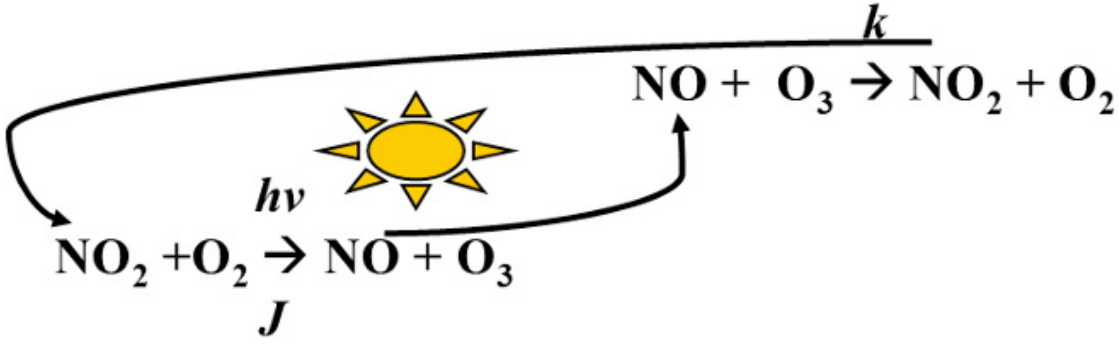
Al pie de la Tabla 2 aparecen los errores de cada método. MAE y RMSE son el error promedio de la columna y el error medio cuadrático, respectivamente. La prueba de la diferencia entre medias demuestra que estas no son significativas. La conclusión es que la red Kohonen es al menos tan eficiente como el trabajo de los expertos, con la ventaja de que el procedimiento es muchísimo más rápido.

Hay que tener en cuenta que la eficiencia del método dependerá de la relación del número de vectores incompletos respecto del total y del número de parámetros incompletos en cada vector. Estos detalles figuran en la bibliografía.

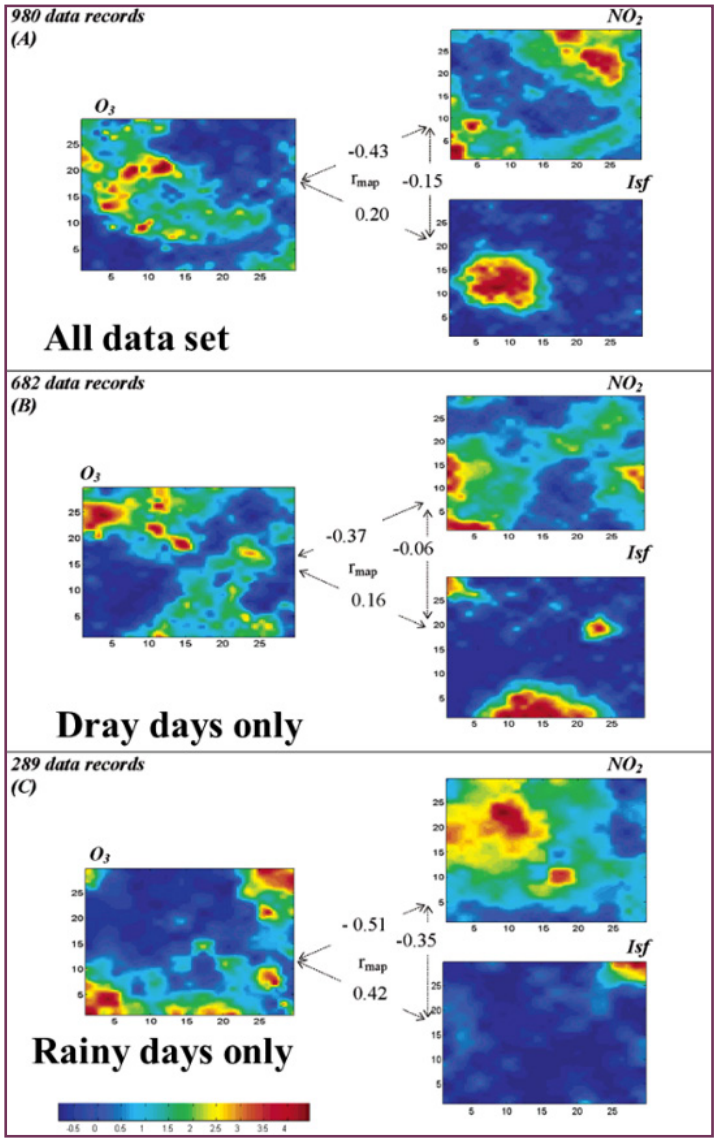
La red Kohonen tiene muchas utilidades diferentes, la mayoría aplicadas a problemas de clasificación, pero otras veces pueden aplicarse a problemas muy distintos. Por ejemplo, el siguiente, aplicado al análisis de una campaña de monitoreo de aire (Ref. 11).

2.4 Campaña de Monitoreo de Aire

En la atmósfera se produce una reacción química entre el óxido nítrico (NO), el dióxido de nitrógeno (NO_2) y el Ozono. La figura inferior ilustra el mecanismo de esta reacción, cuya velocidad de reacción de izquierda a derecha está controlada por la constante química k y la velocidad de la reacción inversa está controlada por la constante J, el factor de irradiación solar. Obviamente, el equilibrio de la reacción, llamado "Estado foto estacionario" depende de las horas del día y del clima.



La red fue programada con las siguientes condiciones:
Red 30x30 rectangular. Training:600 épocas. 14 variables (6 químicas, 8 meteorológicas)



La figura superior muestra cómo han sido las concentraciones de Ozono, Dióxido de nitrógeno y factor de irradiación solar para distintas condiciones climáticas. Observe la escala de color que va desde colores azul obscuro (concentraciones bajas) a rojo (concentraciones altas) y la diferencia entre días secos y lluviosos para tres variables importantes.

Otro ejemplo no aplicado a tareas de clasificación es el control terapéutico de un tratamiento médico (Ref. 12,13). En este trabajo se aplica la técnica de "*missing data*" pero no se lo explica aquí por ser un problema muy específico de la técnica médica.

SOM 3D: Red Kohonen Tridimensional

Un modo aún más eficiente de organizar un mapa de Kohonen es disponerlo en tres dimensiones. Tendremos entonces 26 celdas vecinas a cada celda unidad en la primera vecindad. La celda cúbica central tendría 26 celdas vecinas en la primera vecindad: 6 vecinas por cada cara + 12 vecinas por cada arista + 8 vecinas por cada córner. Pero teniendo en cuenta las vecindades, v, el número total crece rápidamente con v como muestra la siguiente tabla:

Tabla 3. Relación entre el número de vecinas a una celdas según las distintas geometrías y vecindades

Vecindad	Rrectangular		Hexagonal		Cúbica	
v	(8.v)	Total	(6.v)	Total	(2v+1)3-(2v-1)3	Total
1	8	8	6	6	26	26
2	16	24	12	18	98	124
3	24	48	18	36	218	342
4	32	80	24	60	386	728
5	40	120	30	90	602	1330

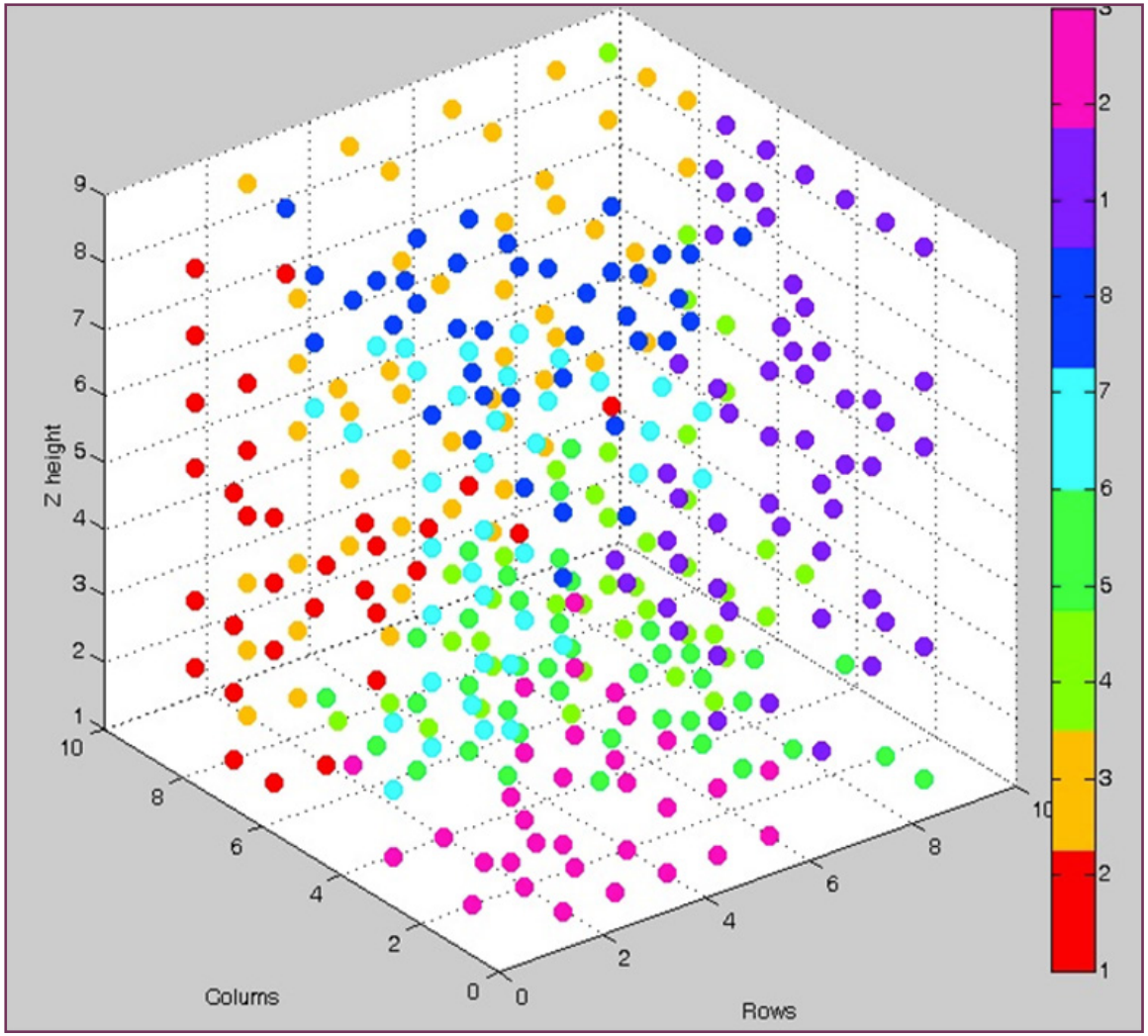
Debido al creciente número de celdas con las vecindades el programa de cálculo debe optimizar la velocidad de procesamiento. Además, el "top map" ahora es un cubo, los objetos aquí son más difíciles de observar que en un rectángulo, por lo tanto, deben proveerse rutinas para la observación del cubo.

También se debe tomar en cuenta la continuidad de los bordes del cubo, así como en 2D el rectángulo se transforma en un toroide, en 3d el cubo se transforma en una geometría hipertoroidea 4D.

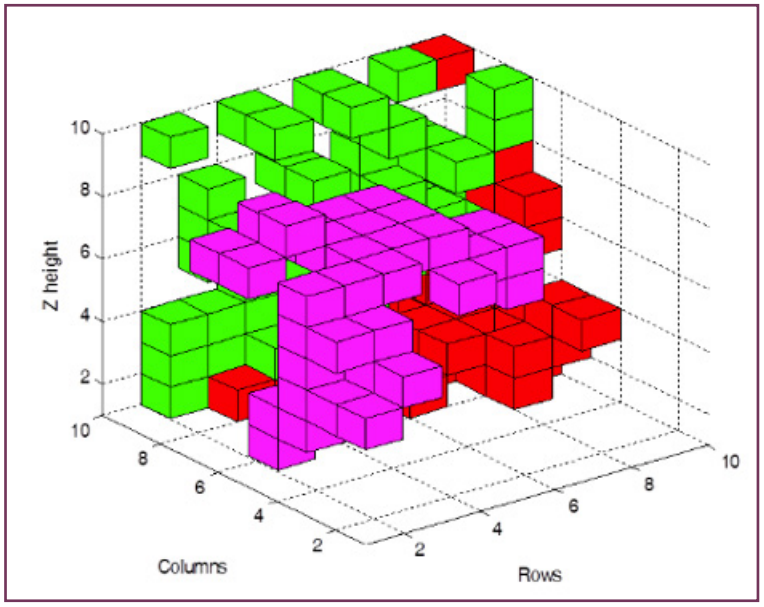
La salida de texto del programa tiene las mismas características que la de 2D.

Para mostrar la eficiencia de esta geometría utilizaremos el mismo ejemplo 1.1 de este capítulo para la representación de una esfera. La distribución espacial de los casquetes de la esfera y todos los datos de entrada a la red son los mismos.

La figura siguiente muestra la salida de una red de 8 celdas por lado, o sea 512 celdas en total. En ella, 8 colores diferencian los casquetes de la esfera, cada punto representa una celda.



En la etapa de entrenamiento, el número de épocas para la geometría cúbica fue 25% menor al de la celda rectangular. Los errores de mezclado, esto es los errores contabilizados cuando un objeto de una clase cae en una celda de otra clase, fueron reducidos en 26,4% en la etapa de entrenamiento y más del 50% en la etapa de predicción.



Para mostrar la separación de las celdas en una forma más clara, en la nueva figura siguiente se muestra el símil del "top map" para las clases 2,3 y 5. En ella se puede ver el espacio que ocupa cada clase individualmente y también las 8 clases a la vez. Pero en este último caso, debido a que los cubos no son transparentes, se superponen, y para verlos todos es necesario rotar la figura.

Los tipos de gráficos de salida que muestran los resultados durante el proceso de entrenamiento, tales como el error de mezclado son en todo similares a aquellos que se presentan para mapas rectangulares (Ref.14).

Referencias

1. Jure Zupan; Johann Gasteiger, Neural Networks in Chemistry and Drug Design. 2nd Edition, Wiley-VCH, Weinheim, 1999.

2. The effect of factor interactions in Plackett-Burman experimental designs. Comparison of BayesianGibbs analysis and genetic algorithms. Magallanes, JF, Olivieri, AC, Chemom. Intell. Lab. Syst. 102 (2010) 8-14.

3. Kennedy, J.; Eberhart, R. (1995). «Particle Swarm Optimization». *Proceedings of IEEE International Conference on Neural Networks* **IV**. pp. 1942-1948. doi:10.1109/ICNN.1995.48896819 de julio de 2011

4. Shi, Y.; Eberhart, R. (1998). «A modified particle swarm optimizer». *Proceedings of IEEE International Conference on Evolutionary Computation*. pp. 69-73.

5. *Waldner, Jean-Baptiste (2008). Nanocomputers and Swarm Intelligence. London: ISTE John Wiley & Sons. p. 225. ISBN 978-1-84704-002-2.*

6. Monmarché Nicolas, Guinand Frédéric and Siarry Patrick (2010). Artificial Ants. Wiley-ISTE. ISBN 978-1-84821-194-0.

7. Magallanes, J.F.; Zupan, J.; Gomez, D.; Reich, S.; Dawidowski, L.; Groselj, N. The Mean Angular Distance among Objects and its Relationships with Kohonen Artificial Neural Networks. (2003) Journal of Chemical Information and Computer Sciences. 43(5):1403-1411.

8. Eigenvector projection and simplified non linear mapping of fatty acid content of Italian olive oil. Michele Forina and Carla Armanino. Annali di chimica 72 (1982), by Socitá Chimica Italiana.

9. Jorge F. Magallanes and Cristina Vazquez Authomatic Clasification of Steels by Processing EDX Spectra with Artifitial Neural Networks. J. Chem. Inf. And Comput. Sci. 38(1998)605-609

10. Laura Folguera, Jure Zupan, Daniel Cicerone, Jorge F. Magallanes. Self-organizing maps for imputation of missing data in incomplete data matrices. Chemometrics and Intelligent Laboratory Systems 143 (2015) 146–151.

11. N Grošelj, J Zupan, S. Reich, L. Davidowski, D. Gomez and J. Magallanes J. Chem. Inf. Comput. Sci. 2004,44,339-346.

12. Jorge Magallanes, Alejandro García-Reiriz, Sara Líberman, Jure Zupan. Kohonen Clasification Applying 'Missing Variables' Criterion to Evaluate the BPA Human-body-concentration Decreasing Profile in Blood-BPA Samples of BNCT Patients. 03-2011. Journal of Chemometrics, **25**,340-348.

13. Alejandro García-Reiriz, Jorge Magallanes, Jure Zupan, Sara Líberman. Applied Radiation and Isotopes. 12-2011. 69,1793-1794. ISSN: 0969-8043.

14. Jorge Magallanes, Ezequiel Morzan. Trabajo presentado en el 9° congreso Argentino de Químicos Analíticos. Río Cuarto, Córdoba. 7 al 10 de noviembre de 2017.

CAPITULO 6

Redes de Más de Una Capa y Algoritmos Genéticos

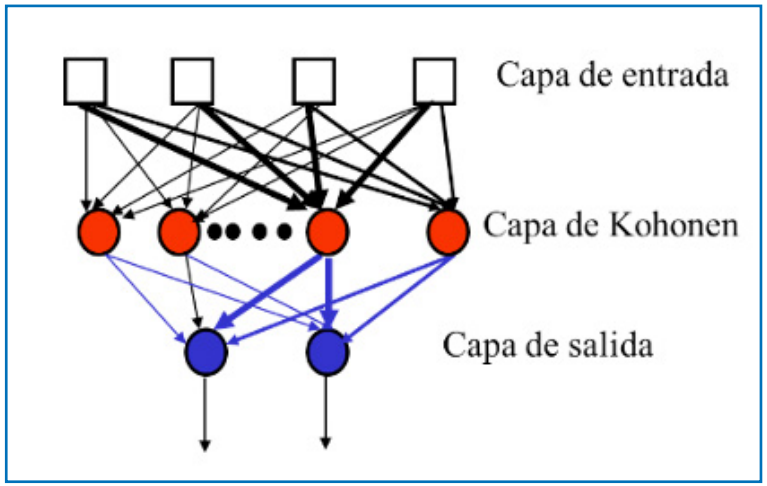
Redes Multicapas

La red de contra-propagación (Counter-Propagation)

Esta red consta de 2 capas y es una extensión de una red Kohonen, sus cualidades, sin embargo, son bien diferentes de esta última (Ref 1). La mayor utilidad de esta red es como 'Look up table' o sea 'buscar información en una tabla'. Imaginemos un proceso complicado de cálculo exacto cuyo proceso computarizado tarda considerable tiempo. Si este cálculo debe usarse repetidas veces o es una subrutina de un programa mayor, entonces tendríamos como resultado un proceso muy lento. En este caso, sería más práctico llevar a cabo **una serie** de cálculos de antemano, entrenar una red de contra-propagación y obtener los resultados del cálculo complejo desde una tabla, lo cual sería mucho más rápido y eficiente. Por ejemplo, cuando calculamos una función trigonométrica o logaritmo, el resultado exacto proviene del cálculo de una serie que, en principio, es infinita. La precisión del cálculo dependerá de cuantos términos de la serie calculemos, de modo que un programa que use repetidamente muchas funciones trigonométricas sería muy lento. Una solución es crear una tabla y usar el ángulo como un índice, luego no tendremos más que ubicar una dirección en lugar de efectuar el cálculo. Y si el cálculo no fuera el de una serie trigonométrica sino otro que depende de múltiples variables, entonces esta red posiblemente fuera el camino más conveniente. Otro ejemplo cotidiano lo tenemos en los juegos animados por computadora: imaginemos la trayectoria que recorre una bola de billar cuando se le ha dado cierto impulso, con determinado 'efecto', que rebota contra las bandas y otras bolas, y a su vez estas también comienzan a desplazarse, etc. Si bien todos estos cálculos pueden ser hechos exactamente por la física, sería imposible hacerlos en tiempo real mientras los observamos en la pantalla. Por lo tanto, puede emplearse una red de contra-propagación para solucionar el problema.

Hay otro tipo de problemas en los cuales esta red puede ser útil. Son aquellos en los cuales el valor que debe ser recuperado es una débil función de la entrada, o dicho de otro modo, cuado puede aceptarse un amplio rango de tolerancia en la respuesta o cuando va a ser usada información corrompida, como se vio en capítulos anteriores.

Esta red no es muy útil para la construcción de modelos, como lo es la ***back propagation***, y no puede dar infinitas respuestas, sino que puede proveer tantas respuestas como hayan sido cargadas durante el entrenamiento. Esto significa que para esta etapa debemos conocer de antemano el resultado de tantos experimentos (ó cálculos) como respuestas queramos.



La capa Kohonen funciona exactamente igual que en la red original. Cuando la neurona ganadora ha sido determinada, se corrigen sus pesos y el de las neuronas de su vecindad, según el método descrito. Pero luego se procede a corregir todos los pesos que conectan a esas neuronas con las de salida, tal como muestra la figura.

El algoritmo para la corrección de pesos de las neuronas de salida es el siguiente:

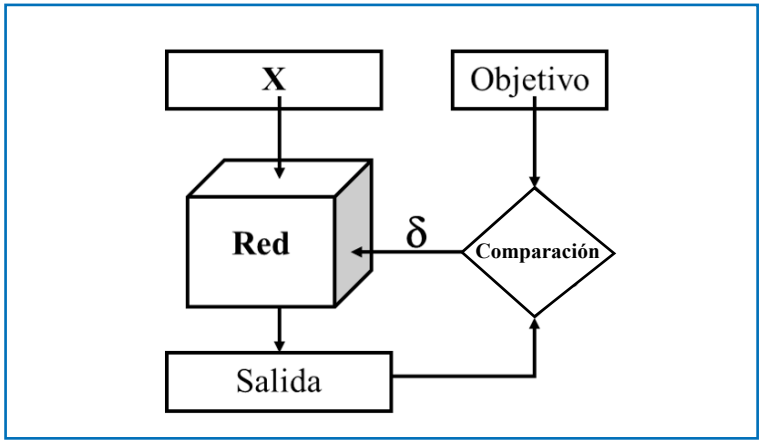
$$c_{ji}^{t+1} = c_{ji}^{t} + \eta(t).D_{cj}.\left(y_i - c_{ji}^{t}\right)$$

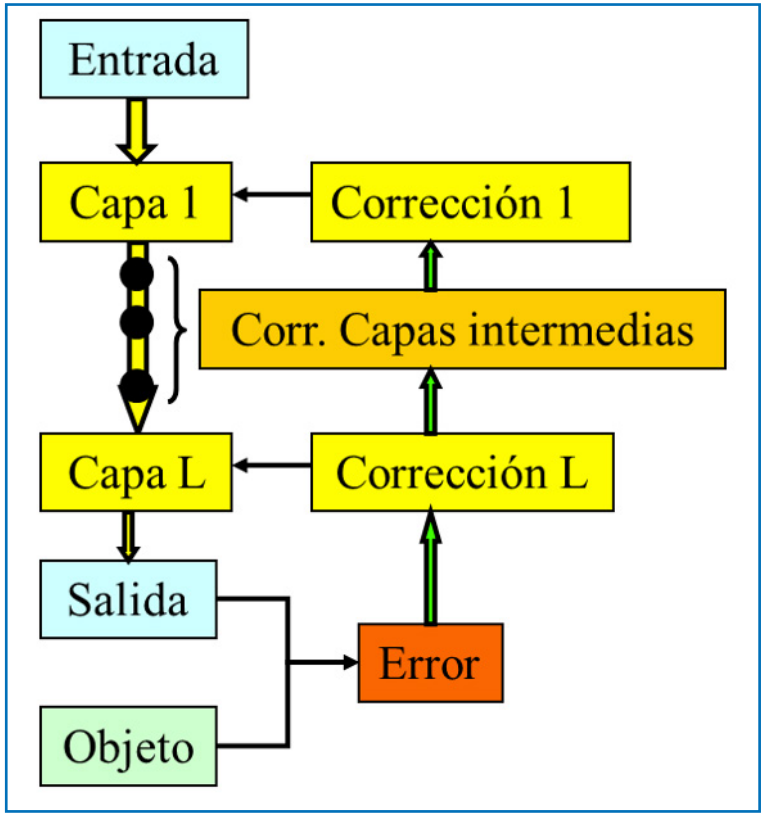
Como antes, c es la neurona ganadora de la capa Kohonen, j es el índice de la neurona de la vecindad que está siendo corregida, i es el índice de todos los pesos que unen la neurona j con la neurona de salida y_i Cada una de las n neuronas de salida representa un componente (una variable) del vector respuesta $\mathbf{Y=(y_1, y_2, ...,y_n)}$.

La red 'Retropropagación de Errores' (Back-Propagation of Errors)

Esta red ofrece muchas potencialidades de uso y por eso es una de las más importantes (Ref. 1). Su aplicabilidad llega a tal punto que un alto porcentaje de problemas son resueltos con ella y muchos usuarios de redes neuronales usan ***back-propagation*** exclusivamente. Sin embargo, es de hacer notar que, para algunas tareas, ésta no es la red más eficiente aunque pueda llevarla a cabo. Por ejemplo, para trabajos de clasificación es mucho más eficiente una red Kohonen, y lo mismo puede decirse de la red de contra propagación para tareas de búsqueda tabuladas.

Esta es una red ideal para trabajos de modelado, entre otros, sus aplicaciones son tan extensas en todos los ámbitos de la ciencia y tecnología que no pueden mencionarse sino a costa de cometer grandes mutilaciones.

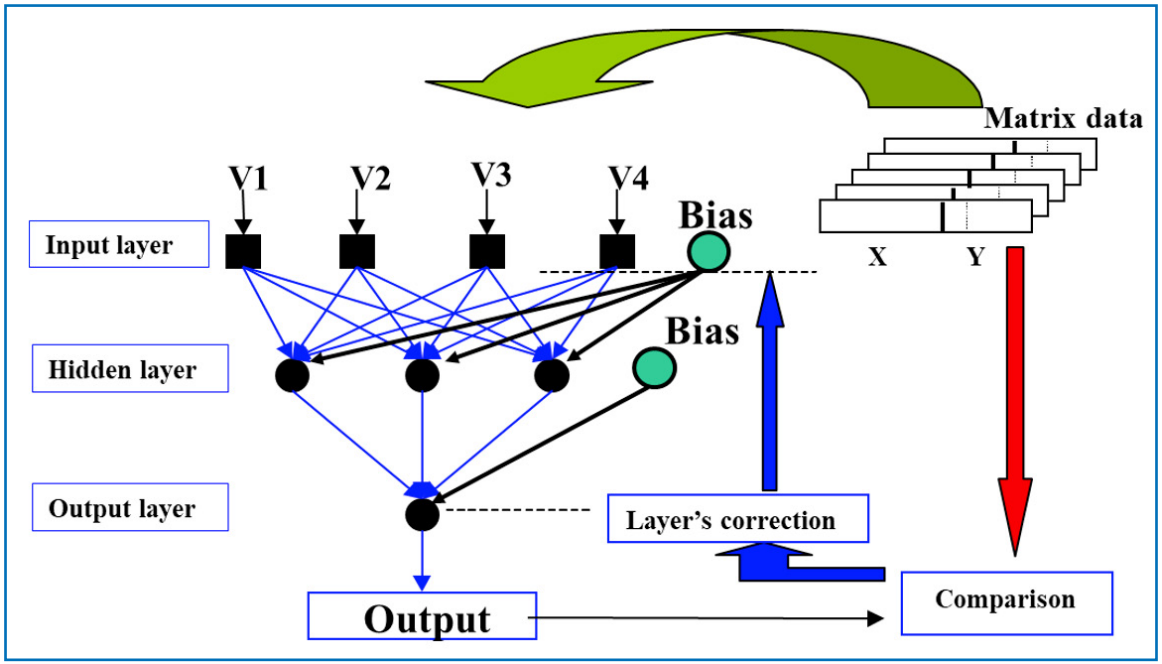




Esta es una red de aprendizaje asistido y su mecánica general de cálculo consiste, como en otros casos, en comparar la salida con la respuesta correcta y autocorregirse a partir de esta diferencia, tal como muestra la figura.

Debido a que esta es una red de multicapas, la autocorrección es más compleja que en casos anteriores porque para la última capa es fácil determinar la diferencia entre la salida y el valor real del objeto pero para las capas más internas no tenemos esa ventaja. La corrección de errores en este caso se fundamenta en la dispersión de los errores producidos en la última capa. Un esquema más detallado de la forma de corrección se ve en la figura siguiente.

Esta red puede tener muchas capas de neuronas pero un mínimo de tres: Una de entrada, una capa escondida (intermedia) y una de salida. El número de capas escondidas puede variarse. La capa de entrada tiene tantas neuronas como variables de entradas tiene el problema. Asimismo, la capa de salida tiene tantas neuronas como respuestas tiene el problema. El número de neuronas en las capas escondidas es variable y deben ajustarse mediante prueba y error.



En el campo de científico, una gran cantidad de problemas de modelado que consisten de un número discreto de variables, pueden resolverse con sólo 3 capas con un número suficiente de neuronas en la capa escondida. Un problema de modelado se resuelve con esta red de la siguiente manera: Un lote de datos con sus correspondientes respuestas se utiliza para entrenar la red. Usualmente esta red requiere un número grande de épocas hasta alcanzar la convergencia. Un segundo lote de datos similar al anterior y que estén dentro del mismo rango de trabajo se utiliza para determinar el error de predicción que se ha alcanzado. Si este error es aceptable, entonces estamos en condiciones de generar una enorme cantidad de respuestas del sistema con sólo simular un conjunto adecuado de experiencias con los valores deseados de las variables. Este conjunto servirá para obtener ***superficies de respuesta*** que nos permiten interpretar el comportamiento del sistema en estudio.

Hay muchos tipos de algoritmos para la corrección de pesos de la red. Aquí, sin entrar en la profundidad del cálculo, mostraremos el que se denomina ***delta rule*** y que consiste en hacer una corrección en los pesos de las neuronas, δ, proporcional al valor de entrada de la variable cada vez que un objeto es introducido en la red, durante la etapa de entrenamiento. La mecánica completa del cálculo sería la siguiente:

1- Se introduce un objeto **X** (x_1, x_2, ...x_m) en la red. A cada uno de los elementos de X los denominaremos **Out^0** (out^0_1, out^0_2,...out^0_m, 1), donde se ha agregado un 1 como bias. El superíndice de Out indica el número de capa, siendo, 0, la capa de entrada, *l*, las intermedias y *last* la última.

2- Propagamos **Out^0** a través de la red obteniendo consecutivamente los **Out^l**. Aquí entran en juego los pesos w^l_{ji} de la capa l y la salida out^{l-1}_i de la capa anterior (que es la entrada de la capa l). La *f* simboliza cualquier función de transferencia, pero aquí ejemplificaremos con la función sigmoidea, donde la sumatoria representa a Net (ver capítulo 5).

$$out^l_j = f\left(\sum_{i=1}^{m} w^l_{ji}.out^{l-1}_i\right)$$

j es el subíndice de la neurona que está siendo corregida, i es el índice de todos los pesos que unen la neurona j con la entrada de la información (proveniente de la capa anterior).

3- Calculamos el factor de corrección, δ^{last}_j para todos los pesos de la capa de salida, utilizando **Out^{last}** y el valor real **Y** para el aprendizaje asistido.

$$\delta^{last}_j = \left(y_j - out^{last}_j\right).out^{last}_j.\left(1 - out^{last}_j\right)$$

4- Se corrigen todos los pesos de la última capa.

$$\Delta w^{last}_{ji} = \eta.\delta^{last}_j.out^{last-1}_i + \mu.\Delta w^{last:t-1}_{ji}$$

Aquí, η, es la velocidad de aprendizaje (*learning rate*) que no tiene el mismo significado que su homónima utilizada en las redes Kohonen. μ es el ***momento*** y tiene un efecto amortiguador para evitar cambios bruscos en los pesos cuando se pasa de un objeto a otro.

5- Se calculan, consecutivamente, los factores de corrección de las capas escondidas desde l=last-1 hasta l=1. r es el número de pesos que tienen las neuronas de la capa l.

$$\delta_j^l = \left(\sum_{k=1}^{r} \delta_k^{l+1} w_{kj}^{l+1} \right) . out_j^l . \left(1 - out_j^l\right)$$

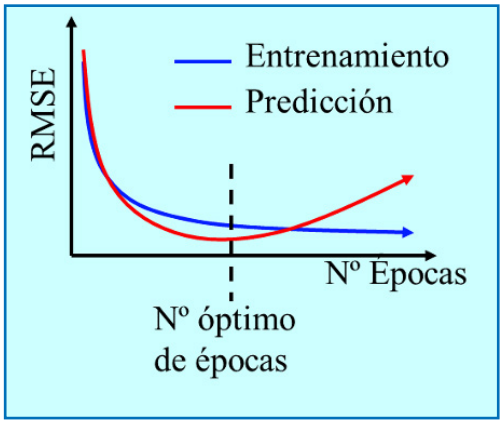
6- Se corrigen todos los pesos de cada capa escondida.

$$\Delta w_{ji}^l = \eta . \delta_j^l . out_i^{l-1} + \mu . \Delta w_{ji}^{l:t-1}$$

7- Cuando se llega a la capa 1, se calcula el Output con f(Net) y se vuelve al punto 1. Se repite el procedimiento 1 a 7 con un nuevo par de entradas **X,Y.**

El control de errores tanto en la etapa de entrenamiento como en la de predicción se realiza calculando el error medio cuadrático (RMSE) entre los valores de salida de la red y las respuestas experimentales. Es importante observar la evolución de estos errores con el número de épocas de cálculo, tal como muestra la figura. El error durante el entrenamiento se aproxima asintóticamente a cero, de modo que al menos en teoría, podríamos achicar este error tanto como quisiéramos.

Pero el error de predicción, después de alcanzar un mínimo, comienza a crecer indefinidamente. El número de épocas de entrenamiento debería estar ubicado en el mínimo de los errores de predicción para evitar el sobre entrenamiento *(overfitting)*. Como se ha señalado, existen muchas variantes de esta red en el tipo de algoritmos de convergencia.



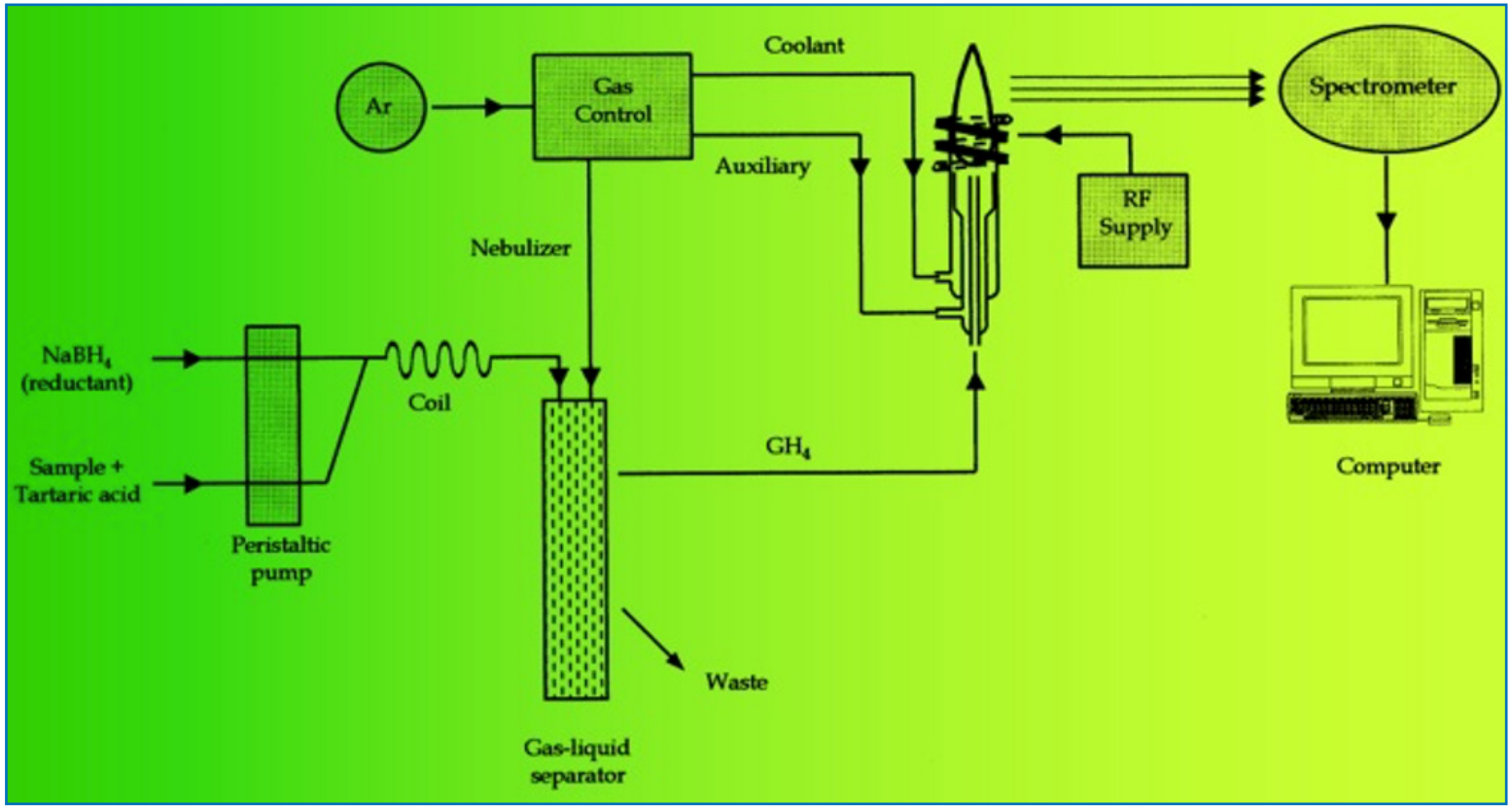
Sin embargo, si se tiene la correcta interpretación del problema que se quiere resolver, las diferencias entre estas variantes serán mínimas. Lo recomendable es dominar bien alguno de los programas de cálculo y lo óptimo es tener un programa propio.

Ejemplos de aplicaciones de la red de retropropagación de errores

Optimización de un sistema químico analítico

Se describirá un ejemplo de modelado y optimización de un sistema químico, pero más allá del problema específicamente químico, este ejemplo puede aplicarse a cualquier sistema instrumental que se desee optimizar.

El objetivo de este trabajo era determinar las condiciones óptimas de trabajo para la determinación de Germanio a niveles de traza con una técnica combinada, Generación de hidruros-ICP-Espectroscopía de emisión atómica (HG-ICP-AES) en muestras ambientales (Ref. 2). Previamente al inicio de este trabajo de optimización, las variables de operación habían sido ajustadas a través del método de una variable a la vez, OVAT. Los factores y respuestas para la optimización se exponen en el cuadro ubicado a la derecha, la figura muestra la disposición experimental del equipo.



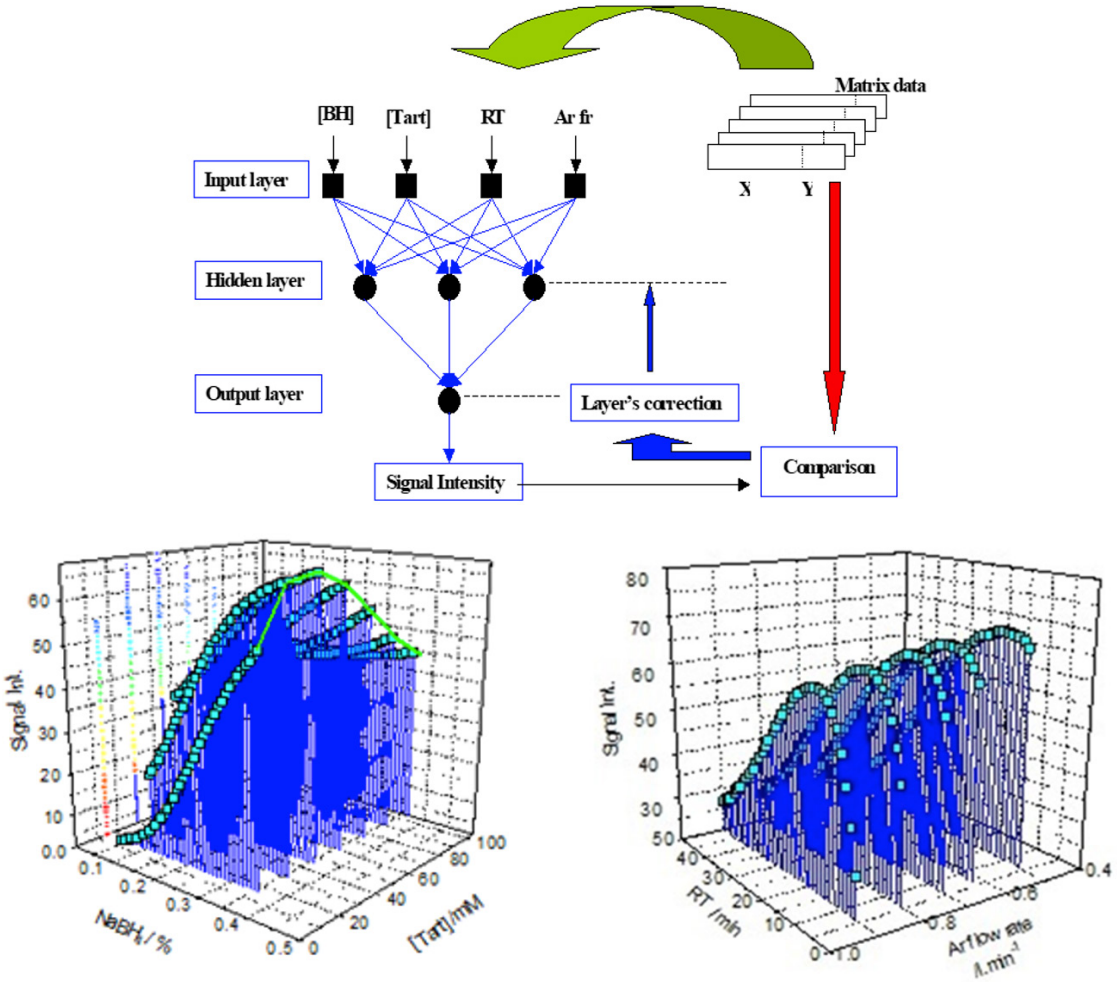
FACTORES:

1-Concentración de borohidruro de sodio ($NaBH_4$), 2-Concentración de ácido Tartárico, 3-Caudal de reactivos, 4-Longitud del tubo reactor (espiral), 5-Flujo de argón.

RESPUESTA:

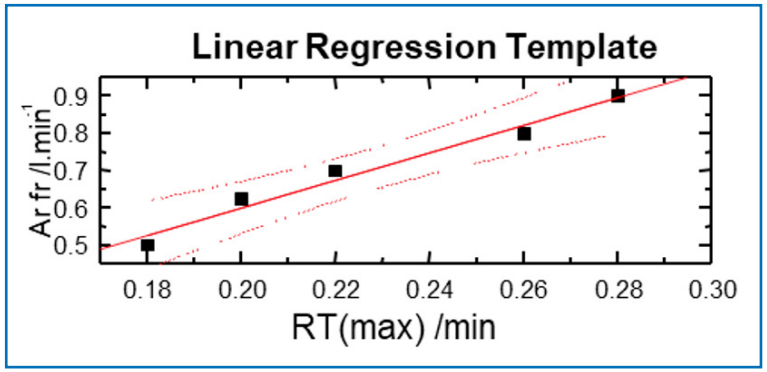
Intensidad de señal

Para estimar el modelo se realizaron 31 experimentos en el marco de un diseño experimental Doehlert (se verá en el capítulo correspondiente). Los datos, hechos por duplicado se utilizaron para entrenar la red BP que se muestra en la figura.



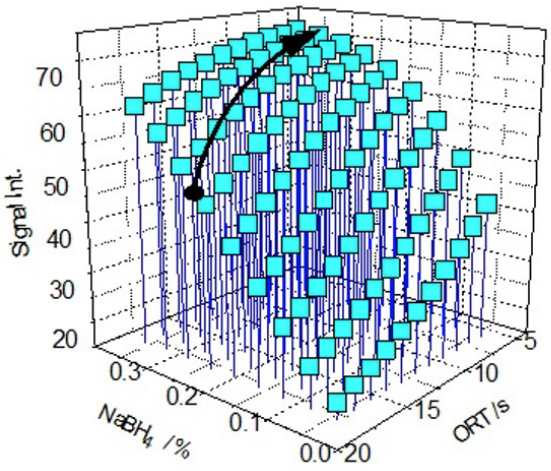
Una vez entrenada la red, se simularon una gran cantidad de experiencias para obtener diferentes superficies de respuesta que nos permitieran comprender el comportamiento del sistema. La figura de la izquierda muestra la relación entre las variables químicas y la de la derecha, entre variables físicas. En esta figura RT es el tiempo de retención en el *coil* de re-

acción (en minutos) que es el cociente entre la long. del tubo reactor x sección del tubo/Caudal de reactivos. En ella puede apreciarse una relación lineal entre los máximos de las curvas acampanadas de RT, demostrando que 'para condiciones de máxima señal' estas variables están fuertemente correlacionadas, como muestra la figura de abajo:



Ahora podemos obtener una superficie de respuesta con estas dos variables optimizadas y combinarla con la variable química de operación más importante, que es la concentración del generador de hidruro

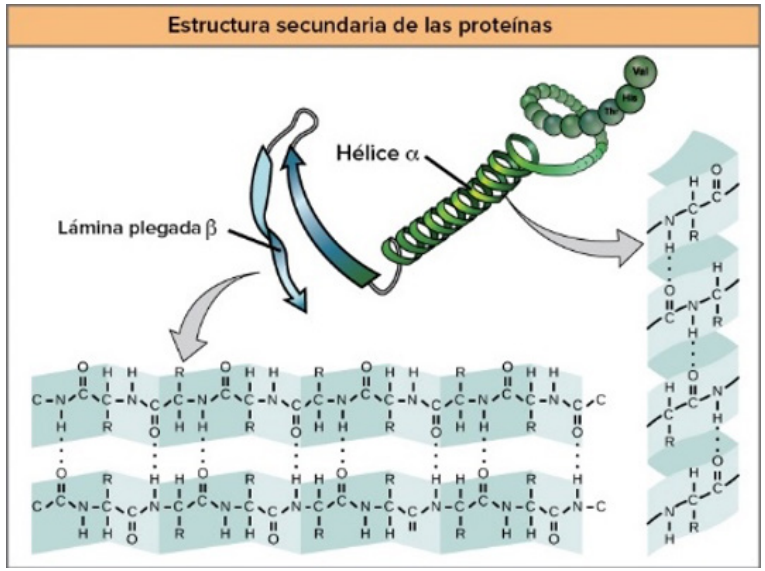
En el gráfico siguiente, la flecha indica el cambio de posición de la operación que se había obtenido por el método OVAT (punto negro) y la que se obtiene por ANN (cabeza de la flecha). La tabla muestra la comparación de resultados obtenidos con tres métodos diferentes. Obsérvese el aumento en intensidad de señal obtenido por redes neuronales (ANN) y la diferencia con otros métodos.



