UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

ANÁLISIS DE ESPECTROS RAMAN MEDIANTE COMPONENTES PRINCIPALES

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIA

P R E S E N T A :

SOFÍA EDITH DÍAZ LÓPEZ

DIRECTOR DE TESIS:
DR. ALFREDO GÓMEZ RODRÍGUEZ
CIUDAD..DE.MÉXICO,.2018

Hoja de datos del jurado

Dr. Raúl Herrera Becerra

Mat. Margarita Elvira Chávez Cano

Dr. Alfredo Gómez Rodríguez

Dr. Fernando Brambila Paz

M. en C. Francisco Pérez Carbajal

AGRADECIMIENTOS

A mis padres, Sofía y Fernando, por brindarme siempre el apoyo en la elección de mi camino, mi madre que siempre ha estado en los momentos más difíciles a pesar de nuestras diferencias y sobre todo por el gran apoyo que me ha dado para poder culminar este trabajo, mi padre que siempre me acompaño en mi vida académica, por creer siempre en mí y brindarme su confianza.

A mi hermano, Fernando Díaz, porque siempre has estado para explicarme cuando algo no lo entiendo, por tu gran apoyo en los momentos más difíciles de la carrera y sobre todo por preocuparte siempre por mí. Eres un gran ejemplo para mí.

A mi abuelita, Hermelinda, por enseñarme a luchar y trabajar para conseguir lo que quiero, por el gran cariño que siempre me dio y sobre todo por los buenos consejos, los extraño.

A mi familia por todo su cariño, especialmente a mis primos, por sus pláticas y risas en la reunión de los lunes, hacen que inicie la semana más agradable.

A Fel por confiar en mí y darme el tiempo necesario para poder culminar este trabajo, pero sobre todo por tu cariño y tus palabras de ánimo que siempre me das.

A mis amigos de la facultad por haber hecho el transcurso de la carrera más fácil.

Al Dr. Raúl Herrera por abrirme las puertas del Instituto de Física, pero sobre todo por confiar en mí y brindarme su apoyo para realizar este proyecto.

Al Dr. Alfredo Gómez por darme la oportunidad de trabajar con él, compartirme sus conocimientos pero sobre todo por su paciencia.

A mi Alma Mater, la Universidad Nacional Autónoma de México, por proporcionarme la mayor parte de mi formación académica. Es aquí donde he conocido a las personas más brillantes tanto profesores como compañeros de clase. Así mismo por proveerme un conocimiento extra aulas e inculcarme la libertad de pensamiento y compromiso social.

ÍNDICE

ÍNDICE DE FIGURAS

RESUMEN

En esta tesis se explora la posibilidad de usar el Análisis de Componentes Principales como auxiliar en la identificación de los espectros Raman usados en laboratorios avanzados de materiales.

En la primera parte se presenta una revisión bastante extensa del método de las componentes principales haciendo énfasis en las características de reducción de la dimensión.

A continuación, se revisa muy brevemente la técnica Raman.

Posteriormente presentamos un sencillo programa MATLAB para procesar los espectros. Entre los temas discutidos se encuentran: la importación de los datos a MATLAB, la estandarización de los datos, el cálculo de la matriz de correlación, su diagonalización, el cálculo de las varianzas y la representación de los datos usando unas cuantas componentes.

En la parte final se presentan los resultados obtenidos con unos cuantos espectros y se discuten y plantean algunas posibilidades para el futuro, tales como hacer bases de datos con cantidades reales de datos. Todo el tiempo se usan espectros reales de diversas sustancias y simulaciones como en otros trabajos publicados.

El resultado principal es que el análisis de componentes principales ayudan a identificar patrones y a descubrir aquellos que, por las razones que sea, se apartan de la norma para una sustancia dada.

CAPÍTULO I

INTRODUCCIÓN

El Análisis Multivariado (AM) [1], es la parte de la estadística y del análisis de datos que estudia, analiza, representa e interpreta los datos que resultan de observar un número $p > 1$ de variables estadísticas sobre una muestra de n individuos. Las variables observables son homogéneas y correlacionadas, sin que alguna predomine sobre las demás. La información estadística en el AM es de carácter multidimensional, por lo tanto la geometría, el cálculo matricial y las distribuciones multivariadas juegan un papel fundamental en ella.

La información constituye una matriz de datos, pero a menudo, en el AM la información de entrada consiste en matrices de distancias o similaridades, que miden el grado de discrepancia entre los individuos.

El análisis de datos multivariados tiene por objeto el estudio estadístico de varias variables medidas en cada elemento de una muestra. Pretende los siguientes objetivos:

1. Reducir el conjunto de variables en unas pocas variables, construidas como transformaciones lineales de las originales, con la mínima pérdida de información.
2. Encontrar grupos en los datos si existen.
3. Relacionar los diversos conjuntos de variables.

Disponer de indicadores tiene varias ventajas: (1) si son pocos podemos representarlos gráficamente y comparar distintos conjuntos de datos o instantes en el tiempo; (2) simplifican el análisis al permitir trabajar con un número menor de variables; (3) si las variables indicadas pueden interpretarse, podemos mejorar nuestro conocimiento de la realidad estudiada.

El análisis multivariante de datos proporciona métodos objetivos para conocer cuántas variables indicadoras, que a veces se denominan factores, son necesarias para describir una realidad compleja y determinar su estructura.

El segundo objetivo es identificar grupos si existen. En muchas situaciones los grupos son desconocidos a priori y queremos disponer de un procedimiento objetivo para obtener los grupos existentes y clasificar las observaciones.

Un tercer objetivo relacionado con el anterior aparece cuando los grupos están bien definidos a priori y queremos clasificar nuevas observaciones.

Para alcanzar estos tres objetivos una herramienta importante es entender la estructura de dependencia entre las variables, ya que las relaciones entre las variables son las que permiten resumirlas en variables indicadoras, encontrar grupos no aparentes por las variables individuales o clasificar en casos complejos. Un problema distinto es relacionar dos conjuntos de variables (esto es correlación canónica).

Algunos métodos del ACP consisten en obtener e interpretar combinaciones lineales adecuadas de las variables observables. Una variable compuesta Y es una combinación lineal de las variables observables con coeficientes $a = (a_1, \dots, a_p)'$

$$Y = a_1X_1 + \cdots + a_pX_p.$$

Si $X = [X_1, \dots, X_p]$ es la matriz de datos, también podemos describir

$$Y = Xa.$$

Si $Z = b_1X_1 + \cdots + b_pX_p = Xb$ es otra variable compuesta, se verifica:

1. $\bar{Y} = \bar{X}a, \quad \bar{Z} = \bar{X}b.$
2. $var(Y) = a'Sa, \quad var(Z) = b'Sb.$
3. $cov(Y, Z) = a'Sb.$

Donde S es la matriz simétrica $p \times p$ de covarianzas muestrales.

Ciertas variables compuestas reciben diferentes nombres según la técnica del análisis multivariado: componentes principales, variables canónicas, funciones discriminantes, etc. Uno de los objetivos del análisis multivariante es encontrar variables compuestas adecuadas que expliquen aspectos relevantes de los datos.

El análisis multivariante puede plantearse a dos niveles. En el primero, el objetivo es utilizar sólo los datos disponibles y extraer la información que contienen. Los métodos encaminados a este objetivo se conocen como **métodos de exploración de datos**, los cuales son: descripción de datos multivariantes, análisis gráfico y datos atípicos, componentes principales, escalamiento multidimensional, análisis de correspondencias y análisis de conglomerados. A un nivel más avanzado, se pretende obtener conclusiones sobre la población que ha generado los datos, lo que requiere la construcción de un modelo que explique su generación y permita prever los datos futuros. En este segundo nivel hemos generado conocimiento sobre el problema que va más allá del análisis particular de los datos disponibles. Los métodos encaminados a este objetivo se conocen como **inferencia estadística**, y son: distribuciones multivariantes, inferencia con datos multivariantes, métodos de inferencia avanzada multivariante, análisis de factores, análisis discriminante, discriminación logística y otros métodos de clasificación, clasificación mediante mezclas de distribuciones y correlación canónica.

El problema de resumir o condensar la información de un conjunto de variables se aborda, desde el punto de vista descriptivo, construyendo unas nuevas variables indicadoras que sinteticen la información contenida en las variables originales. Existen distintos métodos exploratorios para conseguir este objetivo. Con variables continuas, el método más utilizado se conoce como *Componentes Principales*. Las componentes principales nos muestran las dimensiones necesarias para representar adecuadamente los datos. Con ello podemos hacer gráficos de los datos en pocas dimensiones, con mínima pérdida de información, para entender su estructura subyacente.

El análisis de componentes principales puede generalizarse en dos direcciones: la primera cuando los datos disponibles no corresponden a variables sino a similitudes o semejanzas entre elementos. La segunda generalización de componentes principales es para datos cualitativos, que se presentan en una tabla de contingencia. Esta técnica permite además cuantificar de forma objetiva atributos cualitativos.

OBJETIVO Y METAS

A la hora de analizar un espectro Raman de cualquier material o sustancia medido en un determinado laboratorio, surgen distintos obstáculos relacionados tanto con la medida de dicho espectro como con su comparación con los espectros Raman de referencia, es por eso que el principal objetivo a tratar a lo largo de este proyecto es diseñar un método de identificación automático de espectros Raman utilizando la técnica de reducción dimensional basada en el Análisis de Componentes Principales.

Debido a la complicación del proceso de identificación por comparación espectral, se propone diseñar un sistema de reconocimiento de espectros Raman de materiales, automatizando el proceso de identificación, independizándolo, de esta forma, de la subjetividad que pueda introducir el analista con base en su juicio y experiencia.

Para resolver dicho problema se va a diseñar un programa en Matlab, el cual necesita al menos 2 espectros de un mismo material, lo primero que hace el programa es producir una matriz de $n \times p$, donde n es el número de observaciones y p el número de variables (espectros), se le realizan todos los cálculos descritos más adelante a dicha matriz hasta obtener las componentes principales, las cuales al graficarlas nos proporcionan los espectros de referencia con los cuales el investigador va a hacer la comparación de los espectros Raman que esté estudiando con los obtenidos.

Una vez obtenidos dichos espectros de referencia se puede ir generando una base de datos con los distintos materiales para así poder comparar los espectros a estudiar y hacer más fácil la identificación del material presente en dicho espectro.

ESPECTROSCOPIA RAMAN

Esta técnica deriva su nombre del físico C.V. Raman, su principal utilidad es que es una de las técnicas que sirve para obtener la huella digital de diversas sustancias, es decir, permiten identificar los materiales presentes en una muestra.