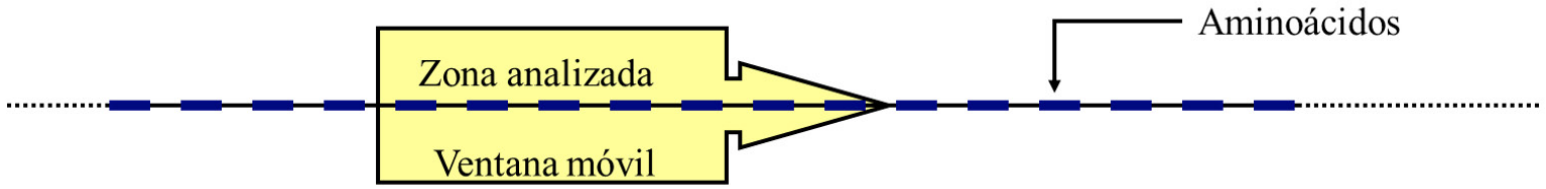
Parameter	OVAT	SIMPLEX	ANN
[$NaBH_4$] (% m/V)	0.25	(0.291) 0.234	0.28
[Tartaric acid] (mM)	50	(91) 91	45
Residence time (min)	0.300	(0.17) 0.28	0.145
Ar flow rate (l min^{-1})	0.7	(0.46) 0.52	0.3
Signal Intensity (Arbitrary Units)	68	87 78 Trials:14 17	119

Estructura Secundaria de Proteínas: Un ejemplo de aplicación de ventana móvil

Las proteínas pueden describirse mediante una estructura primaria constituida por el orden de secuencia de 20 diferentes aminoácidos esenciales. Y además una estructura **determinada por los plegamientos de las cadenas de aminoácidos que se producen justamente debido a la conformación de la estructura primaria.** Se consideran tres tipos diferentes de plegamientos que son: alambres, β-láminas y α-hélices, tal cual muestra el esquema de la figura. Debido a que la estructura secundaria es la que determina las propiedades biológicas de la proteína, su conocimiento es muy importante. (Ref. 1,3)

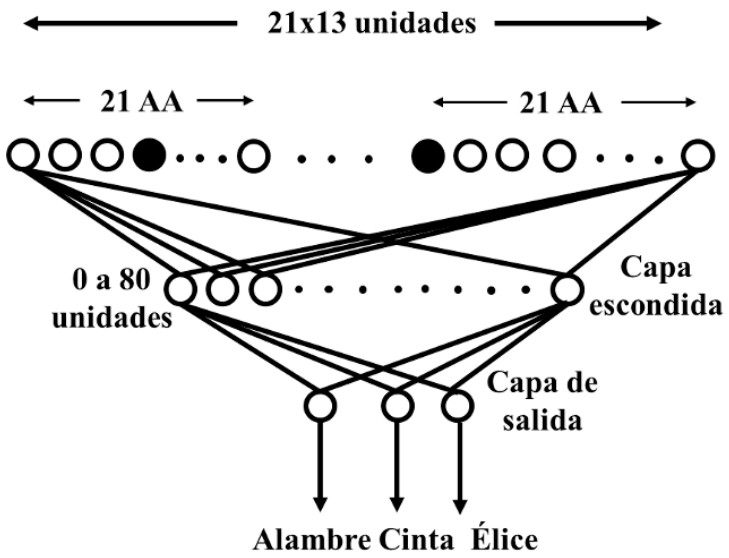


En este problema se intenta determinar la estructura secundaria a partir del conocimiento de la estructura primaria sobre la base de la hipótesis de que la estructura secundaria depende de la estructura de un aminoácido particular y sus vecinos. Entonces, un análisis con ventana móvil, como muestra la figura, podría determinar la estructura de toda la proteína. Para ello el lote de datos para entrenamiento estuvo constituido por 106 proteínas teniendo en total 18.105 aminoácidos



El lote de datos para predicción estuvo constituido por otras 15 proteínas con un total de 3.520 aminoácidos.

Por conveniencia computacional este problema es mejor manejarlo en forma binaria. Los 20 aminoácidos fueron identificados binariamente con una cadena de 21 bits. La cadena consta de sólo un 1 y el resto son ceros. La posición que ocupa el 1 en la cadena identifica el aminoácido. Esta técnica, hemos visto en la red Abbam, se reconoce como ***representación distribuída***.



El tamaño de la ventana es de 13 aminoácidos (uno central y seis vecinos a cada lado). Las 13 variables reales han sido reemplazadas entonces por 21x13=273 entradas binarias. Este es el número de neuronas de entrada que requiere la red. La salida esta formada por tres neuronas, una para cada clase de estructura secundaria. El número más apropiado de neuronas para la capa escondida fue 40. Si incluimos el bias esta red tiene:

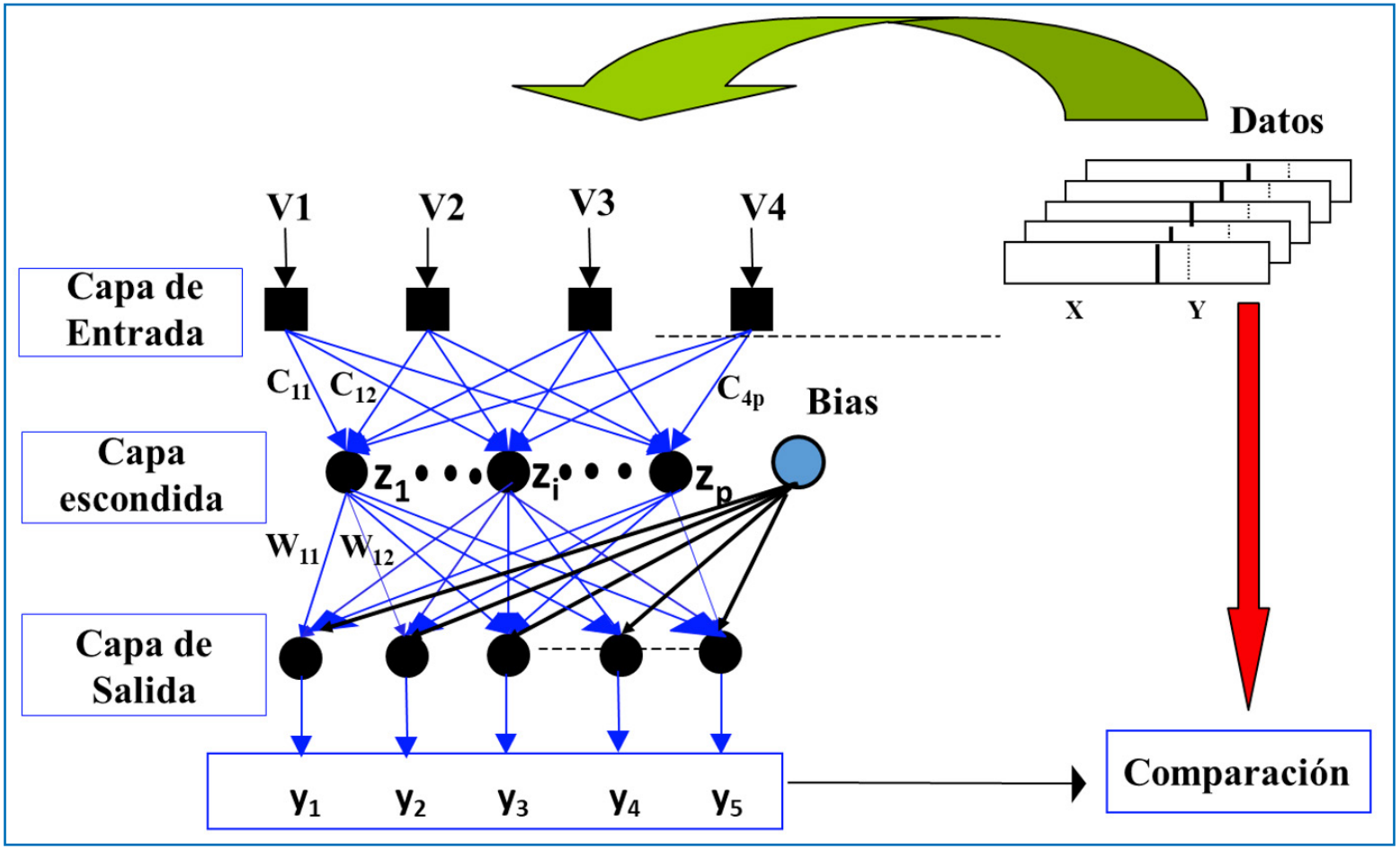
$$(273+1)x40+(40+1)x3=11.083 \text{ pesos w.}$$

Para el entrenamiento, la ventana se coloca al comienzo de la cadena con los primeros 13 aminoácidos, esto dará lugar a la primera respuesta. Luego se mueve la ventana una posición (aminoácido) produciendo la segunda respuesta y así sucesivamente. Al llegar al final de la cadena, cuando haya menos de 13 aminoácidos, la entrada quedaría incompleta. Para evitar el problema se utiliza un aminoácido inexistente como bandera que es el que se identifica con el bit 21. Cuando este está presente a la entrada, la salida no es válida.

La red dio como resultado el 62.7% de respuestas correctas sobre el lote de predicción. Otros métodos de cálculo utilizados anteriormente llegaban sólo al 50-53%. En este tipo de problemas, una mejora del 10% es muy importante.

Red “radial basis functions” (RBF)

Este tipo de red tiene características particulares. Es una red de tres capas, la de entrada, con las mismas funciones que en las otras redes, una capa escondida cuya particularidad es que está compuesta por neuronas con funciones estadísticas y la capa de salida, que es una combinación lineal de las salidas de la capa escondida. Las conexiones entre las capas son del tipo “full conection” (Ref 4-6). La arquitectura se muestra en la figura siguiente.



Supongamos que tenemos un lote de datos ordenados en una matriz **X(n,m)** donde m representa a las variables de entrada $j_1=v_1, j_2=v_2,\ldots j_m=v_m$ y n es el número de datos.

Función de la capa escondida: Esta capa hace el trabajo de las funciones de transferencia; es una transformación no lineal que llamaremos **Θ**, que representa a una función estadística, generalmente Gaussiana.

$$\boldsymbol{\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

Una forma general simplificada, con la misma geometría responde a una expresión:

$$\boldsymbol{\theta(r) = -\frac{e^{r^2}}{\sigma}}$$

Como sabemos, las funciones estadísticas clásicas suelen definirse con los parámetros *N*(μ,σ), donde μ es la media y σ es la desviación estándar de la distribución estadística. Pero en esta ecuación σ puede ser la desviación, la anchura o dilatación de la función de base radial.

El parámetro r en la última función representa la distancia* entre el valor de entrada x(n,j) y μ_i, donde μi,σi , i=1,2, …p son los parámetros de las clases o categorías de las p neuronas de la capa oculta y j es el índice de la variable de entrada (ver en arquitectura de la red). *Recuerde la definición de distancia Euclideana en el capítulo de clusters.

$$\boldsymbol{r(n) = \frac{\left(\sum_{j=1}^{m}\left(x_{j,n} - \mu_{j,i}\right)^2\right)^{1/2}}{\sigma i}}$$

La salida de cada neurona de la capa escondida es entonces:

$$\boldsymbol{Zi(n) = \theta(n) = e^{\frac{\left(\sum_{j=1}^{m}\left(x_{j,n}-\mu_{j,i}\right)^2\right)^{1/2}}{\sigma i}}}$$

La capa de salida:

Como se ha adelantado, esta capa es una combinación lineal de la capa escondida, de modo que para cada neurona de salida y_k, k=1,2…,q es:

$$y_k(n) = \sum_{i=1}^{p} W_{ik}.Z_i(n) + u_k$$

W_{ik}, como en otras redes, son los pesos que conectan las neuronas de la capa escondida con las de salida y u_k es el "umbral" asociado a cada neurona de salida.

En el gráfico de arquitectura de la red se dibujó un bias, sin embargo, este elemento puede estar presente o no dependiendo de la estrategia de programación.

Ajuste de la red:

1) La red se ajusta para establecer los valores de los pesos W_{ik} entre las neuronas de la capa oculta y las de salida

2) La red se ajusta para establecer los valores de μ y σ de las neuronas de la capa oculta. **Métodos de ajuste:** Existen los tipos de aprendizaje, supervisado y no supervisado. Pero el no supervisado es en realidad híbrido (fase no supervisada y fase supervisada).

Aprendizaje híbrido

La parte **no supervisada** determina la posición y desviación de los centros de las funciones estadísticas de las neuronas de la capa escondida.

Para determinar la posición de los centros se utiliza el algoritmo k-means (ver capítulo 3, clusters) o una red neuronal de Kohonen.

Para determinar las desviaciones con el objetivo de evitar solapamientos entre las distribuciones se utilizan medidas de distancias aritmética o geométrica entre las posiciones de los centros. Por ejemplo, se mide la distancia media aritmética entre cada centro y los más próximos o todos, de 1 hasta p.

$$\overline{d_i} = \frac{1}{p}\sqrt{\sum_1^p (\mu_i - \mu_{i\prime})^2}$$

También se puede utilizar la media geométrica entre cada centro y sus 2 vecinos más cercanos.

$$d_i = \sqrt{(\mu_i - \mu_{v1})(\mu_i - \mu_{v2})}$$

En la **parte supervisada** se minimiza el error obtenido a la salida de la red.

$s_k(n)$ es el valor objetivo de los datos de la muestra e y_k el valor de salida calculado por la red.

$$\mathbf{Err} = \frac{1}{\mathbf{n}}\sum_1^{\mathbf{n}} \mathbf{e(n)} \quad e(n) = \frac{1}{2}\sum_{k=1}^{q}\left(s_k(n) - y_k(n)\right)^2$$

Determinación de los pesos W_{ij} y umbral u_k

$$W_{ik}(n) = W_{ik}(n-1) - u_k \frac{\partial e(n)}{\partial W_{ik}} \qquad [1]$$

$$u_k(n) = u_k(n-1) - u_k \frac{\partial e(n)}{\partial u_k} \qquad [2]$$

$$\frac{\partial e(n)}{\partial W_{ik}} = -\big(s_k(n) - y_k(n)\big)\frac{\partial y_k(n)}{\partial W_{ik}} \qquad \frac{\partial y_k(n)}{\partial W_{ik}} = \theta_i(n) \qquad [3]$$

$$\frac{\partial e(n)}{\partial u_k} = -\big(s_k(n) - y_k(n)\big)\frac{\partial y_k(n)}{\partial u_k} \qquad \frac{\partial y_k(n)}{\partial u_k} = 1 \qquad [4]$$

Y reemplazando:

$$W_{ik}(n) = W_{ik}(n-1) + u_k\big(s_k(n) - y_k(n)\big)\theta_i(n) \qquad [5]$$

$$u_k(n) = u_k(n-1) + u_k\big(s_k(n) - y_k(n)\big) \qquad [6]$$

Aprendizaje supervisado

El aprendizaje supervisado es similar, pero no igual, al de una red "back propagation", en el sentido que ésta se realiza mediante un grupo de datos de entrenamiento. Sin embargo, la gran diferencia es que este entrenamiento se realiza con solamente una época de pasaje del lote de datos por la red en lugar de un ajuste por convergencia a través de un número grande de épocas.

Las ecuaciones de ajuste son, por lo tanto, diferentes para esta red:

El ajuste de los pesos, W_{ij}, y de los umbrales u_k se realiza a través de las ecuaciones 1 a 6 dadas. Se deben agregar los ajustes de C_{ij} y d_i.

$$C_{ij}(n) = C_{ij}(n-1) - u\frac{\partial e(n)}{\partial C_{ij}} \qquad \frac{\partial e(n)}{\partial C_{ij}} \cong \frac{e(n) - e(n-1)}{C_{ij}(n) - C_{ij}(n-1)}$$

$$u_k(n) = u_k(n-1) - u_k\frac{\partial e(n)}{\partial u_k} \qquad \frac{\partial e(n)}{\partial u_k} \cong \frac{e(n) - e(n-1)}{u_k(n) - u_k(n-1)}$$

Comentario general

Hasta aquí se han comentado las redes más comunes aplicables a una gran cantidad de problemas. Téngase en cuenta que para cada problema particular habrá alguna/s red más apropiada que las demás, aunque muchas veces se utiliza un tipo de red para todo propósito, como suele ocurrir con la red de retropropagación de errores. Si bien es cierto que esta red se adapta para resolver muchos tipos de problemas, no significa que sea más eficiente que otras en algunos casos. Por ejemplo, para problemas de clasificación, la red Kohonen es usualmente más apropiada que la de retropropagación.

Además, no se ha cubierto exhaustivamente todos los tipos de redes, de las cuales hay muchas variantes. Existen redes, por ejemplo, donde se incluye estadística Bayesiana combinada con alguna arquitectura ya conocida. Pero no nos referimos a ellas aquí ya que no hemos tratado la estadística de Bayes. Tampoco se han incluido todas las redes de retropropagación de errores que se diferencian por el tipo de función de transferencia con que trabajan. Sin embargo, las variantes tratadas hasta aquí permiten resolver una gran cantidad de problemas a todos aquellos que quieran iniciarse en este tema

Algoritmos Genéticos

Algoritmos genéticos (GA en inglés) es una técnica útil para resolver difíciles problemas de optimización numérica por técnicas clásicas. El método fue introducido alrededor de los años 60 (Ref. 7,8). En problemas muy complejos, por lo amplio del espacio de búsqueda, un inconveniente puede surgir debido a la magnitud del tiempo de cálculo.

Este tema requiere una detallada explicación técnica que está fuera del alcance de este libro, pero se verán los lineamientos básicos para comprender sus principios y utilidad y poder luego avanzar hacia bibliografía más completa.

Ejemplos de problemas típicos, entre otros, donde esta técnica se puede aplicar eficazmente son: ajuste de curvas, búsqueda de picos en vectores de registro de tiempo, frecuencia, etc. Un ejemplo típico es la búsqueda de ventanas de longitudes de onda en problemas de calibración multivariada. También se pueden atacar otros problemas diferentes tales como el recorrido óptimo del viajante comercial que debe recorrer varias ciudades sin repetirlas en su paso. O, la selección óptima de términos en la optimización de modelos de regresión lineal.

Principios del método

Hemos dicho que esta técnica, igual que las redes neuronales, pertenece a la clase de *métodos de computación naturales*, o sea, que tratan de imitar los mecanismos de la naturaleza.

En este caso se trata de imitar la selección natural a través del mecanismo de combinación en la codificación de los cromosomas para le evolución de la vida. Esto incluye los siguientes pasos:

- Una técnica de codificación de los candidatos probables para la solución.
- Un proceso de competencia entre estos candidatos.
- Un proceso de recombinación de los candidatos, tal que pueda aparecer una nueva generación de mejores soluciones, que reemplace a la existente.
- Introducción de cambios al azar, que imitan la mutación genética.

Representación de soluciones candidatas

Éstas son combinaciones válidas de parámetros a ser optimizados. Deben ser escaladas y codificadas para el tratamiento computacional, pueden ser variables enteras o reales. La representación más común es la de tipo binario. Por ejemplo, si un parámetro debe adquirir un valor 13, éste estará representado por una cadena de bits1101, sin embargo, si se requiere más precesión, será necesario agregar más bits. Con B bits se determina el número total de diferentes valores, 2^B-1, que pueden ser representados en el rango.

Los parámetros constituyen sub partes de una cadena total que los contiene, generalmente los parámetros tienen el mismo número de bits. Por ejemplo, la cadena:

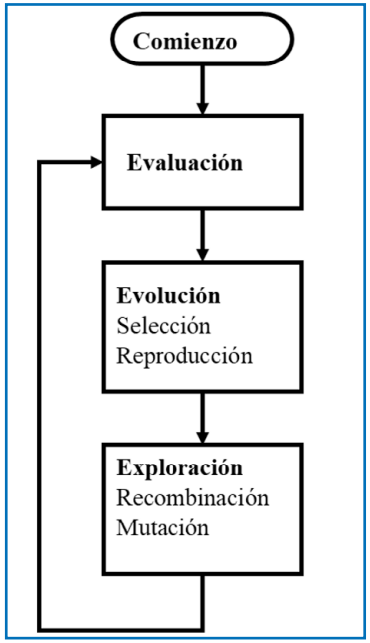
1001011|0100100|10111|1001 constituye una cadena de 4 parámetros.

Muchos usuarios prefieren la representación en números reales porque es más sencilla a simple vista, en cambio la codificación binaria es transparente, no visible, al usuario.

Diagrama de flujo del procedimiento de cálculo

Un modo sencillo de comprender los pasos del procedimiento es representarlos en un diagrama de flujo. Primero se crean **al azar** un número de soluciones candidatas, que puede oscilar comúnmente entre 50 y 100. Esto constituye la *población* de la *primera generación*.

Pueden seleccionarse algunas soluciones candidatas por conocimiento previo del usuario para reducir tiempo de cálculo, pero no es imprescindible.



De esta primera generación se irán sucediendo nuevas generaciones y es de esperar que en la última generación emerja la solución candidata óptima.

En la etapa de evaluación se necesita un valor de ajuste que sirva de control para las soluciones candidatas. Este valor de ajuste debe provenir del conocimiento previo del problema y muchas veces no es sencillo. Pero, supongamos que estamos ante un problema de ajuste de una curva, obviamente no conocemos el mejor ajuste de antemano, pero podemos calcular el error medio cuadrático entre los datos experimentales y la solución candidata. Lo lógico sería que su diferencia no fuera mayor al error experimental de las medidas y este valor podría usarse como función de ajuste.

El algoritmo terminará su ejecución cuando una solución candidata reúna las condiciones determinadas por el algoritmo de ajuste. Esto ocurrirá generalmente después de muchos ciclos de generaciones de cálculo. Podría ocurrir que la solución no alcance valores adecuados, en este caso pueden mejorarse las condiciones iniciales o recurrir a técnicas de corrección durante el cálculo, que están fuera del alcance de esta introducción. Información más descriptiva y detallada pueden consultarse en las referencias (Ref. 9 ,10).

En la **etapa de evolución**, se ponen en juego estrategias propias de la evolución y mejoramiento de la población, en inglés se la denomina *exploitation stage*. La idea básica está inspirada en la teoría de selección natural de Darwin. Una nueva serie de soluciones candidatas es creada desde la población presente. Usualmente el nuevo número de soluciones candidatas es igual al original, Np. El proceso **de selección** es en general como sigue:

Se elige una solución candidata tomada de la población presente de acuerdo a **una estrategia predefinida**, una copia de ella se enlista aparte como el primer miembro de la nueva generación que por ahora **es transitoria**. Se repite el proceso N veces, siempre desde la lista original hasta que la nueva población está completa. Observe que como la población original es siempre la misma, una solución candidata puede ser elegida varias veces.

Existen variadas estrategias predefinidas, cada una con distintas ventajas; no está al alcance de este capítulo describirlas a todas.

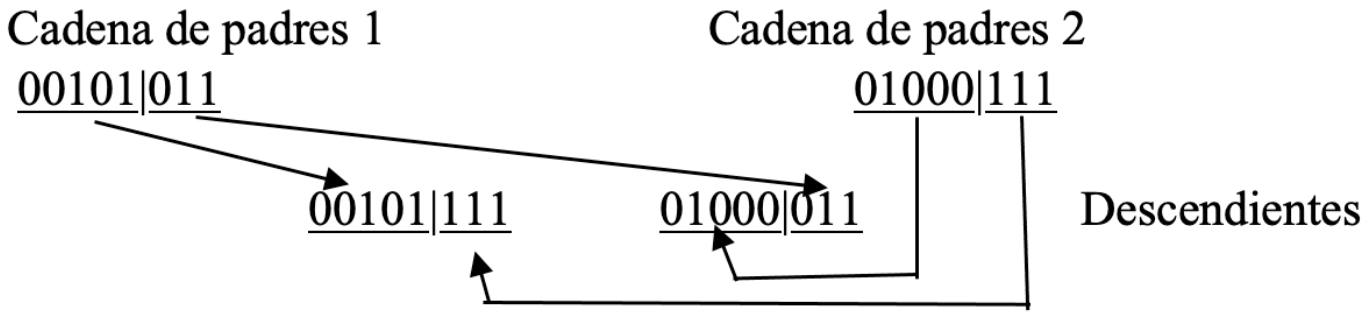
Lo resultante es que, debido a esta estrategia de selección, se ha creado una nueva generación que en promedio es mejor que la previa.

Hasta ahora nos hemos asegurado de haber seleccionado las mejores soluciones de la población original, sin embargo, no ha sido encontrada aún ninguna nueva solución. Esto es necesario para recorrer todo el espacio de búsqueda. Para lograr esto deben introducirse variaciones en la población y al mismo tiempo preservar la información importante de las mejores soluciones.

Para lograr esto se utilizan los operadores genéticos siguientes:

Recombinación (o, Cross-over) y mutación

La recombinación es una imitación de la recombinación de alelos en un cromosoma. En el método más simple, de un punto, las cadenas candidatas se dividen en 2 partes, y de la recombinación surgen 2 descendientes. Por ejemplo, fracciones de dos parámetros de soluciones candidatas son recombinadas para que surjan 2 nuevos descendientes:



Recuerde que re-escalando la información binaria a decimal, tenemos:

00101=5, 011=3, 01000=8, 111=7, los valores de los parámetros no se pierden, de modo que de la cadena 5,3 y 8,7 surgen los descendientes 5,7 y 8,3. Ésta es la más lenta de las recombinaciones.

Del mismo modo existen recombinaciones de 2 puntos.

101|100|11 = 5-4-3	Padres	010|111|00 = 2-7-0
101|111|11 = 5-7-3	Descendientes	010|100|00 = 2- 4-0

Y la recombinación uniforme:

110|101|11= 6-5-3	Padres	101|010|00= 5-2-0
110|000|11= 6-0-3	descendientes	101|101|11= 5-5-3

Ésta recombinación suele ser la que funciona mejor. Se intercambia un número de bits al azar. Éstos no necesariamente deben pertenecer a una misma sub cadena en toda la cadena, a menos que se esté trabajando con números reales, lo que hace mucho más restrictiva la búsqueda espacial que cuando se trabaja en numeración binaria. La característica de esta recombinación es que pueden aparecer descendientes que no estaban presentes en la población original. Existen aún otras variantes de esta recombinación.

En la recombinación siempre pueden aparecer soluciones inválidas. Éstas pueden ser descartadas en la etapa de evaluación. Cuando aparece un número inaceptable de soluciones inválidas significa que el método de recombinación no está funcionando bien y hay que reemplazarlo.

Mutación

Puede ocurrir que, alcanzada alguna etapa, ésta tenga soluciones donde una o más sub-cadenas tengan alguna posición en sus bits donde siempre tenga el mismo valor. En este caso, ningún método de recombinación cambiará el valor de ese bit y el cálculo quedaría encerrado en un sector local del cual no puede salir. Para evitar este problema, la mutación cambia, al azar, el valor de los bits en cualquier parte de la cadena, con una cierta probabilidad, P, usualmente menor a 0.05. Si la probabilidad fuese más alta se corre el riesgo de arruinar ciertas soluciones correctas.

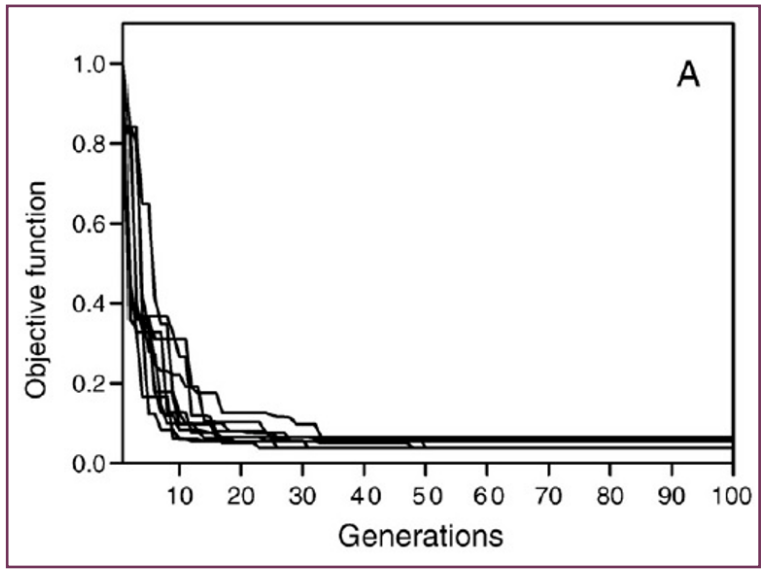
Reemplazo de la generación

Una vez ejecutadas las dos etapas anteriores, recombinación y mutación, la generación actual es reemplazada por la temporaria. Si el número de la población va a permanecer constante, el reemplazo es directo, pero si es menor se requiere una estrategia de reducción adicional, semejante al ya descripto, **proceso de selección**.

Este paso completa un ciclo, denominado una *generación*. Muchos ciclos o generaciones, entre cientos o miles según la dificultad del problema, deben procesarse antes de alcanzar la convergencia suficiente del cálculo.

Control de la evolución de las generaciones

Hay muchos modos de controlar la evolución del cálculo a través de los errores, o que se nos puedan ocurrir sobre control de alguna de las etapas descriptas. Una muy común y general, consiste en representar para un ciclo, la media o la mediana de la población conjuntamente con la mejor solución candidata o “cadena” en función del número de generación.



En la figura se muestra la evolución de la función objetivo o ajuste con las generaciones. En este caso la función es la raíz cuadrada del error cuadrático medio de un modelo de regresión lineal respecto del modelo estándar, RMSE/$RMSE_0$. Se muestran 10 corridas del programa de cálculo (Ref. 11).

Configuración experimental del cálculo

Para ejecutar el cálculo de GA es necesario utilizar programas de computación, de los cuales existen varios. Pero además, durante la explicación del método se han mencionado muchos parámetros que es necesario introducir para iniciar el procedimiento. Por un lado, la ventaja de tener varias opciones de cálculo hace al método muy flexible, mientras que por otro, requiere cierta experiencia poder manejarlos. En este sentido la práctica es semejante a lo que ocurre con las redes neuronales. No existe una metodología práctica para seleccionar de antemano los parámetros iniciales.

Debido a que el método es bastante robusto, es posible que se alcance la solución aunque los parámetros iniciales no sean óptimos, pero al costo de emplear mucho más tiempo de cálculo. Un problema más complicado se presenta cuando no se alcanza la convergencia, porque es difícil darse cuenta de cuáles parámetros deben ser cambiados con el fin de lograrla.

Se ha asumido aquí que los parámetros iniciales se mantienen constantes durante la ejecución del cálculo. Sin embargo, es razonable suponer que los parámetros óptimos pueden ir evolucionando durante la ejecución. Pero el requerimiento de usar parámetros dinámicos de ajuste torna al problema todavía mucho más complejo.

Un ejemplo muy completo de la ejecución de GA puede consultarse en (Ref. 6, pag. 826).

Comportamiento ante múltiples máximos locales

Una característica interesante de GA es su eficiente comportamiento ante máximos locales. Cuando esta situación existe, el hecho de que haya una cantidad de soluciones compitiendo hace que estas soluciones se dividan durante el proceso de cálculo y se puedan analizar varias de ellas. Cuando se desea evitar la presencia de algún máximo local se pueden introducir penalizaciones para impedirlo.

Referencias

1. Jure Zupan; Johann Gasteiger, Neural Networks in Chemistry and Drug Design. 2nd Edition, Wiley-VCH, Weinheim, 1999.

2. JF Magallanes, P Smichowski. Optimization and empirical modeling of HG-ICP-AES analytical technique through artificial neural networks. . Chem. Inf. Comput. Sci. 2001, 41, 3, 824–82. DOI: 10.1021/ci000337k

3. https://es.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/orders-of-protein-structure.

4. Gonzalez Panedo http://www.varpa.org/~mgpenedo/cursos/scx/archivospdf/Tema5-6.pdf

5. Diego Milone y Leonardo Rufiner. http://infofich.unl.edu.ar/upload/226d76eb008637c62114ab584b6cd71210209fe6.pdf.

6. Mark J. L. Orr. https://www.google.com/search?q=Introduction+to+Radial+Basis+Function+Networks+Mark+J+L&rlz=1C1CHBD_esAR902AR902&sxsrf=AOaemvLg5oCQHydxBWga2GdlsS-VqaE-7hw%3A1630507995800&ei=25MvYcGoMMKy5OUPm924qAk&oq=Introduction+to+Radial+Basis+Function+-Networks+Mark+J+L&gs_lcp=Cgdnd3Mtd2l6EAMyBQghEKABMgQIIRAVOgcIABBHELADOgcIIxCuAhAnSgQIQRgAUL5vWLp9YMfmAmgFcAN4AIABoA6IAYkQkgEHMi0xLjgtMZgBAKABAcgBCMABAQ&sclient=gws-wiz&ved=0ahUKEwiBgpLPg97yAhVCGbkGHZsuDpUQ4dUDCA4&uact=5

7. A.S. Eraser, Simulation of genetic systems. J. Theoretical Biol., 2 (1962). 329-346.

8. J.H. Holland, Adaption in natural and artificial systems. University of Michigan Press, Ann Arbor, MI, 1975, Revised Print: MIT Press, Cambridge, MA, 1992.

9. Massart D.L, Handbook Of Chemometrics And Qualimetrics part A 1997. Pag. 805.

10. Darrell Whitley A Genetic Algorithm Tutorial. https://www.cs.jhu.edu/~ayuille/courses/Stat202C-Spring10/ga_tutorial.pdf. Otras citas del mismo trabajo: Darrell Whitley A genetic algorithm tutorial. Publication: Statistics and Computing. Publisher: Springer Nature. Date: Jan 1, 1994. Whitley, D. A genetic algorithm tutorial. Stat Comput 4, 65–85 (1994). https://doi.org/10.1007/BF00175354

11. J.F. Magallanes, A.C. Olivieri / Chemometrics and Intelligent Laboratory Systems 102 (2010) 8–14

TERCERA PARTE

Diseño de Experimentos y Optimización de Modelos

CAPITULO 7

Introducción

Objetivos del Diseño de Experimentos

Objetivo: El diseño experimental propone, por un camino eficiente, obtener la optimización de un proceso o producto con un mínimo de experimentos.

En inglés al diseño de experimentos se lo denomina 'Design of Experiments' (DOE)

Hasta ahora, en las técnicas presentadas se han dado tablas de valores para ejemplificar los resultados obtenidos. Pero ¿Es indiferente que los datos sean obtenidos al azar o sean planificados para obtener mejores resultados?

¿Por qué incumbe a los científicos e ingenieros el diseño experimental? Porque una investigación experimental cualquiera es en sí mismo un procedimiento o proceso. En estos procesos intervienen un número considerable de variables (concentraciones de reactivos, parámetros instrumentales o cualquier otro parámetro vectorial.) y todas ellas deben ser ajustadas para obtener una óptima respuesta (máximo rendimiento, óptima separación señales, mínimo error relativo, entre otras). Avizoramos entonces desde un comienzo que el diseño de experimentos es una técnica orientada a problemas multivariantes, es decir: sistemas que dependen de muchas variables.

Alcances: El diseño experimental involucra dos contextos, a saber:

1- Describir **la serie de experimentos** que será llevada a cabo con la intención de desarrollar un modelo (por ejemplo, un modelo de regresión).

2- En la optimización de procesos o productos se aplica para determinar **la serie de condiciones** que son requeridas para obtener un resultado con características deseables u óptimas.

Para introducirnos en el tema comenzaremos por definir algunos vocablos de uso corriente en diseño experimental.

Definiciones elementales

Factor: es el nombre dado a las variables que son manejables de un modo controlado para estudiar su efecto sobre el proceso o producto y que tienen (o podrían tener) influencia sobre las características estudiadas. Factores posibles son: velocidad, temperatura, pH, caudales de gases o reactivos, tipo de catalizadores, diferencia de potencial, resistencia eléctrica, presión, entre otros tantos.

Respuestas: son las características del proceso o producto a ser optimizado. Son, por lo tanto, variables que describen la 'performance' del sistema.
Respuestas posibles son: rendimiento de un proceso, durabilidad de un producto, calidad, costo del proceso, precisión de un resultado, entre otras muchas.

Ambos **tipos de variables** pueden ser relacionadas como:

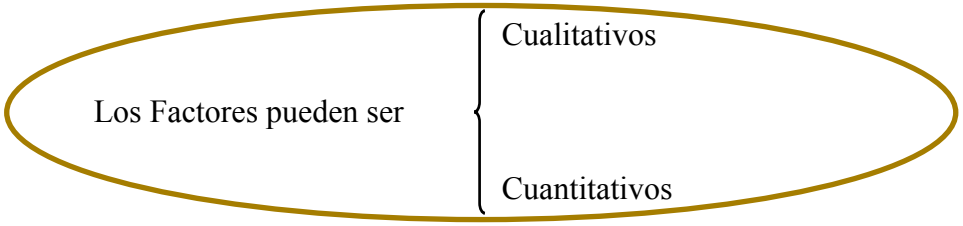
$$y_1, y_2, \ldots\ldots, y_n = \mathbf{f(x_1, x_2, \ldots\ldots x_m)} \quad \text{o,} \quad y = \mathbf{f(x_1, x_2, \ldots\ldots x_m)} \text{ donde: } y_i = \text{respuesta}i,\ x_i = \text{factor}_i.$$

En el primer caso el sistema incluye varias respuestas simultaneas y en el segundo caso sólo una. El primer caso es reducible al segundo ya que no hay inconvenientes en analizar **las respuestas** una a una. Sin embargo, es importante remarcar aquí que no debe confundirse la posibilidad de **analizar una respuesta por vez** con la de **realizar un experimento independiente** por vez. En **la serie** de experimentos estarán presentes **determinadas combinaciones de factores** y todas las respuestas a la vez. Este concepto se aclarará posteriormente con ejemplos específicos.

Función respuesta f( ): es 'el modelo' que relaciona ***la respuesta*** al efecto de ***los factores***. Esta relación puede ser algebraica o gráfica.

'El modelo' es obtenido de los experimentos. El término **'diseño'** significa que estos experimentos no son llevados a cabo de un modo azaroso sino **a través de una planificación considerada cuidadosamente**.

Algo más sobre Factores y Respuestas



Factores cualitativos pueden ser, por ejemplo: el ensayo de distintos solventes (agua, alcohol, acetona, etc.), el ensayo con materiales de distinta resistencia, ensayos con distinta calidad de repuestos o ingredientes, etc.

Factores cuantitativos son aquellos que pueden tomar un valor numérico, por ejemplo, pH, longitud de onda, densidad, resistencia eléctrica, dureza, peso, etc.

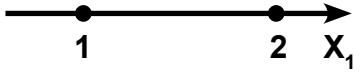
Niveles: Se llama "niveles" a los diferentes valores que toman los factores **en la serie** de experimentos que se llevarán a cabo. Por ejemplo, la resistencia eléctrica podría tomar valores de 100, 200 o 300 Ohm. La definición de los niveles suele ser el primer paso a considerar en el diseño.

Recuerde que cada variable representa una dimensión espacial y que como el DOE es una técnica multivariable usualmente trataremos con más de tres dimensiones, que por supuesto no son representables sino imaginables (o inimaginables) para cada persona.

Cuando se hace un diseño experimental es necesario muestrear este espacio multidimensional bajo ciertas reglas. Concretamente, durante la mayor parte de estos capítulos, se verán las diferentes formas de recorrer este espacio conforme a la característica de los problemas a resolver y a las ventajas y desventajas intrínsecas de cada método. A continuación, se muestran algunos ejemplos aislados del 'mapeo' multidimensional que luego se tratarán detalladamente.

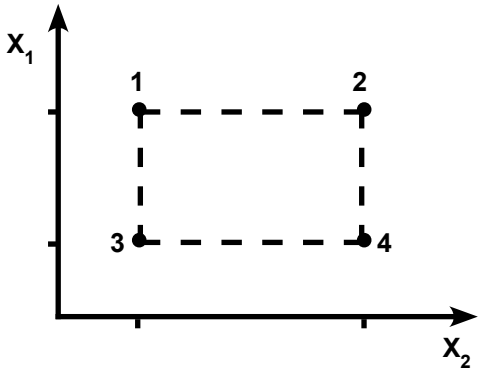
Mapeo del Espacio Multidimensional

a)



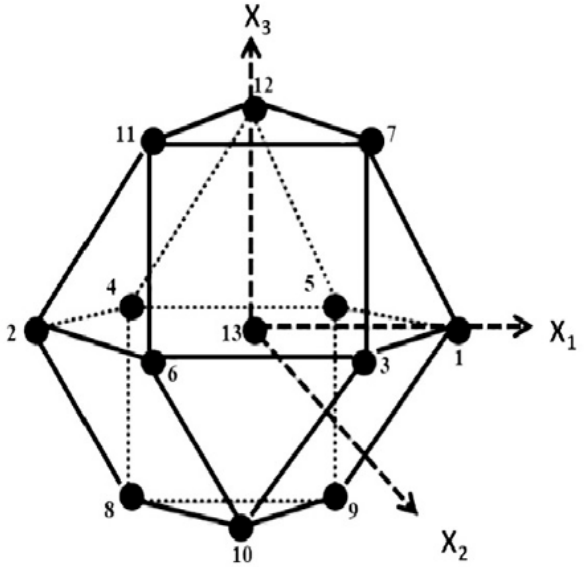
a) Espacio unidimensional: 1 Factor con 2 niveles.

b)



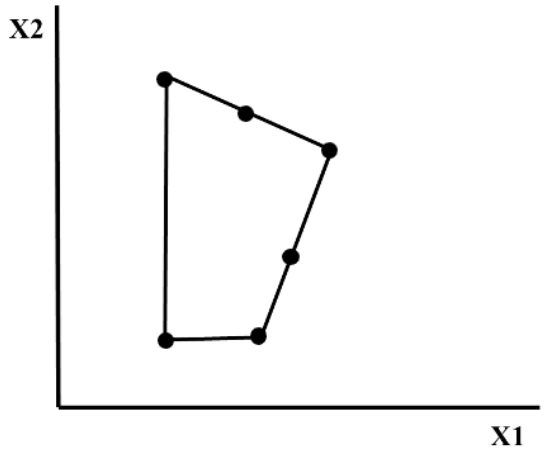
b) Espacio bidimensional: 2 Factores con 2 niveles (4 puntos experimentales).

c)



c) Espacio tridimensional: Diseño Doehlert (Ref. 1) (13 puntos experimentales), 3 factores con 3, 5 y 7 niveles.

d)



d) Espacio bidimensional: Dominio experimental irregular.

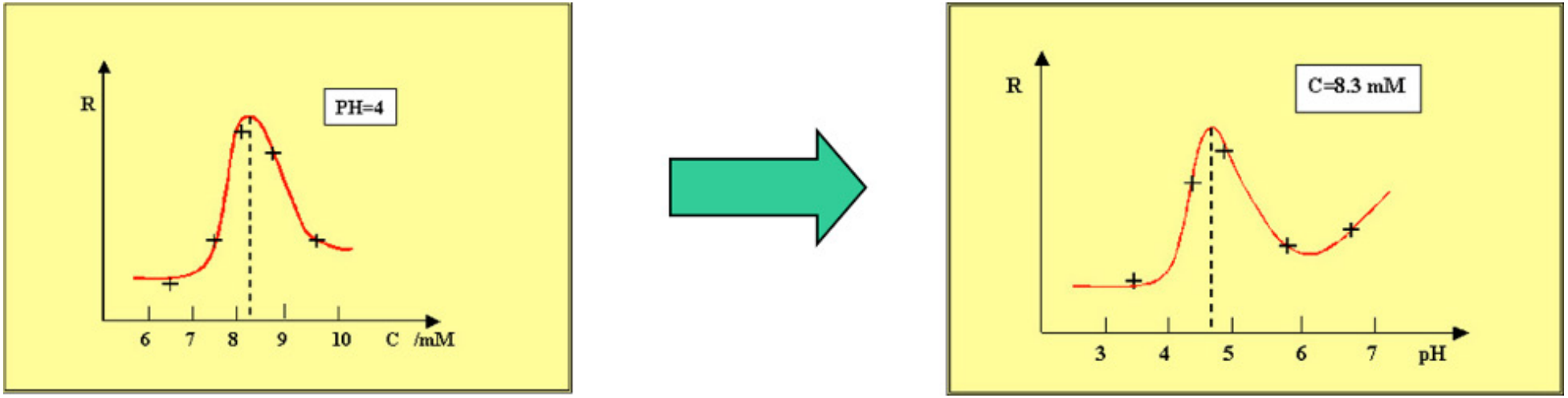
Necesidad del Diseño Experimental

A continuación, describiremos dos ejemplos que justifican la necesidad del DOE en la ciencia y la tecnología.