La idea es iluminar el material con luz láser de un solo color o longitud de onda. Como las moléculas están vibrando pueden absorber parte de la luz láser y luego la re-emiten pero con un cambio en longitud de onda (o de color).
Un espectro típico muestra los cambio en la longitud de onda en el eje horizontal. El eje vertical nos dice qué tanta luz tuvo ese cambio. Ver la figura

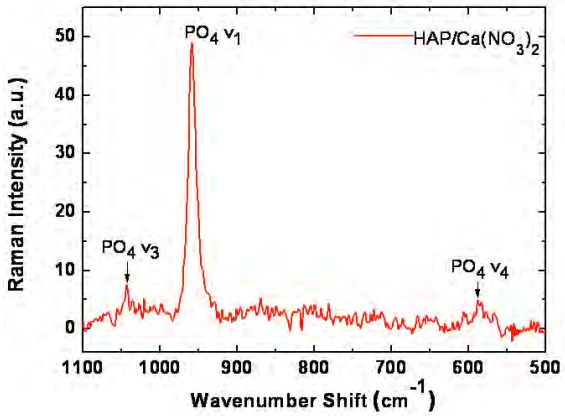

Fig. 1.1. Espectro Raman nitrato. Imagen tomada de muestras de Maricela (Compañera del Instituto de Física)

Un diagrama típico (y muy simplificado) seria

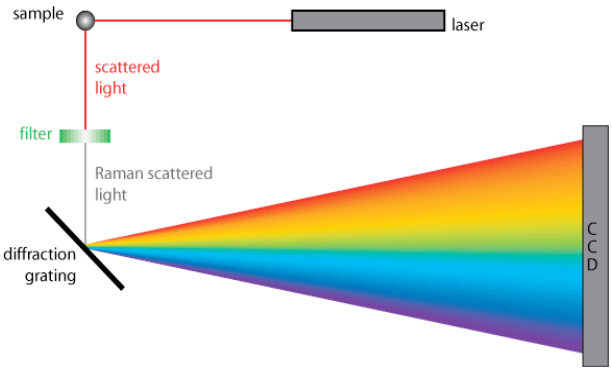

Fig. 1.2. Diagrama de la espectroscopia laser Raman. https://www.doitpoms.ac.uk/tlplib/raman/images/spectrometer_schematic.gif

En donde se ilustra, muy esquemáticamente, cómo un láser ilumina una muestra ("simple") y por algún medio (aquí mediante una rejilla de difracción) la luz que sale en la muestra se analiza en términos de sus longitudes de onda (colores).

La dispersión es la desviación de luz de su dirección original de incidencia. La interacción del vector de campo eléctrico de una onda electromagnética con los electrones del sistema con el que interactúa da lugar a la dispersión de la luz incidente. Tales interacciones inducen oscilaciones periódicas en los electrones del compuesto; por lo tanto, produce momentos eléctricos oscilantes. Esto lleva a tener nuevas fuentes emisoras de radiación, es decir, fuentes que re-emiten radiación en todas las direcciones (la luz dispersada).

Existen dos tipos básicos de dispersión:

> Elástica. Misma frecuencia (longitud de onda) que la luz incidente, llamada dispersión Rayleigh.
> No elástica. Dentro de la inelástica existen dos tipos, una que tiene frecuencia más baja (longitud de onda mayor) y, la que tiene frecuencia más alta (longitud de onda más corta) que la luz incidente.

Es a la luz dispersada no elásticamente a la que se le llama dispersión Raman y, por lo tanto, existen dos tipos de ella: en uno de ellos la luz dispersada tiene menor energía que la luz incidente (la que tiene menor frecuencia) y el efecto se llama dispersión Raman Stokes. En el otro, la luz dispersada tiene mayor energía que la luz incidente, es decir tiene mayor frecuencia que la luz incidente, y se le llama dispersión Raman anti-Stokes. En la dispersión Rayleigh (misma frecuencia) no hay cambio en la energía de la luz incidente.

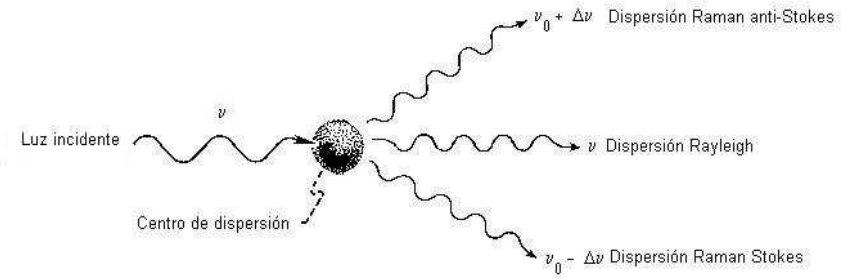

Figura 1.3. Representación esquemática de los tres tipos de luz dispersada.

La dispersión Rayleigh es la más común y los objetos se pueden ver debido a este efecto. Se ha demostrado que la eficiencia de dispersión es inversamente proporcional a la cuarta potencia de la longitud de onda. La intensidad depende de la posición desde la cual se observe este fenómeno.

Comparada con la dispersión Rayleigh, la dispersión Raman es menos común en la vida diaria; sin embargo, es importante para quien esté interesado en los estados vibracionales y rotacionales de las moléculas.

En el proceso Raman intervienen dos fotones de diferentes energías. Esta diferencia de energía es debida a un cambio de estado, rotacional o vibracional de la molécula, causado por la interacción con los fotones. En consecuencia, el análisis de los espectros Raman provee información acerca de propiedades moleculares tales como los modos y tipos de vibraciones.

La intensidad de la luz dispersada depende de los siguientes factores.

1. El tamaño de la partícula o molécula iluminada.
2. La posición de observación. La intensidad dispersada es una función del ángulo con respecto al haz incidente.
3. La frecuencia de la luz incidente.
4. La intensidad de la luz incidente.

CAPÍTULO II

ANTECEDENTES

El primer método para medir la relación estadística entre dos variables es debido a Francis Galton (1822-1911), que introduce el concepto de recta de regresión y la idea de correlación entre variables en su libro *Natural Inheritance*, publicado en 1889 cuando Galton tenía 67 años. Estos descubrimientos surgen en sus investigaciones sobre la transmisión de los rasgos hereditarios, motivadas por su interés en contrastar empíricamente la teoría de la evolución de las especies, propuesta por su primo Charles Darwin en 1859. El concepto de correlación es aplicado en las ciencias sociales por Francis Edgeworth (1845-1926), que estudia la normal multivariada y la matriz de correlación. Karl Pearson (1857-1936), un distinguido estadístico británico autor de la famosa prueba ji-cuadrada que lleva su nombre, obtuvo el estimador del coeficiente de correlación en muestras, y se enfrentó al problema de determinar si dos grupos de personas, de los que se conocen sus medidas físicas, pertenecen a la misma raza. Este problema intrigó a Harold Hotelling (1885-1973), un joven matemático y economista americano, que, atraído por la estadística, entonces una joven disciplina emergente, viaja en 1929 a la estación de investigación agrícola de Rothamsted en el Reino Unido para trabajar con el ya célebre científico y figura destacada de la estadística, R. A. Fisher (1890-1962). Hotelling se interesó por el problema de comparar tratamientos agrícolas en función de varias variables, y descubrió las semejanzas entre este problema y el planteado por Pearson. Debemos a Hotelling (1931) el contraste que lleva su nombre, que permite comparar si dos muestras de poblaciones multivariadas vienen de la misma población. A su regreso a la Universidad de Columbia en Nueva York, Truman Kelley, profesor de pedagogía en Harvard, planteó a Hotelling el problema de encontrar los factores capaces de explicar los resultados obtenidos por un grupo de personas en un test de inteligencia. Hotelling (1933) inventó los componentes principales, que son indicadores capaces de resumir de forma óptima un conjunto amplio de varias variables y que dan lugar posteriormente al análisis de factores. El problema de obtener el mejor indicador resumen de un conjunto de variables había sido abordado y resuelto desde otro punto de vista por Karl Pearson en 1921, en su trabajo para encontrar el plano de mejor ajuste a un conjunto de observaciones astronómicas. Posteriormente, Hotelling generaliza la idea de componentes principales introduciendo el análisis de correlaciones canónicas, que permiten resumir simultáneamente dos conjuntos de variables.

El problema de encontrar factores que expliquen los datos fue planteado por primera vez por Charles Spearman (1863-1945), que observó que los niños que obtenían buenas puntuaciones en un test de habilidad mental también las obtenían en otros, lo que le llevó a postular que eran debidas a un factor general de inteligencia, el factor g (Spearman, 1904). L. Thurstone (1887-1855) estudió el modelo con varios factores y escribió uno de los primeros textos de análisis de factores (Thurstone, 1947): El análisis de factores fue considerado hasta los años 60 como una técnica psicométrica con poca base estadística, hasta que los trabajos de Lawley y Maxwell (1971) establecieron formalmente la estimación y el contraste del modelo de factores bajo la hipótesis de normalidad.

La primera solución al problema de la clasificación es debida a Fisher en 1933. Fisher inventa un método general, basado en el análisis de la varianza, para resolver un problema de discriminación de cráneos en antropología. El problema era clasificar un cráneo encontrado en una excavación arqueológica como perteneciente a un homínido o no. La idea de Fisher es encontrar una variable indicadora, combinación lineal de las variables originales de las medidas del cráneo, que consiga máxima separación entre las dos poblaciones en consideración. En 1937 Fisher visita la India invitado por P. C. Mahalanobis, que había inventado la medida de distancia que lleva su nombre, para investigar las diferentes razas en la India. Fisher percibe enseguida la relación entre la medida de Mahalanobis y sus resultados en análisis discriminante y ambos consiguen unificar estas ideas y relacionarlas con los resultados de Hotelling sobre el contraste de medidas de poblaciones multivariadas.

El Análisis de Componentes Principales (ACP) fue iniciado por K. Pearson en 1901 y desarrollado por H. Hotelling en 1933. Es un método referente a una población, pero W. Krzanowski y B. Flury han investigado las componentes principales comunes a varias poblaciones.

El ACP tiene muchas aplicaciones. Una aplicación clásica es el estudio de P. Jolicoeur y J. E. Mosimann sobre tamaño y forma de animales, en términos de la primera, segunda y siguientes componentes principales. La primera componente permite ordenar los animales de más pequeños a más grandes, y la segunda permite estudiar su variabilidad en cuanto a la forma. Nótese que tamaño y forma son conceptos "independientes".