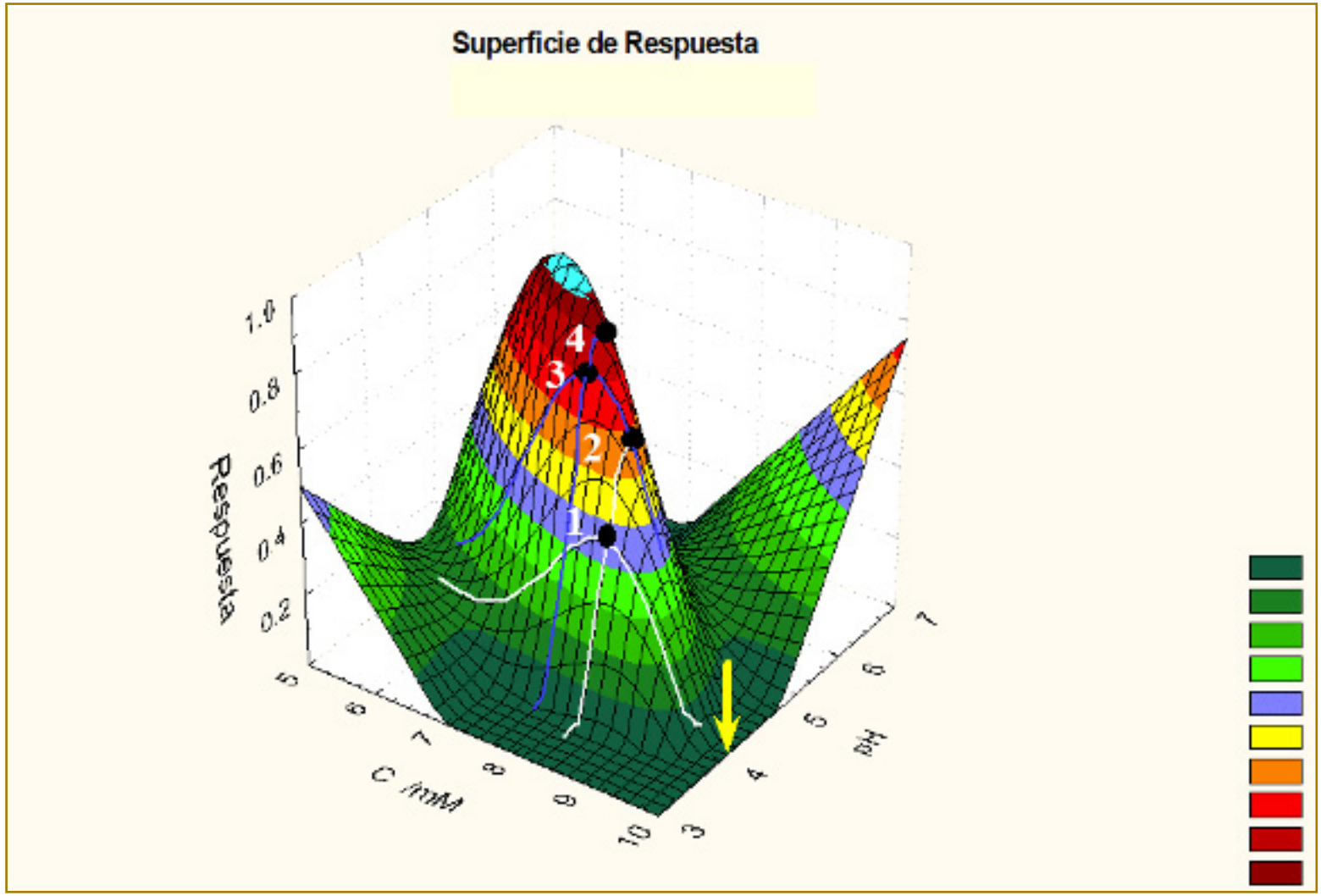
Ejemplo 1- Malas y buenas prácticas de optimización.

Este sencillo ejemplo hipotético es muy común en las tareas del investigador o desarrollador: se pretende optimizar (maximizar en este caso) la intensidad de señal (respuesta) ajustando 2 factores: la concentración de un reactivo y el pH. Un procedimiento **tan habitual como erróneo** es proceder a ajustar los factores (concentración y pH) uno a uno como se muestra en el procedimiento A.

Procedimiento A: El operador fija un valor de pH que le parece adecuado (4 en este caso) y barre el eje de concentración de reactivo con 5 puntos para obtener la máxima respuesta (en C=8.3 mM). Luego, en el siguiente experimento, fija la concentración de reactivo en 8.3 mM y barre el eje de pH con otros 5 experimentos para encontrar un máximo (pH=4.7). Concluye entonces que su punto óptimo de trabajo es **C=8.3 mM y pH=4.7**.



Veamos lo que realmente ocurre representando la respuesta en función de los 2 factores en un gráfico tridimensional.



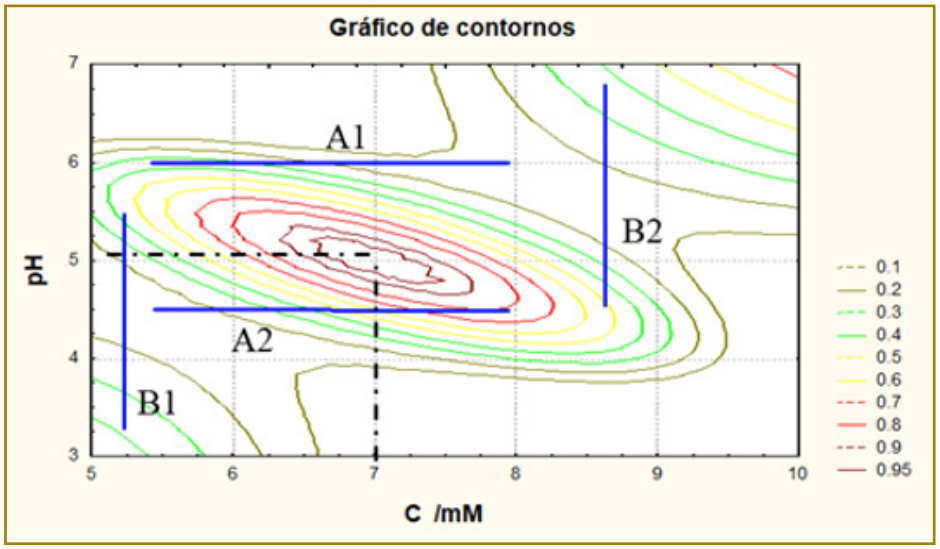
El operador partió desde pH 4, donde indica la flecha y barrió el eje de concentración encontrando un máximo a C=8.3mM en el punto 1. Luego fijó C en 8.3 mM y barrió el pH encontrando un máximo en pH=4.7 (punto 2). Para él este es el punto óptimo de trabajo, sin embargo, observe que el máximo verdadero está muy lejos del punto 2, por lo tanto, todo el

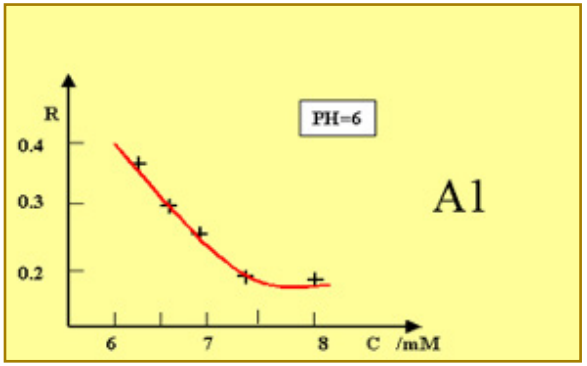
procedimiento es erróneo si lo que se pretendía era alcanzar el máximo. Alguien podría argumentar que si se repiten los ensayos comenzando ahora desde el punto 2 podría acercarse mejor...

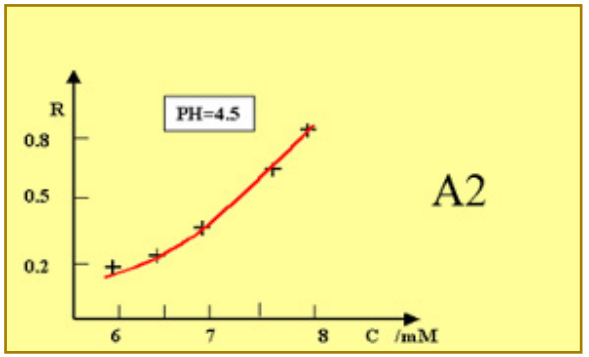
Efectivamente, en una segunda serie de experimentos se alcanzaría el punto 4, si bien está mejor ubicado que el punto 2 aún no se ha alcanzado el máximo. Hasta ahora ya hubiese realizado 20 experimento (5 por cada serie) y quien sabe cuántos **más tendría que realizar para alcanzar el verdadero máximo. Esto significa que este camino es muy ineficiente e** inseguro porque con todas las experiencias efectuadas aún no habríamos descrito la superficie de respuesta que muestra la figura, o sea estaríamos a ciegas sin saber dónde se encuentra el máximo. En nuestro ejemplo el óptimo verdadero se encuentra en C=7 mM y pH=5.1. A primera vista no parece una posición tan errónea comparada con la del punto 2. Sin embargo, la diferencia entre el óptimo y el punto 2 es de ¡60% más de intensidad de señal!

Obsérvese que este sencillo ejemplo incluye sólo 2 factores. Si este número fuese mayor, además de ser imposible representar la superficie de respuesta, la dificultad en encontrar el máximo por esta vía sería muchísimo mayor.

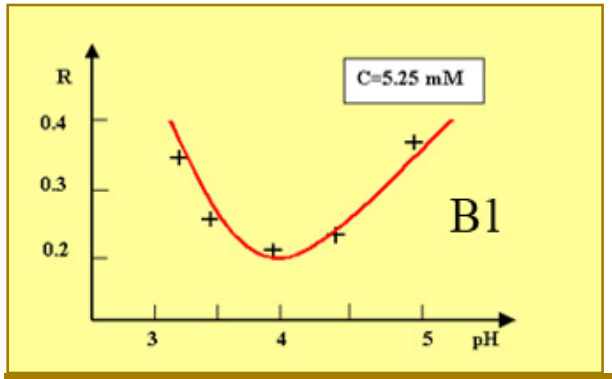
Otra manera de localizar un óptimo a partir de la figura 1 es proyectar la superficie de respuesta sobre el plano que forman los ejes de los factores. Esto genera el **gráfico de contorno** de la figura siguiente. Este gráfico está formado por curvas de igual valor de respuesta, a semejanza de las curvas de altitud en un mapa geográfico. El máximo se ubica en el centro de la elipse menor. Siguiendo con el mal procedimiento de estudiar las variables una a una (en inglés *one variable at time*, 'OVAT'), veamos qué ocurre si el operador decide hacer 2 experiencias a distintos pH fijos y dos a distintas concentraciones (figura siguiente). En el primer caso estaría recorriendo las rectas A1 y A2 de la figura y observaría los respectivos gráficos A1 y A2. Similarmente, en la segunda parte tendría los resultados de la serie B (figura 1). Observemos los resultados de la serie A: en A1 a medida que C aumenta la respuesta baja, pero en A2 es al revés. ¿Qué ocurre? Si observamos la serie B vemos que en B1, para pH creciente entre 4 y 5, la respuesta sube, pero en B2 baja. Estos resultados contradictorios no le permitirían al operador sacar buenas conclusiones sobre su optimización. Se deben al hecho de que los factores C y pH están asociados, fenómeno que veremos más adelante.







Lo que aquí se trata de mostrar es que hay que seguir un camino más racional para conocer el comportamiento de los factores y respuestas.



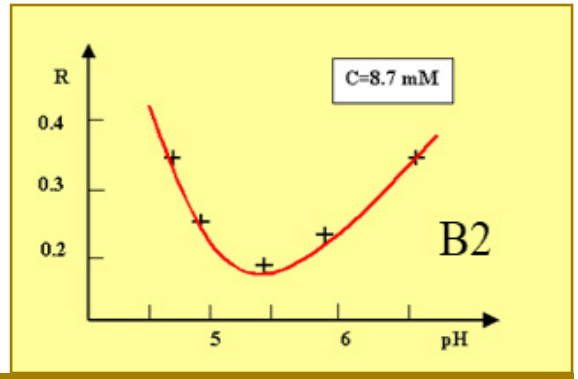


Figura 1

Este camino lo ofrece el diseño de experimentos, con la ventaja adicional de que se economizará tanto en el número de ensayos como en esfuerzo, tiempo y presupuesto.

En el segundo ejemplo veremos una ventaja adicional que se obtiene con la práctica del DOE. Esta ventaja está relacionada con la precisión de los resultados. Por ejemplo, si se ajusta una recta por cuadrados mínimos, ¿la precisión es la misma para cualquier conjunto de puntos elegidos?, ¿Cuántos puntos tomar y donde ubicarlos? La respuesta a estas preguntas también las da el DOE y uno puede imaginarse la importancia de este aspecto en los casos multidimesionales.

Ejemplo 2- La calidad de un modelo.

El cálculo de una regresión lineal múltiple.

> **NOTA**: recuerde que, en operaciones de álgebra lineal, el orden de los factores **altera** el producto. O sea que el orden de los factores **debe ser respetado.**

Supongamos que deseamos hacer una regresión lineal múltiple con m variables, o sea ajustar un modelo η, de la siguiente forma general:

$$\eta = \beta_0 + \beta_1.x_1 + \ldots + \beta_m.x_m \,. \qquad [1]$$

según este modelo, cada observación puede ser representada por:

$$y_i = \beta_0 + \beta_1.x_1 + \ldots + \beta_m.x_m + \varepsilon_i \qquad [2]$$

ε_i: error de la medida

Definimos ahora ciertos vectores y matrices para trabajar con expresiones matriciales.

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_m \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1m} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2m} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{nm} \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

n x 1 (m+1) x 1 n x (m+1) n x 1

En notación matricial podemos ahora expresar [2] como $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$.

La minimización de la suma de cuadrados de los residuos conduce a:

$\mathbf{X^T.X.b = X^T y}$ de donde se puede calcular **b**, el estimador de β.

$b=(X^TX)^{-1}X^Ty$ y también $\hat{y}$, el vector estimador de **y**.

$$\hat{y}=x(x^Tx)^{-1}x^Ty$$

La varianza residual s_e^2 que es una estimación del error experimental σ^2 se calcula como:

$$s_e^2 = \frac{\sum e_i^2}{n-(m+1)} = \frac{\sum (y_i - \hat{y_i})^2}{n-(m+1)} = \mathbf{e^T e} \,/\, (n-(m+1))$$

y_i: respuesta en x_i

y_i^: cálculo para x_i

Finalmente, la matriz de variancia – covariancia de los coeficientes de regresión **V(b)** permite hallar los intervalos de confianza de los βi.

$$\mathbf{V(b) = s_e^2 (X^TX)^{-1}} \qquad \beta i = bi \pm t_{0.025,n-(m+1)} \cdot s_{bi}$$

Los s^2_{bi} constituyen la diagonal principal de la matriz **V(b)**

En el cálculo vemos que se han obtenido los coeficientes bi y sus correspondientes errores t.sbi . Este error aparece en V(b) que a su vez depende de 2 términos s_e^2 y $(X^TX)^{-1}$. s_e^2 es una estimación del error experimental puro, pero $(X^TX)^{-1}$ no depende del valor de la respuesta (X contiene solamente la ubicación de los factores).

Esto significa que los intervalos de confianza **dependen exclusivamente del rango considerado para cada variable (esto es, el dominio experimental)**, de la distribución de éstas sobre el dominio experimental y del número de medidas. En otras palabras...

¡Conclusión importante!

Toda la información requerida para evaluar el efecto del diseño experimental sobre la *calidad de modelo estimado* está presente *antes* que algún experimento haya sido llevado a cabo.

A Modo de Cierre

En las clases sucesivas siempre estaremos rondando, advertida o inadvertidamente, alrededor de la estimación de la matriz $(X^TX)^{-1}$.

Los dos ejemplos mostrados revelan la importancia del diseño de experimentos en la quimiometría y en el campo de la ciencia y tecnología en general. En las clases futuras pasaremos a desarrollar las técnicas de diseño en particular y comprobaremos específicamente algunas consideraciones que, a modo de introducción, se han hecho aquí en forma general.

Es importante aquí darse cuenta con qué tipo de modelos vamos a trabajar. En general existen dos tipos: los modelos deductivos fundamentados en leyes científicas básicas (también llamados modelos mecanicistas o deductivos) y los modelos empíricos basados exclusivamente en las mediciones. Los primeros se aplican a modelos generalmente sencillos y de pocas variables, por ejemplo, los modelos físico-químicos, los segundos, de los que nos ocuparemos nosotros suelen ser modelos complejos dependientes de fenómenos a veces no muy conocidos o de muchas variables. Entre estos últimos es típico estudiar un fenómeno o procedimiento nuevo del que no existen modelos previos de ninguna naturaleza, por ejemplo, un modelo medioambiental. En estos casos, el desarrollo de modelos empíricos puede inspirar al investigador teórico en la tarea de racionalizar relaciones que lo conduzcan a algoritmos matemáticos. Resumiendo, podemos decir...

El diseño experimental se utiliza para desarrollar modelos empíricos. O sea,

en situaciones donde no es posible deducir la función respuesta desde una teoría.

Ahora, si bien una teoría permite conocer de antemano la relación que existe entre una respuesta y cierto factor, raramente provee los coeficientes de tal función.

Referencias

1. My-Kien Tran, Amin Swed, Frank Boury. Preparation of polymeric particles in CO(2) medium using non-toxic solvents: formulation and comparisons with a phase separation method. European journal of pharmaceutics and biopharmaceutics: official journal of Arbeitsgemeinschaft fur Pharmazeutische Verfahrenstechnik e.V DOI:10.1016/j.ejpb.2012.08.005 Corpus ID: 12295762

CAPITULO 8

Diseño Factorial de 2 Niveles

Introducción

Antes de pasar a definir los objetivos del **diseño factorial de 2 niveles** debemos precisar un vocablo aún no definido que es la **Interacción** entre 2 o más factores. En la mayoría de los sistemas experimentales, el efecto de un factor sobre la respuesta depende **del nivel de otro (u otros) factores**. A este fenómeno se lo llama interacción. La interacción entre 2 factores se llama **interacción doble**, entre 3 factores triple y así sucesivamente. Usualmente se investigan hasta las interacciones dobles, las interacciones superiores son usualmente muy difíciles de interpretar y distinguir. Algebraicamente, la consideración de las interacciones aparece en el modelo dentro de términos independientes, por ejemplo, en el siguiente modelo para 2 factores...

$\eta = \beta_0 + \beta_1.x + \beta_2.y + \beta_3.\ x.y$ El término que contiene a 'x.y' estima la interacción entre los 2 factores.

La apreciación de las interacciones nos obliga a aumentar el número de experiencias

necesarias, de modo que **para apreciar todos los efectos incluyendo las interacciones** es necesario hacer algunas experiencias más.

Objetivo: El diseño factorial de 2 niveles "completo" (two-level full factorial design **'FFD'**) se lleva acabo para determinar si ciertos factores o interacciones entre dos o más factores tienen un efecto sobre la respuesta y estimar la magnitud de tales efectos.

El diseño requiere un experimento para cada una de las posibles combinaciones de los 2 niveles y de los ***k* factores** considerados. Ejemplo, 2 factores darán lugar a $2^2=4$ experimentos (8 si se hacen replicados), 3 factores darán lugar a $2^3=8$ experimentos. Y en general, ***k* factores originarán 2^k experimentos** o 2^{k+1} si se consideran los replicados.

La identificación de niveles (cualitativos o cuantitativos) se hace con símbolos, por ejemplo "+,-"; "+1,-1"; "0,1".

Para ejemplificar todo el tratamiento, primero se explicará la mecánica del cálculo a través de un ejemplo hipotético con tres factores. Para analizar resultados numéricos nos guiaremos por datos reales parciales de un estudio de pre-tratamiento de muestras para análisis de Cd por una técnica química (ICP-MS). En las tablas del ejemplo que se dará a continuación conviene observar las siguientes reglas prácticas:

1- Los **factores se indican con letras mayúsculas**.

2- Las tablas están ordenadas según un cierto **'orden de cálculo'**, sin embargo el **orden de los experimentos** debe ser organizado cuidadosamente para evitar el **'bloqueo'***.

3- Cada experimento se identifica con una **letra minúscula para cada factor que tiene un nivel** +.

4- Cuando todos los niveles son '-' el experimento se indica como (1).

***Bloqueo** significa asignar **voluntaria o involuntariamente** un sesgo o tendencia a un lote de experimentos. En este caso debe evitarse la introducción **involuntaria** de un sesgo. Por ejemplo, suponga que por razones de tiempo el lote completo de experimentos se ejecutará en dos días. Y que el primer día se ejecutan todos los experimentos para el factor A con nivel '+' y el segundo día para A '-'. Si por alguna razón uno de los días tuvo condiciones experimentales desapercibidamente diferentes (cambio de un lote de reactivos ligeramente diferente, condiciones de funcionamiento de un equipo, cambio del material volumétrico, etc.) el efecto de esta anomalía se va a manifestar sumándose al efecto de A. Entonces el efecto de A puede aparecer **falsamente** significativo. Luego se mostrarán dos planillas que ejemplifican el punto 2 expuesto.

Estimación de los Efectos Principales

Consideremos la tabla siguiente para analizar la mecánica del cálculo para k=3 factores:

Tabla 1				
Run	A	B	C	Rta.
1	+	+	+	y_1
2	+	+	-	Y_2
3	+	-	+	Y_3
4	+	-	-	Y_4
5	-	+	+	Y_5
6	-	+	-	Y_6
7	-	-	+	Y_7
8	-	-	-	Y_8

Observamos que en las experiencias 1 y 5, con excepción de A, el resto de los factores (B y C) tiene un nivel constante '+'. La diferencia entre y_1 e y_5 estima el efecto de A cuando B y C tienen nivel '+'. El mismo razonamiento puede seguirse para A, con la diferencia entre y_2 e y_6 para los niveles +B y –C.

Analizando toda la tabla, tendríamos en total 4 estimaciones del efecto de A que pueden ser promediadas, dando la expresión:

$$\textbf{Efecto de A} = (\Sigma\textbf{corridas de nivel '+'} - \Sigma\textbf{corridas de nivel '-'})/\ 2^{(k-1)} = \qquad [1]$$

$$= \textbf{Media de corridas de nivel '+'} - \textbf{Media de corridas de nivel '-'.}$$

Para los efectos de B y C pueden hacerse estimaciones semejantes tomando en cuenta los niveles'+' y '-' de cada factor en particular.

Nota 1: Cuando el efecto no es descripto como la diferencia media entre niveles '+' y '-' (llamados a veces niveles extremos), sino como la diferencia entre estos niveles y el valor intermedio 'cero' (llamado a veces valor nominal), debe entonces dividirse el efecto por 2 y el denominador de [1] es 2^k en lugar de $2^{(k-1)}$.

Nota 2: Obsérvese que se obtiene una **estimación promedio**, pero que llevando a cabo **replicados** es posible conseguir una **evaluación estadística** de los factores como veremos más adelante.

Estimación de las Interacciones

La estimación de las interacciones puede hacerse por el mismo camino que antes....

Volviendo a la Tabla 1

Comparemos la evaluación del efecto de A desde **(y_1-y_5)** y desde **(y_3-y_7)**. En ambas **diferencias** el nivel de C es el mismo '+', pero el de B cambia de '+' a '-'. Haciendo la diferencia entre ambas y dividiendo por 2 estimaremos en qué medida el efecto de A es influenciado por B, o sea:

Interacción de B sobre A = [(y_1-y_5) - (y_3-y_7)]/2 = [(y_1+ y_7)-(y_3+ y_5)]/2

También se puede estimar la interacción de **A sobre B** al mismo nivel '+' de C

Interacción de A sobre B = [(y_1- y_3) – (y_5- y_7)]/2 = [(y_1+ y_7) - (y_3+ y_5)]/2

Que es exactamente la misma que la anterior. Entonces para el nivel +C, la interacción entre A y B, escrita como AB o AxB es: **[(y_1+ y_7) - (y_3+ y_5)]/2**. Para el nivel –C puede verificarse que **AxB = [(y_2+y_8) - (y_6+y_4)]/2**. Y promediando las dos interacciones se obtiene la interacción completa entre A y B:

$$\mathbf{AxB = (y_1 + y_2 + y_7 + y_8 - y_3 - y_4 - y_5 - y_6)/4.}$$

No se puede obtener una expresión general para este cálculo y así, el método se vuelve tedioso para interacciones triples o dobles de más de tres factores. Se puede recurrir a una forma mecánica de cálculo con la ayuda de la tabla siguiente como se explica a continuación.

Las tres primeras columnas son idénticas a las de la tabla 1; las otras columnas, para fines computacionales se obtienen por la regla de multiplicación de los signos.

A	B	C	Rta.	AB	AC	BC	ABC
+	+	+	y_1	+	+	+	+
+	+	-	Y_2	+	-	-	-
+	-	+	Y_3	-	+	-	-
+	-	-	Y_4	-	-	+	+
-	+	+	Y_5	-	-	+	-
-	+	-	Y_6	-	+	-	+
-	-	+	Y_7	+	-	-	+
-	-	-	Y_8	+	+	+	-

Entonces sí se puede aplicar también para las interacciones la expresión general:

$$\textbf{Interacción} = (\Sigma \text{ corridas de nivel '+'} - \Sigma \text{ corridas de nivel '-'})/2^{k-1}.$$

Vale tanto para las interacciones dobles como para las triples, tomando las columnas encabezadas con la interacción deseada. Otro método similar para trabajar en forma mecánica para hallar tanto los efectos principales como las interacciones es el **método de Yates**, pero no es tan simple.

Otro modo de estimar las interacciones es mediante los gráficos de interacción (*interaction plot*) (Ref. 1). Estos consisten en graficar, por ejemplo, para las variables, A y B, **los promedios** de los ensayos $B^{+}A^{-}$, $B^{+}A^{+}$, y $B^{-}A^{-}$, $B^{-}A^{+}$. Cuando se conectan los dos primeros puntos entre sí y también los dos restantes se generan dos rectas (Fig. 1)

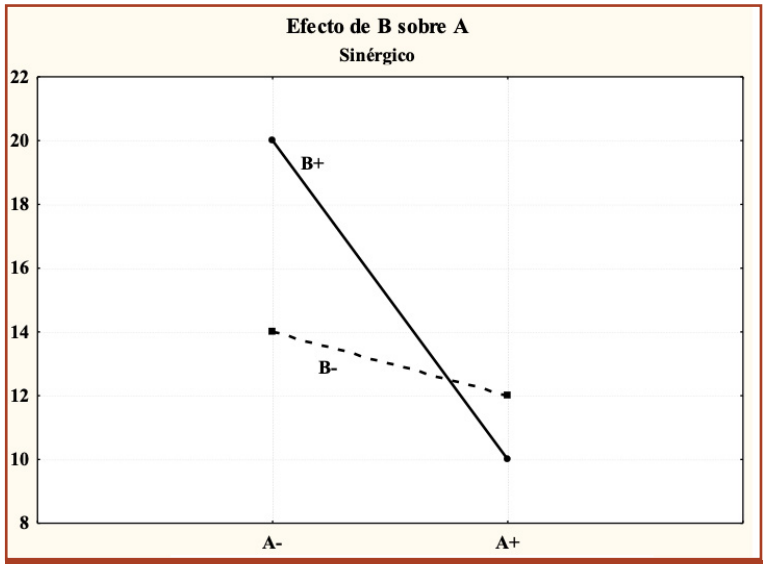


Figura 1: Efecto sinérgico

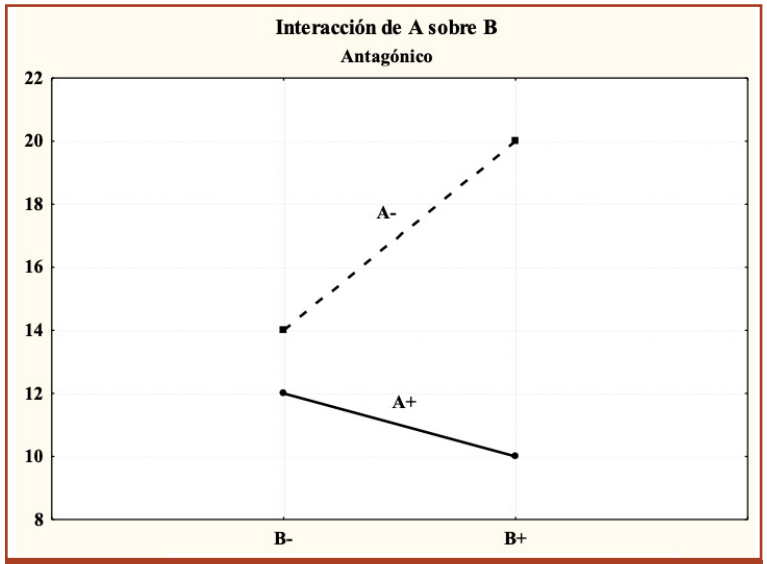


Figura 2: Efecto antagónico

Cuando las rectas tienen pendiente del mismo signo, como en el efecto de B sobre A, significa que hay un efecto sinérgico. Es decir, cuando el promedio de uno de los niveles de las variables baja (o sube) por efecto de la otra, por ejemplo B+; B+A- =20 →B+A+=10, el otro nivel B- hace lo mismo; B-A-=14 →B-A+=12. O sea, el efecto se refuerza. Pero el efecto contrario, A sobre B puede llegar a ser diferente (Fig. 2) y opuesto. Esto revela una información importante porque indica que cuando el nivel de una variable cambia, A-→A+, el efecto sobre el promedio también cambia, por lo tanto, a un nivel intermedio de la variable, su efecto sobre el promedio será nulo.

Esta situación exige un nivel más profundo de investigación, que podría ser entre otros, utilizar modelos de más de dos niveles. Es importante entonces, analizar los gráficos de interacción entre todas las variables: AB, BC, CD, AC, AD y BD para cuatro variables.

Análisis de los Efectos

Para el análisis de los resultados tomaremos el ejemplo del análisis de Cd por ICP-MS mencionado anteriormente.

pH	Caudal Carga	Conc Eluyente	Rta Cd Minimax
7.78	1.81	8.18	0.8169
7.78	1.81	2.82	0.9566
7.78	4.19	8.18	0.7907
7.78	4.19	2.82	0.8918
4.22	1.81	8.18	0.6884
4.22	1.81	2.82	0.7089
4.22	4.19	8.18	0.5653
4.22	4.19	2.82	0.628

Factores: pH=A, Caudal=B, Eluyente=C

Factor y niveles		
Factor	**Nivel Mínimo**	**Nivel Máximo**
pH	4.22	7.78
Caudal de carga	1.81	4.19
Conc. De eluyente	2.82	8.18

Las tablas siguientes muestran los resultados ordenados de dos formas diferentes...

Corrida	A	B	C	Respuesta
a	+	-	-	0.8918
ab	+	+	-	0.9566
abc	+	+	+	0.8169
b	-	+	-	0.7089
bc	-	+	+	0.6884
c	-	-	+	0.5653
ac	+	-	+	0.7907
(1)	-	-	-	0.6280

	A+		A-	
	B+	B-	B+	B-
C+	0.8169	0.7907	0.6884	0.5653
C-	0.9566	0.8918	0.7089	0.6280

El cálculo de los efectos de acuerdo a la ecuación [1] es el siguiente:

Efecto de A= 0.8640-0.6477= 0.2163

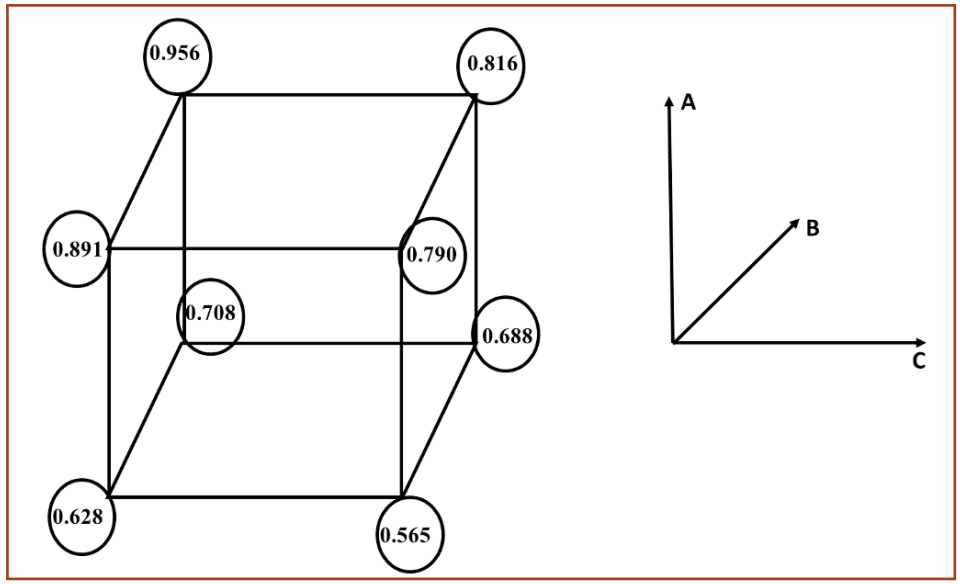
Efecto de B= 0.7927-0.7190= 0.0737

Efecto de C= 0.7153-0.7963=-0.0810

Ahora tenemos calculados los efectos de los factores principales, pero no sabemos cuánto significan cada uno o, dicho de otro modo, no sabemos **si los efectos superan el error de las medidas**. También vemos que el factor C tiene un efecto negativo, veremos más adelante qué es lo que esto significa. Por supuesto podríamos haber hecho el mismo cálculo para las interacciones, pero no nos interesan por ahora.

Significación de los efectos

1-Interpretación visual: Resultados del ejemplo



Los 3 factores están representados sobre un eje espacial, si hubiese un cuarto factor D tendría que estar representado en un cubo aparte para ese factor y sus combinaciones con los otros. En los vértices de cada cubo se inscribe la respuesta correspondiente a esa combinación de niveles (Comprobar con las tablas anteriores).

Este método de inspección es aproximativo pero muy útil sobre todo cuando no se hacen réplicas de las observaciones. Observamos que los valores de las bases del cubo son sistemáticamente menores que los del tope, indicando que el efecto de A es significativo, cuando A crece, la respuesta crece. En el factor B, si se compara la cara frontal con la trasera, ocurre lo mismo que con A. Para el factor C, ocurre que es significativo, pero en distinto sentido, ya que cuando C crece la respuesta baja. O, dicho de otro modo, **la respuesta aumenta cuando C *baja* de nivel**. Para alcanzar conclusiones más seguras debería decidirse un análisis estadístico más completo.

¿Qué hubiera pasado si hubiéramos invertido los niveles en algún factor? Por ejemplo, si hubiéramos elegido 8.18 para el nivel mínimo y 2.12 para el máximo de C.

¡El análisis habría sido igualmente válido! Lo que habría cambiado es que entonces el resultado para C sería: **la respuesta aumenta cuando C *sube* de nivel**. Esto es importante porque si un factor es cualitativo, como se ha dicho en el capítulo anterior, no sabríamos posiblemente como asignar su nivel ya que no tenemos datos numéricos. De modo que la asignación de los niveles a los factores no tiene importancia para el resultado del análisis, si este es correctamente interpretado.

2- 'Rankit method' (Prueba de probabilidad 'Normal')

Este método se fundamenta en que, si las respuestas fueran producto de las desviaciones experimentales **'exclusivamente'**, entonces debieran tener una distribución Normal. Pero los Factores e Interacciones **significativos** se desvían de la normalidad.

Prueba gráfica de Normalidad
Cálculo Rankit

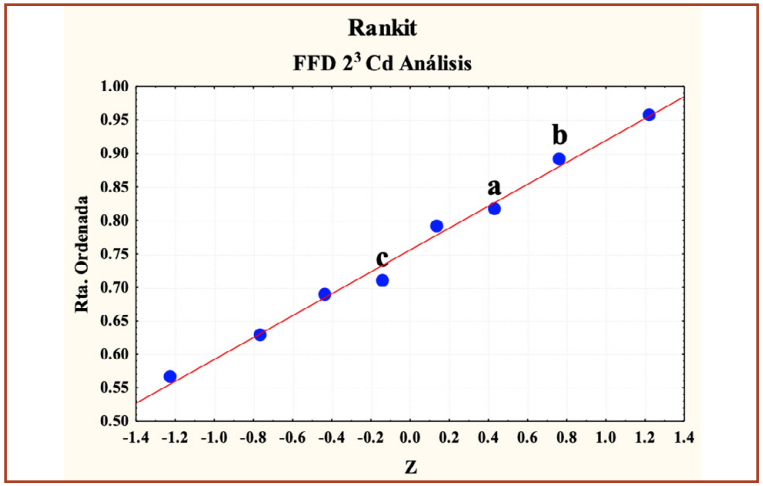
El primer paso para esta prueba consiste en **ordenar los datos** (Ref. 2) (en este caso las respuestas del análisis de Cd) de menor a mayor (o a la inversa).

Orden del diseño	0.8169	0.9566	0.7907	0.8918	0.6884	0.7089	0.5653	0.628
Orden de cálculo	0.5653	0.628	0.6884	0.7089	0.7907	0.8169	0.8918	0.9566

Luego se calcula la ***frecuencia relativa*** *%* =n/(n+1) x100 y se la ordena con los datos como se muestra a continuación: Primer dato:1/9.100=11.11%, segundo 2/9.100=0.22, último 8/9.100=0.89. Téngase en cuenta que el primer y último datos se toman como *las colas* de la distribución, por ejemplo, la primera fracción es de todos los hipotéticos datos menores al primero y en forma similar, para todos los hipotéticos datos mayores a la última fracción. Entonces la tabla total tendrá una casilla más que el número de datos. Luego se busca en una tabla de distribución normal la probabilidad (z) para cada **frecuencia relativa**. El programa *qqplot* de Matlab® hace este cálculo automáticamente.

n/(n+1)%	11.11	22.22	33.33	44.44	…	87.5%	100%
Datos Ordenados	≤0.5653	>0.5653≤0.628	…	…	…	>0.8918≤0.9566	>0.9566
Z	-1.221	-0.7647	…	…	…	0.7647	1.2206

Finalmente se construye un gráfico de los **datos ordenados** Vs. Z



Como se ve, no es tan fácil, en este caso, ver la desviación de los ensayos A y B que no caen más afuera de la recta. Esto se debe, a 2 problemas: Los efectos no son muy marcados, como se ve en la pequeña diferencia numérica y además, no tenemos repeticiones para conocer el error de los datos. Sin embargo, este es un buen método para estimar los factores significativos si no hay replicados.

Cuando no se han hecho replicados de los datos, por otra parte, es un excelente complemento del método de interpretación visual con los cubos.

3- Usando la desviación estándar de los efectos

Nota: para continuar con otros métodos de significación de los efectos ampliaremos los datos experimentales anteriores. Aunque los datos eran parciales pero reales, ahora los ampliaremos hipotéticamente suponiendo que repetimos las mediciones para validar mejor los resultados.

a) Por experimentos duplicados

Si se replican los experimentos, la varianza de experimentos duplicados viene dada por $\mathbf{s^2 = \Sigma\, d_i^2/(2.n)}$ donde $\mathbf{d_i}$ es la diferencia entre los replicados y **n** es el número de pares de datos (N=2.n).

Puede demostrarse que la varianza de los efectos vale $\mathbf{s^2_{efecto}=4.s^2/N.}$ Ahora podemos calcular la significación estadística de los efectos. Observe que $\mathbf{s_{efecto}}$ es común para todos los efectos.

$$\mathbf{Límite = efecto \pm t_{\alpha/2;n} \,.\, s_{efecto}} \qquad \mathbf{[2]}$$

Si el intervalo de confianza no incluye el cero, o lo que es lo mismo, el valor absoluto del efecto es mayor que el intervalo ($\mathbf{|efecto| > t_{\alpha/2;n} \,.\, s_{efecto}}$), el efecto se considera significativo. $\mathbf{t_{\alpha/2;n}}$ se extrae de la tabla de distribución t de Student.

Veamos los resultados para el ejemplo para el análisis de Cd:

Tabla 2: Análisis por experimentos duplicados						
Corrida	A	B	C	D	Resp. 1	Resp. 2
a	+	-	-	-	0.8918	0.8245
ab	+	+	-	-	0.9566	0.9655
abc	+	+	+	-	0.8169	0.798
b	-	+	-	-	0.7089	0.9001
bc	-	+	+	-	0.6884	0.6948
c	-	-	+	-	0.5653	0.7155
ac	+	-	+	-	0.7907	0.5705
-1	-	-	-	-	0.628	0.6338

Diferencia	d_i^2
0.0673	**0.0045293**
-0.0089	**7.921E-05**
0.0189	**0.0003572**
-0.1912	**0.0365574**
-0.0064	**4.096E-05**
-0.1502	**0.02256**
0.2202	**0.048488**
-0.0058	**3.364E-05**

Σd_i^2>>	**0.1126458**
t(0.025,8)>	**-2.306**
Err.>>	**0.01935**

Si comparamos ahora este error con el valor de los efectos calculados previamente con la ecuación [1], vemos entonces que todos los factores **son significativos**.

b) Despreciando las interacciones altas

Para este ejemplo demostrativo necesitamos un diseño FFD de 4 factores (2^4). Se trata estudiar el rendimiento (Rta) de un proceso que depende de (A), X en dilución 1; (B), X en dilución 2; (C) % de catalizador; (D) temperatura.

Casi siempre es difícil entender que significan las interacciones de tres factores. Por esta razón se las considera muchas veces como debidas al error experimental. Se puede, por lo tanto, deducir de ellas una estimación de la desviación estándar de los efectos. Comprobemos el cálculo aplicado a nuestro ejemplo.

Tabla 2B						
Run	Run	Rta	A	B	C	D
1	abcd	0.7	1	1	1	1
2	abc	0.3	1	1	1	-1
3	abd	0.5	1	1	-1	1
4	ab	2	1	1	-1	-1
5	acd	1.3	1	-1	1	1
6	ac	2	1	-1	1	-1
7	ad	1.3	1	-1	-1	1
8	a	4	1	-1	-1	-1
9	bcd	1.7	-1	1	1	1
10	bc	2.3	-1	1	1	-1
11	bd	1.3	-1	1	-1	1
12	b	3.5	-1	1	-1	-1
13	cd	3	-1	-1	1	1
14	c	2.3	-1	-1	1	-1
15	d	2.5	-1	-1	-1	1
16	(-1)	2.5	-1	-1	-1	-1

Efecto de los Factores Principales				
	A	B	C	D
Sum +	1.5125	1.5375	1.7	1.5375
Sum-	2.3875	2.3625	2.2	2.3625
Dif=Efecto	-0.875	-0.825	-0.5	-0.825

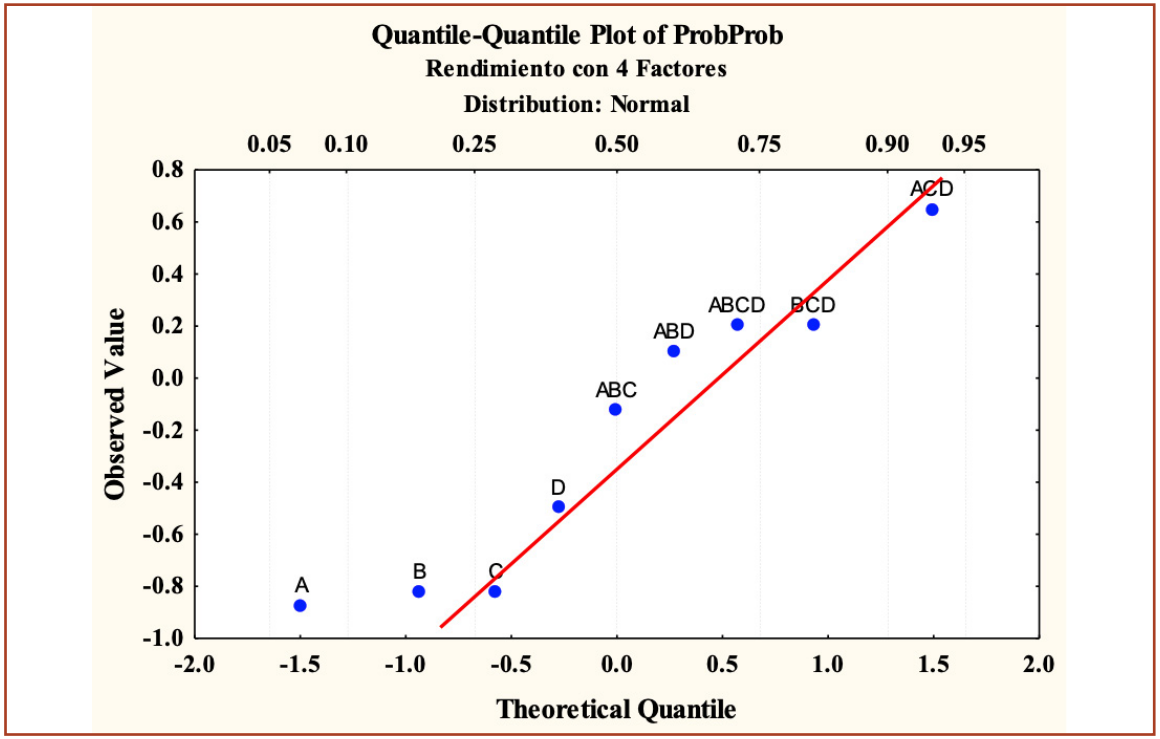
Efecto de las interacciones					
	ABC	ABD	ABCD	BCD	ACD
Sum +	2.05	2.31	1.89	2	2.05
Sum-	1.85	1.67	2.01	1.9	1.85
Efecto	**0.2**	**0.65**	**-0.13**	**0.1**	**0.2**

Estimación de la desviación estándar de los efectos de asociaciones altas		
Factor	Efecto	(Efecto)2
Efec.ABC>	0.2	0.04
Efec.ABD>	0.647	0.419
Efec.ABCD>	-0.125	0.015625
EfecBCD>	0.1	0.01
Efec.ACD.>	0.2	0.04

Cálculo del error límite

Promedio de la suma de cuadrados de los efectos de asoc. **altas**: 0.105
Sefecto=Raiz(0.105)= 0.324
t inv(0.95,5)=2.015
Error límite=$t_{\alpha,n.}$ Sefecto = **0.653**

Como se ve, en este caso, A,B y D son significativos, pero las interacciones altas no. En este método, cuando alguna de las interacciones altas resulta significativa o es cercana a la de los efectos, puede que ningún factor resulte significativo, lo que es sospecha de algún error. Comparemos con un QQ plot del resultado anterior, ya que no tenemos repeticiones. Se observa que para C y D el resultado no coincide.



Se puede ahora interpretar mejor la importancia de poder hacer mediciones con repetición, ya que los resultados son mucho más precisos y seguros

Significancción estimada por ANOVA

Cuando se hace un replicado de los experimentos el análisis de los factores puede hacerse con una tabla ANOVA. Este método puede ser recomendado como el más conveniente, pero significa duplicar las experiencias. Para interpretar los resultados siguientes, para quienes no estén familiarizados con ANOVA, es imprescindible ver cuidadosamente el Anexo de este capítulo.

En la tabla siguiente se muestran los resultados para 3 factores, ANOVA de tres vías, para el ejemplo de análisis de Cd.

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
A	0.07286	1	0.07286	10.35	0.0123
B	0.05165	1	0.05165	7.34	0.0267
C	0.04721	1	0.04721	6.71	0.0321
A*B	0.00001	1	0.00001	0	0.9772
A*C	0.01297	1	0.01297	1.84	0.2118
B*C	0.00242	1	0.00242	0.34	0.5736
A*B*C	0.00535	1	0.00535	0.76	0.4086
Error	0.05632	8	0.00704		
Total	0.2488	15			

Constrained (Type III) sums of squares.

Este resultado confirma el de los métodos gráficos de los cubos, el de Rankit y el de los experimentos duplicados en cuanto a que todos los factores principales son significativos. También muestra que ninguna asociación resulta significativa (incluyendo la de ABC **que parecía** significativa en el método Rankit).

Cuando se trate de interpretar el análisis de los factores es conveniente aplicar **todos los métodos que se tengan a mano** y comprobar la coherencia entre los resultados para despejar las dudas que puedan plantearse entre los métodos y analizar sus causas.

Modelado por Cuadrados Mínimos (importancia de la ortogonalidad)

Para un modelo 2^2, ANOVA conduce a un modelo lineal de efectos fijos del tipo $y = \mu + a + b + (ab) + e$. Los efectos a, b y ab pueden ser asociados a las variables x_1, x_2 y a la interacción entre ambas, obteniéndose un modelo isomorfo,

$$y = b_0 + b_1.x_1 + b_2.x_2 + b_{12}.x_1.x_2 + e \qquad [3]$$

Esta es una ecuación de regresión donde b1 estima a β_1 y β_1 es una medida del efecto de x_1. Para calcular un modelo cualquiera las variables originales deben estar escaladas usualmente en el rango (1,-1), (ver ecuación de escalado [4]).

Desde el punto de vista del cálculo de la regresión, este modelo presenta la ventaja de que las variables no están correlacionadas, son ortogonales (la matriz $(\mathbf{X^T.X})$ es diagonal).

Vamos a verificar esto reemplazando en la tabla 1 los signos '+' por '1' y los signos '-' por '–1'. Como en este ejemplo se consideraban 3 factores podemos construir el simple modelo $y = b_1.x_1 + b_2.x_2 + b_3.x_3$. La matriz $(\mathbf{X^T.X})$ es:

$$\mathbf{X^TX} = \begin{vmatrix} 8 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 8 \end{vmatrix} = 8 \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 8.\mathbf{I}$$

Vimos en el ejemplo introductorio al diseño experimental (Capítulo 7) que: $\mathbf{V}(b) = se^2$. $(\mathbf{X^T.X})^{-1} = s_e^2.\mathbf{I}/8$. Resultando que la covarianza entre los coeficientes de regresión en $\mathbf{V}(b)$ valen cero. O sea que, b_1, b_2 y b_3, los estimadores de β_i, **no están influenciados unos con otros**.

Se ve entonces cuál es la razón fundamental del diseño factorial, **el orden experimental no es caprichoso**, sino que responde a las **ventajas de la ortogonalidad de las variables.** Y también **la necesidad del escalado de variables originales**.

Los límites de confianza de los parámetros de la regresión se pueden calcular también en forma independiente $\beta_i = \mathbf{b_i} \pm \mathbf{t_{n-p}}\ \mathbf{s_{bi}}$ y el coeficiente es considerado significativo si el intervalo no incluye el cero.Recordemos que los $\mathbf{s_{bi}}$ componen la diagonal principal de **V(b)** y por lo tanto, en este cálculo, como en el de ANOVA, los límites de confianza del efecto de las variables se conocen individualmente para cada factor y no en general, lo que significa información más detallada.

Debe tomarse en cuenta que la regresión se lleva acabo con **valores escalados** entre (-1, +1) y que deberán re expresarse los coeficientes como función de las variables originales. Si el sufijo **'o'** indica valores originales de las variables 1,2,...,etc. y el sufijo **'e'** indica valores escalados; y además el sufijo **min** significa valor mínimo y el sufijo **max** significa valor máximo; entonces puede aplicarse la siguiente ecuación en cualquiera de los sentidos.

$$\frac{O_x^o - O_{min}^o}{O_{max}^o - O_{min}^o} = \frac{E_x^e - E_{min}^e}{E_{max}^e - E_{min}^e} \qquad [4]$$

La ecuación final se obtiene reemplazando estas igualdades por x_1, x_2 y x_3 en la ecuación $\eta = \beta_0 + \beta_1.x_1 + \beta_2.x_2 + \beta_3.x._3$, o en una ecuación mayor que considere las interacciones, dejándola en función de los x_i. En los casos donde existen factores de segundo orden puede ocurrir que, al describir el modelo con las unidades originales, las operaciones algebraicas hagan aparecer asociaciones que en el modelo ortogonal no figuran.

Causales de Errores
Efecto de valores aberrantes (o 'outliers')

Estos valores son respuestas erróneas debidas a fallas no deseadas (instrumentales, confusiones, tipeos, etc.) alejadas del valor verdadero y que pueden pasar desapercibidas.

1- Si se está utilizando el método del duplicado para estimar s y s_{efecto}, probablemente la gran mayoría, o todos, los efectos aparecerán como no significativos debido a la sobrestimación de s. Esto es causa de sospecha de la presencia de valores aberrantes.

2- Cuando hay más de 3 factores y se aplica el método de determinar s_{efecto} despreciando las altas interacciones, ocurre que los valores aberrantes conducen a sobrestimar s_{efecto} (por los valores altos que toman las interacciones) y los efectos principales pueden resultar no detectados.

3- Otra causa de error es que unos pocos puntos fueran determinados fuera del rango normal de operación. Si bien estas no son respuestas erróneas, pertenecen a un dominio no investigado.

Efectos de bloqueo y aleatoriedad

Un orden experimental, temporal o espacial, impensadamente tendencioso, conduce a fuentes adicionales de variación **que aparecerán dentro de los efectos.**

Como ejemplo comparemos la tabla 1 con la siguiente, 3, que está ordenada experimentalmente para evitar efectos de bloqueo cuando el trabajo debe desarrollarse durante 2 días. Como se ve, si se parte por la mitad la tabla 1, todas las experiencias del primer día tienen el Factor A con nivel '+' y en el segundo día todas son A '-'. En este caso, cualquier impensada variación experimental entre el primer día y el segundo quedará afectando artificialmente al factor A.

Esto no ocurrirá si las observaciones se disponen como en la tabla 3.

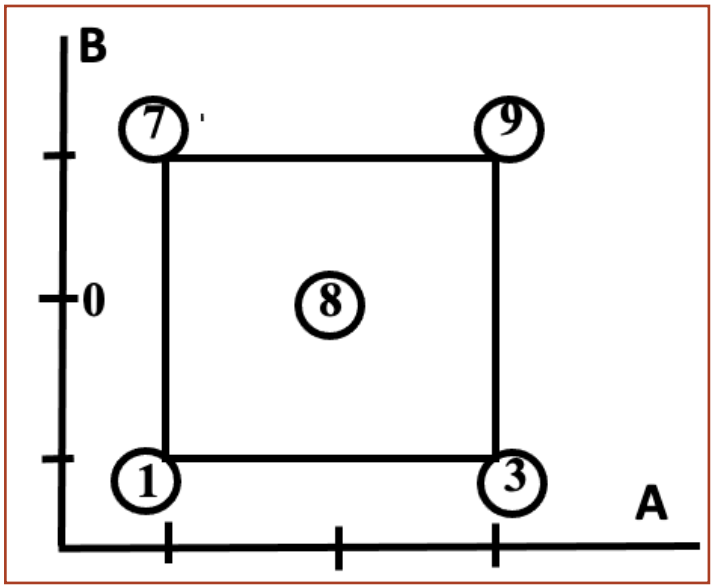
Tabla 3

Bloc día 1			
Exp.	**A**	**B**	**C**
1	-	-	-
2	+	+	-
3	+	-	+
4	-	+	+
Bloc día 2			
5	+	-	-
6	-	+	-
7	-	-	+
8	+	+	+

Investigación de la curvatura

Con diseños factoriales de dos niveles obtendremos solamente **modelos lineales** en el espacio multivariable. Si una variable no sigue una ley lineal en relación a la respuesta, su incidencia será 'promediada' a la de una recta, en un modelo lineal.

Podemos investigar si existe curvatura en alguna de las variables.



Un método es hacer varias mediciones en el punto central del diseño como muestra la figura anterior de análisis visual para dos variables. Esta es una práctica siempre aconsejable para cualquier diseño que se practique porque cumple varios fines.

Una simple inspección visual permite apreciar la presencia de curvatura cuando esta es prominente, pues el centro está lejos de ser aproximadamente el promedio entre el nivel '+' y '–' de B, que debería ser próximo a 5 y no a 8.

Cuando es difícil apreciar a simple vista, también se puede calcular si la curvatura es significativa utilizando la diferencia entre las medias de las esquinas (**e**) y la del punto central (**c**). Comparando el t experimental (ec. 5) con el crítico, t'(ec.6), si $|t| > |t'|$ entonces las medias no son iguales y existe curvatura.

$$t = \frac{\bar{y}_e - \bar{y}_c}{\sqrt{\left(\frac{s_e^2}{n_e} + \frac{s_c^2}{n_c}\right)}}$$

[5]

$$t' = \frac{t_1 \cdot \left(\frac{s_e^2}{n_e}\right) + t_2 \cdot \left(\frac{s_c^2}{n_c}\right)}{\left(\frac{s_e^2}{n_e} + \frac{s_c^2}{n_c}\right)}$$

[6]

t_1 representa el valor crítico para α=0.05 y n_e-1 grados de libertad y lo mismo vale para t_2 con n_c-1 grados de libertad.

Otro método es evaluar si los términos de un modelo cuadrático son significativos o no, es el siguiente:

Supongamos $y = b_0 + b_1.x_1 + b_2.x_2 + b_{12}.x_1.x_2 + b_{11}.x_1^2 + b_{22}.x_2^2.$

Si bien este modelo, para ser evaluado, requeriría más medidas experimentales que las previstas para un modelo lineal, puede demostrarse que:

$$b_{11} + b_{22} = \bar{y}_e - \bar{y}_c$$

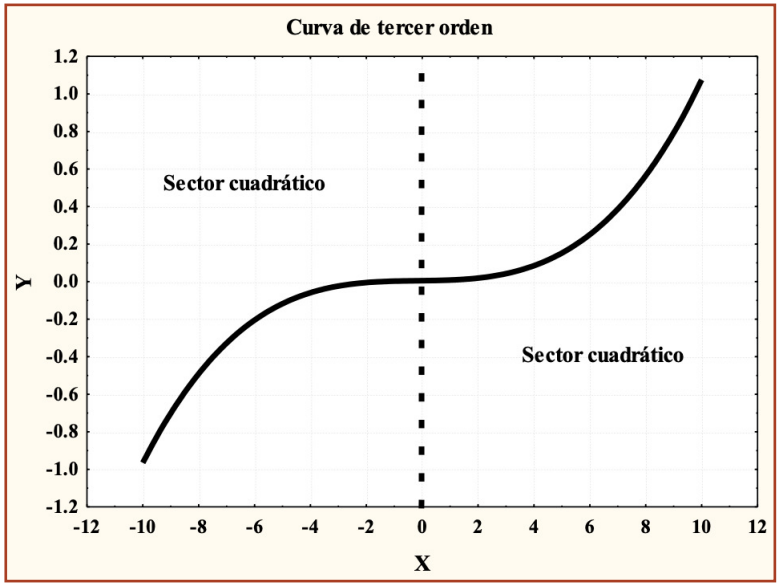
Si b_{11}+ b_{22} es significativamente diferente de cero, puede considerarse que existe curvatura.

Número de términos

Hasta ahora hemos visto el **número de experimentos** necesarios para desarrollar un modelo Full Factorial, y este número depende del número de factores involucrados en el modelo. Sin embargo, no hemos tenido en cuenta hasta ahora el **número de términos** que necesitamos para desarrollar un modelo estadísticamente lineal. Estos modelos son matemáticamente descriptos, en general, como un polinomio. Para un modelo cuadrático de 2 factores, por ejemplo, la expresión general es:

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \sum_{j=1}^{k} \beta_{ij} x_i x_j + e$$

k es el número de factores, *y,* es la respuesta del modelo, β son los coeficientes estimados, x_i y x_j son los factores del modelo y *e* un término que toma en cuenta el error del modelo. La ecuación anterior incluye un término independiente (β_0), k efectos principales (βi), [k(k-1)/2] interacciones de 2 factores (β_{ij}, i≠j) y k términos cuadráticos (βii). Para k factores, **el número mínimo** de términos para desarrollar modelos similares es **(k+1).(k+2)/2**. Cuando las variables son mayores a 2, los términos con interacciones de más de 2 factores son raramente necesarios y difíciles de interpretar. Tampoco es común desarrollar modelos de orden mayor a 2 por más de una razón; en primer lugar, el número de términos crece rápidamente y además, en estos casos conviene achicar el rango del modelo y desarrollarlo en forma dividida en dos modelos de orden 2.



Como se ve en la figura, para representarla con un modelo se requeriría un orden 3. Pero podemos dividirla en 2 sectores de orden 2, lo que resultaría más sencillo y probablemente más preciso cuando están presentes los errores experimentales

Anexo

El análisis de la varianza en la evaluación de modelos

Regresión múltiple

El análisis de una regresión es el proceso mediante el cual se determina cómo una variable "y" (variable dependiente), está relacionada con una o más variables x_1, x_2, ...x_n (variables independientes).

El modelo de regresión más comúnmente utilizado es el modelo de regresión lineal:

$$y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_m x_{jm} + \varepsilon_j \qquad j = 1,2,\ldots,n; \quad i = 1,2,\ldots,m \qquad [1]$$

Donde **n es el número observaciones** (objetos) y **m es el número de parámetros**. Los ε_j's son los errores de predicción del modelo los cuales son independientes e idénticamente distribuidos como una distribución normal $N(0,\sigma)$.

Una de las mayores cualidades de las regresiones múltiples es que posibilitan determinar las propiedades estadísticas de los parámetros estudiados y conformar un modelo.

Varianza

El análisis de la varianza o (ANOVA) es primariamente un método para identificar cuales de los β's en un modelo lineal no son cero (o insignificantes), lo cual equivale a preguntar ¿la presencia de alguno de los factores tiene algún efecto en la respuesta? Las estimaciones de los β's se encuentran por estimación de cálculos de cuadrados mínimos.

Por otro lado, los valores ajustados, son aquellos que en un modelo representan el **valor calculado** de la respuesta acorde a los mínimos cuadrados con respecto a los **valores observados** (y_j). Las diferencias entre los valores de la variable dependiente observados (o medidos) y los valores ajustados, $r_j = y_j - \hat{y}_j$, se denominan residuos (r_j):

Siendo:

$\hat{y}_j = \hat{\beta}_0 + \sum_{i=1}^{m} \hat{\beta}_i x_{ji}$ $\hat{\beta}_i$ son los valores estimados de los β_i en [1].

Los residuos se relacionan con σ^2, la varianza de los errores εj's, de modo que podemos estimar σ^2 a través de los datos por medio de los cuadrados medios (S^2) que en este caso particular estimará la varianza del error:

$$S^2 = \frac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{n-(m+1)} = \frac{RSS}{\nu}$$

Donde n es el número total de datos u observaciones y m es el número de parámetros del modelo. El numerador se denomina comúnmente **suma de cuadrados de los residuos** (RSS) y el denominador son los **grados de libertad de los residuos** (n).

Supongamos que para un determinado conjunto de datos ajustamos un modelo lineal particular, que llamaremos $\Omega 1$. La varianza de los ε_j's, σ^2, puede estimarse entonces de acuerdo a:

$$S_1^2 = \frac{RSS_1}{\nu_1}$$

Donde RSS_1 y ν_1 son la suma de cuadrados de los residuos y los grados de libertad de los residuos, respectivamente. Esta es una estimación válida aun si algunos o todos los βi's son cero.

Si queremos investigar cuales β's son cero, ajustamos un modelo más chico Ω_0, en el cual esos β's y sus correspondientes variables independientes están ausentes. De modo que podemos obtener otra estimación de la varianza para ese modelo Ω_0, de acuerdo a:

$$S_0^2 = \frac{RSS_0}{\nu_0}$$

Donde RSS_0 y ν_0 son la suma de cuadrados de los residuos y los grados de libertad de los residuos respectivamente para el modelo Ω_0.

Si los β's que están ausentes en Ω_0 son realmente cero, entonces ambas estimaciones de σ^2, S_0^2 y S_1^2, son válidas y deben ser aproximadamente iguales. En el otro caso, si algunos de los β's que están ausentes en Ω_0 no son realmente cero sino que son significativos, entonces S_1^2 es aún válida para estimar σ^2 pero S_0^2 tenderá a ser mayor que σ^2 (porque el error será más grande). En principio entonces, podemos comparar S_0^2 con S_1^2. Si S_0^2 es relativamente mayor que S_1^2, entonces algunos de los β's ausentes no son cero y Ω_0 es un modelo insatisfactorio. Si S_0^2 y S_1^2 son similares, esto sugiere que Ω_0 es un modelo satisfactorio.

Una comparación mucho mejor se realiza calculando una tercera estimación de σ^2 de acuerdo a:

$$S_E^2 = \frac{ESS}{\nu_E}$$

Donde $ESS = RSS_0 - RSS_1$, es la **suma extra de los cuadrados** y ν_E los grados extra de libertad calculados como $\nu_E = \nu_0 - \nu_1$ .La ventaja de S_E^2 sobre S_0^2 es que la primera es más sensible a los β's ausentes que no son cero debido a que la diferencia es más sensible.

Las varianzas pueden compararse utilizando una prueba de hipótesis nula ("test de hipótesis", en este caso, el "F test") para analizar si un factor tiene algún efecto en el comportamiento del modelo (o sea, en las respuestas). El "F test" (tabla de Fisher) es uno de los posibles ensayos de hipótesis, definido como $F = S_1^2/S_2^2$, siendo S_1 y S_2 las desviaciones estándar comparadas (con $S_1 > S_2$). Realizando ahora el test de hipótesis, queremos evaluar si: Ω_0 es adecuado o no (o sea $\Omega_0 = \Omega_1$ o $\Omega_0 \neq \Omega_1$). Si el test de hipótesis indicase que Ω_0 no es el adecuado, entonces se utiliza Ω_1.

$$F = \frac{S_E^2}{S_1^2} = \frac{ESS/\nu_E}{RSS_1/\nu_1}$$

Si Ω_0 es un modelo adecuado, F debe ser aproximadamente 1 mientras que si Ω_0 no lo es, F será mayor que el valor crítico de F para una confiabilidad α y (ν_E , ν_1) grados de libertad.

El F test permite comparar un efecto determinado con el residuo y determinar si las distribuciones son similares entre sí o no.

La Tabla ANOVA

Los resultados del análisis de la Varianza se resumen frecuentemente en una tabla típica ANOVA (o MANOVA= múltiple ANOVA), en la cual se resumen los cómputos para dos (o más) tratamientos (o factores) distintos (1 y 2), con niveles l y k respectivamente, y donde ν_r representa los grados de libertad de los residuos y ν_t los grados de libertad totales. En esa tabla se muestran principalmente los parámetros que se describirán a continuación.

Efecto *df*: grados de libertad (ν);
RSS: suma de cuadrados debida al efecto estudiado;
Efecto MS: cuadrados medios (S^2);
F test: test que compara las varianzas;
Nivel p: valor experimental de α. En general y en nuestro caso particular un valor de referencia crítico $p< (\alpha = 0,05)$ significa que el factor es significativo.

Efecto	Efecto *df*	Efecto MS	Test F	Nivel p
Factor 1	ν_1=l-1	$S_1^2=RSS_1/(l-1)$	S_1^2/S_R^2	
Factor 2	ν_2=k-1	$S_2^2=RSS_2/(k-1)$	S_2^2/S_R^2	
Residuos	$\nu_r= \nu_t -[(k-1)+(l-1)]$	$S_R^2=RSS_R/\nu_r$	S_R^2/ν_r	
Total	ν_t=kl-1	RSS_T		

Referencia sobre este tema. (Ref. 3)

Referencias

1. C.F. Jeff Wu and Michel Hamada. Experiments. Planning, Analisis and Parameter Designs Optimization. John Willey & Sons, Inc. USA, .2000.

2. D.L. Massart; B.G.M. Vandeginste; L.M.C. Buydens; S. De Jong; P.J. Lewi and J. Smeyers-Verbeke. Hanbook of Chemometrics and Qualimetrics. Part A.Capítulo 3 pág 63.

3. Michael Berthold and David J. Hand (Eds.). Intelligent Data Analysis. Springer, Berlin Heidelberg New York 2003. Second edition.

Lecturas recomendadas

D.L. Massart; B.G.M. Vandeginste; L.M.C. Buydens; S. De Jong; P.J. Lewi and J. Smeyers-Verbeke. Hanbook of Chemometrics and Qualimetrics. Part A. Capítulo 22. Elsevier, Amsterdam 1997.

Douglas c. Montgomery. Diseño y Análisis de Experimentos. segunda edición.

Editorial Limusa, SA de C.V. Grupo Noriega Editores. Balderas 95, Mexico, D.F.

CAPITULO 9

Diseño Factorial Fraccionario y Diseños Reducidos

Diseño Factorial Fraccionario

Introducción. Hagamos una revisión de un diseño factorial **total** para un número relativamente alto de factores, por ejemplo 7 factores:

Efectos estimados	*N° de efectos*
Valor medio	1
Efectos principales	7
Efectos de 2 Factores	21
“ de 3 “	35
“ de 4 “	35
“ de 5 “	21
“ de 6 “	7
“ de 7 “	1

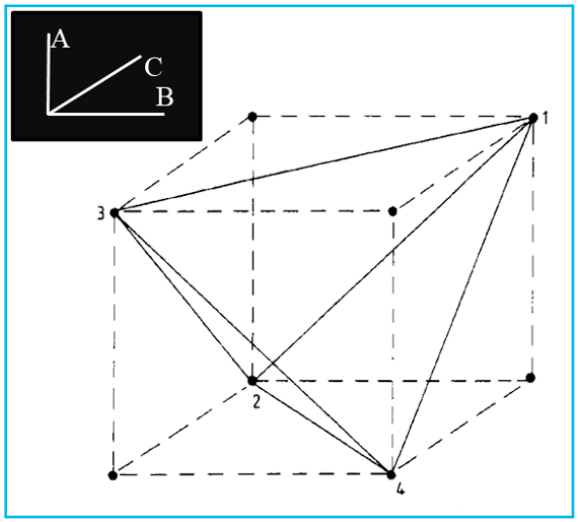
Considerando que, a los *efectos de tres o más factores*, **usualmente**, no se les asigna importancia; se ve que hay aquí gran redundancia de mediciones. Sería posible entonces, **en estos casos**, definir diseños experimentales más chicos (Ref. 1).

Estos nuevos diseños deberían conservar las siguientes características:

1. Deben ser balanceados en niveles.
2. Deben ‘mapear’ el dominio experimental tan bien como sea posible.
3. Deben conservar la ortogonalidad.

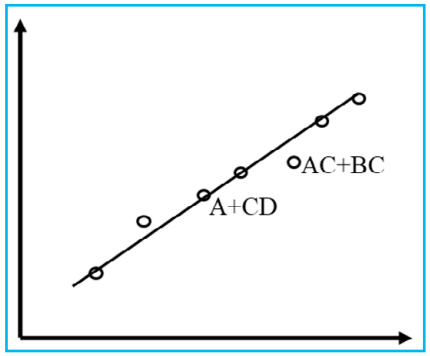
Una forma de hacer esto es dividir el diseño factorial total por 2, 4, ...2^{k-p} y obtener **una fracción** de los experimentos del diseño factorial total. ‘p’ es un entero menor que k.

Observe en la figura la forma en que se 'mapea' el espacio para estudiar 3 factores. Lo que en FFD requiere 8 experiencias ahora con un diseño 2^{3-1} se reduce a la mitad.



La pérdida de información debida al menor número de datos se manifiesta en que los factores que es posible calcular por el método habitual de las sumatorias, están ahora "confundidos", es decir, se suman 2 o más factores en un mismo efecto. Si se suman las expresiones del Efecto A más las del Efecto BCD del diseño factorial para 3 factores (2^{4-1}=8) se ve que:

$$\textbf{Efecto (A+BCD)} = \tfrac{1}{4}\,(y_1 + y_2 + y_3 + y_4 - y_5 - y_6 - y_7 - y_8)$$



Observe que se repite la expresión general Efecto=promedio de los niveles '+' – promedio de los niveles '-'.

Otros pares similares son: y_{media}+ABCD, B+ACD, AB+D, C+ABD, AC+BD, BC+AD, D+ABC.

En un gráfico del tipo "rankit" podemos ver como aparecen los 'efectos sumados' respecto del de los factores individuales. Cuando aparecen cerca del alineamiento general, esto nos indica que no son significativos, diferente de lo que es el efecto de AC+BC. El problema con estos diseños es que cuando los efectos de **factores individuales** se suman a los de

una interacción de orden superior, como por ejemplo los mencionados A+CD, u otros, es difícil distinguir entre el efecto de la interacción y del factor individual.

Se pueden hacer diseños aún más chicos 1/4, 1/8 o en general 2^{k-p}, hasta alcanzar un **diseño fraccional "saturado",** es decir el número mínimo posible de experimentos (recordemos que debe cumplirse siempre que **Número de experimentos ≥ Número de factores**).

Cuanto más fraccionado es el diseño mayor es la cantidad de factores que se confunden entre sí (3, 4, etc.).

Si después de practicar un diseño fraccionario se quieren realizar más experimentos para completar la información se puede aumentar el diseño. Por ejemplo, de un diseño fraccionario "mitad" puede ejecutarse la otra hemi-fracción y obtener un diseño total. En muchos casos, entonces, en que se trabaja con muchos factores, se comienza con diseños factoriales fraccionarios que luego se van aumentando para mejorar la interpretación. Cuando se procede de esta manera se debe tener cuidado en evitar el bloqueo.

Cabe preguntarse ahora ¿cómo encontramos los factores confundidos? Y ¿cómo generamos las plantillas de experimentos? Para hacer esto necesitamos previamente definir los *definidores de contraste* y los *generadores*.

Definidores de Contraste y Generadores

Anteriormente hemos indicado que, para deducir los niveles de una interacción, por ejemplo, AB en una planilla experimental, teníamos que multiplicar la columna de niveles de A por la de B. Si multiplicásemos la columna de un factor cualquiera por si misma obtendríamos siempre una columna de signos '+' exclusivamente. A esta operación la designamos con 'I' y le damos el valor algebraico '1'. Entonces:

$$A^2=B^2=...=I=1$$

Si quisiéramos hacer un diseño *mitad* para cuatro factores, los $16=2^4$ experimentos de un FFD quedarían reducidos a 8. Un diseño mitad debemos expresarlo como $2^{(4-1)}=2^3$. Tomamos la plantilla del FFD para 3 factores (A, B y C) que se corresponde con 8 experiencias y agregamos el cuarto factor D con los niveles obtenidos multiplicando los de A.B.C. Ahora tenemos una planilla completa de **8 experiencias para 4 factores.** Y también tenemos:

$$ABC=D, \text{ luego } \quad A^2=B^2=C^2=D^2=ABCD=I=1$$

Llamaremos a la relación ABCD=I ***definición de contraste***.

La definición de contraste nos sirve para averiguar todos los factores y asociaciones que aparecen confundidos en el diseño. Para ello no hay más que multiplicar a esta con cada uno de los factores y asociaciones. Por ejemplo:

$$A.I=A^2BCD=BCD, \text{ o sea A se confunde con BCD}$$

$$AB.I=A^2B^2CD=CD, \text{ o sea AB se confunde con CD}$$

El resto de factores confundidos para este ejemplo son B=ACD, C=ABD, D=ABC, AC=BD y AD=BC. Ahora tenemos completa nuestra planilla de diseño mitad para 4 factores, que quedaría así:

Experimento	A+BCD	B+ACD	C+ABD	D+ABC
1	+	+	+	+
2	+	+	-	-
3	+	-	+	-
4	+	-	-	+
5	-	+	+	-
6	-	+	-	+
7	-	-	+	+
8	-	-	-	-

Tenga en cuenta que **el encabezamiento de las columnas indica el nivel del factor individual para el experimento**, por ejemplo, el nivel de A para la primera columna expresa, sin embargo, **el efecto observado,** que es A+BCD. Para calcular el efecto de las asociaciones tales como AD+BC, los niveles para las sumatorias se obtienen por el método habitual de multiplicar las columnas, por ejemplo, el efecto de AD+BC se puede obtener multiplicando la primera columna con la cuarta (A,D) o la segunda con la tercera (B,C).

$$A.(D+ABC)=AD+A^2BC=AD+BC \quad y \quad B+(C+ABD)=BC+AD$$

Obsérvese que para generar este diseño hemos **seleccionado** a la asociación ABC para confundirla con D. A la relación D=ABC se la llama ***el generador***. Uno puede elegir distintos generadores con lo que se obtendrán diseños diferentes en los cuales los factores estarán confundidos de manera distinta.

Supongamos que queremos hacer un diseño ¼ para 5 factores, $2^{5-2}=2^3$. Tenemos 5 factores, comenzamos con 3 (A, B y C), parecería lógico sacrificar la asociación ABC y hacer ABC=D. Ahora nos falta un factor más y podríamos elegir sacrificar por ejemplo BC haciendo BC=E, el quinto factor. Debido a que cada generador originará un diseño diferente es necesario designar los diseños según su ***resolución***.

Resolución de los diseños

La resolución **R** depende del generador y por lo tanto de los definidores de contraste; se la expresa en números romanos. **La resolución es igual a la más corta de las definiciones de contraste**, por ejemplo, en un diseño 2^{5-2} podría ser que las más cortas definiciones de contraste fuesen ABC y ADE, entonces la resolución es R=III. Existen ciertas reglas generales que explican la utilidad de R.

- En los diseños de R=III los efectos principales no son confundidos entre sí sino con interacciones de dos factores.

- En los diseños de R=IV los efectos principales no son confundidos entre sí ni tampoco con interacciones de dos factores. Las interacciones de dos factores son confundidas entre sí.

- En los diseños de R=V los efectos principales y las interacciones de dos factores no son confundidos uno con otros.

- Los programas de computación buscan proponer diseños con la más alta resolución posible. Se puede hacer manualmente eligiendo un buen generador, pero no es fácil.

Los diseños factoriales fraccionarios pueden usarse como ***diseños de screening*** para eliminar factores que no son significativos y luego hacer diseños más pequeños. Existe un diseño algo más reducido que los factoriales fraccionarios para ese fin, que veremos a continuación, pero que es más limitado en su aplicación y más difícil de implementar sin información previa.

Diseños de *screening*

Cuando uno tiene que estudiar la/las respuestas de un sistema con varias variables, la primera pregunta que surge es si todas las variables tienen efecto sobre la/las respuestas o no. Ya sabemos que el número de variables incrementa potencialmente (2^k) el número de experimentos necesarios, por lo tanto, una disminución del número de variables achicaría el diseño. Para ello se utilizan los diseños de ***screening***, es decir, un diseño pequeño que permita eliminar todas aquellas variables que no son significativas. El más interesante de estos diseños es el de Plackett-Burman (Ref.2).

Diseño Plackett-Burman

Hemos visto que en los diseños factoriales fraccionarios los efectos principales se confunden con las asociaciones. El diseño Plackett-Burman (PB) se caracteriza porque, **si todas las interacciones son despreciables**, entonces los efectos principales **no están confundidos y tienen las varianzas más pequeñas posibles**. Otra ventaja es el reducido número de experimentos necesarios: Es un diseño experimental para 4xN experimentos (4, 8, 12, ...etc) para un máximo de hasta 4xN-1 factores (3, 7, 11, ...etc.). Por ejemplo 12 experimentos (N=3) para un máximo de 11 factores surgen de la primera línea del diseño

+ + - + + + - - - + -

Las 10 líneas siguientes se agregan por permutación cíclica de ésta (ver el ejemplo) y se agrega al final una línea completa de niveles "-". El diseño puede estar transpuesto, como en el ejemplo, y es igualmente válido.

Los efectos se calculan del modo usual, por ejemplo, para 12 corridas

$$\textbf{Efecto} = 1/6[\Sigma y_{+} - \Sigma y_{-}]$$

Cuando el número de columnas es mayor al necesario para los factores calculados, se pueden dejar como "factores fantasmas", que son factores que no significan nada, pero sirven para calcular $\mathbf{S_{efecto}}$ y determinar la significación estadística de los efectos que interesan.

El siguiente ejemplo se refiere a la optimización de un equipo de cromatografía líquida acoplado a espectrografía de masa (LC-MS) para una determinada familia de compuestos (Ref. 3). Se estudian 8 controles del equipo con un diseño PB de 12 experimentos. Además de los 8 factores estudiados quedan 3 factores fantasmas para el cálculo de los factores significativos. En este caso se utilizó t(0.05,3)= 2.3534 que representa error de 10% para los 3 factores fantasmas (J,E,B).

Run	ESI V	J	DL V	Qa DC V	Qa RF V	T HB	E	T DL	F DG	B	F NG	Respuesta
1	+	-	+	-	-	-	+	+	+	-	+	88840.5
2	+	+	-	+	-	-	-	+	+	+	-	11925.5
3	-	+	+	-	+	-	-	-	+	+	+	667101
4	+	-	+	+	-	+	-	-	-	+	+	82037.5
5	+	+	-	+	+	-	+	-	-	-	+	694476
6	+	+	+	-	+	+	-	+	-	-	-	871456
7	-	+	+	+	-	+	+	-	+	-	-	45578
8	-	-	+	+	+	-	+	+	-	+	-	40823
9	-	-	-	+	+	+	-	+	+	-	+	6915.5
10	+	-	-	-	+	+	+	-	+	+	-	1318695.5
11	-	+	-	-	-	+	+	+	-	+	+	31242
12	-	-	-	-	-	-	-	-	-	-	-	10873.5

Factor	Unidad	Nivel +	Nivel -
ESI V	KV	-4.5	-2.5
J			
DL V	V	-80	-20
Qa DC V	V	-70	-10
Qa RF V	V	60	10
T HB	°C	450	200
E			
T DL	ªC	250	100
F DG	L/min	18	6
B			
F NG	L/min	1.5	0.7

Cálculos	Sum. Cuad.
J,E,B	
Σ Efectos^2	3.09E+10
ΣSe^2/3-->	1.03E+10
$\sqrt{Se^2}$	1.01E+05
Ecritico	2.38E+05

	ESI V	J	DL V	Qa DC V	Qa RF V	T HB	E	T DL	F DG	B	F NG
Efecto--->	3.77E+05	1.29E+05	-4.64E+04	-3.51E+05	5.55E+05	1.40E+05	9.49E+04	-2.95E+05	6.80E+04	7.23E+04	-1.21E+05

Todo efecto mayor (en valor absoluto) que Ecrítico es considerado significativo. Comparando el Ecrit con los valores absolutos de los Efectos se ve que DLV,THB,FDG y FNG no son significativos; J,E y B, aunque son mayores que Ecrit, no significan nada. Como antes, el signo negativo indica que el efecto del factor es contrapuesto al de la respuesta para los valores asignados a los niveles (cuando la respuesta aumenta el factor disminuye, y viceversa).

Una particular aplicación de los diseños de *screening* es la medida de la *robusticidad* de un proceso. Cuando se ha desarrollado un proceso, en este caso un procedimiento de medida como esta técnica químico analítica, uno quiere conocer si pequeños apartamientos de los parámetros del proceso tienen influencia sobre la eficacia de la medida.

A continuación, se dan otras líneas generadoras. Para valores de N mayores el lector debe referirse al trabajo original de Plackett-Burman (Ref. 2).

4xN=8 + + + - + - -

4xN=16 + + + + - + - + + - - + - - -

4xN=20 + + - - + + + + - + - + - - - - + + -

Como se ha dicho anteriormente, si existe alguna interacción, este análisis no es válido, sino que conduce a errores. Sin embargo, cabe preguntarse ¿Hasta qué punto esta afirmación es válida? Y además, ¿Podemos averiguar algo acerca de las interacciones? En efecto, sí es posible, pero hay que recurrir a métodos más avanzados, por ejemplo: los ya mencionados algoritmos genéticos u optimización Bayes-Gibbs. Entonces, es posible averiguar tanto la validez del método como la significación de las interacciones. Quien desee interiorizarse en este punto puede consultar la (Ref. 4).

Diseños Taguchi

Genichi Taguchi (Japón, 1924 - 2012), realizó importantes avances en el Control Estadístico de la Calidad. Los elementos clave de la filosofía de Taguchi son la Función Pérdida de Calidad, la incorporación de los Arreglos Ortogonales al Diseño de Experimentos, el índice Señal/Ruido y el índice de Capacidad de Procesos Cpm. La complejidad de algunos de estos temas está más allá del nivel didáctico de este libro, por lo tanto, nos referiremos aquí exclusivamente al aspecto del Diseño de Experimentos y especialmente a la incorporación de los arreglos ortogonales. Taguchi innovó y simplificó el Diseño

de Experimentos con la introducción de las tablas conocidas como arreglos ortogonales (A.O.), que son una modificación de las matrices de Hadamard (matrices ortogonales nxn) (Ref. 5).

Hay varios ejemplos en la literatura, de diseños que no se podrían haberse llevado a cabo en ciertas aplicaciones industriales con los diseños clásicos, ya que el número de ensayos los hubiera hecho experimentalmente impracticables. Estos fueron factibles gracias a los diseños de Taguchi.

Los diseños Taguchi se caracterizan por ser muy condensados y ortogonales. Son hoy en día, los diseños más reducidos en experimentos. Tienen además la particularidad de poder trabajar con 2 y más niveles y sobre todo, poder introducir factores con diferentes números de niveles en un mismo diseño.

Características generales de los diseños Taguchi

- Sirven para definir cuales factores principales son significativos, o no.
- Estos diseños son los más económicos en cuanto al número de experiencias.
- A diferencia de Placket-Burman, donde todos los factores tienen 2 niveles, aquí pueden tener varios.
- Pueden combinarse algunos con un cierto número de niveles y otros con diferente nivel.
- ¡El diseño no es válido si existen interacciones entre los factores!

Propiedades de los diseños Taguchi

- En las columnas de cada efecto, cada nivel aparece el mismo número de veces, lo que se llama ***balance ortogonal de los vectores***.
- Todos los niveles de los factores independientes son utilizados para llevar a cabo los ensayos.
- Los vectores (factores) columnas son mutuamente ortogonales. Lo que implica que no se puede cambiar el orden de los niveles en las columnas.

Características del cálculo del diseño

El número mínimo de ensayos (corridas) de los diseños se calcula sobre la base de los grados de libertad (GD). Este número mínimo debe ser mayor que los grados de libertad.

Como ejemplo, para un diseño de 3 factores principales con 2 niveles, los GD se calculan multiplicando el número de niveles menos 1, por el número de factores y sumando 1 al total. En este caso GD = (2-1).3+1=4 y se lo designa como **L4 (2^3).**

Pero además el número de ensayos debe ser el mínimo común múltiplo (MCM) del número de niveles que participan del diseño, si no resulta así se deben sumarse corridas para que se cumpla esta condición. Esto se debe cumplir para mantener el *balance ortogonal de los vectores*.

Por ejemplo, para un diseño de 3 variables con dos niveles más cuatro variables con cuatro niveles sería: GD=(2-1).3+(4-1).4+1=16 esto cumple con la condición previa. y se lo designa como **L16 (2^3 4^4).**

Otro ejemplo; un diseño con una variable a 2 niveles más 7 variables a 3 niveles tendría:

GD=(2-1).1+(3-1).7+1=16, la condición MCM>GD no se cumple (MCM=2.3 <16). Entonces, para 2 y 3 el MCM tiene que ser 18 y el diseño sería **L18 (2^1 3^7).**

Taguchi ha elaborado una serie de tablas de diseño que se pueden consultar para evitar el cálculo. Algunas de ellas se dan en el anexo del capítulo.

Conducción de las corridas del diseño

Aunque la matriz de diseño usualmente toma en cuenta sólo las columnas de los factores principales, se pueden agregar a la derecha de la tabla otras columnas para las interacciones (con el método de multiplicar las columnas correspondientes a las variables en cuestión y dependiendo de cómo lo permitan los niveles límite de las columnas) (Ref. 6).

Análisis de los factores

El método más directo para analizar la significación estadística de los resultados es aplicar un análisis ANOVA. También se puede recurrir al siguiente procedimiento para analizar uno cualquiera de los factores:

- Obtenga la media de los valores de las respuestas correspondientes a cada nivel, o sea obtenga la media del nivel 1, 2, etc.

- Obtenga la suma de cuadrados de la desviación de cada media del factor respecto de la *gran media* (media general).

- Si la suma de cuadrados relativa es próxima a cero o insignificante, se puede suponer que ese factor no es estadísticamente significativo.

En el capítulo 11 se verá un método más avanzado para el análisis de modelos multirespuesta sobre la base de los diseños Taguchi.