El ACP puede servir para estudiar la capacidad. Supongamos que el caparazón de una tortuga tiene longitud L, ancho A, y altura H. La capacidad sería $C = L^{\alpha}A^{\beta}H^{\gamma}$, donde α, β, γ son parámetros. Aplicando logaritmos, obtenemos

$$logC = \alpha logL + \beta logA + \gamma logH = \log(L^{\alpha}A^{\beta}H^{\gamma}),$$

que podemos interpretar como la primera componente principal Y_1 de las variables $logL, logA, logH$, y por tanto α, β, γ serían los coeficientes de Y_1. Por medio del ACP es posible efectuar una regresión múltiple de Y sobre $X_1, \dots, X_p$, considerando las primeras componentes principales $Y_1, Y_2, \dots$ como variables explicativas, y realizar regresión de Y sobre $Y_1, Y_2, \dots$, evitando así efectos de colinealidad, aunque las ultimas componentes principales también pueden influir (Cuadras, 1993). La regresión ortogonal es una variante interesante. Supongamos que se quieren relacionar las variables $X_1, \dots, X_p$ (todas con media 0), en el sentido de encontrar los coeficientes $\beta_1, \dots, \beta_p$ tales que $\beta_1 X_1 + \cdots + \beta_p X_p \cong 0$. Se puede plantear el problema como $var(\beta_1 X_1 + \cdots + \beta_p X_p) = mínima$, condicionando a $\beta_1^2 + \cdots + \beta_p^2 = 1$. Es fácil ver que la solución es la última componente principal Y_p.

METODOLOGÍA

Cuando se realiza la espectroscopia Raman de un material o sustancia, éste arroja como resultado una gráfica, la cual consta de una línea continua con múltiples máximos y mínimos en forma de picos, algunos muy pronunciados y otros más pequeños. Cuando son grandes y visibles es fácil identificarlos, pero cuando no es el caso, la identificación se complica mucho.

Para estos últimos varios son los factores que afectan su identificación, siendo el más importante el ruido (existen varios tipos de ruido) generado por el propio aparato principalmente por su detector, o por otras partes electrónicas que componen el equipo Raman.

En este trabajo se plantea una herramienta digital que busca identificar los picos característicos de un material en particular a pesar del ruido que pueda generarse al obtener la gráfica característica en este tipo de instrumentos.

Debido a que las señales que están mezcladas son la señal del material que se está analizando y el ruido del aparato, se utiliza el método de componentes principales, ya que es una herramienta estadística que permite detectar las similitudes y las diferencias entre los datos, obteniendo una nueva expresión simplificada de la gráfica respecto a la expresión inicial.

Basándonos en este método se hace un programa en Matlab, utilizando técnicas estadísticas y algunas otras de algebra lineal que van ayudar a reducir los datos, lo cual al momento de graficar nos arrojará una línea más limpia la cual será única para el material, sustancia o elemento que se esté analizando. Con el programa gráficamente podremos ver picos pronunciados en donde no los había antes, lo cual nos indicará en dónde está la sustancia o elemento presente.

Un determinado material es caracterizado por su Espectro Raman de forma única ya que éste depende de la configuración molecular de dicho material. Así, en el momento de identificar un material desconocido, su Espectro Raman permite reconocerlo de manera inequívoca. Esta identificación se lleva a cabo por comparación entre el espectro desconocido y una serie de espectros conocidos, llamados espectros patrones. De esta forma, se hace necesario disponer de una biblioteca espectral en la que previamente se hayan almacenado los espectros de referencia.

Un Espectro Raman puede ser interpretado como un vector de entre 1000 y 2000 componentes, típicamente, de manera que la biblioteca espectral puede contener dimensiones elevadas. Además, puede estar formado por un gran número de bandas, con lo que el reconocimiento del espectro desconocido por comparación visual resulta ser una tarea complicada y costosa para el analista. Por ello, el objetivo de este proyecto se basa en implementar un sistema automático que agilice el procedimiento de identificación, de forma que este sea más rápido y objetivo, y, puesto que el reconocimiento es realizado de forma automática, minimice la intervención del analista.

Una vez seleccionados los espectros constituyentes de la biblioteca espectral de referencia, se les aplica la técnica de reducción dimensional establecida por el ACP, siguiendo el procedimiento expuesto en el Capítulo III. No obstante, para llevar a cabo dicho análisis es necesario que la biblioteca inicial cumpla una serie de condiciones de homogeneidad en el formato de los datos.

Para ello, se debe asegurar que los espectros compartan un mismo margen de número de onda. Asimismo, se debe independizar al máximo la intensidad de las bandas Raman de las condiciones de medida mediante la normalización

Lo primero que necesitamos para poder utilizar el programa es al menos 2 espectros de un mismo elemento, los cuales deben tener el mismo número de datos ya que cuando se cargan al programa este produce una matriz de $n \times p$ donde n es el número de datos(variables) y p el número de espectros Raman, a dicha matriz se le empieza hacer todo el cálculo necesario para limpiar los datos, lo primero es normalizar los datos, esto se hace obteniendo los máximos y mínimos de los datos luego se hace una diferencia, después una división y así se obtiene la matriz normalizada, una vez terminado se procede a estandarizar las variables para poder realizar el cálculo de la matriz de covarianzas mediante técnicas de estadística, ya obtenida dicha matriz se trabaja en ella, primero se diagonaliza y se calculan los vectores y valores propios mediante cálculos de algebra lineal, se pueden ordenar los valores propios y se calculan las Componentes Principales, es importante decir que una vez obtenidas se selecciona el número de Componentes Principales bajo el criterio de 100% varianza acumulada. Ya obtenidas las Componentes Principales se grafican resultando curvas más fáciles de comparar.

CAPÍTULO III

TEORIA

Definición

El Análisis de Componentes Principales [3] consiste en encontrar transformaciones ortogonales de las variables originales para conseguir un nuevo conjunto de variables no correladas, denominadas Componentes Principales, que se obtienen en orden decreciente de importancia.

Un problema central en el análisis de datos multivariados es la reducción de la dimensión: si es posible describir con precisión los valores de p variables por un pequeño subconjunto $r < p$ de ellas, se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información.

El análisis de componentes principales tiene este objetivo: dadas n observaciones de p variables, se analiza si es posible representar adecuadamente esta información con un número menor de variables construidas como combinaciones lineales de las originales. Por ejemplo, con variables con alta dependencia es frecuente que en un pequeño número de nuevas variables (menos del 20% de las originales) expliquen la mayor parte (más del 80%) de la variabilidad original.

La técnica de componentes principales es debida a Hotelling (1933), aunque sus orígenes se encuentran en los ajustes ortogonales por mínimos cuadrados introducidos por K. Pearson (1901). Su utilidad es doble:

1. Permite representar óptimamente en un espacio de dimensión pequeña, observaciones de un espacio general p −dimensional. En este sentido componentes principales es el primer paso para identificar posibles variables "latentes" o no observadas, que están generando la variabilidad de los datos.

2. Permite transformar las variables originales, en general correlacionadas, en nuevas variables no correladas, facilitando la interpretación de datos.

Método de Componentes Principales

El Análisis por Componentes Principales (ACP) es un método de análisis multivariado [4], que utiliza métodos matemáticos y estadísticos, lo cual es una manera eficaz de suprimir información redundante y concentrar, en solo una o dos imágenes compuestas, la mayor parte de la información de los datos iniciales.

El ACP está entendido sobre todo como herramienta para la reducción dimensional de datos. Dado un conjunto de datos, esta técnica facilita un nuevo espacio de dimensión menor al original, conocido como espacio de los Componentes Principales (PCs del inglés Principal Components), donde los datos se representan de tal manera que se resalta la información. Se trata, así, de una herramienta estadística que permite detectar las similitudes y las

diferencias entre los datos, obteniendo una nueva expresión reducida respecto a la expresión inicial.

Esta característica convierte al ACP en una herramienta muy poderosa en los procesos de clasificación o reconocimiento de patrones, ya que los agiliza y optimiza el resultado. Así, esta técnica se ajusta y adapta a estudios donde se analizan datos espectrales, ya que cualquier tipo de espectro está generado por un gran número de variables (frecuencia, periodo, longitud de onda normalizada, etc.), de manera que pueden resultar ser señales difíciles de interpretar. Por ello, aplicar esta técnica de reducción dimensional permite obtener resultados muy satisfactorios en materia de identificación espectral.

La principal motivación a la hora de aplicar una herramienta de reducción dimensional es conseguir expresar un conjunto de datos definido por n variables como un conjunto de menor dimensión, k, con k<n, pero equivalente en cuanto a contenido informativo. Así, el ACP se encuadra en la categoría de métodos de transformación de variables, puesto que se basa en sustituir las n variables iniciales (Ui) que explican los datos, por otras n variables (ui), no correlacionadas y ordenadas de tal manera que las primeras tienen más relevancia que las últimas.

El principal objetivo del ACP es obtener un espacio k-dimensional, siendo éste un Espacio Vectorial Euclídeano, donde el conjunto inicial de datos se expresa conservando la estructura y la información inicial. Normalmente existe un sistema de coordenadas óptimo para representar cada conjunto de datos, considerando como óptimo aquel donde se puedan diferenciar claramente todas las observaciones. Este sistema de coordenadas es precisamente el que se obtiene mediante la transformación implementada por el ACP. El nuevo sistema de coordenadas es de dimensión igual al inicial, n, pero su configuración permite llevar a cabo a posteriori la reducción de n a k dimensiones.

Ya que la transformación, o cambio de base, asegura obtener un espacio ortogonal, la variabilidad total de los datos se mantiene en ambos espacios (tanto en el original como en el transformado). Esta transformación permite obtener unas nuevas variables generadas de tal manera que las primeras ($i = 1, \dots \dots, k$) tienen más relevancia, es decir, mayor contenido informativo, que las últimas ($i = k + 1, \dots \dots, n$). Esta propiedad permite seleccionar un valor k, de estas nuevas variables, siendo $k < n$, en función de la cantidad de información que se quiera ignorar, de forma que finalmente se consigue convertir un conjunto de datos definido por n variables, Ui, a uno descrito por un valor inferior k, de variables, ui. El ACP se comporta, como una técnica de selección de variables, puesto que se selecciona un subconjunto de las nuevas variables para la representación de los datos iniciales. Esta reducción debe asegurar la mínima pérdida de información y aportar la menor distorsión posible.

Las Componentes Principales, u_i, resultan de la combinación lineal de las variables originales, U_i, tal y como se muestra a continuación:

$$u_i = c_{1i}U_1 + \cdots + c_{ji}U_j + \cdots + c_{ni}U_n$$

donde los coeficientes c_{ji}, conocidas como loads, representan la aportación o peso de la variable U_j a la variable u_i. Es importante destacar que las variables u_i son las variables, no correlacionadas entre sí, que eliminan la redundancia que pueda existir entre las variables iniciales U_j. A partir de los coeficientes se obtiene una matriz de transformación:

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix}$$

donde c_{ji} es el peso de la variable U_j respecto a la variable u_i, con $i = 1, \dots, n$ y $j = 1, \dots, n$. Esta matriz cuadrada explica la relación lineal entre las nuevas variables, PC_i, y las antiguas, los números de onda normalizados. En concreto, cada una de sus columnas define las coordenadas de una Componente Principal en el espacio de las variables u_i.

La ventaja de este nuevo sistema de coordenadas es que se conoce la porción de la varianza total (v_{total}) que aporta cada uno de sus ejes (PC_i), y que estos están ordenados en función de esta aportación, de mayor a menor. También resulta posible valorar qué porcentaje de v_{total} se tiene en cuenta si sólo se consideran k Componentes Principales para definir los datos. De esta forma, se puede calcular la varianza acumulada cada vez que se tiene en cuenta un PC de más.

Es en el momento de escoger el valor k de Componentes Principales a contemplar cuando se produce la reducción dimensional, es decir, cuando se decide cuantos PCs son suficientes para representar los datos de manera fiable. Cuando k Componentes Principales concentran gran parte o la totalidad de la varianza original (v_{total}), los datos tratados se pueden expresar en el espacio definido por estas k nuevas variables sin que la pérdida de información sea crítica. De esta manera se considera diferenciar los datos y se consigue una reducción dimensional muy considerable. Una vez seleccionado el número de PCs a contemplar se obtiene la matriz reducida de cambio de base:

$$C_{red} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{21} & c_{22} & \cdots & c_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nk} \end{pmatrix}$$

Así, se considera un conjunto inicial de referencia de P elementos, contemplado, por ejemplo, el criterio de concentrar el 100% de la varianza de las variables originales, se obtienen k=P-1 Componentes Principales.

Una vez obtenida la matriz reducida de transformación, el conjunto inicial puede ser expresado en el nuevo espacio generado por los PCs aplicando el cambio de base definido por C sobre la matriz que recoge las variables originales, siendo ésta X:

$$S = X \cdot C$$

donde cada fila de S (S_i) corresponde a una de las observaciones iniciales expresada en el espacio de los Componentes Principales. Puesto que el objetivo a alcanzar es el de la reducción de dimensionalidad, la proyección de las variables originales en el nuevo espacio definido por los PCs se realiza mediante:

$$S_{red} = X \cdot C_{red}$$

de donde resulta, suponiendo un conjunto inicial formado por P observaciones:

$$S_{red} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1k} \\ s_{21} & s_{22} & \cdots & s_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{P1} & s_{P2} & \cdots & s_{Pk} \end{pmatrix}$$

Esta matriz, por filas, permite conocer cuáles son las coordenadas (conocidas como scores) de cada una de las observaciones en el espacio generado por las nuevas coordenadas. Es decir, la matriz S es equivalente a la matriz X pero definida en el espacio de PCs.

Como paso previo a la proyección del conjunto inicial en el espacio definido por los PCs, las variables originales deben ser pretratadas, existiendo dos posibilidades: centrar las variables iniciales o bien estandarizarlas (o tipificarlas). La elección de realizar un centrado o una estandarización de variables genera dos tipos de ACPs: el ACP no normado y el ACP normado, respectivamente.

Sea X la matriz inicial de las observaciones, consideradas por filas:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{P1} & x_{P2} & \cdots & x_{Pn} \end{pmatrix}$$

El centrado se basa en la sustracción del valor medio de las variables originales (o columnas de la matriz X):

$$x_{cij} = x_{ij} - \bar{x}_j$$

donde, para un espectro de n variables:

$$\bar{x}_j = \frac{1}{P}\sum_{i=1}^{P} x_{ij}$$

obteniendo, así, la matriz X_c:

$$X_c = \begin{pmatrix} x_{c11} & x_{c12} & \cdots & x_{c1n} \\ x_{c21} & x_{c22} & \cdots & x_{c2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{cP1} & x_{cP2} & \cdots & x_{cPn} \end{pmatrix}$$

En cambio, la estandarización, además de centrar las variables, les impone varianza unitaria:

$$x_{estij} = \frac{x_{ij} - \bar{x}_j}{\widehat{\sigma_j}}$$

siendo:

$$\widehat{\sigma_j} = \sqrt{\frac{\sum_{i=1}^{P}(x_{ij} - \bar{x}_j)^2}{P}}$$

generando la matriz X_{est} :

$$X_{est} = \begin{pmatrix} x_{est11} & x_{est12} & \cdots & x_{est1n} \\ x_{est21} & x_{est22} & \cdots & x_{est2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{estP1} & x_{estP2} & \cdots & x_{estPn} \end{pmatrix}$$

La principal diferencia entre ambos ACPs se hace patente en el espacio de los PCs generado. Mientras el ACP no normado pondera las variables originales a la hora de crear el espacio de PCs, otorgando un mayor peso a ciertas variables, en los rangos donde existe mayor dispersión, el ACP normado uniformiza el valor de todas las variables a la hora de generar este nuevo espacio.

En el caso del proceso de Espectros Raman, el conjunto inicial a considerar es el constituido por los espectros de referencia.

Para aplicar el ACP es necesario adaptar los datos que se requieren reducir a la nomenclatura matemática requerida. En concreto, se deben expresar los datos en forma de matriz, donde las filas sean las expresiones de los datos en sí (observaciones, espectros de las muestras contempladas) y las columnas las variables (número de onda normalizados). Los Espectros Raman, desde el punto de vista de la formulación, son vectores de *n* coordenadas definidos en lo que se conoce como "espacio de señal de los números de onda normalizados":

$$E_i=[e_{i1},.....,e_{ij},.....,e_{in}]$$

donde e_{ij} son las intensidades Raman del espectro E_i para un número de onda normalizado (desplazamiento Raman) U_j.

La secuencia de pasos a seguir para realizar el ACP sobre una librería espectral de referencia es la siguiente:

1. Escoger los P espectros patrones o de referencia (de n variables) a utilizar como librería espectral, creando una matriz *E* (de coordenadas e_{ij}):

$$E = \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{22} & \cdots & e_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ e_{P1} & e_{P2} & \cdots & e_{Pn} \end{pmatrix}$$

2. Con el objetivo de minimizar problemas asociados con la comparación de espectros de diferentes intensidades, que puedan sufrir modificaciones por distintos factores, como por ejemplo, diferencias en el tiempo de medida, potencia láser, concentración, etc., los espectros son normalizados mediante la expresión:

$$e_{ij_{norm}} = \frac{e_{ij} - e_{ij_{min}}}{e_{ij_{max}} - e_{ij_{min}}} \equiv x_{ij}$$

donde $e_{ij_{norm}} \equiv x_{ij}$ es la intensidad normalizada para cada valor, e_{ij} es la intensidad original, y por último, $e_{ij_{min}}$ y $e_{ij_{max}}$ son las intensidades mínimas y máximas de cada espectro, respectivamente.

Así, se hace imprescindible homogeneizar dicha librería espectral de referencia, obteniendo la matriz *X* (de coordenadas x_{ij}):

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{P1} & x_{P2} & \cdots & x_{Pn} \end{pmatrix}$$

3. En función de qué ACP se desea llevar acabo (si ACP no normado o ACP normado), en este punto se procede a pretratar las variables iniciales, existiendo dos posibilidades: *Centrar las variables originales para el ACP no normado o bien estandarizarlas (o tipificarlas) para el ACP normado*.

Se obtiene la matriz X_c para el ACP no normado:

$$X_c = \begin{pmatrix} x_{c_{11}} & x_{c_{12}} & \cdots & x_{c_{1n}} \\ x_{c_{21}} & x_{c_{22}} & \cdots & x_{c_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{c_{P1}} & x_{c_{P2}} & \cdots & x_{c_{Pn}} \end{pmatrix}$$

En cambio, se obtiene la matriz X_{est} para el ACP normado:

$$X_{est} = \begin{pmatrix} x_{est_{11}} & x_{est_{12}} & \cdots & x_{est_{1n}} \\ x_{est_{21}} & x_{est_{22}} & \cdots & x_{est_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{est_{P1}} & x_{est_{P2}} & \cdots & x_{est_{Pn}} \end{pmatrix}$$

La decisión entre utilizar un ACP u otro es crucial a la hora de tratar determinadas librerías espectrales, puesto que dependiendo de la forma de los espectros patrones, como se verá más adelante, un ACP se adapta mejor o peor.

4. El siguiente paso es el cálculo de la matriz de covarianzas de X_c o de X_{est}, según la elección determinada entre centrado o estandarización de variables originales, obteniendo la matriz X_{cov} :

$$X_{cov} = cov\left(X_{c/est_i}, X_{c/est_j}\right) = E\left[\left(X_{c/est_i} - E\left(X_{c/est_i}\right)\left(X_{c/est_j} - E\left(X_{c/est_j}\right)\right)\right] =$$

$$= \begin{pmatrix} E[(X_{c/est_1} - E(X_{c/est_1})(X_{c/est_1} - E(X_{c/est_1})] & \cdots & E[(X_{c/est_1} - E(X_{c/est_1})(X_{c/est_n} - E(X_{c/est_n})] \\ E[(X_{c/est_2} - E(X_{c/est_2})(X_{c/est_1} - E(X_{c/est_1})] & \cdots & E[(X_{c/est_2} - E(X_{c/est_2})(X_{c/est_n} - E(X_{c/est_n})] \\ \vdots & \ddots & \vdots \\ E[(X_{c/est_n} - E(X_{c/est_n})(X_{c/est_1} - E(X_{c/est_1})] & \cdots & E[(X_{c/est_n} - E(X_{c/est_n})(X_{c/est_n} - E(X_{c/est_n})] \end{pmatrix}$$

5. Se debe proceder a la diagonalización de dicha matriz de covarianzas, obteniendo sus valores propios (*vap's*) y sus vectores propios (*vep's*). Puesto que la matriz X_{cov} tiene dimensión $n \times n$, se obtienen n *vep's* y n *vap's*. La información proporcionada por cada *vap* es extremadamente interesante ya que dicho valor coincide con la varianza de las variables originales que contempla su respectivo *vep* (convertido en PC).

6. En seguida, se ordenan de forma descendiente dichos *vep's*, según el valor de sus respectivos *vap's*, obteniendo los Componentes Principales por orden de importancia, generando así la matriz de cambio de base, C.

7. En este punto, se selecciona el número de PCs $(k < n)$ suficiente para caracterizar los datos de manera fiable. Para ello existen diversos criterios basados en la obtención de un determinado valor de varianza mínima Puesto que se desea la mínima pérdida de información, se selecciona como criterio que el valor mínimo de varianza a contemplar sea el del 100%, generando así la matriz de cambio de base reducida C_{red} .