Diseños de 2 niveles

L4 (2^3)

Run	Columns		
	1	2	3
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1

L8 (2^7)

Run	Columns						
	1	2	3	4	6	7	8
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

L16 (2^15)

Run	Columns														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
3	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2
4	1	1	1	2	2	2	2	2	2	2	2	1	1	1	1
5	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2
6	1	2	2	1	1	2	2	2	2	1	1	2	2	1	1
7	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1
8	1	2	2	2	2	1	1	2	2	1	1	1	1	2	2
9	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
10	2	1	2	1	2	1	2	2	1	2	1	2	1	2	1
11	2	1	2	2	1	2	1	1	2	1	2	2	1	2	1
12	2	1	2	2	1	2	1	2	1	2	1	1	2	1	2
13	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1
14	2	2	1	1	2	2	1	2	1	1	2	2	1	1	2
15	2	2	1	2	1	1	2	1	2	2	1	2	1	1	2
16	2	2	1	2	1	1	2	2	1	1	2	1	2	2	1

L12 (2^11)

Run	Columns										
	1	2	3	4	5	6	7	8	9	10	11
1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	2	2	2	2	2	2
3	1	1	2	2	2	1	1	1	2	2	2
4	1	2	1	2	2	1	2	2	1	1	2
5	1	2	2	1	2	2	1	2	1	2	1
6	1	2	2	2	1	2	2	1	2	1	1
7	2	1	2	2	1	1	2	2	1	2	1
8	2	1	2	1	2	2	2	1	1	1	2
9	2	1	1	2	2	2	1	2	2	1	1
10	2	2	2	1	1	1	1	2	2	1	2
11	2	2	1	2	1	2	1	1	1	2	2
12	2	2	1	1	2	1	2	1	2	2	1

Diseños de 3 niveles

L9 (3^4)

Run	Columns			
	1	2	3	4
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

L27 (3^13)

Run	Columns												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	2	2	2	2	2	2	2	2	2
3	1	1	1	1	3	3	3	3	3	3	3	3	3
4	1	2	2	2	1	1	1	2	2	2	3	3	3
5	1	2	2	2	2	2	2	3	3	3	1	1	1
6	1	2	2	2	3	3	3	1	1	1	2	2	2
7	1	3	3	3	1	1	1	3	3	3	2	2	2
8	1	3	3	3	2	2	2	1	1	1	3	3	3
9	1	3	3	3	3	3	3	2	2	2	1	1	1
10	2	1	2	3	1	2	3	1	2	3	1	2	3
11	2	1	2	3	2	3	1	2	3	1	2	3	1
12	2	1	2	3	3	1	2	3	1	2	3	1	2
13	2	2	3	1	1	2	3	2	3	1	3	1	2
14	2	2	3	1	2	3	1	3	1	2	1	2	3
15	2	2	3	1	3	1	2	1	2	3	2	3	1
16	2	3	1	2	1	2	3	3	1	2	2	3	1
17	2	3	1	2	2	3	1	1	2	3	3	1	2
18	2	3	1	2	3	1	2	2	3	1	1	2	3
19	3	1	3	2	1	3	2	1	3	2	1	3	2
20	3	1	3	2	2	1	3	2	1	3	2	1	3
21	3	1	3	2	3	2	1	3	2	1	3	2	1
22	3	2	1	3	1	3	2	2	1	3	3	2	1
23	3	2	1	3	2	1	3	3	2	1	1	3	2
24	3	2	1	3	3	2	1	1	3	2	2	1	3
25	3	3	2	1	1	3	2	3	2	1	2	1	3
26	3	3	2	1	2	1	3	1	3	2	3	2	1
27	3	3	2	1	3	2	1	2	1	3	1	3	2

Diseños con factores de diferentes números de niveles

L8 (2^4 4^1)

Run	Columns				
	1	2	3	4	5
1	1	1	1	1	1
2	2	2	2	2	1
3	1	1	2	2	2
4	2	2	1	1	2
5	1	2	1	2	3
6	2	1	2	1	3
7	1	2	2	1	4
8	2	1	1	2	4

L16 (2^12 4^1)

Run	Columns												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	2	2	2	2	2	2	2	2	1
3	2	2	2	2	1	1	1	1	2	2	2	2	1
4	2	2	2	2	2	2	2	2	1	1	1	1	1
5	1	1	2	2	1	1	2	2	1	1	2	2	2
6	1	1	2	2	2	2	1	1	2	2	1	1	2
7	2	2	1	1	1	1	2	2	2	2	1	1	2
8	2	2	1	1	2	2	1	1	1	1	2	2	2
9	1	2	1	2	1	2	1	2	1	2	1	2	3
10	1	2	1	2	2	1	2	1	2	1	2	1	3
11	2	1	2	1	1	2	1	2	2	1	2	1	3
12	2	1	2	1	2	1	2	1	1	2	1	2	3
13	1	2	2	1	1	2	2	1	1	2	2	1	4
14	1	2	2	1	2	1	1	2	2	1	1	2	4
15	2	1	1	2	1	2	2	1	2	1	1	2	4
16	2	1	1	2	2	1	1	2	1	2	2	1	4

L16 (2^9 4^2)

Run	Columns										
	1	2	3	4	5	6	7	8	9	10	11
1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2	2	2	1	2
3	2	2	2	1	1	1	2	2	2	1	3
4	2	2	2	2	2	2	1	1	1	1	4
5	1	2	2	1	2	2	1	2	2	2	1
6	1	2	2	2	1	1	2	1	1	2	2
7	2	1	1	1	2	2	2	1	1	2	3
8	2	1	1	2	1	1	1	2	2	2	4
9	2	1	2	2	1	2	2	1	2	3	1
10	2	1	2	1	2	1	1	2	1	3	2
11	1	2	1	2	1	2	1	2	1	3	3
12	1	2	1	1	2	1	2	1	2	3	4
13	2	2	1	2	2	1	2	2	1	4	1
14	2	2	1	1	1	2	1	1	2	4	2
15	1	1	2	2	2	1	1	1	2	4	3
16	1	1	2	1	1	2	2	2	1	4	4

L16 (2^6 3^4)

Run	Columns								
	1	2	3	4	5	6	7	8	9
1	1	1	1	1	1	1	1	1	1
2	1	1	2	2	2	2	1	2	2
3	2	2	1	1	2	2	1	3	3
4	2	2	2	2	1	1	1	4	4
5	2	2	1	2	1	2	2	1	2
6	2	2	2	1	2	1	2	2	1
7	1	1	1	2	2	1	2	3	4
8	1	1	2	1	1	2	2	4	3
9	1	2	2	2	2	1	3	1	3
10	1	2	1	1	1	2	3	2	4
11	2	1	2	2	1	2	3	3	1
12	2	1	1	1	2	1	3	4	2
13	2	1	2	1	2	2	4	1	4
14	2	1	1	2	1	1	4	2	3
15	1	2	2	1	1	1	4	3	2
16	1	2	1	2	2	2	4	4	1

L16 (2^3 4^4)

Run	Columns						
	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	2	2	1	2	2	2
3	2	1	2	1	3	3	3
4	2	2	1	1	4	4	4
5	2	2	1	2	1	2	1
6	2	1	2	2	2	1	2
7	1	2	2	2	3	4	4
8	1	1	1	2	4	3	3
9	1	2	2	3	1	3	4
10	1	1	1	3	2	4	3
11	2	2	1	3	3	1	2
12	2	1	2	3	4	2	1
13	2	1	2	4	1	4	2
14	2	2	1	4	2	3	1
15	1	1	1	4	3	2	4
16	1	2	2	4	4	1	3

L18 (2^1 3^7)

Run	Columns							
	1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1	1
2	1	1	2	2	2	2	2	2
3	1	1	3	3	3	3	3	3
4	1	2	11	1	2	2	3	3
5	1	2	2	2	3	3	1	1
6	1	2	3	3	1	1	2	2
7	1	3	1	2	1	3	2	3
8	1	3	2	3	2	1	3	1
9	1	3	3	1	3	2	1	2
10	2	1	1	3	3	2	2	1
11	2	1	2	1	1	3	3	2
12	2	1	3	2	2	1	1	3
13	2	2	1	2	3	1	3	2
14	2	2	2	3	1	2	1	3
15	2	2	3	1	2	3	2	1
16	2	3	1	3	2	3	1	2
17	2	3	2	1	3	1	2	3
18	2	3	3	2	1	2	3	1

Referencias

1. Massart D.L, Handbook Of Chemometrics And Qualimetrics part A. Chapter 23. 1997

2. R.L. Plackett, J.P. Burman, The design of optimal multifactorial experiments, Biometrika 33 (4) (1946) 305–325.

3. Mariela Soledad Espinosa, Laura Folguera, Jorge Federico Magallanes, Paola Alejandra Babay. Exploring analyte response in an ESI-MS system with different chemometric tools. Chemometrics and Intelligent Laboratory Systems 146 (2015) 120–127.

4. Jorge F. Magallanes, Alejandro C. Olivieri. The effect of factor interactions in Plackett–Burman experimental designs. Comparison of Bayesian-Gibbs analysis and genetic algorithms.

5. Alicia B. Hernandez, María de la Paz Guillón, Liliana A. García. Investigación Operativa. Año XXIII-N° 37. 65-83. Mayo 2015.

6. Kaushik & Singhal, Cogent Engineering (2018), 5: 1467196. https://doi.org/10.1080/23311916.2018.1467196

CAPITULO 10

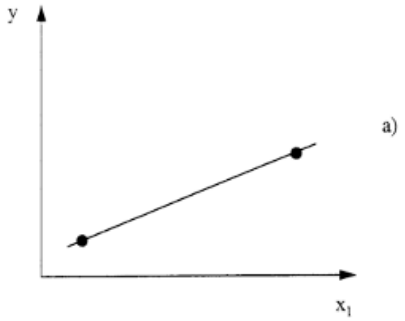
Diseños Multinivel

Introducción

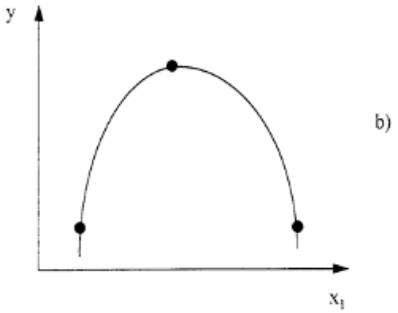
Con 2 niveles para cada factor, como hemos trabajado hasta ahora, solo se podrán describir rectas, planos o hiperplanos en el espacio multivariable, según el número de factores involucrados. O, dicho de otra forma, con 2 niveles podemos plantear modelos de primer orden solamente.

Las relaciones entre cada variable y la respuesta podrían ser mayores que el orden 1, algunas como aquellas donde interviene el pH podrian ser de tipo sigmoideo. Y aún podríamos estar ante la presencia de relaciones estadísticamente no lineales, que requieren de técnicas especiales.

En muchos casos habrá relaciones curvilíneas entre las variables y la/las respuestas.



Cálculo con 2 niveles



Cálculo con más de 2 niveles

Modelos cuadráticos

Repasemos las ecuaciones que describen modelos de segundo orden para 2 y 3 variables:

$$\eta = \beta_0 + \beta_1.x_1 + \beta_2.x_2 + \beta_{12}.x_1.x_2 + \beta_{11}.x_1^2 + \beta_{22}.x_2^2 \quad .$$

El diseño conduce a la estimación de los coeficientes β_i a través de la ecuación:

$$\hat{y} = b_0 + b_1.x_1 + b_2.x_2 + b_{12}.x_1.x_2 + b_{11}.x_1^2 + b_{22}.x_2^2. \qquad [1]$$

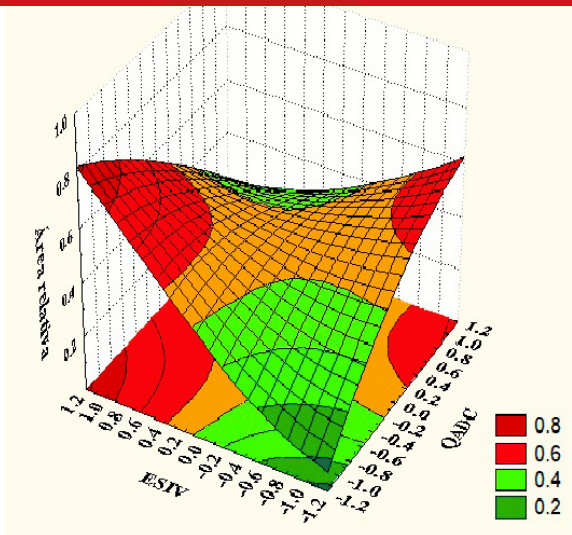
Para 3 variables:

$$\hat{y} = b_0 + b_1.x_1 + b_2.x_2 + b_3.x_3 + b_{12}.x_1.x_2 + b_{13}.x_1.x_3 + b_{23}.x_2.x_3 + b_{11}.x_1^2 + b_{22}.x_2^2 + b_{33}.x_3^2 + + b_{123}.x_1.x_2.x_3. \qquad [2]$$

Como vimos en la introducción del capítulo 8, la precisión de los coeficientes bi para estimar los βi depende del diseño experimental. Con estos coeficientes uno puede estimar **ŷ** como función de las variables **x** y construir "superficies de respuesta" o "gráficos de contorno".

La figura siguiente muestra una superficie de respuesta típica para 3 variables y orden 2; en su parte inferior su respectiva proyección, que constituye el diagrama de contorno.

Superficie de Respuesta y Contornos



Criterios de Calidad del Diseño

Vimos en el ejemplo introductorio del capítulo 8 que el cálculo de los coeficientes **b** en las ecuaciones [1] y [2] es:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}.\mathbf{X}^T.\mathbf{y}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \qquad b = \begin{bmatrix} b_0 \\ b_1 \\ . \\ . \\ . \\ b_m \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{11}^2 & x_{12}^2 & x_{11}x_{12} \\ 1 & x_{21} & x_{22} & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & x_{n1}^2 & x_{n2}^2 & x_{n1}x_{n2} \end{bmatrix}$$

X se deriva de la "matriz de diseño" **D** que **contiene las combinaciones de niveles a los cuales se deberían llevar a cabo los experimentos para los factores en cuestión**.

También habíamos visto que la varianza de los parámetros de la regresión (es decir la calidad del modelo) viene dada por:

$$\mathbf{V(b)} = s_e^2(\mathbf{X}^T\mathbf{X})^{-1} \qquad s_e^2 = \Sigma e_i^2/(n-p) \qquad [3]$$

La varianza está determinada por $\mathbf{X}^T\mathbf{X}$ y **X** depende de los valores elegidos $x_1, x_2, ..., x_n$.

La situación ideal es aquella en que la matriz **X** es ortogonal, porque en este caso no habrá correlación entre los factores y, además, la estimación del vector **b** será la mejor. Cuando esto ocurre la matriz $(\mathbf{X}^T\mathbf{X})^{-1}$ es diagonal.

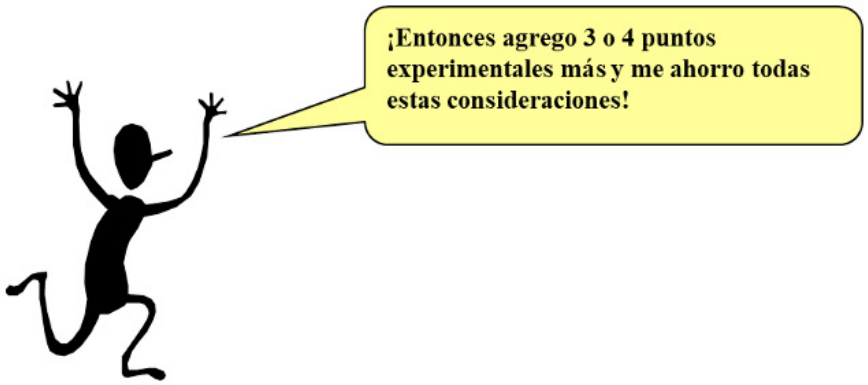
Nótese que $(\mathbf{X}^T\mathbf{X})^{-1}$ no incluye información acerca de la respuesta.

O sea, toda la información requerida para evaluar el efecto del diseño sobre la calidad del modelo estimado, **está presente antes que cualquier experimento sea llevado a cabo.**

Muchas veces la situación anterior no es posible y entonces se tratará de aproximarse a esta situación tanto como sea posible.

Puede demostrarse que la estimación de **b** es la mejor cuando el determinante de $(\mathbf{X}^T\mathbf{X})^{-1}$ es mínimo, y como los determinantes determinan geométricamente volúmenes en el espacio multidimensional, esto significa que el volumen de la unión de los intervalos de confianza de los b_i es mínimo. Como $\det((\mathbf{X}^T\mathbf{X})^{-1}) = 1/\det(\mathbf{X}^T\mathbf{X})$, el $\det(\mathbf{X}^T\mathbf{X})$ debe ser máximo.

En general, como el $\det(\mathbf{X}^T\mathbf{X})$ se incrementa con el número de puntos, se concluye que cuanto mayor sea número de puntos, la estimación será mejor.



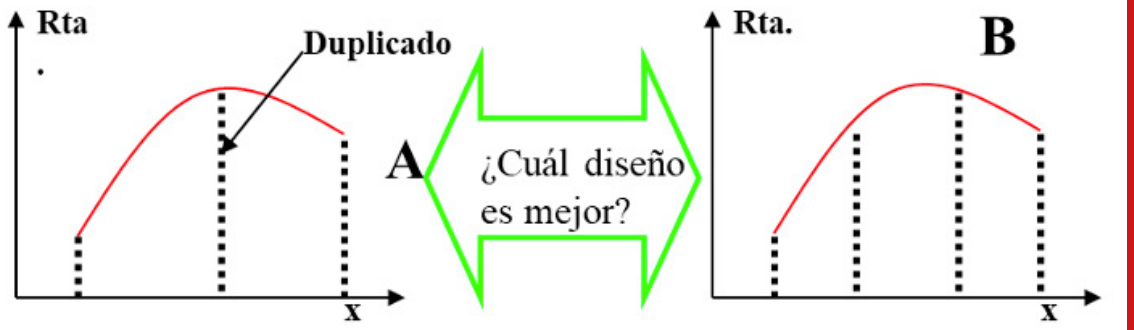
¿Hasta qué punto será cierta o no la afirmación que hace este aspirante a quimiometrista?

Más adelante lo averiguaremos.

Para comparar la calidad de diseños con igual número de puntos uno puede comparar det($\mathbf{X^TX}$) para saber cuál será mejor. Un diseño experimental se llama **D-óptimo cuando comparando su det($\mathbf{X^TX}$) con otro de igual número de puntos, resulta ser el mayor.**

Los diseños factoriales estudiados de 2 niveles (totales y fraccionarios) son

D-óptimos para modelos de primer orden y además son ortogonales.



Para describir una parábola, estos dos diseños con igual número de puntos pueden calcular **un mismo modelo con distinta precisión**

Cuando se necesita comparar diseños **con diferentes números de puntos experimentales**, puede usarse el criterio M-óptimo.

$$\textbf{M-óptimo}=\textbf{Máx}[\det(\mathbf{X^TX})/n] \qquad n=\text{N° de puntos de cada diseño} \qquad [4]$$

Otro criterio es el **A-óptimo**; este compara diseños con igual cantidad de puntos y toma en cuenta la traza de la matriz $(\mathbf{X^TX})^{-1}$.

$$\textbf{A-óptimo} = \textbf{Mín}[\text{tr}(\mathbf{X^TX})^{-1}] \qquad [5]$$

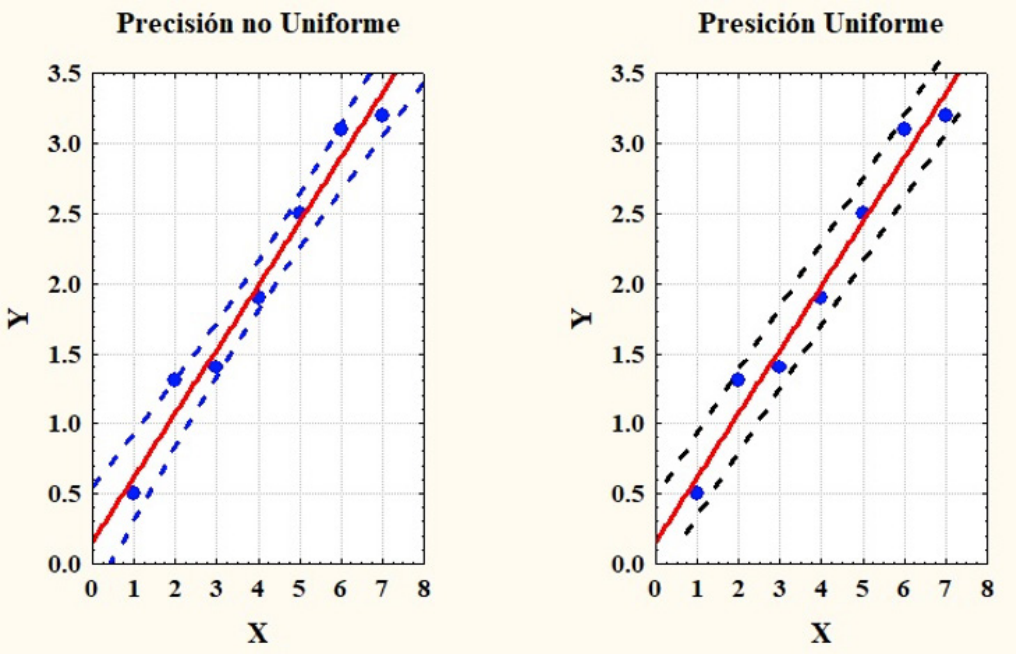
Se fundamenta en la consideración de la ec. [3] y el hecho de que S_e^2 está relacionado con el error experimental y no con el diseño.

Definiciones y Criterios Adicionales

Se dice que un diseño es "rotable" cuando observando los puntos experimentales desde el centro geométrico del diseño, la varianza es constante en cualquier dirección. **Nuevamente, los diseños FFD y Fraccional FD de 2 niveles son "rotables" y ortogonales.**

Una condición necesaria para que los diseños sean "rotables" es que los experimentos estén rigurosamente ubicados sobre una esfera o (hiper) esfera. Sin embargo, no todos los diseños esféricos son "rotables".

Por un adecuado ordenamiento del número de puntos centrales, también es posible obtener un diseño para el cual la precisión de las respuestas predictivas sea similar sobre el dominio experimental completo. Tales diseños se llaman "de precisión uniforme".



La "función varianza", $d(x_i)$, es una medida de la incertidumbre de las respuestas predictivas.

$$Var(\hat{y}_i)=\mathbf{x_i\,(X^TX)^{-1}.x_i^T}.\ s_e^2=\mathbf{d(x_i).}\ s_e^2,$$

$d(x_i)= x_i\,(X^TX)^{-1}.x_i^T$ es la **función varianza** en el punto (o ensayo) i.

Si bien se pretende que $Var(\hat{y}_i)\rightarrow$**mín,** usualmente se considera aceptable $d(x_i)=1$, o sea, $Var(\ddot{y}_i)= s_e^2$ lo que significa que la incertidumbre de la predicción es igual a la incertidumbre experimental.

Los diseños para un mismo número de puntos experimentales son considerados

G-óptimos cuando el máximo de la función varianza **en toda la región experimental** es el menor de todos.

$$\mathbf{G\text{-}óptimo=mín[máx\ d(x_i)]_R \quad R=\ región\ experimental} \qquad \mathbf{[6]}$$

Para comparar diseños con diferente cantidad de puntos 'n' se puede computar el **G-eficiencia**.

$$\mathbf{G\text{-}eff=p/(d(x)_{máx}.n)}$$

Donde **p** es el número de coeficientes del modelo y **d(x)** es el máximo valor de **$d(x_i)$**.

Nótese que, **para evaluar la calidad de los diseños con los criterios establecidos, éstos tienen que ser lineales en el sentido de la regresión (polinomios)**. En los casos en los cuales estas restricciones no pueden ser fácilmente aplicadas se deberá seguir otro criterio, y para éstos diseños la distribución de puntos deberá ser igualmente espaciada. Tales diseños, como veremos por ejemplo el diseño Doehlert, se llaman uniformes (o de celda uniforme).

Diseños Simétricos Clásicos

Éstos diseños son altamente simétricos y la mayoría de ellos califica muy bien con los criterios descriptos más arriba. El dominio experimental que ellos describen puede ser (hiper) esférico o (hiper) cúbico. Se debe considerar el dominio experimental que se quiere describir exactamente y tomar buen cuidado de no extrapolar fuera de la región descripta cuando se hacen predicciones.

Puede ser útil hacer mediciones experimentales adicionales por varias razones. Se puede, por ejemplo, replicar el punto central en el diseño para tener una idea del error experimental. La replicación de puntos experimentales permite la **valida-**

ción del modelo. Y la medida de puntos adicionales, distintos de los del diseño experimental, dentro de su rango, permite la **validación de la performance de las predicciones.**

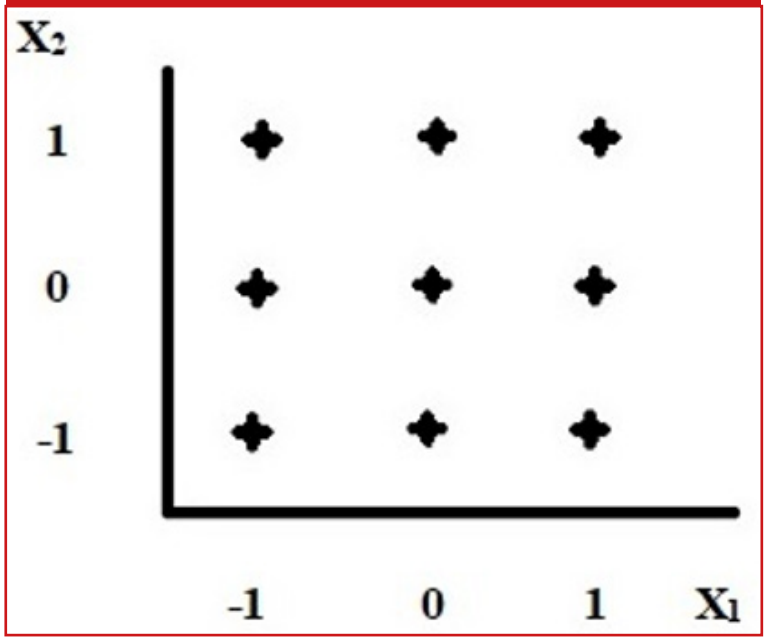
Diseño factorial de 3 niveles

El diseño factorial completo de 3 niveles (3^k) es conveniente para modelos cuadráticos, pero excepto para pequeños k, el diseño requiere muchos experimentos.

Plantilla de diseño factorial 3^2

Experimento	x1	x2
1	-1	-1
2	-1	0
3	-1	1
4	0	-1
5	0	0
6	0	1
7	1	-1
8	1	0
9	1	1

Distribución espacial en 2 dimensiones del diseño 3^2



Si se quieren hacer mediciones experimentales adicionales, puede calcularse que, para mejorar el diseño, un buen D-óptimo se obtiene replicando las cuatro esquinas (n=13).

El diseño factorial de 3 niveles es el único diseño multinivel que **es completamente ortogonal. No es "rotable",** lo que se puede comprobar a simple vista porque la distancia desde el centro a alguno de los ángulos no es igual a la del centro hacia la mitad de un lado (o simplemente, desde el centro, los puntos no se pueden ubicar sobre una circunferencia).

A fin de comprobar la ortogonalidad, analicemos las matrices para el diseño de la tabla anterior para el modelo de la ecuación [1].

$$\mathbf{X}= \begin{vmatrix} 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 0 & 0 & 1 & 0 \\ 1 & -1 & 1 & -1 & 1 & 1 \\ 1 & 0 & -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{vmatrix}$$

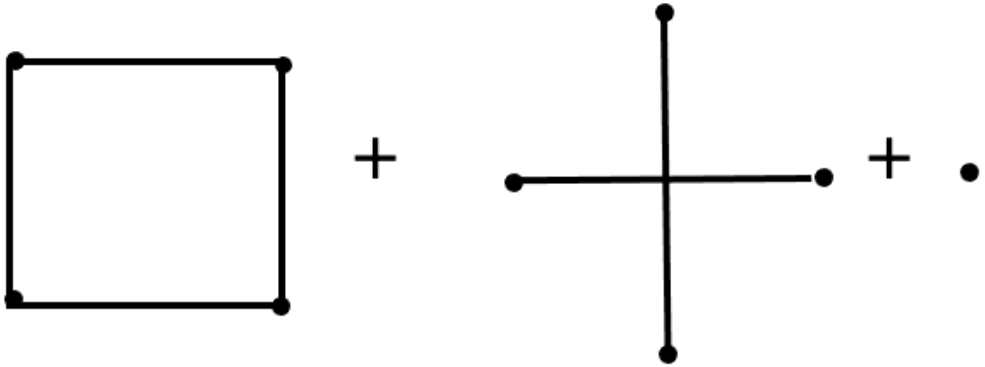
$$\mathbf{X^t.X} = \begin{vmatrix} 9 & 0 & 0 & 0 & 6 & 6 \\ 0 & 6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 \\ 6 & 0 & 0 & 0 & 6 & 4 \\ 6 & 0 & 0 & 0 & 4 & 6 \end{vmatrix}$$

Obsérvese que no hay ortogonalidad, en la primer columna de $\mathbf{X}^T\mathbf{X}$ (los dos seis), pero esta está ligada a β_0, otros factores distintos de cero fuera de la diagonal principal aparecen en la quinta y sexta columnas, relacionadas con covariancia de los términos cuadráticos, sin embargo por tener medias de columnas iguales, ésta se desvanece.

Como se ha anticipado, este diseño dista de ser eficiente cuando el número de variables crece levemente. Para apenas 4 variables el diseño requeriría ¡81 experimentos! Por lo tanto, debemos estudiar otros diseños que no son tan ideales como este, pero mucho más económicos.

Diseño Central Compuesto ("Central Composit")

Para mejorar la economía de experimentos se ha propuesto el diseño Central Compuesto. Veamos un ejemplo para 2 factores:

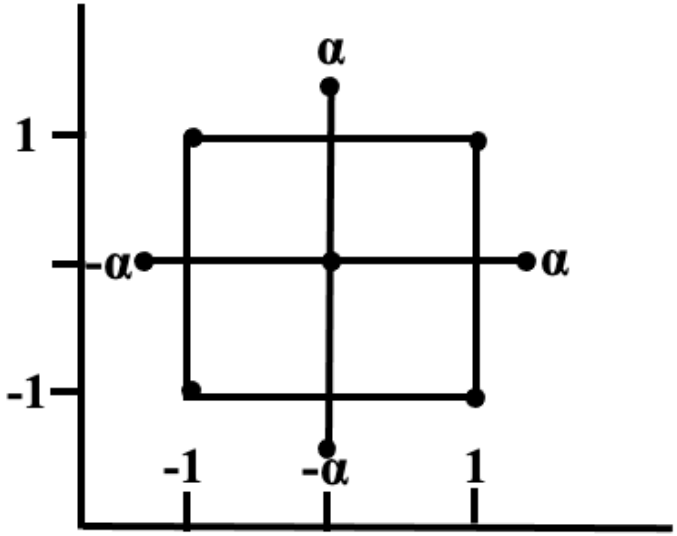


El diseño consiste de 3 partes:

-Un diseño factorial de 2 niveles. Número de córners, $n_c=2^k$ puntos con niveles –1 y +1.

-Un diseño 'estrella', $n_s=2.k$ puntos con niveles $-\alpha$ y $+\alpha$.

-Puntos centrales n_0. Con nivel 0 para todas las variables.

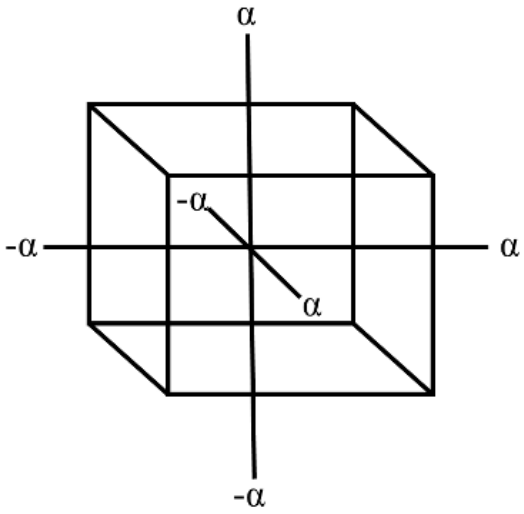


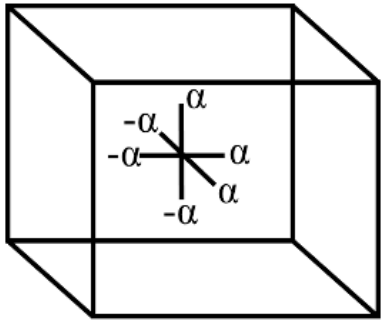
Como se ve, cada factor se encuentra en 5 niveles posibles $-\alpha$, -1, 0, +1, $+\alpha$, y el número total de experimentos, n, es mucho menor que el de un FFD.

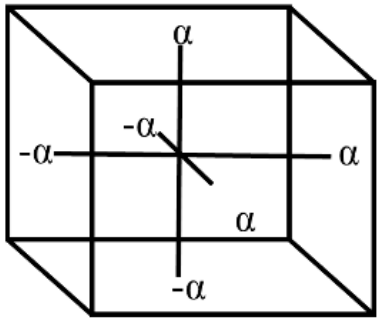
$n=2^k + 2.k +1$

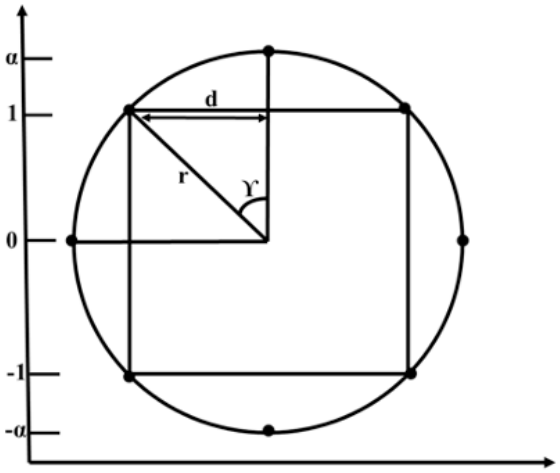
Los 3 tipos de Central Composit

- ***Central Composit Circuncribed (CCC)*** **:** **$|\alpha|>1$**
- ***Central Composed Inscribed (CCI)*** **:** **$|\alpha|<$Límites**
- ***Central Composit Face-Centered (CCF)*:** **$|\alpha|=1$**









Como se ve en la figura de la izquierda, para 2 factores, en el CCC $r = \alpha = 2^{1/2}$. Todos los puntos caen sobre un círculo y el diseño resulta "rotable". Más generalmente, puede comprobarse fácilmente que para más factores la rotabilidad se alcanza cuando $\alpha = n_c^{1/4}$.

El punto central es a menudo replicado por razones prácticas:

1- Con eso se puede tener una idea inmediata del error experimental.

2- Cuando el diseño se bloquea, comparando el punto central con cada bloque se puede saber si ocurrió un efecto de bloqueo, el cuál debe cuidarse.

3- Con un adecuado valor de n0 se puede alcanzar la cuasi-ortogonalidad y en este caso se alcanza también la *rotabilidad*.

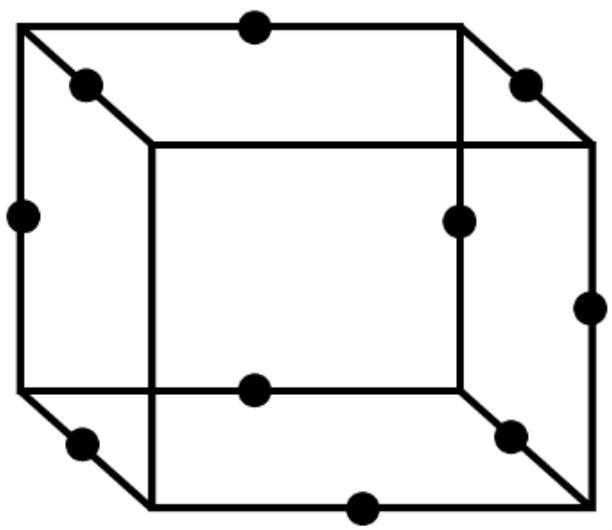
k	α para rotabilidad (en CCC)	n_0 para ortogonalidad y rotabilidad	n_0 para precision uniforme
2	1.40	8	5
3	1.68	9	6
4	2.00	12	7
5	2.38	17	10
5*	2.00	10	6
6	2.83	24	15
6*	2.38	15	9

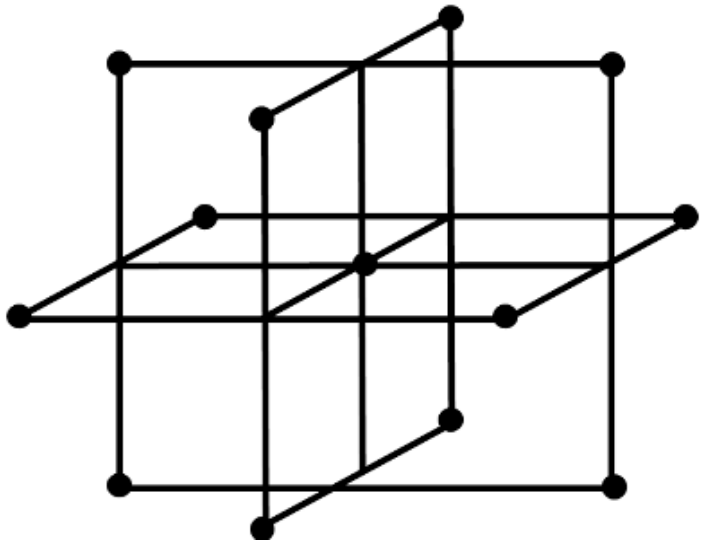
Tabla para seleccionar el número de puntos centrales, n_0. (Ref. 1)
k= número de factores
*: para diseño factorial fraccionario

El diseño Box-Behnken

En las figuras siguientes se muestra este diseño para 3 variables. Es un diseño 'rotable' que puede interpretarse geométricamente marcando los puntos medios de las aristas de un cubo más un punto central, o como el intercalado de 3 diseños factoriales de 2 niveles más un punto central. Debe tenerse cuidado con el volumen abarcado por el diseño porque éste es esférico y no comprende las esquinas del cubo. Es un diseño económico porque requiere solo 13 medidas para 3 factores. Cada factor tiene 3 niveles. Generalizando para k factores, el número de experiencias sería como mínimo, si no consideramos replicados en el punto central, 4 veces el número de pares de variables más un punto central, o sea:

$$N = 4.\left(\frac{k!}{(k-2)!2!}\right)+1 \qquad [7]$$





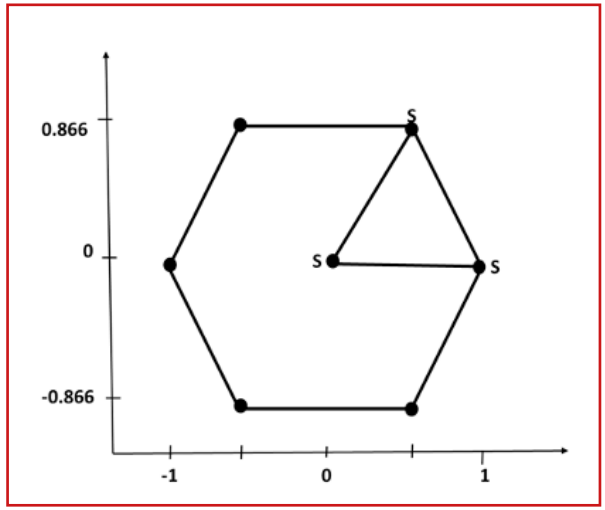
Dimensiones del diseño Box-Behnken

Diseños de celda uniforme (*Uniform Shell Design*)
Diseño Doehlert: Descripción general

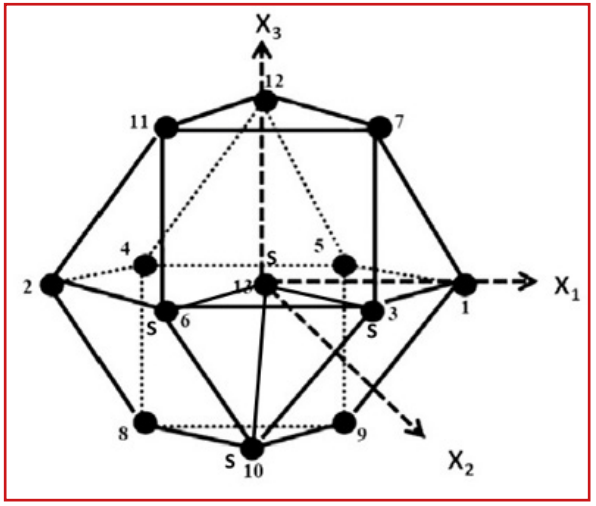
Éste es un diseño con características muy interesantes: Describe un dominio experimental esférico (o hiperesférico para más de 3 variables) y asegura la uniformidad en el llenado del espacio multivariable. Comparado con el *central composit*, para tres variables, el diseño Doehlert requiere menos experimentos y comparado con Box-Bhnken, el mismo número. Pero para un número superior de factores el diseño Doehlert es siempre más económico.

Para 2 variables el diseño requiere 7 puntos que se generan a partir de un SIMPLEX en el espacio considerado un triángulo equilátero en este caso, (puntos S en la figura). El resto de los puntos se obtiene restando todos los puntos entre sí (sin res-

tar el 0,0). Para k=3 se debe obtener el punto equidistante de los 3 puntos del triángulo original, pero considerando ahora 3 dimensiones, se obtiene un tetraedro (puntos S) y se prosigue con el método de las restas; en este caso obtendremos 13 puntos.



Puntos S	
X1	X2
0	0
1	0
0.5	0.866



Para *d* dimensiones se puede agregar un punto a los de dimensión (*d*-1) de acuerdo a la siguiente ecuación:

$$\frac{1}{2}, \frac{1}{2\sqrt{3}}, \frac{2}{\sqrt{6}}, \ldots, \frac{1}{\sqrt{\{2(d-1)(d-2)\}}}, \frac{1}{\sqrt{2d(d-1)}}, \frac{\sqrt{(d+1)}}{\sqrt{2d}} \qquad [8]$$

Cuando se quiere agregar una dimensión más, **las coordenadas** del primer punto de esta nueva dimensión se obtienen reemplazando la última coordenada del punto obtenido para la dimensión anterior por dos nuevas obtenidas **de los dos últimos términos de la ecuación [8]** (con lo que estaremos agregando una dimensión más). **El resto de las coordenadas quedan igual que antes para este punto**. A los puntos **de la dimensión anterior** se les agrega, al final, una columna de ceros en la nueva dimensión. **El resto de los puntos** para esta nueva dimensión se obtienen por el método de las restas explicado anteriormente.

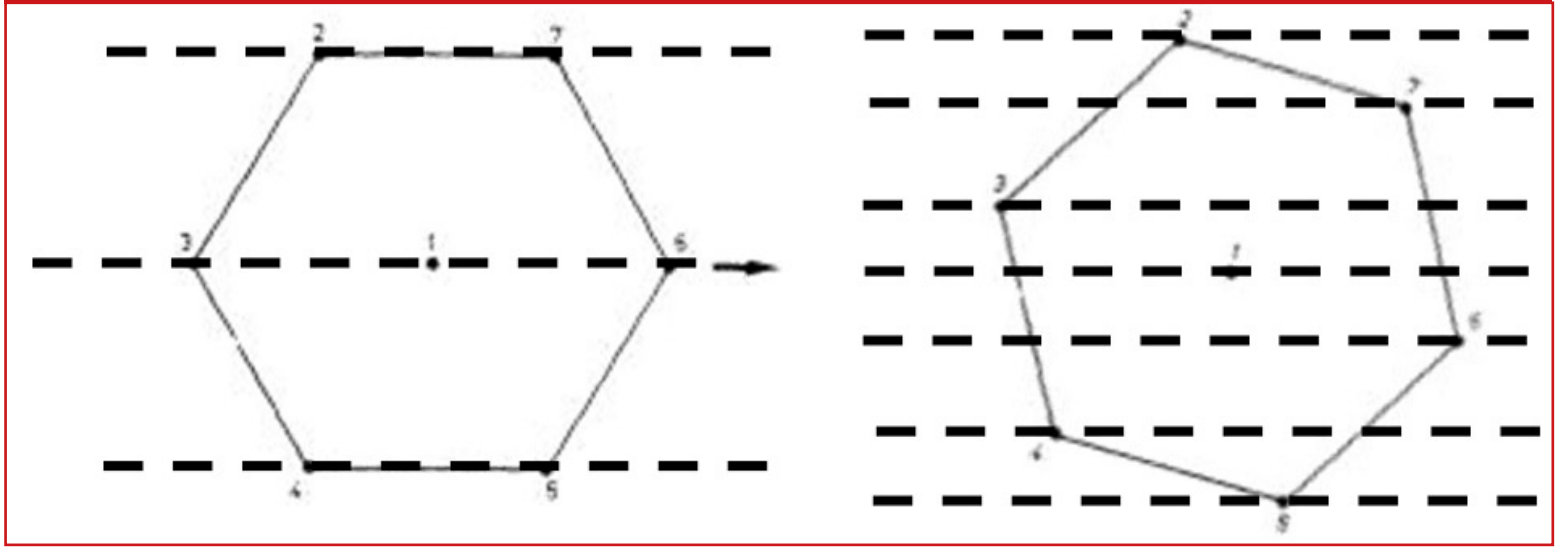
En la tabla siguiente se indica el modo de agregar las coordenadas del primer punto en la dimensión subsiguiente partiendo de dos dimensiones, y los valores correspondientes calculados con el algoritmo [8].

		*: nueva coord. 3D	#: nueva coord. 3D	□: nueva coord. 3D	
2 dimensiones	0.5	0.86602	0	0	0
3 dimensiones	0.5	0.28868 *	0.81650 *	0	0
4 dimensiones	0.5	0.28868	0.20413 #	0.79057 #	0
5 dimensiones	0.5	0.28868	0.20413	0.15812□	0.77460 □

El número total de puntos experimentales para k factores es $\mathbf{k^2 + k + 1}$. Los puntos son equidistantes del centro y están uniformemente distribuidos en el espacio. Para 3 dimensiones los puntos caen sobre la superficie de una esfera y para más dimensiones sobre una hiperesfera. A causa de su uniformidad y de cubrir una celda esférica se llama también a este diseño "de celda uniforme" como lo han propuesto sus autores.

Una de las características del diseño Doehlert es que los factores tienen diferentes números de niveles, para más de tres factores, el número de niveles es 3 para el primero de ellos, 5 para el siguiente y 7 para todos los restantes. De modo que uno es libre de asignar un mayor número de niveles a los factores que necesitan más descripción o asignar menor número de niveles a los factores dificultosas de cambiar experimentalmente. Si se necesitasen más de tres niveles en ese factor particular se puede rotar ligeramente el sistema y obtener más niveles hasta un máximo de 7.

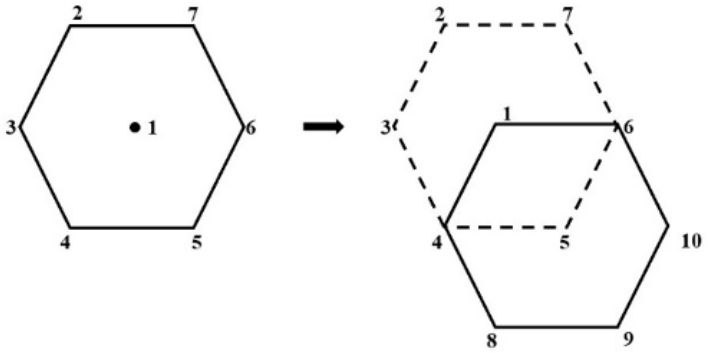
Cambio de 3 a 7 niveles por rotación



Extensión del área de muestreo por traslación

Este diseño también es muy eficiente en el mapeo del espacio, en este ejemplo para 2 dimensiones se pueden seguir agregando hexágonos para cubrir un área experimental mayor aprovechando las medidas previas (cosa que no ocurre en otros diseños y entonces se pierde la tanda de mediciones anteriores).

Observe en la figura siguiente que, para extender el área de muestreo en cualquiera de las direcciones de las variables, conservando el diseño original, sólo se necesitan 3 experimentos adicionales (8, 9 y 10).



Agregado de factores

También se pueden diseñar las experiencias dejando fuera a un factor dudoso y agregarlo en caso que haga falta con algunas medidas adicionales y sin desechar las anteriores.

Veamos cómo se hace esto. En las siguientes tablas vemos planillas de diseños Doehlert para 2 factores y para 3 factores. Ya sabemos que en el primer caso necesitamos 7 experimentos y en el segundo, 13. Observando la tabla para 3 factores vemos que para el tercer factor tenemos 6 niveles distintos de cero. Si cambiamos estos niveles fijándolos en cero, el diseño se reduce al de dos niveles (13 – 6 = 7 experimentos). Luego, si queremos comprobar el efecto del tercer factor, agregamos estos 6 puntos con niveles 0.817 y –0.817 y tendremos la planilla para 3 factores completa provechando los ensayos del diseño de 2 factores. Por supuesto el mismo tratamiento se puede extender a otro número de factores sucesivos.

Diseño Doehlert para 2 Factores		
Run	v1	v2
1	1	0
2	-1	0
3	0.5	0.866
4	-0.5	0.866
5	0.5	-0.866
6	-0.5	-0.866
7	0	0

Diseño Doehlert para 3 Factores			
Run	V1	V2	V3
1	0	0	0
2	1	0	0
3	0.5	0.866	0
4	0.5	0.289	0.817
5	-1	0	0
6	-0.5	-0.866	0
7	0.5	-0.287	-0.817
8	0.5	-0.866	0
9	0.5	-0.289	-0.817
10	0	0.577	-0.817
11	-0.5	0.866	0
12	-0.5	0.289	0.817
13	0	-0.577	0.817

Diseños Crosier

Este diseño es también de celda uniforme. Se parece en mucho al diseño Doehlert y tiene sobre él algunas ventajas y desventajas.

El número de corridas para k niveles es uno más que el de Doehlert, o sea k^2+k+2, pero esto es debido a que repite 2 veces el centro (todas las variables en nivel 0).

A diferencia de Doehlert, en este diseño **todos los factores tienen 7 niveles**, también ocurre que, para cualquier punto del diseño, todas las permutaciones cíclicas de los niveles de los factores también están presentes en el diseño. Cada factor tiene la misma serie de niveles. Otra particularidad es que, para cualquier punto del diseño, existe también su negativo. Una ventaja comparativa de este diseño es que para k par el diseño puede bloquearse en dos bloques ortogonales, siendo uno de ellos el negativo del otro. Para k impar no existen permutaciones cíclicas de grupos que incluyan a sus propios negativos. En cada bloque se repite el centro del diseño, por esta razón hay 2 corridas iguales con nivel cero.

El método de construcción del diseño es el siguiente:

Para cualquier k, el diseño se genera con todas las permutaciones cíclicas de los siguientes 4 puntos;

$$\mathbf{(-1,1,0,0\ldots,0),\ (-a,b,b,\ldots,b),\ (a,-b,-b,\ldots,-b)\ y\ (0,0,\ldots,0)}$$

$$\mathbf{Donde\ a=(k-1-(k+1)^{1/2})\ /k\quad y\quad b=(1+(k+1)^{1/2})/k}$$

Los puntos suspensivos indican la repetición del nivel según el número de factores considerados. Note que a+b=1.

Observe que hay 2 casos especiales: para k=3 a=0 y b=1, en consecuencia, el diseño tiene sólo 3 niveles; en el otro caso, para k=8, a=b=0.5 y en consecuencia hay sólo 5 niveles.

Simplicial Shell Designs

Este diseño ha sido desarrollado también por Crosier. Comparte muchas propiedades con el Uniform Shell Design (USD), por ejemplo, contiene el mismo número de puntos $N=k+k^2+n_0$ donde n_0 es el número de puntos centrales repetidos. A diferencia de USD, el diseño simplicial puede ser ortogonalmente bloqueado para cualquier k.

El diseño simplicial para k>4 tiene 9 niveles, excepto para k=7, que tiene rotaciones de 3, 7 y 9 niveles. Las secuencias k=11, 15, 19,…, (7+n.4), donde n es un entero mayor o igual a 1, tienen rotación de 3 y 9 niveles. Ejemplos de diseños Crosier:

Diseño Crosier USD de 4 factores				
Punto	A	B	C	D
1	0	0	1	-1
2	0	0	-1	1
3	0	1	0	-1
4	0	1	-1	0
5	0	-1	1	0
6	0	-1	0	1
7	1	0	0	-1
8	1	0	-1	0
9	1	-1	0	0
10	-1	0	1	0
11	-1	0	0	1
12	-1	1	0	0
13	0.8090	0.8090	0.8090	-0.1910
14	0.8090	0.8090	-0.1910	0.8090
15	0.8090	-0.1910	0.8090	0.8090
16	-0.1910	0.8090	0.8090	0.8090
17	-0.8090	-0.8090	-0.8090	0.1910
18	-0.8090	-0.8090	0.1910	-0.8090
19	-0.8090	0.1910	-0.8090	-0.8090
20	0.1910	-0.8090	-0.8090	-0.8090
21	0	0	0	0

Diseño Crosier Simplicial de 4 factores					
Punto	A	B	C	D	Bloque
1	-1.0000	-1.0000	0.3820	0.3820	1
2	0.3820	-1.0000	-1.0000	0.3820	1
3	0.3820	0.3820	-1.0000	-1.0000	1
4	-1.0000	0.3820	0.3820	-1.0000	1
5	-1.0000	0.3820	-1.0000	0.3820	1
6	0.3820	-1.0000	0.3820	-1.0000	1
7	-0.5729	0.8090	0.8090	0.8090	1
8	0.8090	-0.5729	0.8090	0.8090	1
9	0.8090	0.8090	-0.5729	0.8090	1
10	0.8090	0.8090	0.8090	-0.5729	1
11	0.0000	0.0000	0.0000	0.0000	1
12	1.0000	1.0000	-0.3820	-0.3820	2
13	-0.3820	1.0000	1.0000	-0.3820	2
14	-0.3820	-0.3820	1.0000	1.0000	2
15	1.0000	-0.3820	-0.3820	1.0000	2
16	1.0000	-0.3820	1.0000	-0.3820	2
17	-0.3820	1.0000	-0.3820	1.0000	2
18	0.5729	-0.8090	-0.8090	-0.8090	2
19	-0.8090	0.5729	-0.8090	-0.8090	2
20	-0.8090	-0.8090	0.5729	-0.8090	2
21	-0.8090	-0.8090	-0.8090	0.5729	2
22	0.0000	0.0000	0.0000	0.0000	2

	Diseño Crosier Simplicial de 6 factores					
Punto	A	B	C	D	E	F
1	-1	-1	0.2616	0.2616	0.2616	0.2616
2	0.2616	-1	-1	0.2616	0.2616	0.2616
3	0.2616	0.2616	-1	-1	0.2616	0.2616
4	0.2616	0.2616	0.2616	-1	-1	0.2616
5	0.2616	0.2616	0.2616	0.2616	-1	-1
6	-1	0.2616	0.2616	0.2616	0.2616	-1
7	-1	0.2616	-1	0.2616	0.2616	0.2616
8	0.2616	-1	0.2616	-1	0.2616	0.2616
9	0.2616	0.2616	-1	0.2616	-1	0.2616
10	0.2616	0.2616	0.2616	-1	0.2616	-1
11	-1	0.2616	0.2616	0.2616	-1	0.2616
12	0.2616	-1	0.2616	0.2616	0.2616	-1
13	-1	0.2616	0.2616	-1	0.2616	0.2616
14	0.2616	-1	0.2616	0.2616	-1	0.2616
15	0.2616	0.2616	-1	0.2616	0.2616	-1
16	-0.654	0.6076	0.6076	0.6076	0.6076	0.6076
17	0.6076	-0.654	0.6076	0.6076	0.6076	0.6076
18	0.6076	0.6076	-0.654	0.6076	0.6076	0.6076
19	0.6076	0.6076	0.6076	-0.654	0.6076	0.6076
20	0.6076	0.6076	0.6076	0.6076	-0.654	0.6076
21	0.6076	0.6076	0.6076	0.6076	0.6076	-0.654
22	0	0	0	0	0	0

Izquierda, bloque 1; derecha, bloque 2

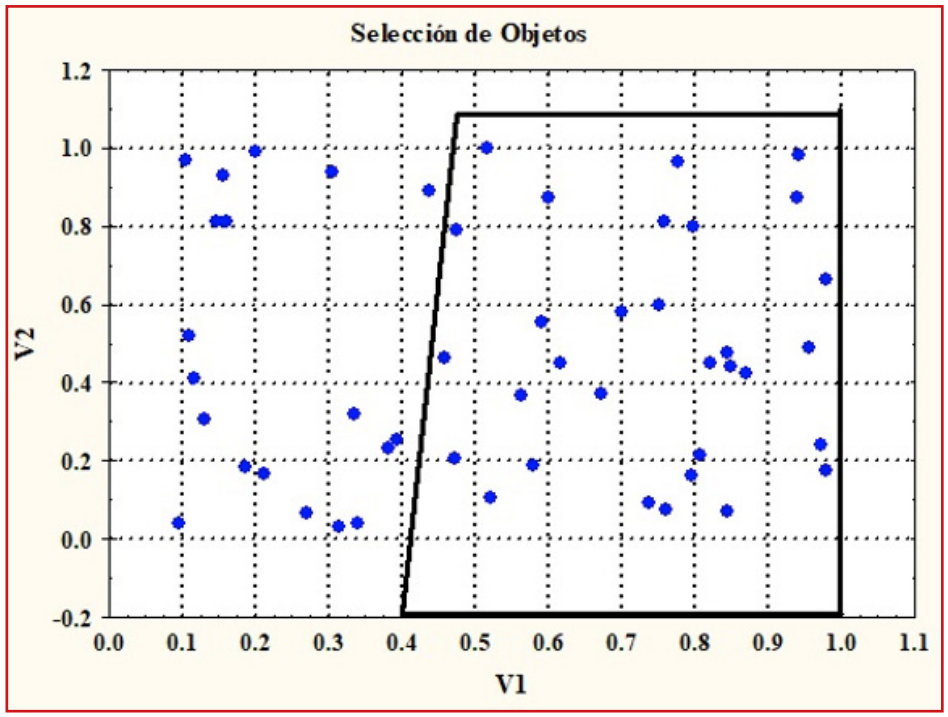
23	1	1	-0.2616	-0.2616	-0.2616	-0.2616
24	-0.2616	1	1	-0.2616	-0.2616	-0.2616
25	-0.2616	-0.2616	1	1	-0.2616	-0.2616
26	-0.2616	-0.2616	-0.2616	1	1	-0.2616
27	-0.2616	-0.2616	-0.2616	-0.2616	1	1
28	1	-0.2616	-0.2616	-0.2616	-0.2616	1
29	1	-0.2616	1	-0.2616	-0.2616	-0.2616
30	-0.2616	1	-0.2616	1	-0.2616	-0.2616
31	-0.2616	-0.2616	1	-0.2616	1	-0.2616
32	-0.2616	-0.2616	-0.2616	1	-0.2616	1
33	1	-0.2616	-0.2616	-0.2616	1	-0.2616
34	-0.2616	1	-0.2616	-0.2616	-0.2616	1
35	1	-0.2616	-0.2616	1	-0.2616	-0.2616
36	-0.2616	1	-0.2616	-0.2616	1	-0.2616
37	-0.2616	-0.2616	1	-0.2616	-0.2616	1
38	0.654	-0.6076	-0.6076	-0.6076	-0.6076	-0.6076
39	-0.6076	0.654	-0.6076	-0.6076	-0.6076	-0.6076
40	-0.6076	-0.6076	0.654	-0.6076	-0.6076	-0.6076
41	-0.6076	-0.6076	-0.6076	0.654	-0.6076	-0.6076
42	-0.6076	-0.6076	-0.6076	-0.6076	0.654	-0.6076
43	-0.6076	-0.6076	-0.6076	-0.6076	-0.6076	0.654
44	0	0	0	0	0	0

Diseños asimétricos

En los diseños simétricos los niveles de las variables pueden ser *determinados* por el experimentador. En los diseños asimétricos los niveles deben *seleccionarse* desde un lote de datos experimentales preexistentes

Puede haber casos en que no es posible llevar a cabo alguno de los diseños simétricos clásicos mostrados precedentemente. Una de las causas puede ser debido a que alguna de las combinaciones de niveles (generalmente en los extremos) no es experimentalmente posible. Otro caso se presenta cuando los datos ya vienen determinados por mediciones o parámetros preexistentes.

Por ejemplo, supongamos que se desea investigar cierta propiedad biológica a partir de propiedades fisicoquímicos de un determinado número de compuestos químicos prototipo. El valor de estas propiedades ya viene determinado y por lo tanto lo único que podemos hacer es elegir el grupo de compuestos con la mejor calidad de diseño D. Lotes de muestras de este tipo pueden ser muy grandes y se necesita hacer una selección de muestras para calcular un modelo. El problema es cómo elegir las muestras.



Supongamos un ejemplo de solo 2 variables, como el de la figura. El área experimental a muestrear está dentro del trapecio irregular que contiene 80 de los 100 datos totales. Si queremos tomar 10 muestras de entre las 80 para calcular su D no podemos utilizar el método de la fuerza bruta porque el número total, N, de grupos de 10 objetos sería más de 1 billón:

$$N = \frac{80!}{70!.10!} = 1\,646\,492\,110\,120$$

Por lo tanto, se deben usar estrategias para resolver este problema. Una de ellas puede ser trazar una red de líneas, como las cuadrículas que están punteadas en la figura y elegir las que están bien cerca de sus cruces. Pero como vemos, los objetos difieren mucho entre sí de su distancia a los nodos. Los siguientes métodos son mejores.

Algoritmos de mapeo uniforme

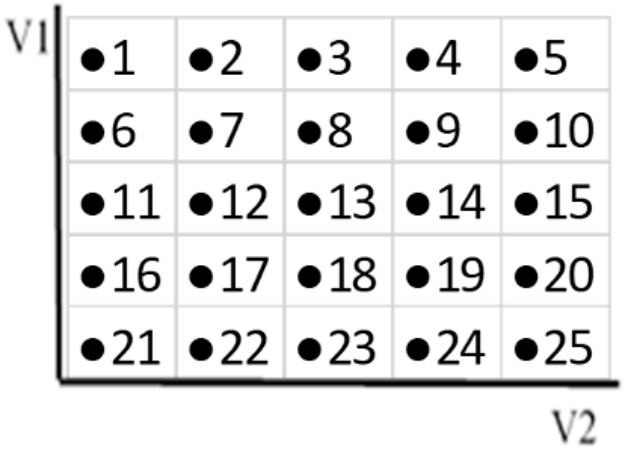
En casos donde los diseños clásicos no son aplicables o en aquellos donde el D *optimality* y criterios relacionados no son fácilmente computables (por ejemplo, en modelos no lineales), aún queda la estrategia de operar con algoritmos de espaciado uniforme. Ya vimos que el diseño Doehlert es del tipo de celda uniforme para abarcar un espacio esférico o hiperesférico. Pero para espacios asimétricos habrá que recurrir a otra estrategia.

Algoritmo de Kennard y Stone

Uno desearía cubrir todo el espacio de los factores tan uniformemente como fuera posible, asegurándose al mismo tiempo que los puntos experimentales estén tan lejos uno del otro, también como sea posible. Un algoritmo que conduce a este objetivo es el de **Kennard y Stone** (Ref. 2,3), que consiste en tomar el máximo de **la distancia mínima entre cada punto seleccionado y todos los otros**. Para ello se utiliza la distancia Euclideana, ya conocida, donde k representa a los factores e, i y j identifican a los puntos. En los casos donde se hayan llevado a cabo experiencias piloto uno puede incluir éstos puntos iniciales para el cálculo. Si el diseño se comienza desde cero se selecciona la distancia más grande entre todos los pares de puntos.

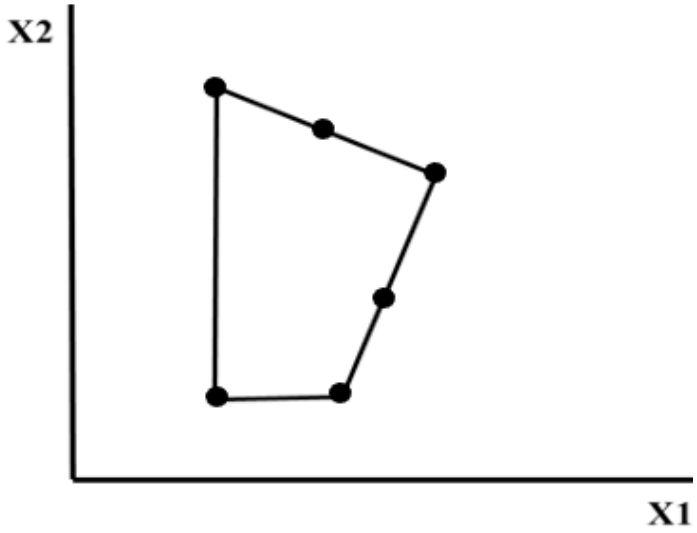
$$D_{ij} = \left[\sum_{l=1,k} (X_{i,l} - X_{j,l})^2 \right]^{1/2} \qquad \mathbf{d_{selec} = máx\ (D_{ij})}$$

Como comprobación, se muestra un ejemplo para un diseño simétrico, de modo de apreciar el resultado, pero ya se dijo que este método se aplica a diseños no simétricos.



La primera selección de los puntos, x, en la figura resultaría en la selección de los puntos 1 y 25 (ó 21 y 5). Ahora entran puntos adicionales computando para cada candidato (i_0) la distancia con los puntos previos (i) , $\mathbf{d_{selec} = i_0(máx(min(d_{i,i0}))}$.

O sea, uno mide todas las distancias entre los candidatos i_0 y los ya seleccionados i, y determina **a cuál de los i** está más cerca ($\min(d_{i,i0})$). Entre éstos, se elige aquel para cuál la distancia es máxima. Para la figura, esto resultaría en que los primeros 4 puntos serían 1, 25, 5, 21. Si se eligiera un quinto punto se recalcula de nuevo y se seleccionaría el 13. Una serie más larga sería 1, 25, 5, 21, 13, 3, 11, 15, 23, 19, etc. Obsérvese que para un desarrollo simétrico como el de la figura, los primeros 4 puntos forman un diseño factorial 2^2, los primeros 5 puntos forman un diseño factorial 2^2 centrado, si se toman los primeros 9 tenemos un 3^2 o *central composite* centrado en las caras.

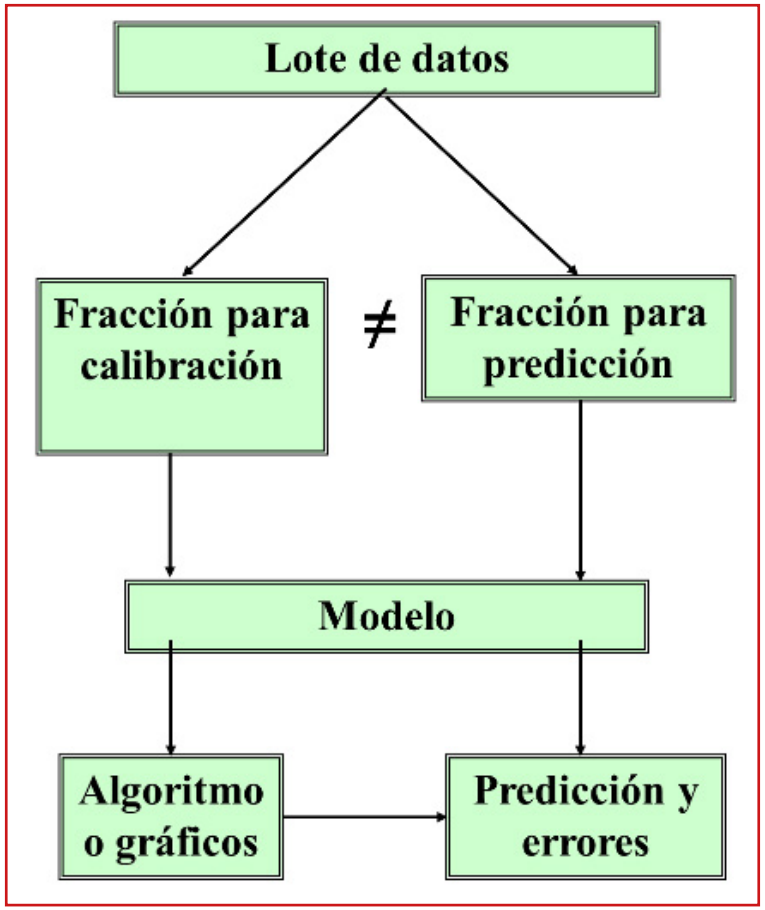


La figura de la izquierda muestra el resultado de unos pocos puntos para un diseño asimétrico y se ve que éstos tienden a ubicarse en los bordes del espacio de objetos

Algoritmo de mapeo de Centroide Progresivo

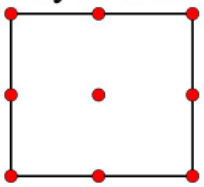
Este algoritmo no es de mapeo uniforme, sino que mapea con preferencia los bordes y el centro del espacio, aunque éste sea no simétrico e irregular (Ref. 4).

Repasemos en la figura siguiente, los pasos para la construcción de un modelo de un sistema en estudio.



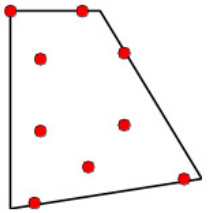
Criterios de selección de la fracción de calibración

Bordes y centro: criterio seguido por la mayoría de los diseños *simétricos*.



Ventajas: ortogonalidad, rotabilidad

Mapeo uniforme: criterio seguido por la mayoría de los diseños *asimétricos*.



Ventaja: Más eficiencia

Mecánica del cálculo del mapeo de centroide progresivo

El mecanismo comienza, como para Kennard-Stone, eligiendo los 2 puntos más alejados entre sí, o los 2 que elija el operador. Luego continúa así:

1- calcula el centroide de los puntos elegidos.

2- El punto más alejado del centroide es el nuevo punto elegido.

3- Se calcula el nuevo centroide.

4- Se repiten los pasos 2 y 3 hasta alcanzar un número de puntos igual al de un *full factory* de 2 niveles para el número de variables tratadas. Si este número es mayor que el número de objetos deseados se reduce directamente al número de variables tratadas.

5- Si el número de los objetos necesarios son mayores a los de un *full Factory*, Se elige el punto más cercano al centroide de todos los puntos.

6- Se continúa repitiendo los pasos 2 y 3 hasta completar los puntos deseados.

Este diseño es aplicable a espacios simétricos o no simétricos con mejores valores de D óptimo que los del método Kennard-Stone y además, para los espacios no simétricos, resulta en un descenso importante en los errores de predicción del modelo resultante.

Comparación de criterios

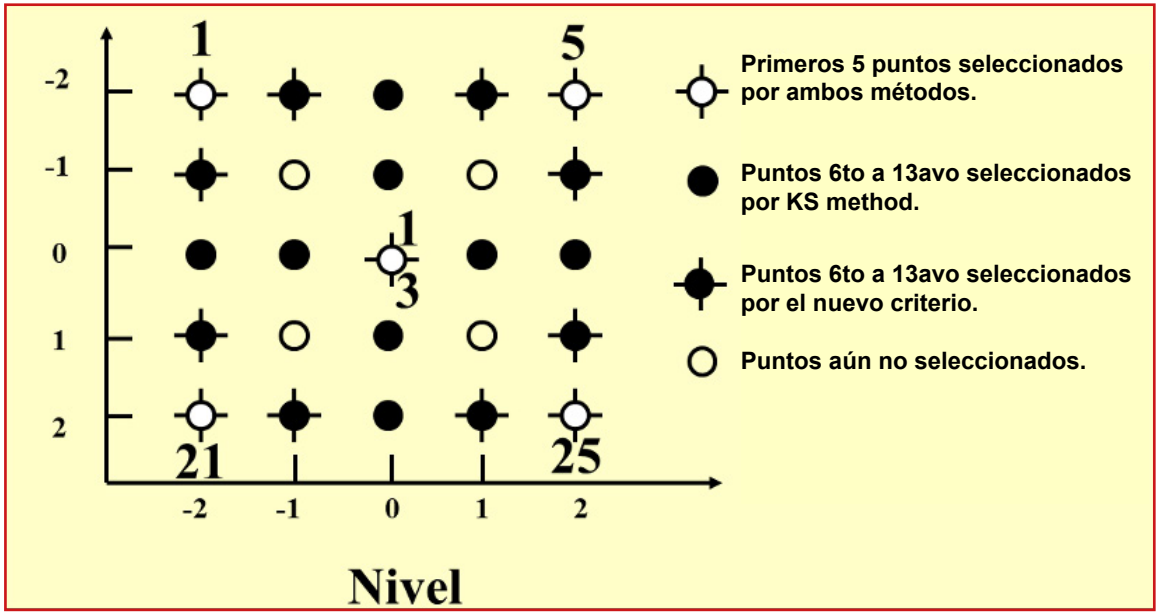
Kennard Stone:

Mapeo uniformemente espaciado.

Mapeo de centroide progresivo:

Bordes y centro:

La figura siguiente muestra las coincidencias y diferencias entre el método de Kennerd-Stone y el de centroide progresivo sobre el ejemplo de los 25 objetos.



Para facilitar la elección de un diseño particular se muestra una **tabla para seleccionar el número de ensayos en función del número de variables**.

	Comparación del Nº de experiencias N según el diseño experimental y el Nº de factores k					
			I = Nº de niveles			
			N			
k	Full Factorial (I=2)	Full Factorial (I=3)	Central Composit (I=5)	Box-Benken I=3	Doehlert (I=3, 5, 7)	Placket Burman (I=2)
2	4	9	9	5	7	4
3	8	27	15	13	13	4
4	16	81	25	25	21	8
5	32	243	43	41	31	8
6	64	729	77	61	43	8
7	128	2187	143	85	57	8
8	256	6561	273	113	73	12
9	512	19683	531	145	91	12
10	1024	59049	1045	181	111	12
11	2048	177147	2071	221	133	12
12	4096	531441	4121	265	157	16
13	8192	1594323	8219	313	183	16
14	16384	4782969	16413	365	211	16
15	32768	14348907	32799	421	241	16

Los diseños no deben ser elegidos por el único principio de “el menor número de experiencias”. Si bien este es un objetivo deseable debe tenerse siempre en cuenta las características del sistema que se va a estudiar y las cualidades de las variables presentes.

Referencias

1. E. Morgan, K.W. Burton and P. Church, Chemom. Intell. Lab. Syst., 5 (1989) 283-302.

2. R.W. Kennard and L.A. Stone, Computer-aided design of experiments. Technometrics, 11 (1969)137-148.

3. D.L. Massart; B.G.M. Vandeginste; L.M.C. Buydens; S. De Jong; P.J. Lewi and J. Smeyers-Verbeke. Hanbook of Chemometrics and Qualimetrics. Parts A Pag. 727. Elsevier, Amsterdam 1997.

4. Jorge F. Magallanes. A new uniform mapping algorithm for sample selection. J. Chemometrics, 23, (2009) 132-138.

5. Ronal B Crosier. Edgewood Reasearch Developent& Engeneering Center. AD- A270 656. ERDEC-TR-104. August 1993. https://apps.dtic.mil/sti/pdfs/ADA270656.pdf

CAPITULO 11

Diseño de Bloqueo y Optimización de Modelos Multivariables

Diseño de Bloqueo

Una introducción a los problemas de bloqueo se ha dado en el capítulo 8. Ahora avanzaremos hacia el diseño del bloqueo cuando existen más de 2 niveles. Por facilidad, reemplazaremos la denominación de los factores con números en lugar de letras mayúsculas.

Supongamos que diseñamos un experimento del tipo 2^3 y queremos ejecutarlo en **2 bloques**. Tenemos 7 grados de libertad (2^3-1) para calcular todos los parámetros del modelo, pero tendremos que sacrificar alguno para diseñar el bloqueo. Sacrificamos, por ejemplo, el usualmente menos importante, o sea la asociación 1-2-3. Entonces utilizamos uno de los niveles de la columna 1-2-3 para seleccionar las corridas de uno de los bloques, por ejemplo, el nivel (-) y el otro nivel para organizar las corridas del otro bloque (+).

Proyecto de Bloqueo en 2 etapas para un Diseño 2^3								
Run	1	2	3	12	13	23	123	Bloque
1	-	-	-	+	+	+	-	**1**
2	-	-	+	+	-	-	+	**2**
3	-	+	-	-	+	-	+	**2**
4	-	+	+	-	-	+	-	**1**
5	+	-	-	-	-	+	+	**2**
6	+	-	+	-	+	-	-	**1**
7	+	+	-	+	-	-	-	**1**
8	+	+	+	+	+	+	+	**2**

Si quisiéramos dividir el mismo diseño en **4 bloques** tenemos que proceder de modo similar. Para esquematizar el procedimiento general digamos que tenemos un diseño 2^k y que queremos dividirlo en un número par de bloques (obligatorio para diseños de 2 niveles), expresados como 2^q. En este ejemplo tendríamos un diseño 2^3 (k=3) con un bloqueo en cuatro bloques 2^2 (q=2) de **tamaño 2^{k-q} (2 ensayos para cada bloque)**. Ahora definimos q variables de bloqueo: B_1 y B_2; necesitamos entonces sacrificar 2 efectos de los 7 disponibles, elegimos por ejemplo B_1=12 y B_2=13.

Proyecto de Bloqueo en 4 etapas para un Diseño 2^3								
Run	1	2	3	B1=12	B2=13	23	123	Bloque
1	-	-	-	+	+	+	-	4
2	-	-	+	+	-	-	+	3
3	-	+	-	-	+	-	+	2
4	-	+	+	-	-	+	-	1
5	+	-	-	-	-	+	+	1
6	+	-	+	-	+	-	-	2
7	+	+	-	+	-	-	-	3
8	+	+	+	+	+	+	+	4

Para seleccionar las 2 corridas de cada bloque desarrollamos el diseño $2^{q=4}$ como muestra la tabla siguiente, donde los números en negrita indican la combinación de los niveles de los **bloques** B1 y B2 que definen las corridas.

	B1	
B2	-	+
-	**1**	**3**
+	**2**	**4**

Ahora bien, al elegir B_1=12 y B_2=13 hemos confundido los efectos de bloque con estas interacciones, es decir, ambos efectos aparecerán sumados. Además, se debe considerar la interacción $B_1.B_2$=12.13=23 lo que implica que también estas interacciones estarán confundidas.

Supongamos que en lugar de haber hecho las elecciones anteriores (B_1=12 y B_2=13) elegimos B_1=123 y B_2=13. A primera vista parece mejor porque hemos sacrificado 123, una interacción de tercer orden en lugar de una de segundo orden. Sin embargo, la interacción $B_1.B_2$=123.13=2, que significa que hemos confundido un efecto de bloque con un factor principal, además de las interacciones 13 y 123. Debido a que confundir un factor principal es menos conveniente, este último es un diseño de bloques menos eficiente que el primero.

Las deducciones y discusiones anteriores se hacen sobre la base de una suposición fundamental:

Las interacciones bloque-tratamiento son despreciables

Esto significa que no debería existir interacción entre algún bloque y un factor, de lo contrario sería muy complicado estimar los efectos principales. Por ejemplo: habíamos tomado en el ejemplo anterior B_1=12, esto implica otras 2 relaciones B1.1=2 y B1.2=1; Si hubiese una interacción entre el bloque B_1 y el factor 1, esto se confundiría con el factor 2.

Para revelar todas las interacciones entre bloques y tratamientos tenemos el siguiente ejemplo:

Supongamos que tenemos un diseño 2^5 (k=5) y queremos dividirlo en 8 bloques = 2^3 bloques (q=3), eligiendo B_1=35, B_2=125 y B_3=1345. Multiplicando los bloques entre sí completamos los 7 grados de libertad (2^q-1).

$$\mathbf{B_1B_2=123,\ B_1B_3=14,\ B_2B_3=234,\ B_1B_2B_3=24}$$

El lado derecho del conjunto de igualdades, desde B1=35 hasta B1B2B3=15 muestra las interacciones que son confundidas con los efectos de bloqueo. Debe tenerse cuidado en elegir adecuadamente los bloques para tratar de que queden confundidos con el mínimo número de interacciones de segundo orden. Por ejemplo, si hubiéramos elegido B_1=125, B_2=245 y B_3=1234, las interacciones confundidas con los bloques hubieran resultado ser **125,245,1234,14, 345, 135,3.**

Aquí hay solo una interacción de segundo orden confundida, pero hay una interacción con un factor principal, por lo tanto, el primer diseño es mejor. Observe, que el **orden** de las interacciones B1, B2 y B3 es distinto en los dos diseños.

Optimización de Modelos Multivariables

Selección de modelos e inferencia

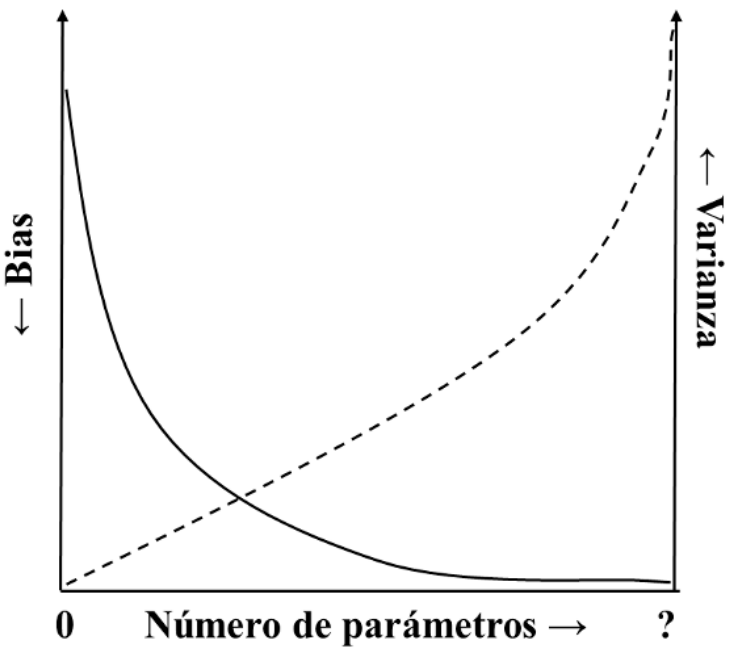
Usualmente utilizamos un modelo de regresión lineal para interpretar muchos de los sistemas que deseamos estudiar, comprender e interpretar hasta tener un algoritmo matemático que represente el comportamiento del sistema influenciado por una serie de variables y parámetros tal como:

$$y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_m x_{jm} + \varepsilon_j \qquad j = 1,2,\ldots,n; \quad i = 1,2,\ldots m \qquad [1]$$

Ya Hemos visto en el capítulo 8 un método de estadística clásica para seleccionar los términos significativos y parámetros que ajustan el modelo bajo estos criterios.

Sin embargo, este punto tiene nuevas e importantes consideraciones que permiten mejorar el ajuste e interpretación del modelo.

La introducción al tema comienza con el ***principio de parsimonia***, que introduce un concepto que incluye una variedad de explicaciones. Aquí mostraremos sólo el concepto básico que nos permitirá avanzar en el tema de la selección de modelos.



El gráfico muestra el sentido opuesto que tienen el '*bias*' y la varianza. Ambos están asociados al número de términos o parámetros introducidos en la ecuación 1. Mientras que el bias disminuye cuando los términos se incrementan, la varianza del modelo aumenta. Este equilibrio puede también interpretarse en términos de *underfitting* o sea insuficiente número de términos para el bias y *overfitting*, sobredimensionado número de términos para la varianza.

De modo que de esto dependerá un bajo ajuste del modelo o un error grande de su varianza, por lo tanto, se debe lograr un equilibrio entre bias y varianza. De eso trata el **principio de parsimonia**, los modelos parsimoniosos alcanzan un apropiado intercambio en el compromiso de estos dos efectos. Una descripción más completa y detallada sobre este y los temas siguientes pueden verse en la (Ref. 1). Comenzaremos a explicar nuevos conceptos para acercarnos al tratamiento de este problema.

Las Teorías de Cuadrados Mínimos y "*Likelihood*"

Likelihood significa probabilidad, pero conservaremos el término en inglés para no confundirlo con la definición clásica de probabilidad, ya que aquí tendrá un significado específico.

Consideremos otra vez el modelo de regresión lineal expresado en la ecuación 1. Los términos del error ε_j son considerados usualmente como una distribución normal al azar con media 0 y varianza constante σ^2, $N(0, \sigma)$. En cuadrados mínimos (CM) las estimaciones de β_0 a β_i (i=1, 2, …, m) son aquellas que minimizan $\Sigma\varepsilon_j^2$, de ahí su nombre de 'cuadrados mínimos'. Todos los cursos básicos de estadística y decenas de libros han explicado este método. Sin embargo, los métodos de la teoría *Likelihood* (TL) son mucho más generales. Aunque esta teoría es el paradigma de los practicantes de estadística frecuentista (clásica) y Bayesiana hay todavía muy poca literatura dirigida a los cursos de grado y posgrado [1-3].

El planteo inicial de la TL comienza con la probabilidad de distribución de un modelo, g, dados los parámetros θ y la descripción de un modelo probable. O sea, el modelo, g, describe la probabilidad de distribución de los datos, x, dados los parámetros θ y una descripción específica de un modelo, denotada por

$$g(x|\theta,\text{modelo}), \qquad [2]$$

que se lee 'g, de x dados θ y el modelo'. Por ejemplo, si se tiene una serie de datos y se sospecha que corresponden a una distribución de Poison, se calcula la distribución de g, dados los datos, x, y los parámetros de la ecuación de Poison. Si la distribución obtenida, g, representa a los datos experimentales, quedará probada la hipótesis de que éstos corresponden al modelo propuesto.

El punto clave para este cálculo es que tanto el modelo (Poison en este ejemplo) como los parámetros θ, son conocidos de antemano y por lo tanto pueden ser dados. Pero para muchos modelos científicos, ni el modelo, ni los parámetros son conocidos. Sin embargo, invirtiendo el razonamiento, pueden colectarse datos, tal que los parámetros pueden ser estimados siempre que un buen modelo pueda ser encontrado o asumido.

La ***function likelihood*** es una función de los parámetros, θ (que es usualmente un vector), dados los datos, x, y el modelo específico, g. O sea:

$$\mathcal{L}(\theta|x,\text{modelo}) \qquad [3]$$

Observe la diferencia en el orden de las variables, comparando con la función [2] y además, que ésta es una función en que lo único desconocido es θ, lo demás es conocido (x) o asumido (el *modelo*). De modo que lo único que cambia es lo que es conocido o dado. Entonces en la función likelihood, los datos son los observados y el modelo es asumido (dado) y el interés recae en estimar los parámetros desconocidos.

Si seguimos con la convención anterior de representar los datos empíricos con x, y a un *modelo aproximante* con g, entonces la escuación [3] se convierte en:

$$\mathcal{L}(\theta|x,g) \qquad [4]$$

La cual debe leerse como "La probabilidad de un valor numérico particular del parámetro desconocido θ (usualmente un vector), dados los datos, x, y un modelo particular, g. Ahora se puede computar la *likelihood function* **(LF)** para obtener varios valores del parámetro θ y guardar el que es **la más probable estimación de θ**, dados los datos y el modelo. Este es el concepto de Fisher de ***maximun likelihood estimation*** **(ML)**.

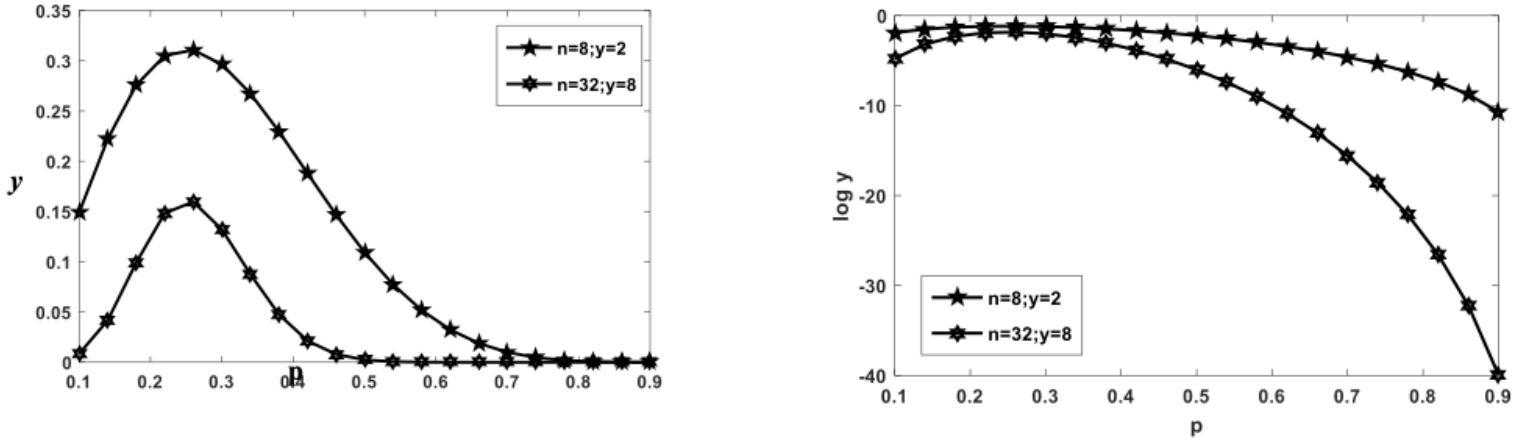
Likelihood es la base de las aproximaciones Bayesianas a la inferencia estadística. Al mismo tiempo es la base de la inferencia estadística, a diferencia de cuadrados mínimos que puede interpretarse solo como un caso limitado y aunque ésta sea muy útil en muchas aplicaciones, no es fundacional en estadística moderna (Ref. 1).

Por razones prácticas, en muchas aplicaciones se utiliza el logaritmo natural de LF :

$$\ln\left(\mathcal{L}(\theta|x,g)\right)$$

o expresiones reducidas tales como $\ln\left(\mathcal{L}(\theta|datos,modelo)\right)$; o solo $\ln\left(\mathcal{L}(\theta)\right)$. Las figuras siguientes muestra cómo cambia la agudeza de la likelihood a medida que el número de datos aumenta. Usaremos para el ejemplo el modelo de la distribución binomial (ecuación [5]) para ver el resultado de arrojar una moneda y contar el número de 'caras' y 'secas'. En el ejemplo, la moneda **no es equilibrada**, ya que n/y ≠ 0.5; *n* es el número de eventos e *y* la distribución de los datos, por lo tanto, el máximo no coincide con 0.5.

$$\mathcal{L}(p|y,n,modelo) = \binom{n}{y} p^y (1-p)^{n-y} \qquad [5]$$



A la izquierda se muestran dos curvas para el modelo *distribución nominal*. Los puntos y líneas indican la probabilidad, p, en función de la distribución de los datos, *y*, y el número de eventos, n, (ver ecuación [5]). A la derecha se muestra el gráfico para el logaritmo natural de *y*. Se observa en ambas figuras que las distribuciones con n=32 datos e y=8 se aproximan al máximo en forma más aguda que lo hace el caso para n=8 e *y*=2. Observe en estos casos la altura de las ordenadas y que la relación n/y es la misma. Vea también que ML=p_{max}≈0.27.

Para cerrar en general este punto, se debe prestar atención a que, tanto para modelos lineales o no lineales, cuando existe en un modelo una distribución normal de los residuos, hay una estrecha relación entre *cuadrados mínimos* (CM) y *maximun likelihood estimation* (ML). Sin embargo, Likelihood y los métodos Bayesianos relacionados, permiten extenderse a gran cantidad de otras clases de modelos. Si se dispone de cálculo computacional es posible explotar estas ventajas para el campo científico experimental (Ref. 1).

Consideraciones Críticas Acerca de “el modelo”

En la gran mayoría de las ciencias fácticas (o empíricas), para explicar un fenómeno a partir de una cantidad limitada de datos, cabe hacerse la siguiente pregunta: ¿Es alcanzable un “***verdadero modelo***”, entendiendo por éste a un modelo que represente la “total realidad”?

La respuesta a esta pregunta es, no, no es alcanzable. Podríamos elegir a la física como la ciencia fáctica más cercana a lograr semejante éxito. Sin embargo, durante muchos siglos se creyó que la teoría de la gravedad de Newton era la verdadera representación de este fenómeno, pero en el siglo XX aparecieron contradicciones que pusieron en crisis a la física, hasta que Albert Einstein las elucidó con una teoría más amplia, la de la relatividad. La física se caracteriza porque sus modelos implican un número bajo de variables, como en el ejemplo que se acaba de dar. Pero en la mayoría de las ciencias fácticas, el número de parámetros es generalmente mucho mayor; fenómenos tales como los que abarcan, por ejemplo, la biología, el ambiente o las ciencias sociales, que necesitarían una enorme cantidad de parámetros y otra enorme cantidad de datos para **aproximarse** a la realidad. Un modelo que considerase por ejemplo unos 100 parámetros podría lograr esta **aproximación** a la realidad, pero ¿Sería entendible, o explicable?

La situación es que entonces la "***total realidad***" será siempre elusiva en las ciencias fácticas, por lo tanto, representamos los hechos con modelos limitados que no la representarán en su totalidad.

Sin embargo, para el desarrollo de la teoría, llamaremos "función f" a la que representa la *total realidad*; aunque en la práctica no es alcanzable, servirá de referencia para la teoría. Un ejemplo de este proceder es la ley de los gases ideales en química; no existe ningún gas ideal perfecto pero la ley sirve como referencia a su proximidad. La función f se considera **fija** y la que varía es la g.

La "Distancia Entre Dos Modelos" o Teoría Kullback–Leibler y la Información

Supongamos que tenemos dos modelos "f" y "g" sean, en este caso, totalmente conocidos, por ejemplo, modelos matemáticos.

La información Kullback–Leibler (K-L) entre los modelos f y g se define para funciones continuas como

$$I(f,g) = \int f(x) \cdot ln\left(\frac{f(x)}{g(x|\theta)}\right) dx \qquad [6]$$

La notación *I(f,g)* denota **"la información perdida cuando g es utilizada para aproximar f"** (la integral suele ser multidimensional). Una forma práctica y equivalente de interpretación es considerar a *I(f,g)* **la distancia de g a f.** Por supuesto, podríamos buscar un modelo aproximante que pierda tan poca información como sea posible, lo que es equivalente a aproximar *I(f,g)* sobre g, pero no es lo usual. *I(f,g)* es siempre positiva, o igual a cero si *f(x)=g(x)*.