8. Una vez obtenida dicha matriz de cambio de base, se expresa la librería espectral en el nuevo espacio obtenido, el espacio de los PCs, de dimensión inferior al original, mediante el cambio expresado por $S_{red} = X_c \cdot C_{red}$, o bien, en su caso, por $S_{red} = X_{est} \cdot C_{red}$:

$$S_{red} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1k} \\ s_{21} & s_{22} & \cdots & s_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{P1} & s_{P2} & \cdots & s_{Pk} \end{pmatrix} =$$

$$\begin{pmatrix} x_{c/est_{11}} & x_{c/est_{12}} & \cdots & x_{c/est_{1n}} \\ x_{c/est_{21}} & x_{c/est_{22}} & \cdots & x_{c/est_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{c/est_{P1}} & x_{c/est_{P2}} & \cdots & x_{c/est_{Pn}} \end{pmatrix} \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{21} & c_{22} & \cdots & c_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nk} \end{pmatrix}$$

Se pueden considerar, pues, tres tipos de coordenadas para los espectros que forman parte de un ACP: sus coordenadas en las variables originales normalizadas $\left(x_{ij}\right)$, sus coordenadas en las variables centradas o estandarizadas $\left(x_{c_{ij}}\ o\ x_{est_{ij}}\right)$, y, por último, sus coordenadas en el nuevo sistema de referencia (S(i)). Las matrices que recogen estas coordenadas son respectivamente $X, X_c\ o\ X_{est}\ y\ S$.

La siguiente tabla muestra las características más destacables de estos tres sistemas de coordenadas para el caso espectral:

Matriz	$\boldsymbol{X}$	$\boldsymbol{X_c}$	$\boldsymbol{X_{est}}$	$\boldsymbol{S}$
Media de las columnas	$\bar{x}_j$	0	0	0
Varianza de las columnas	σ_j	σ_j	1	λ_j
Correlación entre columnas	$Corr(j,j')$	$Corr(j,j')$	$Corr(j,j')$	0

Así, dado un conjunto de datos, el ACP genera un sistema de coordenadas que depende de este conjunto analizado, donde se define la dirección del primer eje en la dirección de la máxima varianza de los datos, la dirección del segundo eje, perpendicular al primero, como la que maximiza la varianza de los datos, y así sucesivamente hasta obtener n ejes ortogonales. Estos ejes, conocidos com Componentes Principales o PCs, definen un nuevo espacio de dimensión n obtenido mediante una transformación del espacio inicial.

CAPÍTULO IV

METODOLOGÍA

Planteamiento del problema

Supongamos que se dispone de los valores de p −variables en n elementos de una población dispuestos en una matriz X de dimensión $n \times p$, donde las columnas contienen las variables y las filas los elementos. Supondremos que previamente hemos restado a cada variable su media, de manera que las variables de la matriz X tienen media cero y su matriz estimada de covarianzas vendrá dada por $1/n\ X'X$.

El problema que se desea resolver es cómo encontrar un espacio de dimensión más reducida que represente adecuadamente los datos. El problema puede abordarse desde tres perspectivas equivalentes.

a) Enfoque descriptivo

Se desea encontrar un subespacio de dimensión menor que p tal que al proyectar sobre él los puntos conserven su estructura con la menor distorsión posible.

b) Enfoque estadístico:

Representar puntos p dimensionales con la mínima pérdida de información en un espacio de dimensión uno es equivalente a sustituir las p variables originales por una nueva variable, z_1, que resuma óptimamente la información. Esto supone que la nueva variable debe tener globalmente máxima correlación con las originales o, en otros términos, debe permitir prever las variables originales con la máxima precisión. Esto no será posible si la nueva variable toma un valor semejante en todos los elementos, la condición para que podamos prever con la mínima pérdida de información los datos observados, es utilizar la variable de máxima variabilidad. En general, la componente $z_r (r < p)$ tendrá varianza máxima entre todas las combinaciones lineales de las p variables X originales, con la condición de estar no correlada con las $z_1, \cdots, z_{r-1}$ previamente obtenidas.

c) Enfoque geométrico

El problema puede abordarse desde un punto de vista geométrico con el mismo resultado final. En varias dimensiones tendremos elipsoides y la mejor aproximación a los datos es la proporcionada por el eje mayor del elipsoide. Considerar los ejes del elipsoide como nuevas variables originales supone pasar de variables correladas a variables ortogonales.

Obtención de las componentes principales

La obtención de las CP puede realizarse por varios métodos alternativos:

1. Buscando aquella combinación lineal de las variables que maximiza la variabilidad. (Hottelling).

2. Buscando el subespacio de mejor ajuste por el método de los mínimos cuadrados. (Minimizando la suma de cuadrados de las distancias de cada punto al subespacio). (Pearson).

3. Minimizando la discrepancia entre las distancias euclídeanas entre los puntos calculados en el espacio original y en el subespacio de baja dimensión. (Coordenadas principales, Gower).

4. Mediante regresiones alternadas (métodos Biplot).

Propiedades de las componentes

Los componentes principales como nuevas variables tienen las propiedades siguientes:

1. Conservan la variabilidad inicial: la suma de las varianzas de los componentes es igual a la suma de las varianzas de las variables originales, y la varianza generalizada de los componentes es igual a la original.
   Comprobemos el primer punto. Como $Var(z_h) = \lambda_h$ y la suma de las raíces características es la traza de la matriz:

   $$tr(S) = Var(x_1) + \cdots + Var(x_p) = \lambda_1 + \cdots + \lambda_p$$

   por tanto $\sum_{i=1}^{p} Var(x_i) = \sum \lambda_i = \sum_{i=1}^{p} Var(z_i)$. Las nuevas variables z_i tienen conjuntamente la misma variabilidad que las variables originales, la suma de varianzas es la misma, pero su distribución es muy distinta en los dos conjuntos.
   Para comprobar que los componentes principales también conservan la Varianza generalizada, valor del determinante de varianzas y covarianzas de las variables, como el determinante es el producto de las raíces características, tenemos que, llamando S_z a la matriz de covarianzas de los componentes, que es diagonal con términos λ_i:

   $$|S_x| = \lambda_1 \dots \lambda_p = \prod_{i=1}^{p} Var(z_i) = |S_z|.$$

2. La proporción de variabilidad explicada por un componente es el cociente entre su varianza, el valor propio asociado al vector propio que lo define, y la suma de los valores propios de la matriz.

En efecto, como la varianza del componente h es λ_h, el valor propio que define el componente, y la suma de todas las varianzas de las variables originales es $\sum_{i=1}^{p} \lambda_i$, igual como acabamos de ver a la suma de las varianzas de los componentes, la proporción de variabilidad total explicada por el componente h es $\lambda_h / \sum \lambda_i$.

3. Las covarianzas entre cada componente principal y las variables X vienen dadas por el producto de las coordenadas del vector propio que define el componente por el valor propio:

$$Cov(z_i; x_1, \dots x_p) = \lambda_i a_i = (\lambda_i a_{i1}, \dots, \lambda_i a_{ip})$$

donde a_i es el vector de coeficientes de la componente z_i.
Para justificar este resultado, vamos a calcular la matriz $p \times p$ de covarianzas entre los componentes y las variables originales. Esta matriz es:

$$Cov(z, x) = 1/nZ'X$$

y su primera fila proporciona las covarianzas entre la primera componente y las p variables originales. Como $Z = XA$, sustituyendo

$Cov(z, x) = 1/nA'X'X = A'S = DA'$,

donde A contiene en columnas los vectores propios de S y D es la matriz diagonal de los valores propios. En consecuencia, la covarianza entre, por ejemplo, el primer componente principal y las p variables vendrá dada por la primera fila de $A'S$, es decir $a_1'S$ o también $\lambda_1 a_1'$, donde a_1' es el vector de coeficientes de la primera componente principal.

4. Las correlaciones entre un componente principal y una variable X es proporcional al coeficiente de esa variable en la definición del componente, y el componente de proporcionalidad es el cociente entre la desviación típica del componente y la desviación típica de la variable.

Para comprobarlo:

$$Corr(z_i; x_j) = \frac{Cov(z_i x_j)}{\sqrt{Var(z_i)Var(x_j)}} = \frac{\lambda_i a_{ij}}{\sqrt{\lambda_i s_j^2}} = a_{ij}\frac{\sqrt{\lambda_i}}{s_j}$$

5. Las r componentes principales (r<p) proporcionan la predicción lineal óptima con r variables del conjunto de variables X.

Esta afirmación puede expresarse de dos formas. La primera demostrando que la mejor predicción lineal con r variables de las variables originales se obtiene utilizando las r primeras componentes principales. La segunda demostrando que la

mejor aproximación de la matriz de datos que puede construirse con una matriz de rango r se obtiene construyendo esta matriz con los valores de los r primeros componentes principales.

6. Si estandarizamos los componentes principales, dividiendo cada uno por su desviación típica, se obtiene la estandarización multivariante de los datos originales.

   Estandarizando los componentes Z por sus desviaciones típicas, se obtienen las nuevas variables

   $$Y_c = ZD^{-1/2} = XAD^{-1/2}$$

   donde $D^{-1/2}$ es la matriz que contienen las inversas de las desviaciones típicas de las componentes. La estandarización multivariante de una matriz de variables X de media cero se define como:

   $$Y_s = XAD^{-\frac{1}{2}}A'$$

   y ambas variables están no correlacionadas y tienen matriz de covarianzas identidad. Se diferencian en que unas pueden ser una rotación de las otras, lo que es indiferente al tener todas las mismas varianzas. Por tanto, la estandarización multivariante puede interpretarse como:

   (1) obtener los componentes principales;
   (2) estandarizarlos para que tengan todos la misma varianza.

   La transformación mediante componentes principales conduce a variables no correladas pero con distinta varianza, puede interpretarse como rotar los ejes de la elipse que definen los puntos para que coincidan con sus ejes naturales. La estandarización multivariante produce variables no correladas con varianza única, lo que supone buscar los ejes naturales y luego estandarizarlos. En consecuencia, si estandarizamos los componentes se obtiene las variables estandarizadas de forma multivariante.

Número de componentes principales

Presentamos algunos criterios para determinar el número $m < p$ de componentes principales.

1. Criterio del porcentaje

El número m de componentes principales se toma de modo que P_m sea próximo a un valor especificado por el usuario, por ejemplo el 80%. Por otra parte, si la representación de

$P_1, P_2, \ldots, P_k, \ldots$ con respecto de k prácticamente se estabiliza a partir de un cierto m, entonces aumentar la dimensión apenas aporta más variabilidad explicada.

2. Criterio de Kaiser

Obtener las componentes principales a partir de la matriz de correlaciones R equivale a suponer que las variables observables tengan varianza 1. Por lo tanto una componente principal con varianza inferior a 1 explica menos variabilidad que una variable observable. El criterio, llamado Kaiser, es entonces:

Retenemos las m primeras componentes tales que $\lambda_m \geq 1$, donde $\lambda_1 \geq \cdots \geq \lambda_p$ son los valores propios de R, que también son las varianzas de las componentes. Estudios de Montecarlo prueban que es más correcto el punto de corte $\lambda^* = 0.7$, que es más pequeño que 1.

Este criterio se puede extender a la matriz de covarianzas. Por ejemplo, m podría ser tal que $\lambda_m \geq v$, donde $v = tra(S)/p$ es la medida de las varianzas. También es aconsejable considerar el punto de corte $0.7 \times v$.

3. Test de esfericidad

Supongamos que la matriz de datos proviene de una población normal multivariante $N_p(\mu, \Sigma)$. Si la hipótesis

$$H_0^{(m)}: \lambda_1 > \cdots > \lambda_m > \lambda_{m+1} = \cdots = \lambda_p$$

es cierta, no tiene sentido considerar más de m componentes principales. En efecto, no hay direcciones de máxima variabilidad a partir de m, es decir, la distribución de los datos es esférica. El test para decidir sobre $H_0^{(m)}$ está basado en el estadístico ji-cuadrado y se aplica secuencialmente: Si aceptamos $H_0^{(0)}$ no hay direcciones principales, pero si rechazamos $H_0^{(0)}$, entonces repetimos el test con $H_0^{(1)}$. Si aceptamos $H_0^{(1)}$ entonces $m = 1$, pero si rechazamos $H_0^{(1)}$ repetimos el test con $H_0^{(2)}$, y así sucesivamente. Por ejemplo, si $p = 4$, tendríamos que $m = 2$ si rechazamos $H_0^{(0)}, H_0^{(1)}$ y aceptamos $H_0^{(2)}: \lambda_1 > \lambda_2 > \lambda_3 = \lambda_4$.

4. Criterio del bastón roto

Los valores propios suman $V_t = tr(S)$, que es la variabilidad total. Imaginemos un bastón de longitud V_t, que rompemos en p trozos al azar (asignando $p - 1$ puntos uniformemente sobre el intervalo $(0, V_t)$) y que los trozos ordenados son los valores propios $l_1 > l_2 > \cdots > l_p$. Si normalizamos a $V_t = 100$, entonces el valor esperado de l_j es

$$E(L_j) = 100 \times \frac{1}{p}\sum_{i=1}^{p-j}\frac{1}{j+i}$$

Las m primeras componentes son significativas si el porcentaje de varianza explicada supera claramente el valor de $E(L_1) + \cdots + E(L_m)$. Por ejemplo si $p = 4$, los valores son:

Porcentaje	$E(L_1)$	$E(L_2)$	$E(L_3)$	$E(L_4)$
Esperado	52.08	27.08	14.58	6.25
Acumulado	52.08	79.16	93.74	100

Si $V_2 = 93.92$ pero $V_3 = 97.15$, entonces tomaremos sólo dos componentes.

Se han sugerido distintas reglas para seleccionar el número de componentes a mantener:

(1) Realizar un gráfico de λ_i frente a i. Comenzar seleccionando componentes hasta que los restantes tengan aproximadamente el mismo valor de λ_i. La idea es busca un "codo" en el gráfico, es decir, un punto a partir del cual los valores propios son aproximadamente iguales. El criterio es quedarse con un número de componentes que excluya los asociados a valores pequeños y aproximadamente del mismo tamaño.

(2) Seleccionar componentes hasta cubrir una proporción determinada de varianza, como el 80% o el 90%. Esta regla es arbitraria y debe aplicarse con cierto cuidado. Por ejemplo, es posible que un único componente de "tamaño" recoja el 90% de la variabilidad y sin embargo pueden existir otros componentes que sean muy adecuados para explicar la "forma" de las variables.

(3) Desechar aquellos componentes asociados a valores propios inferiores a una cota, que suele fijarse como la varianza media, $\sum \lambda_i/p$. En particular, cuando se trabaja con la matriz de correlación, el valor medio de los componentes es 1, y esta regla lleva a seleccionar los valores propios mayores que la unidad. De nuevo esta regla es arbitraria: una variable que sea independiente del resto suele llevarse un componente principal y puede tener un valor propio mayor que la unidad. Sin embargo, si esta incorrelada con el resto puede ser una variable poco relevante para el análisis, y no aportar mucho a la comprensión del fenómeno.

Interpretación de las componentes

Componentes de tamaño y forma

Cuando existe una alta correlación positiva entre todas las variables, el primer componente principal tiene todas sus coordenadas del mismo signo y puede interpretarse como un promedio ponderado de todas las variables. Se interpreta entonces como un factor global de "tamaño". Los restantes componentes se interpretan como factores de "forma" y típicamente tienen coordenadas positivas y negativas, que implica que contraponen unos grupos de

variables frente a otros. Estos factores de forma pueden frecuentemente escribirse como medias ponderadas de dos grupos de variables con distinto signo y contraponen las variables de un signo a las del otro.

Representación gráfica

La interpretación de los componentes principales se favorece representando las proyecciones de las observaciones sobre un espacio de dimensión dos, definido por parejas de los componentes principales más importantes. La representación habitual es tomar dos ejes ortogonales que representen los dos componentes considerados, y situar cada punto sobre ese plano por sus coordenadas con relación a estos ejes, que son los valores de los dos componentes para esa observación. Por ejemplo, en el plano de los dos primeros componentes, las coordenadas del punto X_i son $z_{1i} = a_1' X_i \; y \; z_{2i} = a_2' X_i$.

La interpretación se favorece representando en el mismo plano además de las observaciones las variables originales. Esto puede hacerse utilizando como coordenadas su coeficiente de correlación con cada uno de los ejes. El vector de correlaciones entre el primer componente y las variables originales viene dado por $\lambda_1^{1/2} a_1' D$, donde D es una matriz diagonal cuyos términos son las inversas de las desviaciones típicas de cada variable. La matriz de correlaciones R_{cv} entre los p componentes y las p variables tendrá como filas los términos $\lambda_j^{1/2} a_j' D$ y puede escribirse

$$R_{cv} = \Lambda^{1/2} A D$$

donde A es la matriz de vectores propios, $\Lambda^{1/2}$ es la matriz diagonal con términos $\sqrt{\lambda_i}$ y en el análisis normado como las variables se estandarizan a varianza unidad las correlaciones serán simplemente $\Lambda^{1/2} A$.

Es importante recordar que las covarianzas (o correlaciones) miden únicamente las relaciones lineales entre las variables. Cuando entre ellas existan relaciones fuertes no lineales el análisis de componentes principales puede dar una información muy parcial de las variables.

ANÁLISIS NORMALIZADO O CON CORRELACIONES

Los componentes principales se obtienen maximizando la varianza de la proyección. En términos de las variables originales esto supone maximizar:

$$M = \sum_{i=1}^{p} a_i^2 s_i^2 + 2 \sum_{i=1}^{p} \sum_{j=i+1}^{p} a_i a_j s_{ij}$$

con la restricción $a'a = 1$. Si alguna de las variables, por ejemplo la primera, tiene una varianza s_1^2, mayor que las demás, la manera de aumentar M es hacer tan grande como podamos la coordenada a_1 asociada a esta variable. En el límite si una variable tiene una varianza mucho mayor que las demás el primer componente principal coincidirá muy aproximadamente con esta variable.

Cuando las variables tienen unidades distintas esta propiedad no es conveniente: si disminuimos la escala de medida de una variable cualquiera, de manera que aumenten en magnitud sus valores numéricos, el peso de esa variable en el análisis aumentará, ya que:

(1) su varianza será mayor y aumentará su coeficiente en el componente, a_i^2, ya que contribuye más a aumentar M;
(2) sus covarianzas con todas las variables aumentarán, con el siguiente efecto de incrementar a_i.

Cuando las escalas de medida de las variables son muy distintas, la maximización dependerá decisivamente de estas escalas de medida y las variables con valores más grandes tendrán más peso en el análisis. Si queremos evitar esta problema, conviene estandarizar las variables antes de calcular los componentes, de manera que las magnitudes de los valores numéricos de las variables X sean similares.

La estandarización resuelve otro posible problema. Si las variabilidades de las X son muy distintas, las variables con mayor varianza van a influir más en la determinación de la primera componente. Este problema se evita al estandarizar las variables, ya que entonces las varianzas son la unidad, y las covarianzas son los coeficientes de correlación. La ecuación a maximizar se transforma en:

$$M' = 1 + 2\sum_{i=1}^{p}\sum_{j=i+1}^{p} a_i a_j r_{ij}$$

siendo r_{ij} el coeficiente de correlación lineal entre las variables ij. En consecuencia la solución depende de las correlaciones y no de las varianzas.

Los componentes principales normados se obtienen calculando los vectores y valores propios de la matriz R, de coeficientes de correlación. Llamando λ_p^R a las raíces características de esa matriz, que suponemos no singular, se verifica que:

$$\sum_{i=1}^{p} \lambda_i^R = traza(R) = p$$

Las propiedades de los componentes extraídos de R son:

1. La proporción de variación explicada por λ_p^R será:

$$\frac{\lambda_p^R}{p}$$

2. Las correlaciones entre cada componente z_j y las variables X originales vienen dados directamente por $a_j'\sqrt{\lambda_j}$ siendo $z_j = Xa_j$.

Cuando las variables X originales están en distintas unidades conviene aplicar el análisis de la matriz de correlaciones o análisis normado. Cuando las variables tienen las mismas unidades, ambas alternativas son posibles. Si las diferencias entre las varianzas de las variables son informativas y queremos tenerlas en cuenta en el análisis no debemos estandarizar las variables. Por el contrario, si las diferencias de variabilidad no son relevantes podemos eliminarlas con el análisis normado. En caso de duda, conviene realizar ambos análisis, y seleccionar aquel que conduzca a conclusiones más informativas.

CAPITULO V

RESULTADOS

Los espectros, tal y como salen del aparato Raman, pueden salvarse en multitud de formatos propios del instrumento. Para nuestros propósitos el más relevante es el .cvs que puede manejarse con Excel y se puede importar fácilmente en matlab.

Una vez en matlab los hemos convertido en archivos .mat, que están en un formato propio del programa.

Supondremos, pues, que todos nuestros espectros están en formato .mat.

El programa principal es

```
1. load('HH1.mat')
2. load('HH2.mat')
3. load('HH3.mat')
4. load('HH4.mat')
5. load('HH5.mat')
6. kk=4;
7. E=[HH1';HH2';HH3';HH4';HH5'];
8. [M,N]=size(E);
9. X=zeros(M,N);
10. for v=1:M
11. Imin=min(E(v,:));
12. Imax=max(E(v,:));
13. dif=Imax-Imin;
14. X(v,:)=((E(v,:)-Imin)/dif);
15. end
16. Xneu=X;
17. sumita=0;
18. for k=0:N-1
19. for i=1+(k*M):M+(k*M)
20. sumita=sumita+X(i);
21. end
22. sumita=sumita/M;
23. vAr=sqrt(var(X(:,k+1)));
24. for i=1+(k*M):M+(k*M)
25. X(i)=(X(i)-sumita)/vAr;
26. end
```

```
27. sumita=0;
28. end

29. c=cov(X);

30. [V,D]=eig(c);
31. d2=diag(D);
32. [dd2,I2]=sort(d2,'descend');
33. V2=V(:,I2);
34. V22red=V2(:,1:kk);
35. newmat2=Xneu*V22red;% o X?
36. plot(newmat2')
37. Varianzaporpc=dd2/sum(dd2)
38. figure
39. plot(E')
40. ET=HH6';
41. Imin=min(ET);
42. Imax=max(ET);
43. dif=Imax-Imin;
44. XT=(ET-Imin)/dif;
45. newmatT=XT*V22red;
46. figure
47. plot(newmatT')
```

como puede verse comienza con las instrucciones para cargar los espectros a analizar, en este caso las instrucciones relevantes son las de las primeras cinco líneas. Cuando son muchos los archivos podemos escribir rutinas para cargarlos como la que se muestra a continuación:

```
for h=1:36
load(['A',num2str(h)])
end
```

que carga los archivos A1, A2, ..., A36.