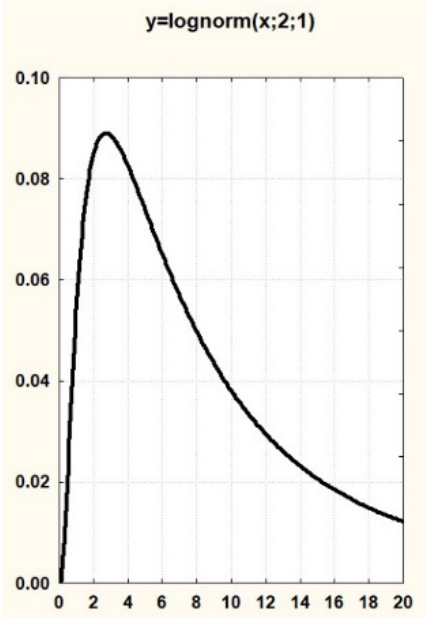
Kullback y Leibler dedujeron una medida de la información que resultó ser el negativo de la entropía de Boltzmann, desde entonces referida como la *distancia o información Kullback–Leibler* y también como *entropía relativa*.

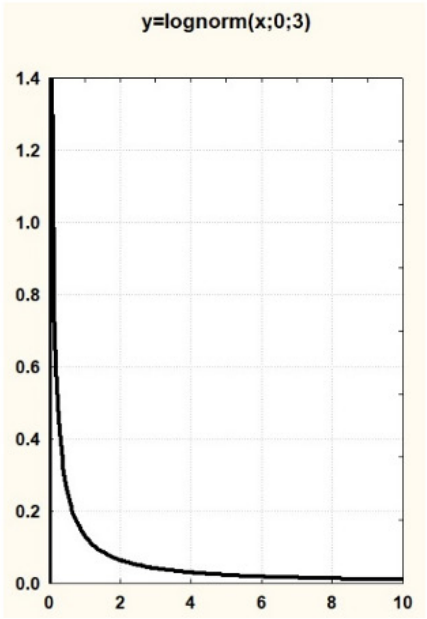
En el caso de las distribuciones discretas la ecuación cambia debido a que la integral de la ecuación [6] se transforma en una sumatoria:

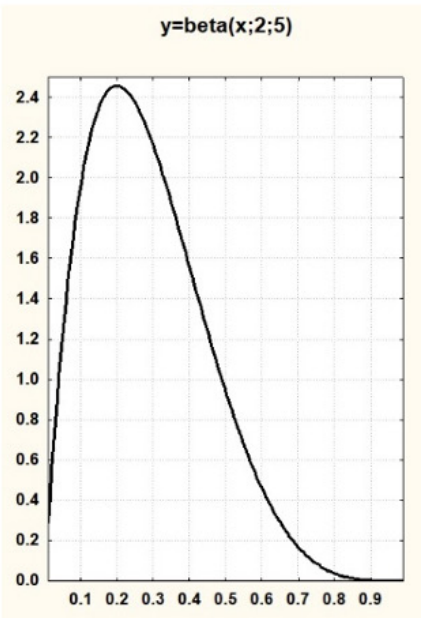
$$\boldsymbol{I(f,g) = \sum_{i=1}^{k} p_i \,.\, ln\left(\frac{p_i}{q_i}\right)} \qquad [7]$$

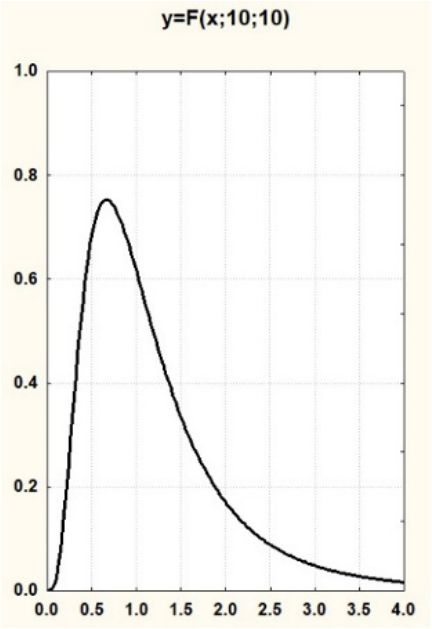
En los casos discretos donde $0 < p_i < 1$ y $0 < q_i < 1$ será $\Sigma\, p_i = \Sigma\, q_i = 1$; entonces $\boldsymbol{f}$ y $\boldsymbol{g}$ se corresponden con p_i y q_i respectivamente.

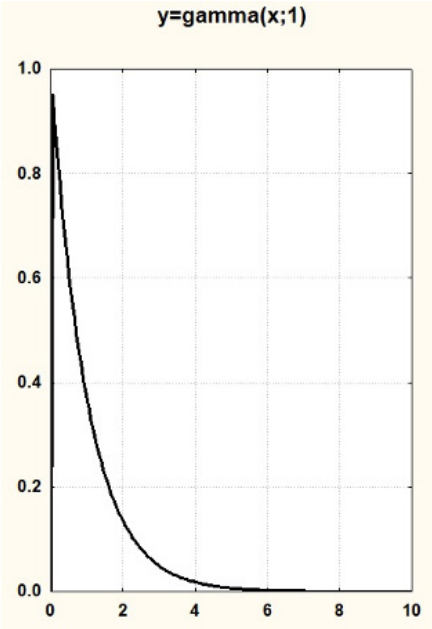
Veamos qué representa *I(f,g)* gráficamente en cuanto a la "**distancia de g a f**". Las figuras siguientes son representaciones de distribuciones de diferentes modelos, cada una con diferentes parámetros. Nos referiremos a ellas contándolas en el sentido de lectura: 1-Lognormal (a), 2-Lognormal (b), 3-Beta, 4-Fisher, 5-Gamma, 6-Student, 7-Normal. A simple vista, algunas de ellas parecen muy similares, por ejemplo, la 6- Student y la 7-Normal; otras parecen muy diferentes, por ejemplo, la 1-Lognormal (a) y la 5-Gamma. Sin embargo, hay que observar dos cosas: Por un lado, que las escalas sean similares o sea que las curvas se superpongan y por otro lado hay que tener en cuenta que la distribución depende de los parámetros de cálculo.

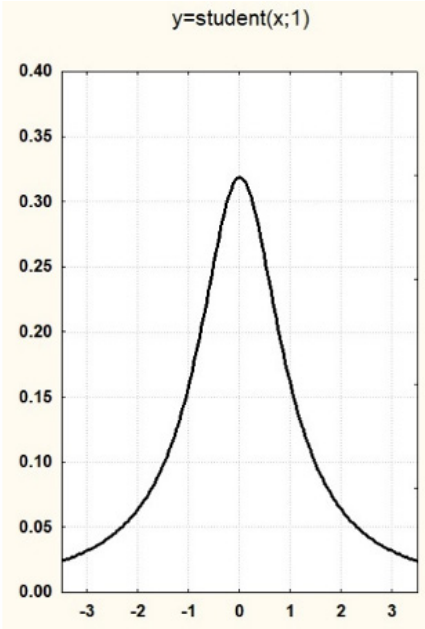

y=lognorm(x;2;1)
0.10
0.08
0.06
0.04
0.02
0.00
0 2 4 6 8 10 12 14 16 18 20

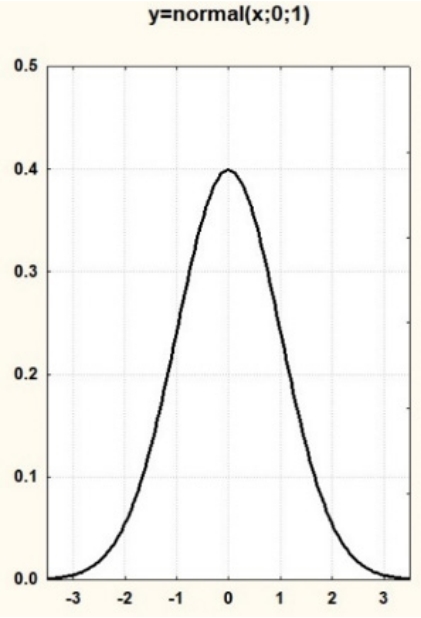

y=lognorm(x;0;3)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0 2 4 6 8 10


y=beta(x;2;5)
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9


y=F(x;10;10)
1.0
0.8
0.6
0.4
0.2
0.0
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0


y=gamma(x;1)
1.0
0.8
0.6
0.4
0.2
0.0
0 2 4 6 8 10


y=student(x;1)
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-3 -2 -1 0 1 2 3



Por ejemplo, 2-Lognormal (b) se aproxima a 5-Gamma mucho más que la 1-Lognormal (a). Pero la comparación precisa se hace con el cálculo de *I(f,g)*; cuanto menor sea *I(f,g)*, menor será la distancia entre las distribuciones o, lo que es lo mismo, menor será la perdida de información al aproximar *g(x)* a *f(x)* y más iguales serán las distribuciones.

Recordemos que formalmente, el cálculo de K-L requiere el conocimiento de la verdadera distribución *f* y el conocimiento de todos los parámetros del modelo *g(i)*, con lo que K-L no puede ser computada en problemas del mundo real. Sin embargo, si se requiere únicamente una distancia relativa, este requerimiento necesita menos exigencia.

Dado que *I(f,g)* (ecuación [6]) puede ser escrita en forma equivalente como,

$$\boldsymbol{I(f,g) = \int f(x).\ln\big(f(x)\big).dx - \int f(x).\ln\big(g(x|\theta)\big).dx}$$

Note que $\int f(x)\,dx$ es la esperanza estadística (o matemática) respecto de f, que llamaremos Ef. Entonces la distancia K-L puede ser expresada como la diferencia entre dos esperanzas estadísticas,

$$\boldsymbol{I(f,g) = E_f.\ln\big(f(x)\big) - E_f.\ln\big(g(x|\theta)\big)} \qquad [8]$$

La esperanza estadística ***Ef.ln(f(x))*** es una constante, **C**, que depende de una distribución verdadera, pero desconocida. Entonces ***Ef.ln(g(x|θ))*** se transforma en una ***distancia relativa*** entre *f y g*. Como consecuencia, si tenemos dos modelos g_1 y g_2 que estiman idéntica *f* y $I(f,g_1) < I(f,g_2)$, la constante **C** se resta y desaparece, por lo que $I(f,g_1)$ es mejor modelo que $I(f,g_2)$. Si bien no conocemos la distancia absoluta entre las g_1-*f* y g_2-*f* (lo que significa que no conocemos cuán buenas son) al menos sabemos cuál de las dos es el mejor modelo. Este concepto es esencial para comprender nuestro objetivo último sobre la aplicación práctica de este tema.

Ahora bien, dado un modelo paramétrico estructural habrá un único valor de θ (recordemos que θ puede ser un vector) que minimiza la distancia K-L, o sea $I(f,g)$. Este valor depende del verdadero valor *f* y de la estructura del modelo *g*. Entonces, existe un *verdadero valor* de θ que conforma la estimación ML; al que llamaremos θ_0. Con lo cual, si uno conoce de alguna manera que el modelo *g* es el mejor modelo para K-L, entonces ML de $\hat{\boldsymbol{\theta}}$ estimará θ_0. Esta cualidad de $g(x)|\theta_0$ como minimizador de la distancia K-L para todo θ es la base de uno de los pilares del ***criterio de información de Akaike***.

Sobre la base de los dos últimos conceptos, Akaike desarrolló en 1973 una demostración rigurosa para este ***criterio de información,***

$$\boldsymbol{AIC = -2\ln\left(\mathcal{L}(\hat{\theta}|y)\right) + 2K} \qquad [9]$$

donde K es el número de parámetros estimables del modelo.

(Para ver la demostración consulte la (Ref. 1)). Con esta ecuación, en lugar de tener una medida de la distancia directa K-L entre dos modelos, uno obtiene una *estimación* de la *esperanza* de la distancia relativa entre el modelo ajustado y el mecanismo verdadero (desconocido, y que podría ser de dimensión infinita) que generaron los datos observados. El término $\left(\mathcal{L}(\hat{\theta}|y)\right)$ es el valor en el máximo de ln-likelihood. En aplicaciones prácticas se calcula AIC para todos los modelos candidatos y se selecciona el que tenga menor valor.

Debe tenerse muy en cuenta que AIC no indica el valor absoluto de la calidad del modelo porque es una medida comparativa entre dos modelos. Entonces, más allá del valor AIC que tengan los modelos comparados, ambos pueden ser malos o buenos.

AIC puede resultar desviada si el número de parámetros del modelo es muy grande en comparación al tamaño de la muestra. Sugiura et.al. (Ref. 4) desarrollaron una corrección para estos casos denominada AICc, donde n es el número de datos de la muestra.

$$AICc = AIC + \frac{2K(K+1)}{n-K-1} \qquad [10]$$

Generalmente se recomienda la aplicación de AICc cuando la relación n/K es menor a 40 (Ref.1). También se desarrollaron fórmulas de corrección (menores) para los casos en que existe "sobredispersión de los datos" (Ref. 1 pág. 67) (los datos están fuera de cualquiera de las distribuciones conocidas).

La comparación con el caso de "Cuadrados Mínimos (LS)"

Volveremos ahora a tratar los modelos de regresión lineal (ver ecuación [1]) y el principio de parsimonia bajo los principios que hemos desarrollado hasta aquí.

En los típicos modelos de regresión lineal se estiman los coeficientes βi (i=0,…,m) para los términos del modelo por el método de la estadística clásica. Desde el punto de vista del principio de parsimonia el problema es estimar el mejor número de términos (K) entre todos los que han sido propuestos en principio. Usualmente, en cuadrados mínimos esto se hace calculando R^2 y eligiendo el más próximo a 1. Veremos ahora cuál es la relación entre cuadrados mínimos y el ***criterio de información de Akaike*** que hemos presentado.

Como en este caso particular estamos evaluando **un mismo modelo** que depende de los βi, en este caso θ será el vector que contiene los βi. AIC puede ser fácilmente calculado desde los resultados de cuadrados mínimos como:

$$AIC = n.(\ln(\hat{\sigma}^2)+2K$$

Siendo $\hat{\sigma}^2 = \sum \frac{\hat{\varepsilon}_i^2}{n}$, (la ML estimación de σ^2) y $\hat{\varepsilon}_i$ la estimación de residuos para el particular candidato del modelo. Debe tenerse cuidado en calcular σ^2 con la expresión dada aquí y no utilizar una salida computada de modo difrrente. Para los cálculos **LS** se debe tener en cuenta que K **es el número total de parámetros estimados tomando en cuenta el término independiente β_0 y además σ^2** porque desde el punto de vista de la teoría likelihood (que es una teoría más general) σ^2 no es estimado en ella.

El menor valor de AIC (o AICc) indicará el mejor modelo. En algunas ocasiones R^2 no coincidió con AIC, pero este último resultó más correcto.

Deseabilidad
Combinación de múltiples respuestas

Si tenemos un sistema con varias respuestas donde una es de primordial importancia, pero sujeta a restricciones impuestas por las otras respuestas, decimos que estamos ante una optimización restringida. Por ejemplo, supongamos que la producción de cierto producto lácteo depende de 5 factores, y para optimizar el producto de acuerdo al consumo actual se necesita que tenga determinada densidad, pH, cantidad de suero lácteo y tenor graso. Todas estas variables (Respuestas) dependen de **los factores** de producción. Por lo tanto, para optimizar el producto se necesita optimizar los factores para una deseada **combinación de respuestas**. Esta técnica fue introducida por Derringer, G. y Suich, R (Ref. 6).

La estrategia consiste en balancear cada respuesta a un valor de deseabilidad di. El **promedio geométrico** de las funciones de deseabilidad d_1, d_2, …d_i. es $\mathbf{d=[d_1.d_2….d_i]^{1/i}}$ **[11]**; que es la deseabilidad total; $0\leq d \leq 1$.

Las d_i pueden estar *pesadas* (wi.di) con pesos wi con las condiciónes $0\leq w_i\leq 1$ y $\Sigma w_i=1$. Observe que si algún $d_i=0$, entonces d=0, o sea que el resultado es inaceptable.

Hay tres tipos de funciones de deseabilidad, a saber: Caso 1, 'la respuesta se desea ajustar a un valor objetivo'; caso 2, 'la respuesta se desea ajustar a un valor mínimo' y caso 3, 'la respuesta se desea ajustar al valor máximo posible (teóricamente infinito)'.

Caso 1: La respuesta se desea ajustar a un valor objetivo 't'. Por simplicidad denominaremos d, en lugar de di, a cada respuesta ŷ que se desea controlar dentro del rango del modelo.

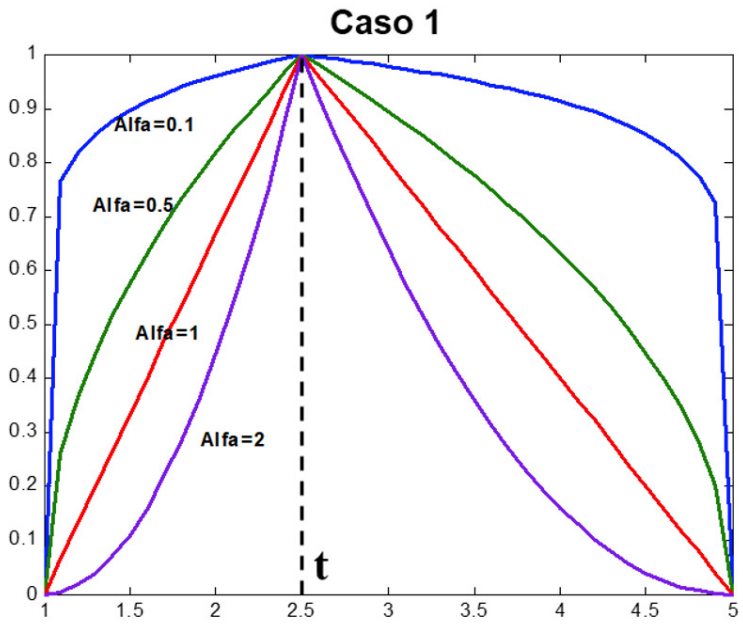
$$d = \begin{cases} \left|\dfrac{\hat{y}-L}{t-L}\right|^{\alpha_1}, & L \le \hat{y} \le t, \\ \left|\dfrac{\hat{y}-U}{t-U}\right|^{\alpha_2}, & t \le \hat{y} \le U, \end{cases}$$

Se eligen L y U tales que el resultado es inaceptable si:

$\hat{y} < L$ o $\hat{y} > U$, o sea

$U \ge y \ge L$

En la figura de la derecha podemos ver cómo el efecto de α modifica el modo de acercarse al valor t, según se desee regular un acercamiento suave o agudo, que estará determinado por la tolerancia alrededor del valor t.

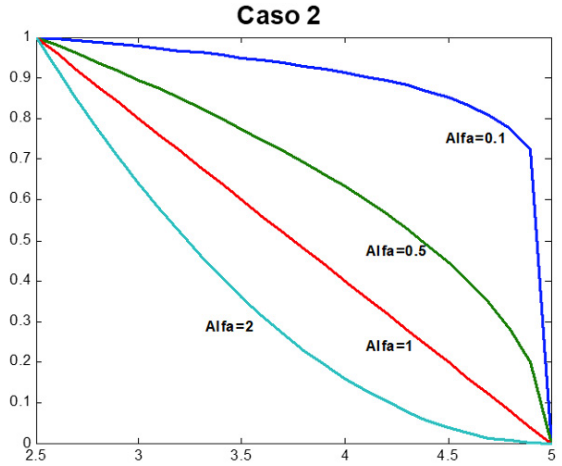


Caso 2: La respuesta se desea ajustar a un valor mínimo

a es el valor más pequeño posible para ŷ**.**

d=0 si y > U.

$$d = \left| \frac{\hat{y} - U}{a - U} \right|^{\alpha}, \qquad a \le \hat{y} \le U,$$

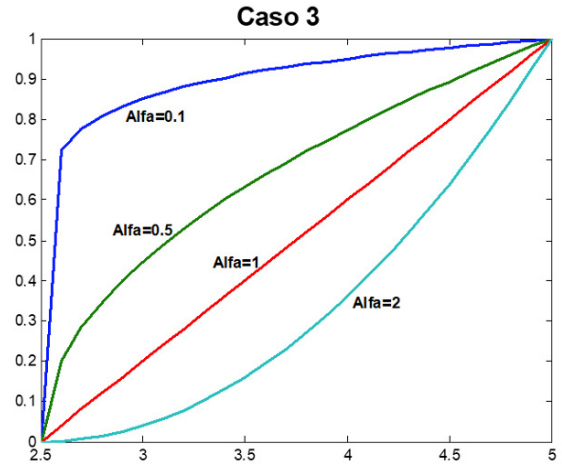


Caso3: La respuesta se desea ajustar al valor máximo posible (teóricamente infinito)

d=0 para y < L

d=1 para y>U

$$d = \left| \frac{\hat{y} - L}{U - L} \right|^{\alpha}, \quad L \le \hat{y} \le U,$$



En algunos casos experimentales los límites L y U son difíciles de establecer en la formulación anterior debida a Derringer y Suich.

Pueden definirse otras funciones de tipo exponencial para evitar este problema, que se muestran a continuación.

Las constantes c y α pueden usarse para ajustar la escala y la forma de la función

Caso 1:

$$\mathbf{d}= \begin{cases} \boldsymbol{exp}[-\mathbf{c_2}|\hat{y} - \boldsymbol{t}|^{\boldsymbol{\alpha_2}}] & -t \le \hat{y} \le \infty, \\ \boldsymbol{exp}[-\mathbf{c_1}|\hat{y} - \boldsymbol{t}|^{\boldsymbol{\alpha_1}}] & -\infty < \hat{y} \le t, \end{cases}$$

Caso 2:

$$\mathbf{d}=\boldsymbol{exp}[-\mathbf{c}|\hat{y} - \boldsymbol{a}|^{\boldsymbol{\alpha}}] \qquad -a \le \hat{y} \le \infty,$$

Caso 3

$$\mathrm{d} = 1 - \frac{\exp[-c\hat{y}^{\alpha}]}{\exp[-cL^{\alpha}]}, \qquad L \le \hat{y} < \infty \qquad \mathrm{d} = 0 \ \text{para y} < \mathrm{L}$$

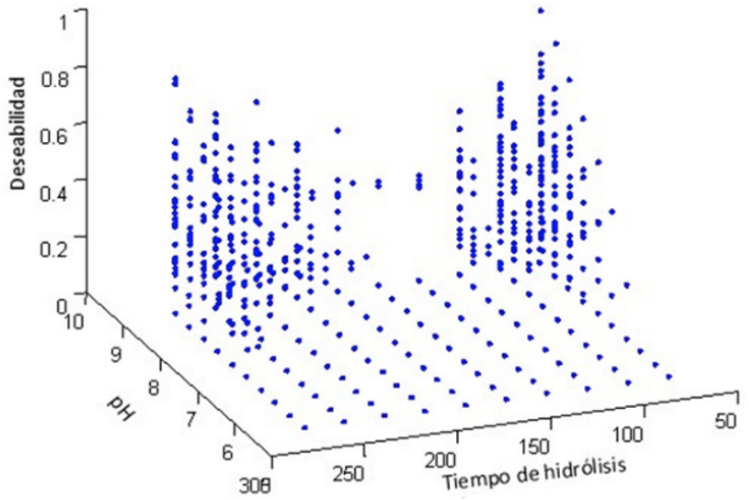


Figura 1

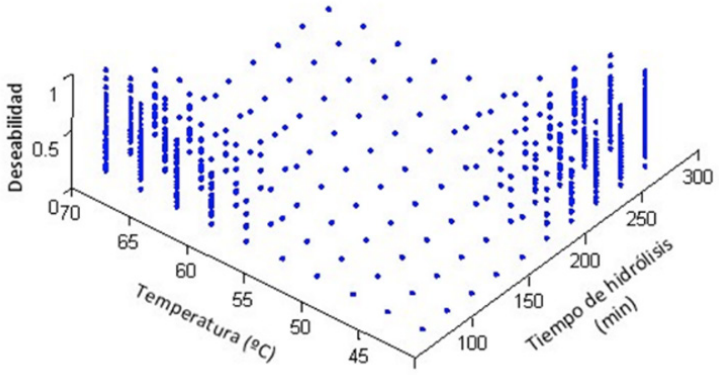


Figura 2

Las figuras 1 y 2 ejemplifican el uso de la deseabilidad para el caso de un modelo de hidrolizado de queso de cabra para producción de diversos productos lácteos a partir de 5 factores (ingredientes y condiciones de operación). El modelo predice 11 respuestas bromatológicas y **cada producto deseado** necesita una combinación específica de respuestas y por lo tanto una determinada combinación de los factores (Ref. 5).

De las muchas combinaciones posibles de factores para analizar la deseabilidad, se muestran aquí, solo 2 para la producción de un aditivo para panificación. En la primera se observa que para una zona de pH de entre 8.5 y 9.5 hay 2 opciones posibles de rangos de tiempo de hidrólisis: 50-100 y 200-275 minutos. Pero en la figura 2 se observa que para las temperaturas de trabajo, la única posibilidad es un rango de 200-275 minutos para el tiempo de hidrólisis. Por lo tanto, el rango 50-100 queda descartado. De esta forma se pueden ir ajustando las sucesivas respuestas para obtener un determinado producto. Obviamente se puede aplicar directamente la ecuación [1] para obtener directamente la deseabilidad total, pero cuando los factores son muchos, el resultado final puede ser una deseabilidad muy baja.

Optimización no Algorítmica Sobre la Base de la Teoría Grey

En 1982 Deng Julong (Ref. 7) enunció las bases de la teoría del *sistema grey* (GS) que ha resultado de interés en muchas ramas de la ciencia y tecnología. Un trabajo introductorio sobre la teoría fue publicado en el mismo año por él mismo (Ref. 8). En éste se menciona una larga lista de aplicaciones en ecología, economía, medicina, geología e hidrología, entre otras.

Un sistema grey es uno en el cual no toda la información es conocida. Yendo desde **negro**, una total carencia de información sobre calidad y cantidad, hasta **blanco**, la información completa; sobre un rango continuo de grises (grey) de información parcial. Un concepto importante del cual derivan las ecuaciones que se dan a continuación es el de ***entropía de la información***, que no está directamente relacionada con la entropía física. En este campo ***la entropía es una medida de la cantidad de información*** (ver **Teoría Kullback–Leibler** en este capítulo).

Entre toda la bibliografía existente hoy en día, se ha publicado un review con una breve descripción de la filosofía y reglas algebraicas básicas (Ref. 9).

Grey relational analysis (GRA), el cual deriva de *Grey relational space* (GRS), es un algoritmo que apunta a resolver interrelaciones entre múltiples variables dependientes e independientes (Ref. 9-12). El tema ha crecido en importancia debido a que puede calcular un modelo no algorítmico en lugar de un análisis por regresión; aún es posible de utilizar como una herramienta analítica especialmente en casos donde los datos son insuficientes (Ref 3,4).

Pasos para el cálculo de GRA:

1- Autoescalado minimax (o minimin) de factores y respuestas

$$\xi_i(k) = \frac{\Delta_{min} + \zeta.\Delta_{max}}{\Delta_{0,i}(k) + \zeta.\Delta_{max}}$$

k =1,...n. n=número de parámetros

i =1,...m. m=número de datos experimentales

$\Delta_{0,i}$=Secuencia de desviación entre factores y respuestas

$$\Delta_{0,j} = \|x_0^*(k) - x_i^*(k)\|, \quad i \neq j$$

$$\Delta_{min} = min^+.min^+\left\|x_0^*(k) - x_j^*(k)\right\|$$

$$\Delta_{max} = max^{+}.max^{+}\left\|x_0^*(k) - x_j^*(k)\right\| \qquad ^{+} \equiv \forall j \in i \forall k$$

$x_0^*(k)$= Secuencia de respuestas

$x_i^*(k)$ = Secuencia de factores

$$\xi_i(k) = \frac{\Delta_{min} + \zeta.\Delta_{max}}{\Delta_{0,i}(k) + \zeta.\Delta_{max}} \qquad [1]$$

ζ coeficiente de ajuste (rango 0-1) para sensibilizar las diferencias entre las series escaladas de factores y respuestas. Usualmente se le da el valor 0.5.

Debido a que Δ_{min} y Δ_{max} toman los valores 0 y 1 respectivamente, la ec. 1 se Reduce a;

$$\xi_i(k) = \frac{0 + \zeta(1)}{\Delta_{0,i}(k) + \zeta(1)} = \frac{0.5}{\Delta_{0,i}(k) + 0.5}$$

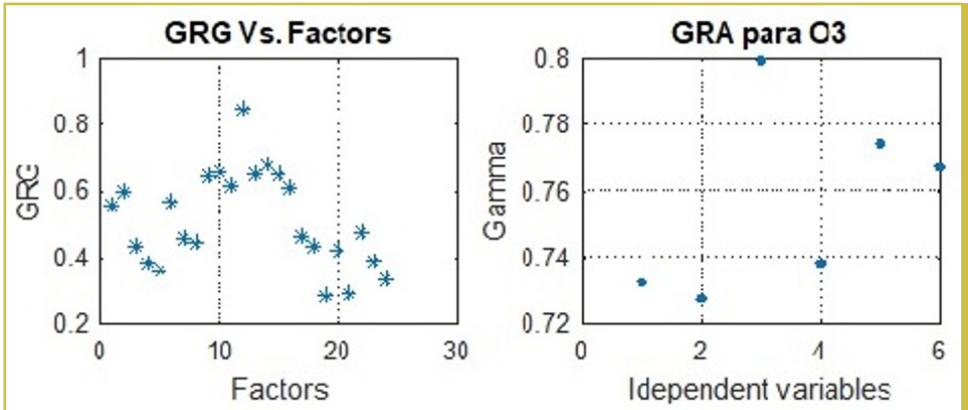
El grado relacional de GRA (GRG) se calcula con ec.(2); GRG se utiliza para establecer el grado de influencia que los factores ejercen sobre la respuesta. Por ejemplo, si: $\gamma(x_0, x_i) > \gamma(x_0, x_j)$ entonces el elemento x_i está más cerca de la respuesta que x_j. γ_i tiene rango 0-1 donde 1 significa máxima relación con la respuesta.

$$\gamma_i = \frac{1}{n}\sum_{k=1}^{n} \xi_i\,(k) \qquad [2]$$

Generalmente γ_i >0.9 indica una marcada influencia del factor; γ_i >0.8 indica una relativamente marcada influencia; γ_i >0.7 una influencia notable y γ_i < 0.6 una influencia despreciable.

Veamos un ejemplo de cálculo para un estudio medioambiental de monitoreo del aire ya mencionado en el capítulo 5. Por el tipo de problema, los datos vienen dados y no puede practicarse un diseño factorial: tomaremos sólo 1 día de muestreo (promedio de 10 minutos cada 24 horas) de la base de datos que tiene 6 factores, a saber: 1, Hora del día; 2, altura de la capa de nubes más baja; 3, grado de nubosidad; 4, dirección, y 5 velocidad del viento por sectores angulares y 6, factor de irradiación solar UV. La respuesta elegida para analizar este ejemplo es la concentración máxima, promedio de 10 minutos, de Ozono (O_3).

Se muestran los resultados del cálculo para analizar los efectos de los factores para la condición O_3 máximo.



Resultado para O3 hacia máximos valores.

Este resultado muestra que para GRA todos los factores están en el rango "influencia notable" γ_i >0.7, particularmente el factor 3, "grado de nubosidad" (alto), está en el límite con "relativamente marcada influencia". En GRG se observa que la medición N°12 tiene máximo valor, esa medición tiene los siguientes parámetros relativos:

Factor	hora	Altura. de la capa de nubes más baja	Grado de nubosidad	Dirección del viento	Intensidad del viento	Factor solar de radiación UV
Valor	12	9/9 (máxima)	7/8	NE	8/23	0.300 (máx. del día)

Claramente, esas son las condiciones para el máximo valor de O3: Mediodía, máxima altura de la capa de nubes, alto grado de nubosidad, viento NE con baja intensidad del viento y máximo factor de radiación solar UV para el día. Observe que estas simples conclusiones fueron obtenidas con apenas 24 muestras.

Volvemos ahora en este tópico al análisis y la optimización de sistemas experimentales multirespuesta. Hasta ahora estaba establecido que un método eficiente para optimizar sistemas multirespuesta simultáneas es el de deseabilidad, introducido por Derringer y Suich.

Mostraremos otro desarrollo de la teoría Grey, denominado *Grey relational analysis*, (GRG). A diferencia de la deseabilidad, éste es un cálculo no algorítmico para optimizar los resultados de un DOE con respuestas múltiples.

El *Grey relational coeficient*, (GRC), está expresado como $\xi_i(k)$ en la ec. 1. Para todos los ensayos del diseño, GRC se toma con la misma ecuación $\xi_i(k)$.

El GRG introduce un cálculo de pesos para GRC; por ejemplo, para 2 respuestas r1 y r2, GRG vale:

$$\gamma_i = W_1 . \xi_{ir1}(k) + W_2 . \xi_{ir2}(k) \qquad [3]$$

Donde W_1 y W_2 son los pesos fijos asignados para toda la columna de ensayos, con la condición $\Sigma W = 1$. El más alto valor de GRG de la columna de ensayos indica que éste es el valor más próximo al idealmente normalizado.

Los factores pueden evaluarse mediante un análisis Taguchi para lo cual el DOE debe ser un diseño Taguchi. Veremos que la eficacia de la evaluación coincide con un análisis ANOVA. Los diseños Full Factor pueden utilizarse porque cumplen con las exigencias de los diseños Taguchi, pero son más extensos que éstos.

Para verificar las características de GRG describiremos el cálculo con datos tomados del problema 9-7 de Montgomery D. (Ref. 11). El problema original es para resolver un DOE 3^3, pero aquí lo utilizaremos para los fines de GRG e incluso agregaremos una respuesta ficticia con objetivos didácticos. La tabla siguiente muestra la planilla de datos:

Tabla de diseño 3^3 - GRG						
0	**Hombre**	**Botella**	**Estante**	**R1**	**R2**	**Riesgos**
1	1	1	1	3.45	3.36	1.9259
2	1	1	0	4.14	4.19	0.4869
3	1	1	-1	5.8	5.23	0.9259
4	1	0	1	4.07	3.52	0.5895
5	1	0	0	4.38	4.26	0.2096
6	1	0	-1	5.48	4.85	0.4124
7	1	-1	1	4.2	3.68	0.4068
8	1	-1	0	4.26	4.37	0.3338
9	0	-1	-1	5.67	5.58	0.6926
10	0	1	1	4.8	4.4	0.0014
11	0	1	0	5.22	4.7	0.1461
12	0	1	-1	6.21	5.88	1.883
13	0	0	1	4.52	4.44	0.101
14	0	0	0	5.15	4.65	0.0975
15	0	0	-1	6.25	6.2	1.9944
16	0	-1	1	4.96	4.39	0.0149
17	0	-1	0	5.17	4.75	0.1104
18	0	-1	-1	6.03	6.38	1.4214
19	-1	1	1	4.08	3.65	0.5742
20	-1	1	0	3.94	4.08	0.806
21	-1	1	-1	5.14	4.49	0.0913
22	-1	0	1	4.3	4.04	0.2892
23	-1	0	0	4.53	4.08	0.0947
24	-1	0	-1	4.99	4.59	0.0232
25	-1	-1	1	4.17	3.88	0.4459
26	-1	-1	0	4.86	4.48	0.0005
27	-1	-1	-1	4.9	4.9	0.0001

El problema consiste en optimizar la descarga y acomodamiento de tres tipos distintos de botellas de 32 onzas (factor B) en tres tipos de estanterías (factor C) por medio de tres diferentes trabajadores (factor A).

Niveles: Escalados a 1, 0 y -1:

A. Hombres

B. Plástico, vidrio de 28 mm y vidrio de 38 mm.

C. Estantes permanentes, vitrinas cerradas al final de pasillo y heladeras para bebidas.

La respuesta es el tiempo (minutos) empleado para acomodarlas (2 replicados, R1 y R2). Hasta aquí el problema original. Para comprobar GRG se agregó una tercera respuesta (ficticia) para medir el riesgo del trabajo tanto para el cuidado de la mercadería como del personal.

Primero analizaremos el caso para las respuestas R1 y R2, ambas para optimizar como 'mínimos' y veremos cuál es la combinación de niveles para obtener el tiempo mínimo empleado. La tabla siguiente muestra los valores de GRG para las respuestas R1, R2 y su suma. El máximo GRG total se alcanza en el primer ensayo. El resultado es obvio en este caso cuando se observa la tabla de diseño con los tiempos R1=3.45 y R2=3.36. En la siguiente figura 3 se muestra el análisis Taguchi,

Se puede ver el valor de GRG para cada factor y en cada nivel. Debajo están los valores numéricos del análisis. Además, figura el ranking de los factores (factor 3>factor 1>factor 2) que es coincidente con el análisis ANOVA del diseño.

Run	R1	R2	total GRG
1	0.5	0.5	1
2	0.2745	0.2609	0.5355
3	0.1317	01632	0.2948
4	0.2877	0.425	0.7126
5	0.2373	0.2508	0.4881
6	0.1463	0.11891	0.3354
7	0.2642	0.3695	0.6336
8	0.2545	0.2364	0.491
9	0.1373	0.1449	0.2822
10	0.1918	0.2328	0.4246
11	0.1609	0.2017	0.3626
12	0.1167	0.1322	0.2489
13	0.2199	0.2281	0.448
14	0.1654	0.2063	0.3716
15	0.1154	0.1209	0.2363
16	0.1787	0.234	0.4127
17	0.1641	0.1973	0.3614
18	0.1228	0.1154	0.2382
19	0.2857	0.3788	0.6645
20	0.3158	0.2786	0.5944
21	0.166	0.2225	0.3885
22	0.2485	0.2856	0.5341
23	0.2188	0.2786	0.4973
24	0.1765	0.2121	0.3885
25	0.2692	.0.3177	0.5869
26	0.1867	0.2236	0.4103
27	0.1875	0.1852	0.3727

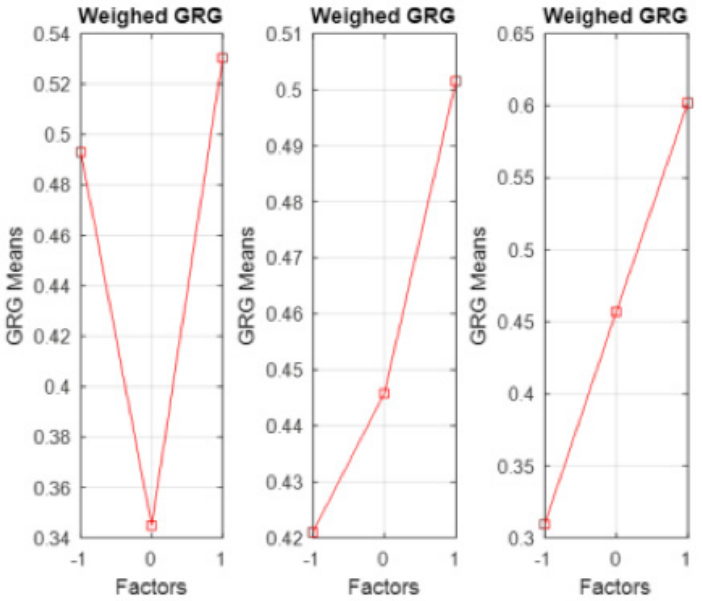


Figura 3

Taguchi Análisis

Factor 1	MEANS	Factor 2	MEANS	Factor 3	MEANS
-1	0.49303	-1	0.42099	-1	0.30951
0	0.34492	0	0.44579	0	0.45691
1	0.53036	1	0.50153	1	0.6019

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	0.17315	2	0.08658	11.09	0.0006
X2	0.03062	2	0.01531	1.96	0.1669
X3	0.38472	2	0.19236	24.63	0
Error	0.15617	20	0.00781		
Total	0.74467	26			

Constrained (Type III) sums of squares.

FACTORS	Deltas/Rank
1	0.1854
2	0.0805
3	0.2924

La predicción del valor medio de GRG, $\gamma_{pred,}$ en el óptimo nivel de los parámetros fijados es 0.7216.
Esta predicción, dentro de los límites de 95% de confianza, puede chequearse con unos pocos experimentos adicionales al diseño original.

$$\gamma_{Pred} = \gamma_m + \sum_{i=1}^{n} (\gamma_i - \gamma_m) \qquad [4]$$

γ_m es la media de los γ; n es el número de parámetros (respuestas). Ver en Ref. 12 el desarrollo de esta ecuación.

En una prueba adicional se decide agregar una respuesta que toma en cuenta **los riesgos** de producir deterioros en la presentación y protección de la mercadería y de los operarios. Esta respuesta debe tender también a un valor mínimo, tal como los tiempos.

Se repite el cálculo para observar las diferencias con el anterior. Los factores de peso para W serán ahora 0.25 para R1 y R2 (son replicados) y 0.5 para Riesgo.

Run	GRG R1	GRG R2	GRG Riesgo	GRG Total
1	0.25	0.25	0.1185	0.6185
2	0.1373	0.1305	0.2757	0.5434
3	0.0658	0.0816	0.1963	0.3437
4	0.1438	0.2125	0.2519	0.6082
5	0.1186	0.1254	0.3703	0.6144
6	0.0732	0.0945	0.296	0.4637
7	0.1321	0.1847	0.2977	0.6145
8	0.1273	0.1182	0.321	0.5665
9	0.0686	0.0725	0.2318	0.3728
10	0.0959	0.1164	0.4989	0.7112
11	0.0805	0.1008	0.4019	0.5832
12	0.0583	0.0661	0.1206	0.245
13	0.1099	0.114	0.4278	0.6518
14	0.0827	0.1031	0.43	0.6158
15	0.0577	0.0605	0.1154	0.2335
16	0.0894	0.117	0.4879	0.6943
17	0.082	0.0986	0.4222	0.6029
18	0.0614	0.0577	0.1481	0.2672
19	0.1429	0.1894	0.2552	0.5874
20	0.1579	0.1393	0.213	0.5102
21	0.083	0.1112	0.4339	0.6281
22	0.1243	0.1428	0.3371	0.6042
23	0.1094	0.1393	0.4317	0.6804
24	0.0882	0.106	0.4814	0.6757
25	0.1346	0.1588	0.2865	0.058
26	0.0933	0.1118	0.4997	0.7048
27	0.0938	0.0926	0.5	0.6864

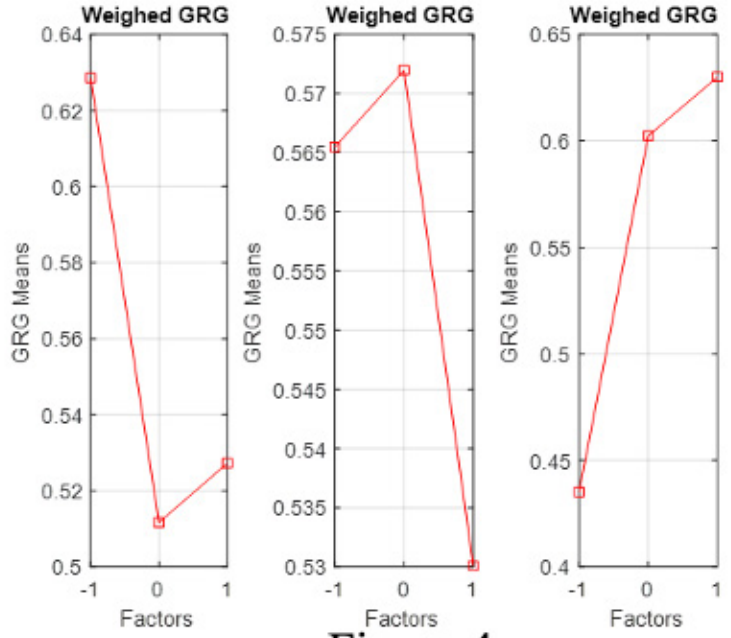


Figura 4

Factor 1	MEANS	Factor 2	MEANS	Factor 3	MEANS
-1	0.62857	-1	0.56547	-1	0.43513
0	0.51167	0	0.57197	0	0.6024
1	0.5273	1	0.53009	1	0.63001

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	0.0725	2	0.03625	2.99	0.0729
X2	0.00914	2	0.00457	0.38	0.6903
X3	0.20015	2	0.10008	8.26	0.0024
Error	0.24218	20	0.01211		
Total	0.52397	26			

Constrained (Type III) sums of squares.

FACTORS	Deltas/Rank
1	0.1169
2	0.0419
3	0.1949

$\gamma_{pred} = 0.7189$

Se observa que ahora el máximo GRG corresponde al ensayo 10. En la tabla de diseño vemos que en este ensayo los tiempos han aumentado un poco pero el riesgo está en el mínimo.

También se observa que el ranking de los factores queda como antes factor 3>factor 1>factor 2.

Lo que ha cambiado significativamente son los GRG medios de los niveles de cada factor (figuras 3 y 4). Comparando entre las corridas 1 y 10, se ve que el nivel del factor 1 (hombre) ha cambiado, como consecuencia el trabajador optimizado cuando se toma en cuenta el riesgo, es el de nivel 0 y no el 1. También hay cambios en los GRG medios de los factores 2, botellas, y 3, estantes, pero no cambian significativamente para la optimización (comparar run 1 con run 10 en la planilla de diseño).

La conclusión es que lo óptimo es sacrificar algo el tiempo mínimo y con un aumento del 35% se logra el menor riesgo y se estima el tiempo óptimo para la operación (4.6minutos, run 10). También permite elegir el mejor trabajador para la tarea, que emplea el tiempo óptimo (nivel 0 del factor 1), no confundir con la media del factor 1 que recae en el nivel -1.

Cabe aclarar que con ensayos para la validación se pueden obtener conclusiones adicionales como la distribución de residuos y su proximidad al valor de γ_{pred}. No poseemos estos datos, pero quién quiera revisar un ejemplo práctico puede consultar la referencia 8.

Referencias

1. Kenneth P. Burnham and David R. Anderson. Model Selection and Multimodel Inference. Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA. Second edition 2002.

2. J.C.F. Wu, M. Hamada, Experiments, Planning, Analysis, and Parameter Designs Optimization, Wiley-Interscience Publications, New York, 2000

3. D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, Amsterdam, 1997.

4. Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods* **A7**, 13–26.

5. Sonia Barberis , Mauricio Adaro, Héctor Sturniolo and Jorge Magallanes. Functional properties of goat cheese protein hydrolysed. Evaluation by artificial neural network. International Journal of Advances in Computer Science & Its Applications Volume 6 : Issue 1 [ISSN 2250-3765].