También conviene poner todos los archivos en un solo archivo E como el de la línea 7 del programa(al contar líneas estaré ignorando líneas en blanco), ya que es necesario tener los espectros cargados en matlab en forma de matriz para realizar el ACP.

Uno de nuestros espectros (el HH1) tiene la apariencia

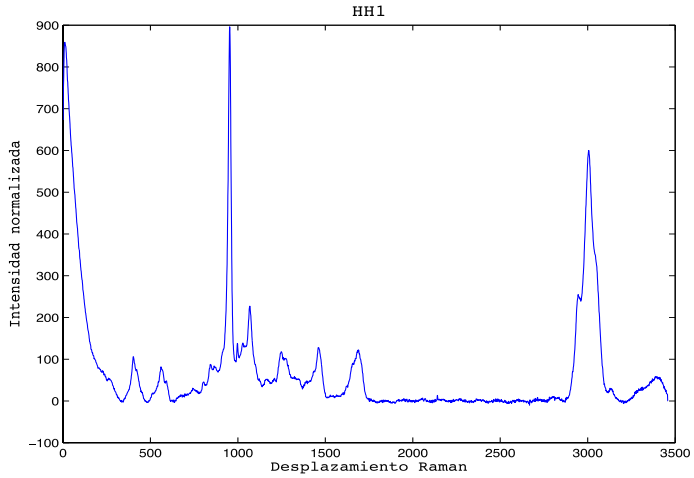

Fig. 5.1. Espectro Raman HH1

Como podemos observar tiene picos que son muy visibles pero entre 1700 -2800 no es fácil distinguir la señal del ruido.

En la línea 10 se procede a la normalización de los espectros para minimizar problemas asociados con la comparación de espectros de diferentes intensidades. En este paso tenemos una matriz con los datos ya normalizados pero del mismo tamaño.

A partir de la línea 17 se estandarizan las variables, ya que además de centrar las variables les impone varianza unitaria, esto se logra restando el promedio (centrado) y dividiendo entre la varianza.

Una vez obtenida dicha matriz estandarizada procedemos al cálculo de nuestra matriz de covarianzas, este es un comando que ya está dado en matlab es muy sencillo y nos ahorra el tener que hacerlo manual.

A partir de la línea 30 realizamos la diagonalización y obtención de los vectores propios y valores propios de nuestra matriz de covarianzas, también es un comando muy sencillo que está dado en Matlab.
La matriz D tiene los valores propios de la matriz c. El comando d2=diag(D) extrae los elementos en la diagonal de D.
A continuación lo que procede es ordenar los vectores propios, esto sería muy tardado si intentáramos una ordenación "ingenua" pero afortunadamente Matlab también cuenta con una rutina de ordenación rápida, se trata de sort que en la forma [dd2,I2]=sort(d2,'descend') ordena los valores propios en orden decreciente, la variable I2 es muy útil pues nos da la permutación de los datos de d2 que produce dd2. Por ejemplo si en algún caso I=[3,1,4] esto querría decir que el valor más grande era el tercero, seguido del primero etc.

Pero una vez ordenados los vectores propios en la línea 37 empezamos el cálculo de las PCs, una vez obtenidos hacemos la selección del número de PCs con el criterio de 100% de varianza acumulada, como podemos ver en este caso se utilizarán las primeros cuatro PCs porque es hasta donde se acumula toda la varianza requerida.

PC	varianza por PC	varianza acumulada
1	55.90%	55.90%
2	22.85%	78.75%
3	13.52%	92.27%
4	7.73%	100%

Tabla 5.1. Varianza por PC y varianza acumulada

Estos datos se calculan a partir de los valores propios ordenados, la varianza total es la suma de todos los elementos de dd2 por lo que cada varianza estará dada por los elementos de dd2 divididos entre la varianza total.

El meollo del método está en que ahora es preciso re-expresar los espectros originales. Digamos que cada espectro tiene las coordenadas en R^{3458} de un vector, pensemos que esta base es la canónica "can".
Pero tras la diagonalización tenemos la base dada por los vectores propios "eig". Si representamos al espectro como un vector columna de 3458 por 1 entonces

$$[v]_{eig} = M^{can}_{eig}\,[v]_{can}$$

que nos dice que las coordenadas de un vector v en la base eig estarán dadas por la matriz de cambio de base de can a eig multiplicando a las coordenadas de v en la base can. Pero la matriz de cambio de base M^{eig}_{can} es precisamente la V22red del programa. Como en nuestro programa estamos usando renglones todo el tiempo tomamos la transpuesta de la ecuación anterior y

$$[v]^T_{eig} = [v]^T_{can} M^{Tcan}_{\ eig}$$

Y

$$[v]^T_{eig} = [v]^T_{can} V22red$$

Así, el ACP genera un espacio de k=4, esta reducción permite una expresión simplificada de los espectros patrones en un espacio donde sus variables, no correladas entre sí, permiten la detección de similitudes y diferencias entre espectros de una forma considerablemente más sencilla, lo cual convierte al ACP en un proceso óptimo para el reconocimiento espectral. En otros casos (a considerar más adelante) tomaremos "cuatro" como referencia y mostraremos cuatro componentes planteando la hipótesis heurística (pero falsificable y cambiable) de que con cuatro componentes podremos lograr la meta de ayudarnos a identificar y clasificar los espectros.

Si se procede a representar la librería de espectros patrones conjuntamente en ambos espacios, se puede comprobar que, efectivamente, los Espectros Raman son más fáciles de diferenciar en su representación en el espacio de PCs que en el espacio original. Se evidencia, así, que la reducción dimensional que permite el Análisis por Componentes Principales facilita de forma considerable la inspección visual de los espectros.

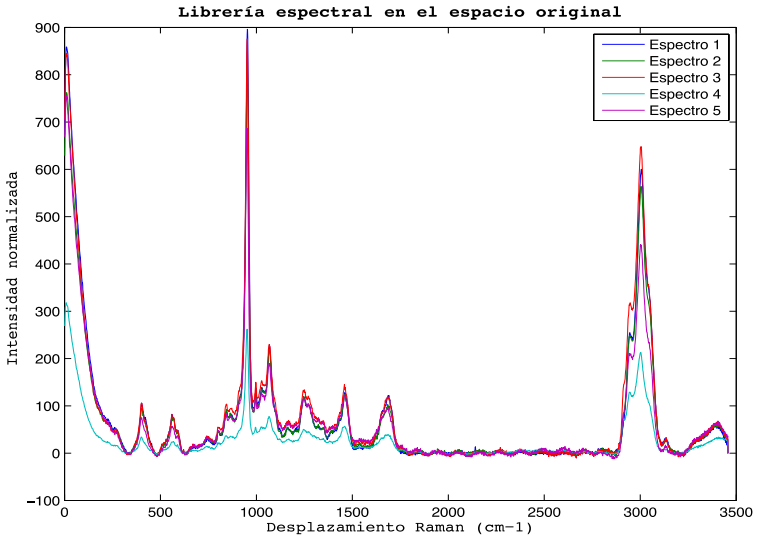

Fig. 5.2Librería espectral en el espacio original

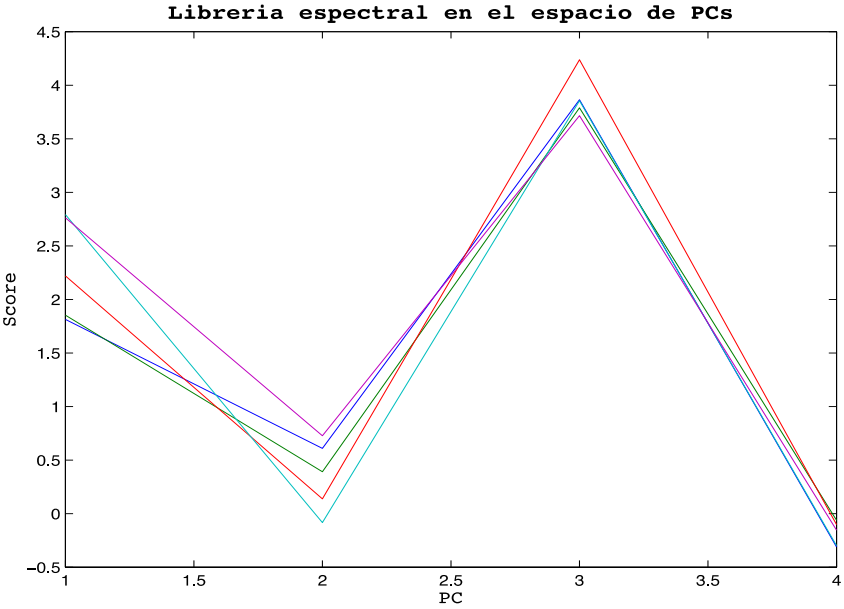

Figura 5.3. Librería espectral en el espacio de PCs

Ahora tomaremos otro espectro, en principio similar a los anteriores pero que visualmente se ve un poco diferente, vea la siguiente figura:

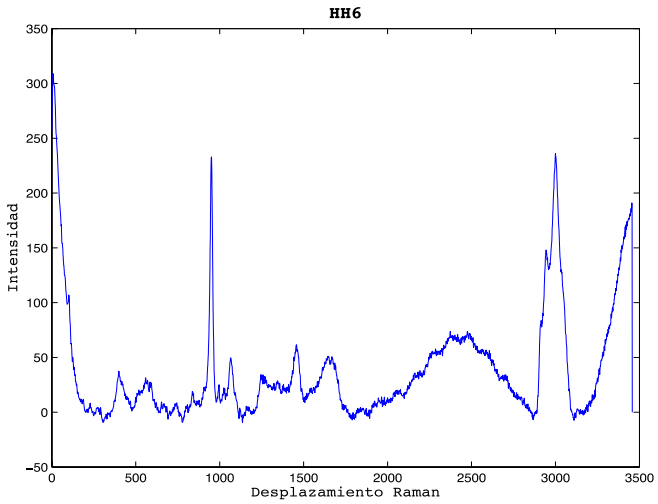

Fig.5.4.Espectro Raman HH6

Si expresamos este espectro en la eigenbase obtenida para los anteriores (de la misma substancia) y presentamos las cuatro componentes, obtenemos la figura 5.5 que, si se superpone con las de todos los espectros, muestra claramente la diferencia. Nuestra técnica puede, en principio, ser de ayuda para discriminar.