6. Derringer, G. y Suich, R. 1980, "Simultaneous Optimization of Several Response Variables" *Journal of Quality Technology*, 12 (4): 214-219.

7. Deng Julong (1982a). Control problems of Grey Systems, Systems and Control Letters,5,288-94.

8. The Journal of Grey Systems 1 (1989) 1-24.

9. Qiaoxing LI, Yi Lin, Review paper: A Briefing to Grey Systems Theory. Journal of Systems Science and Information Apr., 2014, Vol. 2, No. 2, pp. 178–192.

10. Fuel Processing Technology 87 (2006) 123 – 127.

11. Douglas c. Montgomery. Diseño y Análisis de Experimentos. segunda edición. Pág. 388. Editorial Limusa, SA de C.V. Grupo Noriega Editores. Balderas 95, Mexico, D.F.

12. Kaushik & Singhal, Cogent Engineering (2018), 5: 1467196. https://doi.org/10.1080/23311916.2018.1467196

EPÍLOGO

A modo de advertencia

Años atrás yo terminaba el curso con un chiste tomado de una revista de habla inglesa utilizado como ejemplo de advertencia sobre lo que hay que tener en cuenta cuando se trabaja en quimiometría. Se aplican una variedad de técnicas que dan resultados numéricos o gráficos a los problemas que se quieren resolver.

Pero nunca hay que olvidar comparar esos resultados con una interpretación coherente con el sistema bajo estudio. Observe la figura y continúe con el ejemplo.

¡Es asombroso!. De acuerdo a los resultados de las pruebas, los efectos colaterales de la nueva droga son más benéficos que el valor terapéutico en lo que fue propuesta.

Unos pocos años después la realidad superó a la ficción. La historia es que se estaba buscando una droga efectiva contra el colesterol; en las primeras pruebas clínicas resultaba que la droga no era tan efectiva. Por lo tanto, se continuó aumentando las dosis para intentar mejores resultados. Pero llegó un momento en que los efectos secundarios comenzaron a manifestarse en forma sobresaliente, sin mejorar el objetivo buscado. La conclusión final fue que el efecto secundario era más importante que lo buscado y pasó a ser el principal. La droga en cuestión era Sildenafil, que comercialmente se conoció después mundialmente como Viagra.

CUARTA PARTE

Prácticas

Indicaciones para proceder con la parte práctica

Los enunciados de las prácticas aparecen editados a continuación. Tanto las prácticas resueltas como las herramientas (programas y datos) para realizar las prácticas se encuentran en el siguiente link:

https://aargentinapciencias.org/introduccion-a-la-quimiometria-o-infometria/

Lo más recomendable para aquellos que quieran entrenarse en quimiometría, es que traten de resolver las prácticas sólo con la ayuda de la teoría. Luego, pueden contrastar los resultados con los de la práctica resuelta. Es necesario contar con los programas comerciales Word®, Excel® y Matlab® para desarrollarlas. También se incluye un archivo con las herramientas comprimidas que puede descomprimirse con Winrar® u otro semejante.

PRÁCTICA 1

Lenguajes de Programación, Escalado de Datos y Varianza Covarianza

Se ha explicado en el prólogo del libro que para la comprensión de los temas y más aún para la ejecución de las prácticas es imprescindible tener conocimientos de Windows®, Office®, Matlab®, estadística básica, operaciones de álgebra lineal y lenguajes de programación.

Lenguajes de Programación

Existen una variedad de lenguajes de programación imprescindibles para ejecutar comandos, programas y programar cálculos matemáticos. Los más utilizados actualmente son Fortran©, Matlab® y Python®.

Utilizaremos aquí el lenguaje Matlab®, que resulta, por diversas razones, muy práctico para nuestros objetivos. Entre los más importantes es que hay una gran cantidad de programas ya desarrollados en este lenguaje.

Quienes no se han iniciado en este lenguaje pueden hacerlo bajando de Internet alguno de los manuales o libros, pero para las aplicaciones de este libro no se requiere la enorme cantidad de recursos que Matlab® tiene, bastará con un manual sencillo. Se indican 2 en las referencias, la primera es más sencilla que la segunda. Es suficiente avanzar hasta donde comienza la enseñanza de programación, sin incluirla. Varios comandos sencillos de Matlab se indicarán en las prácticas para su fácil acceso. <nombre>

Escalado de Datos

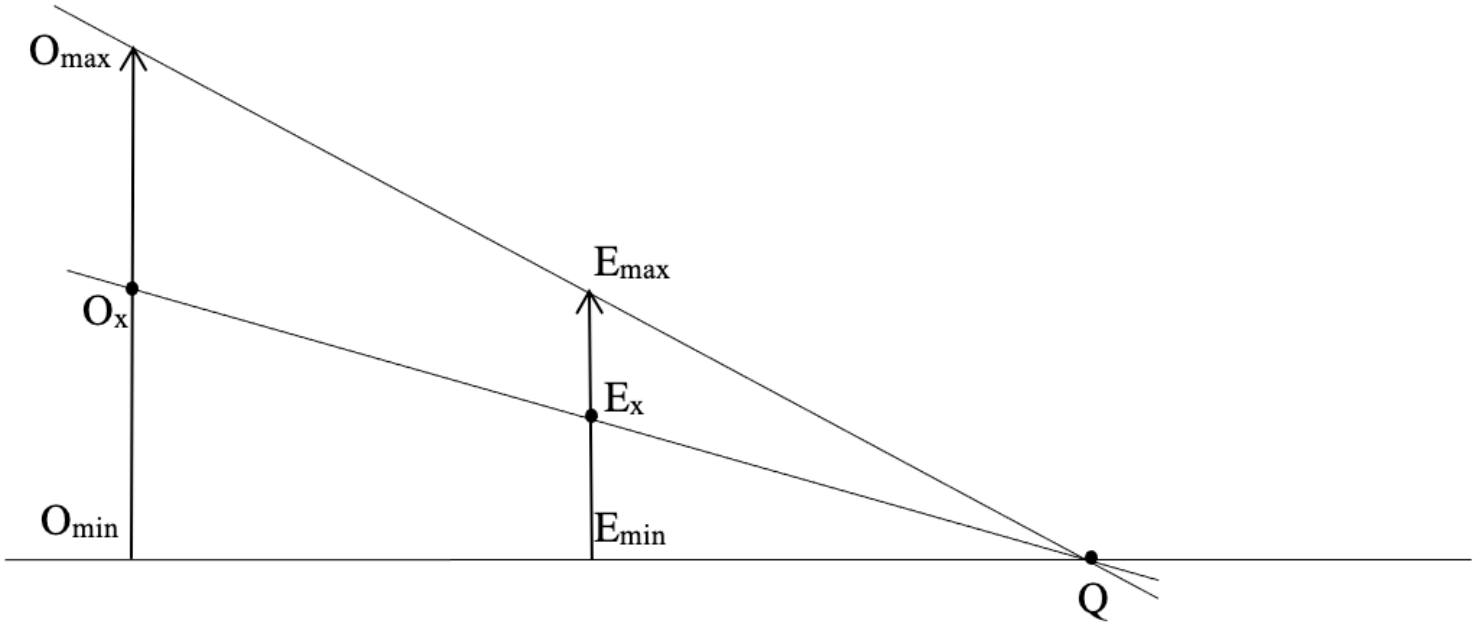
Por razones prácticas, el escalado y des escalado se aplica casi a todos los problemas y técnicas que se tratan en quimiometría, por lo tanto, es fundamental comprender y manejar esta práctica básica, correctamente.

La razón que asiste a esta operación es que los vectores objeto de cálculos, en la gran mayoría de los casos están compuestos por parámetros de distintas unidades, por ejemplo: temperatura, presión, altura, velocidad, etc. Sus valores numéricos, por lo tanto, pueden diferir mucho en magnitud, aún incluso dentro de un mismo tipo de variable, por ejemplo, un problema puede estar tratando con un parámetro de distancias en metros y otro en milímetros.

En las operaciones que se llevan a cabo en los cálculos puede ocurrir que aquellos parámetros con números grandes ‘pesen’ más que aquellos con números pequeños, desbalanceando así la importancia que puede tener cada parámetro.

El escalado entonces lleva a cabo la tarea de igualar (o a veces destacar) el ‘peso’ de todos los parámetros.

No cualquier escalado de las variables es posible, la escala original de los datos y cualquier otra escala de preferencia **deben ser proporcionales entre sí, las distancias relativas entre los puntos deben conservarse**. De lo contrario se introducirán valores que no ‘representan’ a los datos originales. Para asegurarnos esta condición, veamos cómo proceder.



Si queremos ubicar los puntos de la escala O en la escala E, consideremos los triángulos rectángulos Omax-Omin-Q; Emax-Emin-Q, Ox-Omin-Q y Ex-Emin-Q. Entre ellos todos sus elementos son proporcionales, por lo tanto, podemos establecer la siguiente relación.

$$\frac{O_x^o - O_{min}^o}{O_{max-}^o O_{min}^o} = \frac{E_x^e - E_{min}^e}{E_{max-}^e E_{min}^e} \qquad [1]$$

[0] Significa dato original y [e] significa dato escalado.

La fórmula puede usarse en forma directa y luego inversa para escalar y des escalar datos de un vector a otro.

En Matlab, cuando quiera tener presente más de un gráfico sin que se superponga uno nuevo al anterior, tipée *figure* antes de iniciar un nuevo gráfico.

Ejercicio 1: Hacer un gráfico tridimensional de los objetos con la matriz Elicoide.xls (Excel).

Para ello utilice MATLAB®. Copie cada columna de la matriz Elicoide (variables x, y, z) en forma de vectores en Matlab y rotúlelos con distintos nombres, por ejemplo Vx, Vy, Vz. Luego ejecute los siguientes comandos en Matlab en forma sucesiva:

plot3(**Vx,Vy,Vz**);

Observe que los comandos en Matlab, tales como 'plot3' deben escribirse **siempre en letras minúsculas**

title('Original');

xlabel('X');

ylabel('Y');

zlabel('Z');

grid on

El gráfico aparecerá en pantalla, puede girarlo a voluntad con el botón izquierdo del ratón después de hacer 'click' sobre la ventanita *rotate 3D* [↻]. Imprima el gráfico y no lo cierre, solo minimícelo.

Ejercicio 2. Haga un **escalado en columnas entre 0 y 1** (mínimax) de la matriz original y vuelva a obtener el gráfico, y calcular los parámetros de comparación. Para este escalado use la fórmula general dada [1] (despejando Ex^e) para cada columna X, Y y Z originales:

Ejercicio 3: Obtenga la matriz **"normalizada vectorialmente en columnas"** de la matriz original **Elicoide** (use preferentemente MATLAB). Repita el gráfico y calcule la covariancia y correlación de **la matriz escalada** para comparación con **la matriz original**.

Nota: vector normalizado y=norm (y,2) : Es la raíz cuadrada de la suma de cuadrados de los elementos de y.

Ejercicio 4: A) En la pantalla de Excel obtenga la matriz **CC="centrada en columnas"** de la matriz **Elicoide** y repita el gráfico con estas nuevas columnas. Titúlelo como 'Centrado en columnas' y guárdelo. Pregunta: ¿Qué diferencias y similitudes encuentra entre los 2 gráficos?

Ejercicio 5: Calcular la **matriz de Variancia-covariancia** (**VC**) y la **matriz de correlación** (**MC**) de la matriz original **Elicoide** (use MATLAB con los comandos siguientes).

Comando **cov(M)** para calcular la matriz **VC** de **M**.

Comando **corrcoef(M)** para calcular la matriz de correlación de **M**.

Aunque todavía no se ha incorporado el concepto de *distancia*, que se hará más adelante en el libro, úselo con el sentido común de distancia en este caso.

Calcule la **distancia Euclideana estandarizada** entre los primeros 5 puntos **de la matriz original** con la función Distan. Lea las instrucciones del programa con el comando *help* Distan en Matlab.

Haga los mismos cálculos para **la matriz centrada en columnas.**

Divida entre sí 'punto a punto' ambas matrices de distancia, observe el resultado.

Nota: para dividir punto a punto en una operación con matrices debe anteponer un punto antes del operador, por ejemplo, sean:

$$A = \begin{vmatrix} 1 & 2 \\ 1 & 3 \end{vmatrix} \text{ y } B = \begin{vmatrix} 1 & 0 \\ 2 & 2 \end{vmatrix}$$

A*B multiplicará la matriz A por la B con las reglas del álgebra lineal. Y el resultado será

$$C = \begin{vmatrix} 5 & 4 \\ 7 & 6 \end{vmatrix}$$

A.*B multiplicará **los elementos** de la matriz A por **los elementos** de la matriz B. y el resultado es

$$C = \begin{vmatrix} 1 & 0 \\ 2 & 6 \end{vmatrix}$$

Ejercicio 6: En la página Excel obtenga la matriz "**autoescalada"** de **Elicoide**. Vuelva a hacer el gráfico y titúlelo como 'autoescalado'. Calcule otra vez la matriz de VC y la MC y las distancias entre puntos para la matriz autoescalada.

Preguntas: ¿Qué diferencia encuentra con los gráficos anteriores? ¿Qué diferencia encuentra entre las matrices VC y MC y distancias calculadas?

Ejercicio 7- Represente en un único gráfico la variabilidad de <u>todas las variables</u> de la base Río (Excel) que es una base parcial real de datos de agua de Río.

A) Use el comando *Boxplot(M,'PARAM1',val1)* lea el comando tipeando *help boxplot* de Matlab done M es una matriz de datos. El gráfico debe permitir comparar claramente entre si el comportamiento de todas las variables.

B) Repita el gráfico con la matriz Río **autoescalada** y vea las diferencias.

Nota: puede usar el programa Autoescale para autoescalar (consulte el *help*).

FIN

Referencias

1. http://webs.ucm.es/centros/cont/descargas/documento11541.pdf

2. http://dea.unsj.edu.ar/control2/matlab%20para%20ingenieros.pdf

PRÁCTICA 2

Componentes Principales y Análisis de Factores

Problema 1: La planilla de datos Agusup es un lote de 84 muestras de aguas residuales tomadas en 6 sitios diferentes de una cuenca hídrica. Cada muestra contiene 7 variables químicas. Es un lote reducido con el fin de ser útil a la práctica. La primera columna es la del número de muestra, la segunda incluye el sitio de toma. Trataremos de ver algunas conclusiones sobre este lote, extraíbles mediante CP. Para ello obtendremos los gráficos de los PC's, eigenvalues y scores.

Pase los datos a Matlab (sin copiar las 2 primeras columnas).

a- Autoescale la matriz de datos y llámela por ejemplo **MA**. Calcule la matriz de correlación de la matriz de datos originales con el comando *corrcoef(MA)* de Matlab®.

b- Calcule los eigenvectors y eigenvalues de la matriz de correlación con el comando **[V D]=eig(<*Name*>)** o **[V D]=eigh(<*Name*>)** según la versión del programa que esté usando (tipee 'help eig o eigh' para saber cómo funciona el comando.

c- Construya una tabla mostrando, **en orden decreciente**, el % con que contribuye cada **eigenvalue** a la varianza total de la muestra.

Observe que en algunas versiones antiguas de Matlab las columnas de la matriz diagonal de eigenvalues está ordenada en valores **crecientes** de izquierda a derecha. Como **las columnas** de la matriz de **eigenvalues** se corresponden con las de los **eigenvectors,** la última columna es la de PC1, la anteúltima la de PC2, etc. Conviene ordenar ambas matrices invirtiendo el orden de las columnas. (O interpretar el resultado en el orden PC3, PC2, PC1 que no es lo acostumbrado ni lo que sigue en la práctica).

Observe que con sólo 2 PC's se recupera más del 85% de los datos, con lo que alcanza para representar el sistema.

d- Construya un gráfico con los 'pesos' o 'loadings' para PC1 vs. PC2. Use el comando Matlab:

plot(L(:,2),L(:,1),'o')

xlabel('PC2')

ylabel('PC1')

grid on

Cada **fila** de **la matriz de eigenvectors** es una variable (en el orden que se le dio en la matriz de datos). Identifique cada punto del gráfico anterior con la variable correspondiente. (habilite el comando *Plot edit toolbar* en la ventana *view* del gráfico Matlab y luego introduzca el texto con **Tbox**). Guarde este gráfico.

Observe si los resultados obtenidos con el gráfico se corresponden con la matriz de correlación de **MA** (tenga en cuenta el origen de coordenadas del gráfico).

e- Calcule los scores **S** de la matriz con la matriz de datos autoescalada **MA**. **S=MA*V**

Construya un gráfico 'biplot' de **S** para PC1-PC2 adicionando la primera columna de datos de la planilla Agusup. Para ello, una vez que realizó el gráfico de los scores (en forma similar al de los loadings), genere un vector con la numeración de las muestras (o cópielo) de la siguiente manera:
Num=1:84;
NUM=int2str(Num) este comando es para transformar los números en un vector de texto.
Ahora tipée *gnames(NUM),* entonces en el gráfico aparecerá un cursor con el que tiene que *barrer* toda la superficie del mismo. Automaticamente, los datos quedarán identificados con su correspondiente número (en este caso, el orden de la columna 1 de la planilla, o sea los sitios)

Observe si existe o no una estructura de datos (agrupamientos). En el caso que vea agrupamientos, trate de relacionarlos con la la muestra y los sitios de toma (columna 2 de la planilla).

¿Quedan los datos agrupados de acuerdo al origen de las muestras?

Problema 2: El archivo MuestP1 (Ecxel) contiene datos de muestras de aire registrados en el Parque Palermo de Buenos Aires de concentraciones de NOx, SO_2 y PM_{10}, junto con datos meteorológicos y horarios. Excluya los datos horarios, deje sólo las variables químicas y meteorológicas (5 en total). **Se pregunta ¿Existe una relación entre la dirección del viento y el nivel de concentración de SO_2 en aire?**

Pase los datos a Matlab con algún nombre, por ejemplo **M**. Autoescale los datos y calcule los componentes principales de la matriz autoescalada con el comando

[pc, zscores, pcvars] = princomp(M), o pca(M) en lugar de princomp, en las versiones más nuevas de Matlab. Luego represente los scores con los siguientes comandos:

figure

scatter(zscores(:,1),zscores(:,2));

xlabel('Pc1');

ylabel('Pc2');

Nota: Para esto adjunte a la matriz zscores las columnas SO2 y Dir. Viento, dele un nuevo *nombre.* Puede hacerlo con comandos Matlab o en Excell.

En Matlab C=[A B] adjunta las matrices o vectores A y B para formar la matriz C.

Ordene los resultados para SO_2 de menor a mayor con el comando *sortrows(nombre,COL) COL*=columna de SO2 **(lea con cuidado el help de Matlab)** y clasifíquelos en 3 rangos, luego grafique los scores PC1 vs. PC2 con el nuevo orden mediante los comandos Matlab: plot o scatter como en el ejercicio anterior (ayúdese con el help de Matlab si no los domina). Luego repita la operación para la dirección del viento.

Con los gráficos de scores para SO2 y Dir.Viento conteste la pregunta del problema.

Problema 3: El archivo FA-CdCl (Ecxel) contiene una matriz de datos de n filas x m columnas. Las n filas representan 22 potenciales fijos dentro de un rango de barrido de polarografía de pulso diferencial. Las m columnas representan 19 soluciones con una concentración constante de un catión (Cd^{++}) y concentraciones variables de Cl^{-}, manteniendo constante la fuerza iónica. Se pretende averiguar **qué número de especies complejas se forman** entre los iones cadmio y cloruro utilizando Análisis de Factores. Para ello transfiera la matriz de datos a Matlab y asígnele un nombre, por ejemplo, **X,** luego ejecute el siguiente comando para hacer el análisis de factores:

```
LAMBDA = factoran(X, 6)
```

Esto significa que estamos dando la posibilidad de que se formen hasta 6 especies de complejos. Observe las columnas de LAMBDA, que muestran los valores de probabilidad (entre 0 y 1) para que esa columna (factor) sea significativo (más cercanos a 1). Aquellas columnas que no contengan valores altos de probabilidad, no constituyen factores probables.

Según este análisis ¿Qué número de especies se forman entre Cd^{++} y Cl^{-}?

Con las columnas significativas, haga un gráfico en función de las 19 soluciones (eje X) y obtendrá la distribución de especies según la concentración de ligando, identifique cada especie en el gráfico.

La parte teórica de este problema figura en la Bibliografía del curso para quién le interese (no es necesario para la práctica).

Fin

PRÁCTICA 3

Clusters

Comandos Matlab para crear gráficos de dendrogramas jerárquicos

Dada una matriz $\mathbf{X}_{nxp}$ de n observaciones y p variables, obtenga el vector Y=pdist(X, 'distance') (ayúdese con help *pdist* para seleccionar 'distance'). Luego use **Z**=linkage(**Y,<Method>**) y dendrogram(**Z,P**) para obtener el dendrograma.

Ejemplo: dendrogram(Z,0,'Labels',st), (P=0) construye el dendrograma para más de 30 ramas e identifica cada objeto con el vector 'st' que es un string con el ID de los objetos.

Ayúdese con el 'help *función*' para saber que hacen estas funciones

En Matlab, cuando quiera tener presente más de un gráfico sin que se superponga uno nuevo al anterior, tipée *figure* antes de iniciar un nuevo gráfico.

Ejercicio Nº 1: El archivo **Calidad de aguas** (Excel) contiene mediciones fisicoquímicas de muestras de aguas superficiales tomadas en pozos de muestreo en la zona de interés hidrológico. Queremos averiguar si las características de cada pozo permite diferenciarlos entre sí.

1. Copie la base de datos en Matlab (sin las columnas de parámetros e ID) y atribúyale un *nombre*, por ejemplo M
2. Trasponga la matriz para que las variables sean los pozos, la llamamos MT.
3. Obtenga el gráfico de un dendrograma jerárquico con distancias euclideanas y 'average' linkage.

   Ayuda: siga las instrucciones del encabezado de la práctica.

   Observe si el cluster clasifica los pozos.

4. Vuelva a calcular el cluster, esta vez con k-means. Vea el *help kmeans* para interpretar los resultados del siguiente comando Matlab *[IDX, C, SUMD, D] = kmeans (Nombre, 4)*. Aplíquelo.

5. Obtenga el gráfico silueta con el comando *silhouette,* S = silhouette(X, CLUST,'distance') donde distance= Euclidean. **Vea el *help silhouette* e interprete el gráfico.**

6. Observe si ambos métodos de clustering son coincidentes o no.

Ejercicio Nº 2: Retomamos el problema 1 de la práctica 2 (Planilla Agusup). Trataremos de clasificar las aguas superficiales respecto de la contaminación 'alta', 'media' y baja' para los niveles de Uranio. Como hemos visto en el problema 1 de la práctica 2, hay un grupo de objetos agrupados que tienen bajo contenido de U y no interesan ahora. En la planilla Excell filtrar todos los datos con contenido de U<1 y eliminarlos. Queda una matriz de 38x7 (sin la columna 'Muestra' y 'Sitio'). La columna Sitio debe convertirse a un string con el comando **int2str(Sitio)** para usarlo en el dendrograma.

Luego obtener el dendrogama con distancias Euclideanas, likage 'Average' y el gráfico dendrogram. Observe si se obtiene la clasificación buscada.

Ejercicio Nº3: Retomaremos el problema de las "Flores de Iris" tratado en la teoría mediante componentes principales. El archivo es "fishers-irises Data". Trate de clasificar mediante clusters jerárquicos con distancia 'euclidean' primero y luego repítalo con distancia 'cosine' en la sentencia pdist. En ambos casos utilice Linkage 'Average'. Siga las instrucciones similares a las del ejercicio 2. Observe cuál de las 2 es la clasificación correcta y explique por qué.

Ejercicio Nº 4: El archivo "clasificación de aceros" contiene las intensidades relativas de 11 elementos componentes de distintos tipos de aceros. Estas intensidades provienen de espectros de FRx dispersivo en energía y están expresadas como **cuentas del elemento x/cuentas de Fe.** Téngase en cuenta que los datos son cuentas relativas y no concentraciones. Queremos ver si se pueden clasificar aceros a través de un modelo de clusters evitando el cálculo de concentraciones y su posterior búsqueda en una base de datos y reemplazando este procedimiento habitual por otro automático. El archivo contiene 3 familias diferentes de estándares de aceros (ver la columna Tipo), la serie 836 a 841 es de aceros rápidos, la serie 845 a 850 es de aceros inoxidables y las series 1161 y 1167 corresponde a aceros dulces. Hay 19 aceros en total y sobre cada uno se han hecho 5 mediciones componiendo una matriz de datos de 95 x11. Vea si puede clasificar las distintas familias con todos sus integrantes.

Ejercicio Nº 5: El archivo “Mahalanobis” contiene una tabla de datos medioambientales. La columna TSP contiene concentraciones medidas ($\mu g/m^3$) de material particulado suspendido total. La otra columna contiene la frecuencia de ocurrencia diaria de vientos del norte y noroeste (como %). Haga un gráfico X-Y de estos datos e inspeccione si observa clusters. Observe que además de un claro cluster lineal, hay otro que consta de un ‘nucleo’. También hay un punto aparentemente tan alejado de un cluster como del otro. Para definir en cuál cluster ubicarlo utilice las distancias de Mahalanobis.

FIN

PRÁCTICA 4

Calibración Univariada y Multivariada

Nota: Cuando utilice las funciones Matlab dadas para la práctica, asegúrese de que Matlab esté dirigido a la dirección donde se copiaron. Un modo sencillo de asegurarse esto es incluir todas las funciones dentro de la carpeta *work* de Matlab.

Problema 1: Con los datos de "Tabla de datos para el capítulo 4" obtenga una recta de calibración para concentración de Carbazepina a 245 nm. Utilice los datos *Absorbancias de muestras para Calibración.* Haga el cálculo con el agregado de un término independiente para la ordenada en el origen de coordenadas. Luego estime el error sobre las concentraciones calibrantes y el error sobre un lote de predicción utilizando las concentraciones de Carbazepina de la tabla con *mezclas para predicción*. Compare estos resultados con los de Bencidina, dados en la teoría, y obtenga conclusiones.

Ayuda: guíese por el cálculo de la clase teórica para Bencidina. Utilice alguno de los dos métodos: clásico o inverso a su elección.

Problema 2: Con los datos de "Tabla de datos para el capítulo 4" calcule los errores **de calibración** de **Iodobenceno, p-Aminotolueno y Bencidina** mediante el método de la regresión lineal múltiple con 3 sensores: λ =225, λ =235 y λ =290.

Calcule los errores absolutos y relativos, compárelos con los obtenidos para carbazepina en el problema 1. Luego calcule los errores de predicción con la tabla de absorbancias para este fin. Saque conclusiones.

Ayuda: Utilice el método inverso sin necesidad de utilizar datos centrados para el cálculo de b, de la teoría en *Regresión Lineal Múltiple: La Ventaja de la Multidetección*.

Problema 3: Mediante Regresión por Componentes Principales (PCR) calcule los espectros de todos **los componentes puros** y haga un gráfico de éstos a partir de las 16 mezclas de los calibrantes de la "Tabla de datos para el capítulo 4". Haga un gráfico de los espectros individuales. Calcule las concentraciones estimadas de todos los componentes en las mezclas utilizando los 5 factores y compare los resultados (errores) con los del problema 2.

Ayuda: Después de calcular la SVD reduzca la matriz Σ a la dimensión 5x5. Siga el procedimiento de la teoría.

Problema 4: Estime otra vez las concentraciones de todos los calibrantes de la "Tabla de datos para el capítulo 4" pero utilizando ahora pls1BETA de la carpeta de datos, también con 5 factores. Compare los resultados (errores) con los de los problemas 2 y 3. Observe el % de variancia acumulada y estime si se pueden usar menos de 5 factores.

Ayuda: Utilice el programa pls1BETA, que debe estar cargado en el workspace de Matlab. Recuerde que debe calcular los componentes de a uno y no todos a la vez.

Problema 5: Los datos de este problema provienen de la Ref. 1 de la teoría y son utilizados con el permiso del Profesor R. G. Brereton. Lea cuidadosamente la información que contiene la Tabla 3 (de la clase teórica y revisión de la tabla). Estime la concentración de 3-hidroxipiridina utilizando PLS1 trilineal. Haga el cálculo utilizando primero 5 y luego 10 factores.

Ayuda: utilice el programa trilPLS1 en Matlab.

Fin

PRÁCTICA 5

Redes Neuronales tipo Kohonen

En algunos de los problemas presentados previamente se aplicaron métodos que fallaron para la clasificación de los objetos. En esta práctica se dan los resultados teniendo en cuenta que éstos varían según los parámetros que se utilicen en el programa (es decir el resultado no es determinista). Sin embargo, en todos los casos, la clasificación obtenida con Kohonen es correcta y supera en mucho las de los métodos anteriores. Téngase en cuenta que este es un método no lineal.

Analice cuidadosamente sus respuestas en los archivos de salida, tanto de la etapa de training como la de predicción.

Los programas a los cuales se refiere la práctica pueden tener el nombre mencionado o el mismo con terminación VC ej. trainkohonen11 o trainkohonen11VC. La terminación VC indica que esta versión está compilada, o sea que una vez cargada en Matlab® arranca automáticamente con un click.

Problema 1: En este ejercicio se intentará clasificar aceros con los mismos datos de la práctica anterior, pero utilizando una red neuronal tipo Kohonen.

1- Cree un directorio (o carpeta) exclusivo para esta práctica, copie dentro de él todos los archivos de la práctica.

Utilizaremos los programas trainkohonen11, trainkohonen10 y predkohonen10. Lea el help de los programas para saber cómo ejecutarlos. Utilice los 2 archivos: trainingAceros.txt y predictAceros.txt que se construyeron a partir del archivo “Clasificaciín de aceros”. Comience con el programa Trainkohonen11.

2- Compare la clasificación obtenida con este método y la obtenida por ‘clustering’. Intente dar una explicación a los diferentes resultados obtenidos.

Directivas: observe los archivos *Training.txt* y *Predict.txt*. Para que el programa los ejecute correctamente, éstos tienen 2 columnas de números enteros, adicionales a las de las variables. La primera que sigue a las variables identifica a la *clase* a la que pertenece el objeto (importante cuando se utiliza el método como supervisado). Cada objeto de la columna (que aparecerá en top map) se identifica con un *símbolo* que se representa con un número de código ASCII.

La última columna es la identificación del objeto con un número, que debe ser único y no puede repetirse en la columna.

Estos datos deben estar cargados en el workspace de Matlab para ejecutar el programa.

Problema 2: Se repite el enunciado del problema sobre Aguas superficiales planteado en la práctica nº2 y también la plantilla de datos.

La planilla de datos Agusup es un lote de 84 muestras de aguas residuales tomadas en 6 sitios diferentes de una cuenca hídrica. Cada muestra contiene 7 variables químicas. Es un lote reducido con el fin de ser útil a la práctica. La primera columna es el sitio de toma de muestra, la segunda incluye la campaña de toma.

Intentaremos ahora clasificar las muestras con los 7 parámetros experimentales utilizando una red Kohonen. La intención es clasificar estas aguas con un grado de contaminación 'alto, 'medio' y 'bajo'. Observe que se escalaron los datos con el programa *Minimax01* y trabajaremos con el archivo de TrainAguas.txt para **clasificar** sin hacer predicciones. En estos problemas no suele haber un número suficiente de datos para la etapa de predicción, debido a que las campañas para muestreo son complicadas y costosas. Coloque el archivo de *training* TrainAguas.txt en la carpeta work de Matlab.

Corra el programa trainkohonen10.m y Compare los resultados con los de la práctica 2.

Problema 3: El archivo 'Data_base_50_missing_data 26 variables' es una colección de datos de aguas superficiales que contiene 270 registros de 26 variables. En ella, los datos con valores -9999 son datos faltantes codificados para estimarlos.

El archivo 'MatSinHuecos' son los vectores completos de la matriz anterior. Y el archivo 'LosHuecos' son los vectores que contienen los huecos. Las 2 últimas columnas de estos archivos son el código ASCII de la clase de objeto y el código de identificación del mismo.

Corra el programa trainkohonen11 (versión compilada) para entrenar la red con la matriz MatSinHuecos. Unavez entrenada, prediga el valor de los huecos con la matriz LosHuecos. Compare los valores de los datos faltantes con los de la primera matriz Data_base_50_missing_data 26 variables.

Fin

PRÁCTICA 6

Modelado y optimización mediante ANNs

Esta práctica se orienta a demostrar que con una fracción de los datos totales de un archivo es posible representar un lote mucho mayor de datos simulados dentro del mismo espacio multidimensional y dentro del mismo rango. En este caso tanto por comprobación visual como por precisión del resultado.

1- Abra el archivo **OpyMod** (Excel). Notará que hay 500 objetos en 3 columnas, las dos primeras son parámetros experimentales y la tercera es la señal de Respuesta.

2- No haga, por ahora, un gráfico tridimensional de la figura completa. Observe el comportamiento de las variables calculando la matriz de correlación o mejor aún haciendo una matriz gráfica de correlación entre las variables, tal cual figura en la **teórica de correlación** (¿hay algún modelo posible a simple vista?).

3- Investigue las características de las variables y juzgue si es necesario un escalado de alguna/s de ellas para poder trabajar con una red 'back propagation' con funciones de transferencia sigmoideas.

4- Trataremos de modelar y buscar el punto óptimo de trabajo (un máximo) sobre la superficie de respuesta.

5- Cuando el lote de datos esté listo, divida el archivo en dos partes, aproximadamente 20% y 80% (el primero para **predicciones durante el training** y el otro para entrenamiento) con extensión **.txt** (sin encabezamiento de variables). ¿Dará lo mismo dividir el archivo de cualquier manera? Lo veremos en la tercera parte del libro.

6- Lea el help de Backprop. Corra el programa **Backprop** (Matlab®) y entrene la red con distintas arquitecturas y parámetros hasta encontrar un error de predicción aceptable. Recién cuando esté conforme con el resultado reserve el archivo de salida (NetData), que se guarda automáticamente, usualmente ubicado en la carpeta bin de Mtlab®, para hacer predicciones futuras. Tenga en cuenta que si hace otra nueva ejecución del programa el archivo NetData será el de la última ejecución, por lo tanto si quiere reservar una salida cualquiera antes de volver a ejecutar, cámbiele el nombre, pero no la extensión.

7- Arme un archivo de datos **nuevo** con **variables de entrada** v1 y v2 en el mismo rango que usó para el entrenamiento (simule unos 700-1000 objetos).

Haga el *load* de NetData.mat. Luego haga **una predicción** de la respuesta con la ANN usando el mismo programa (pero con la opción **predecir solamente**) para el nuevo archivo de datos.

8- Desescale la/las variables/s, que hubiera escalado antes, a su valor original. **Esto no es estrictamente necesario.** Pueden hacerse las comparaciones con los datos escalados.

9- Haga un gráfico tridimensional de los datos originales y de los datos obtenidos por predicción (ambos desescalados o no). Use el programa *plotNonuniformSurf* (vea el help).

10- Finalmente busque la posición del máximo para ambos archivos. ¿Qué conclusiones saca del ejercicio?

FIN

PRÁCTICA 7

Ejercicios de Diseño Factorial Total (FFD) de dos niveles

En el primer problema se trata de comprobar la utilidad del diseño de experimentos cuando todas las variables (o algunas) son de tipo cualitativo, o sea, no tienen valores numéricos.

Problema Nro. 1. Para estudiar la influencia de tres variables sobre la preparación de una pintura se hicieron las siguientes experiencias: Se prepararon 8 ensayos de acuerdo a un diseño factorial 2^3. Los tres factores eran: Utilizar monómero 'MA' o 'MB, utilizar con acelerador de polimerización 'CA', o sin 'SA' y la tercera, utilizar carga colorante blanca 'CC' o no 'SC'. Se quieren conocer 3 respuestas: a) La pintura 'seca' antes de ½ hora o después, b) La pintura se adhiere sobre cemento o no, c) Se adhiere sobre material plástico o no.

Arme la la planilla de diseño y sus respuestas con el archivo excel "Ej_1_FF", obtenga los resultados utilizando al menos 2 métodos de evaluación del diseño y conteste las siguientes preguntas:

1- ¿Cómo Influyen las variables en la rapidez de secado de la pintura, en la adherencia al cemento y al plástico?

2- ¿Qué relación hay **entre los niveles** de las variables y **el efecto que producen**?

3- ¿Existen asociaciones entre las variables?

4- ¿Qué información se hubiera perdido indefectiblemente si en lugar de usar este diseño experimental se hubiese utilizado el método OVAT?.

Ayuda: para hacer los cálculos, otorgue a las respuestas valores numéricos sencillos tales como si = 1 y no = -1.

Problema Nro. 2. Un diseño fue llevado a cabo para estudiar la calidad de un producto industrial de panadería. Se seleccionaron cuatro factores a dos niveles (2^4) y se hizo un estudio con 2 replicados. Factores: A=Temperatura de cocción, B=Tiempo de cocción, C=Cantidad de levadura y D=relación Harina/Agua.

Organice la tabla del archivo “Ejercicio 2 FFD” y calcule los efectos principales y las interacciones, además, haga un esquema gráfico de los resultados. Luego estime el nivel de significación por el método de las desviaciones estándar usando experimentos duplicados, y también con el método del *rankit.*

Haga los gráficos de interacción (*interaction plot*) de A contra B y de B contra A y compárelos. Observe si hay efectos sinérgicos o antagónicos.

Haga un corto informe sobre las conclusiones acerca de la influencia de las variables y las interacciones.

Problema Nro. 3: Ejercicio de *“screening”* mediante Plackett-Burman

El archivo “Plackett-Burman” (Excel) contiene los datos para resolver un problema de ajuste de un instrumento, en este caso de cromatografía líquida combinada con espectrometría de masa (LC_MS) aplicado a la determinación de Npnonilfenol, cuyo funcionamiento depende de 8 variables. Se pretende hacer un modelo de optimización del instrumento, pero dado la alta cantidad de factores se hace primero un diseño de *screening* para saber cuáles son los factores más importantes.

Tome la planilla de diseño para 11 variables, y cambie los signos “+” por “1” y los signos “–” por ceros. Las variables “extra” (fantasmas), J, E y B, se reservan para el cálculo del error crítico.

Resuelva el diseño por el método Plackett-Burman de las sumas y resta de niveles + y -.

¿Cuáles variables resultan significativas para plantear entonces un modelo de cálculo? No tenga en cuenta los resultados de las variables fantasmas.

Fin de la práctica

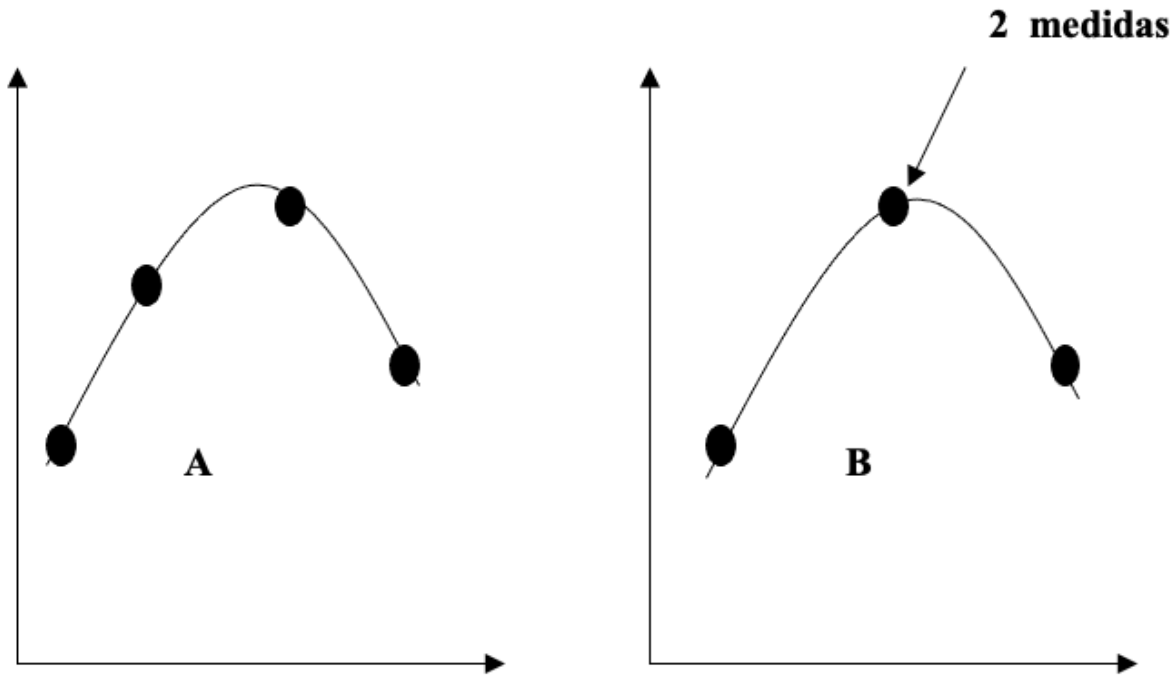
PRÁCTICA 8

Cualimetría y Quimiometría
ANOVA y Cálculo de Modelos

Problema 1

En la práctica 7, el problema 2 para estudiar la calidad de un producto industrial de panadería se resolvió por varios métodos (*Archivo: Práctica 7/práctica resuelta/ Ejercicio 2*). Sin embargo, no todos son exactamente coincidentes. Para agregar un método más de control resuelva el problema mediante el cálculo de ANOVA.

Problema 2:



Compruebe cuál de los dos diseños A, ó B es más eficiente para describir una función parabólica de dos variables. Ayuda: suponga, por ejemplo que el alcance de la abscisa es de 0 a 1. Calcule para ambos casos el D óptimo para un modelo parabólico.

Problema 3: Supongamos que queremos estimar el siguiente modelo

$$y=b_0 + b_1.x_1 + b_2.x_2 + b_{11}.x_1^2 + b_{22}.x_2^2 + b_{12}.x_1.x_2$$

Tenemos 2 factores, si lo estimamos con un diseño de 3 niveles, tendríamos la siguiente matriz **X**.

$$X = \begin{bmatrix} 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 0 & 1 & 0 & 0 \\ 1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 0 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Interprete el significado de cada columna. Calcule el M óptimo para este diseño. Luego reemplace 4 filas de la matriz por otras tantas, eligiendo valores al azar de x_1, x_2 dentro del rango 1,-1 (pero sin incluir los valores -1, 0 o 1). Del mismo modo agregue 3 experiencias más. Ahora compare los M-óptimos y los G-efficiency.

¿Qué conclusión saca de los resultados obtenidos?.

Problema 4

Abra el archivo 'Problema 4 -Datos (Excell)'.

Éstos datos fueron obtenidos de un estudio de potenciales de corrosión dependiente de la concentración de aniones y la temperatura.

El modelo que relaciona la variable de respuesta, Y=Vss, con los 4 factores x es:

$$Y=\beta_0+\beta_1.x_1+\beta_2.x_2+\beta_3.X_3+\beta_{4x4}. \qquad [1]$$

Copie los datos del archivo en Matlab y ejecute el programa modlin_multiv_V9 que figura en la práctica.

Luego genere una matriz de datos objeto (unos 50-100 datos) para las variables x dentro del rango de cada una de ellas y calcule el valor de **Y** estimado para cada objeto con la ecuación [1] y los **β significativos** calculados por el programa modlin_multiv.

Finalmente obtenga la superficie de respuesta con el programa plotNonuniformSurf para las respuestas calculadas (eje z) y las variables significativas (x e y) (sin la ordenada al origen, ya que es una constante).

FIN

PRÁCTICA 9

Cualimetría y Quimiometría
Modelos multirrespuesta

Problema 1: Se quiere ajustar la preparación de una crema dermatológica, si bien está definida la composición de los componentes terapéuticos, se quieren optimizar cuatro factores: Temperatura de la operación de mezclado (°C); % de agua; tiempo de mezcla (minutos) y velocidad de rotación del mezclador (rpm).

Se requiere comprobar la **máxima estabilidad** del preparado (respuesta R1) y la **mínima coloración** por efecto de la luz (R2). Las escalas de las respuestas están en unidades arbitrarias.

Se propone un diseño Taguchi de 4 niveles (L9 3^4) con repetición; 18 experiencias en total y un análisis del tipo Taguchi GRG.

Problema 2: Repetiremos el problema 4 de la práctica 8. Este problema tenía una contradicción en los resultados, que era la siguiente:

'Note que hay una contradicción entre los dos modos de validar el modelo: para la estadística clásica, según Rcuad, el primer modelo es el mejor (mayor R) y la validación mediante Lack of fit indica que el modelo ajustado es el correcto. Aunque en este caso se trate del mismo modelo.'

Resuelva el problema con un criterio más avanzado, el de Akaike, ejecutando el programa modlin_miltiv_V9. Observe cuál es el mejor modelo según este criterio y con cuál modo de evaluación coincide.

Fin

Introducción a la Quimiometría (o Infometría)

Para Científicos e Ingenieros

Este libro surge como resultado de muchos años de docencia en el dictado de un curso de posgrado sobre el tema. La necesidad de estos conocimientos hizo que el curso se dictara anualmente en forma regular y presencial, pero también se dictó en forma *full time* en universidades del interior del país e incluso en el exterior del país. También hubo una ocasión de dictado a distancia para ocho países sudamericanos simultáneamente. Durante toda esta experiencia he tenido alumnos, la mayoría de ellos profesionales, o próximos a recibirse, en carreras tan diversas como geología, farmacia, matemática, medicina, medioambiente, alimentación, física y biología entre otras; donde rara vez los químicos fueron mayoría.

Esto es debido a que la quimiometría tiene cierta universalidad y técnicas iguales o similares se aplican, por ejemplo, en psicometría, econometría y biometría entre otras. De allí la extensión del título a *Infometría*, que suena más apropiado para resolver problemas de interpretación y análisis de la información en muchas áreas de la ciencia e ingeniería.

ISBN 978-987-48617-2-6

Jorge Federico Magallanes