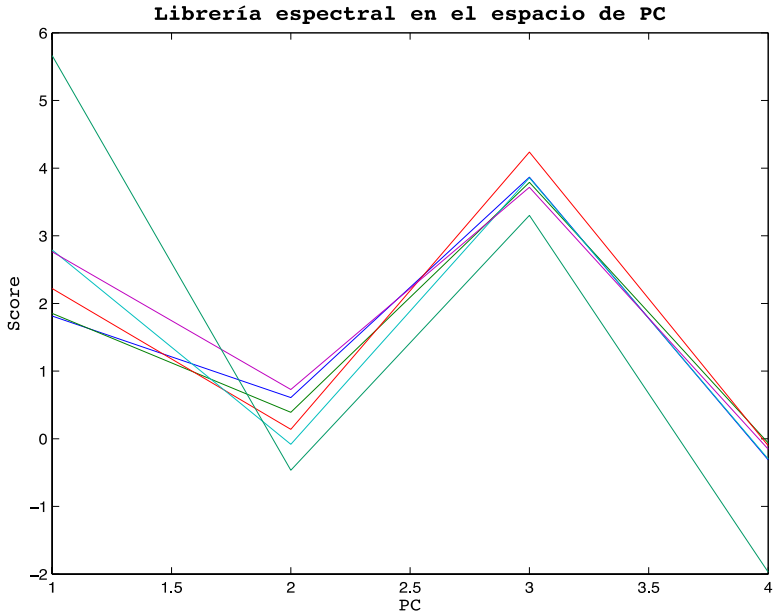

Fig. 5.5. Espectros superpuesto

Como se puede observar en la figura 5.5 el espectro HH6 (verde) es claramente diferente a los otros espectros, esto ayuda al analista a identificar visiblemente si la muestra que se analizó con el Raman contiene solo un elemento o está presente algún otro elemento.

APLICACIÓN A UN CASO REAL.

Ahora corremos nuestros programas a una serie de 36 espectros de silicio.

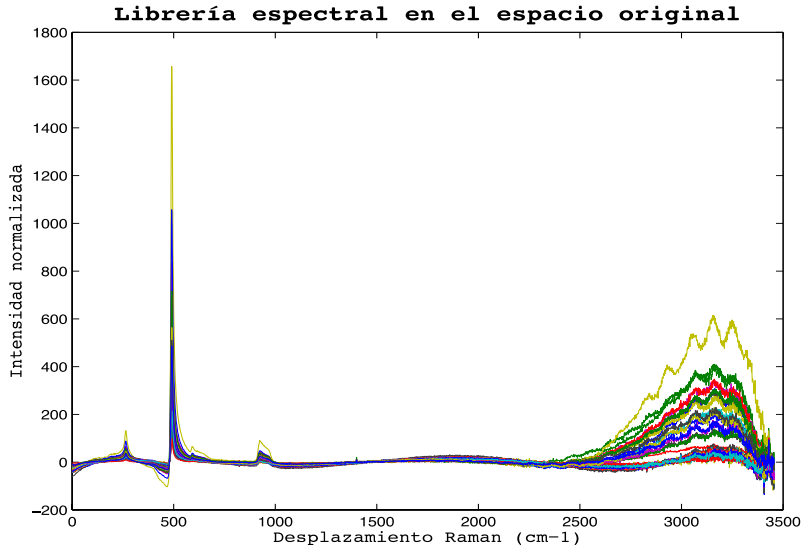

Fig. 5.6. Librería espectral en el espacio original (silicio)

Tal y como se espera, todas las gráficas son similares; pero la contribución de cuatro componentes a la varianza total ya no es del 100%, la varianza acumulada sería en 35 componentes, aunque solo vamos a considerar las cuatro primeras componentes porque se acumula el 99% de varianza, el otro 1% gráficamente no es de interés ya que no muestra cambios visuales.

La figura 5.7 muestra los resultados obtenidos. De la naturaleza de la muestra se espera la gran similitud que se observa.

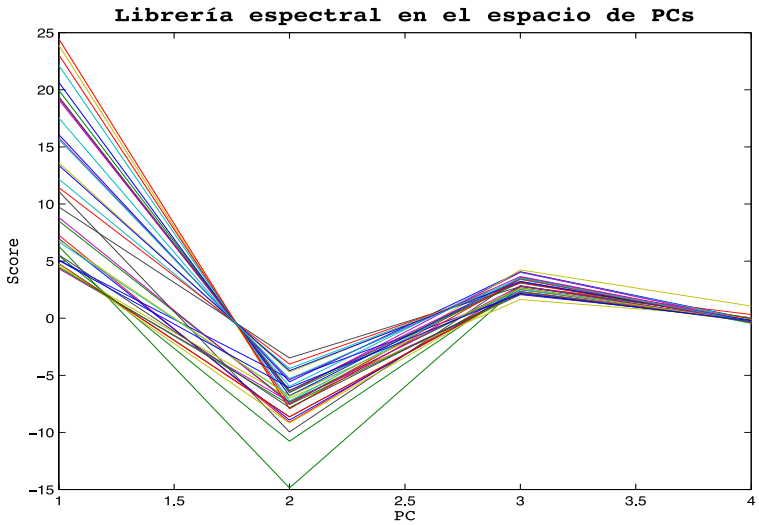

Fig.5.7 Librería espectral en el espacio de PCs (4 componentes)

Como puede apreciarse hay bastante dispersión en los datos. Los investigadores y técnicos que tomaron los espectros nos indican que, en efecto, había en las muestras diversas zonas con contenidos variables de silicio. Las figuras muestran dos morfologías típicas para las regiones de donde provienen los espectros.

ANALISIS DE COMPONENTES INDEPENDIENTES

1. El problema de la separación de fuentes.

Consideremos el problema de la fiesta de cóctel (en inglés cocktail party problem). Supongamos que estamos en una fiesta y deseamos conversar con una persona en particular. El problema es que al mismo tiempo hay muchas otras conversaciones e incluso música ruidosa. Lo que desearíamos es poder aislar la conversación e ignorar todas las demás señales.

Una descripción gráfica útil sería:

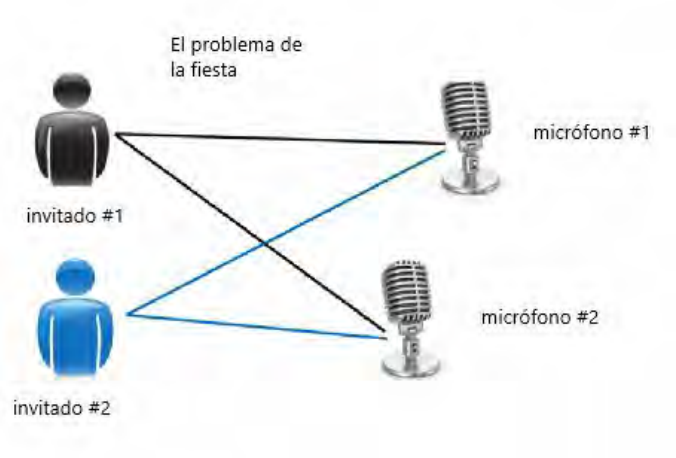

Fig. 5.8. Representación del problema de la fiesta

Pensemos en dos fuentes, convencionalmente denotadas como s1 y s2 (porque en inglés fuente se escribe source) y que usamos dos micrófonos (m1 y m2). Claro, cada micrófono recibirá las dos señales, aunque con diferentes intensidades. El problema es cómo separar las señales que se hallan mezcladas. Éste es el problema del análisis de componentes principales (ICA por sus siglas en inglés).

El siguiente diagrama muestra cómo se mezclan las señales, supondremos que la señal de la fuente s1 llega a M1 con intensidad (volumen) a1 y a M2 con volumen a2 en tanto que la señal de la fuente s2 llega a M1 con intensidad (volumen) b1 y a M2 con volumen b2.

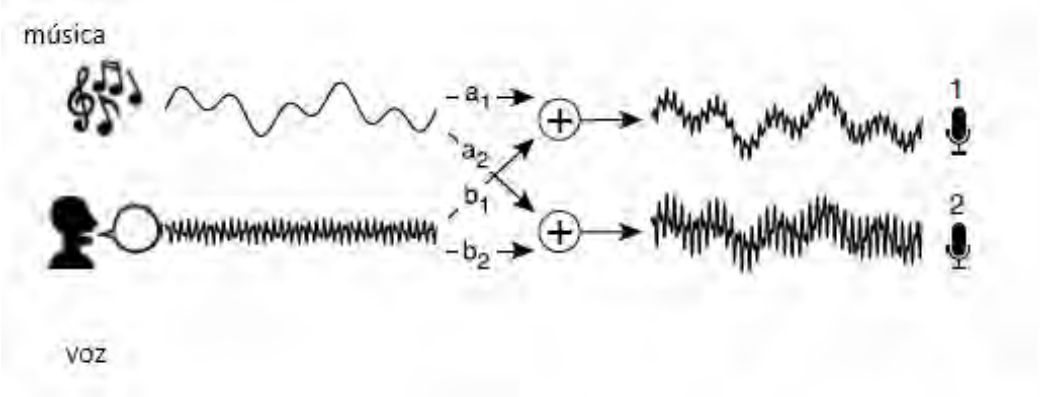

Fig. 5.9. Diagrama de la mezcla de señales

2. Planteamiento del problema

Vamos a juntar las señales en un vector s

$$s = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{bmatrix}$$

donde m es el número de fuentes (que en los ejemplos anteriores era 2). Debemos pensar que cada s_i es a su vez un vector renglón con las n muestras de la señal.

Se define ahora una matriz

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{km} \end{bmatrix}$$

donde a_{ij} indica la contribución de la fuente i al micrófono j (o, en general, el sistema físico que produce la mezcla), k es el número de micrófonos.

Se define x como

$$x = A_s$$

En otras palabras, x es la señal ya mezclada. Conociendo x (es lo que percibimos en el salón de la fiesta) si conociéramos A (pero no es el caso en general) podríamos recuperar s.

La forma habitual, en la literatura, de empezar a plantear el problema es definiendo dos nuevas variables $\hat{s}$ y W relacionadas mediante

$$\hat{s} = W_x$$

y queremos encontrar la $\hat{s}$ que más se parezca a s (en un sentido a ser especificado luego). Es decir, esperamos que $\hat{s} \cong s$. Si A fuera invertible tendríamos que $W = A^{-1}$ y que $s = \hat{s}$. Lo más común es que la relación entre A y W sea del tipo pseudoinversa (ver apéndice).

3. Primeros pasos

Pensemos ahora en la descomposición en valores singulares de A

$$A = U\Sigma V^T$$

(ver de nuevo el apéndice).

Una consecuencia importante de esta descomposición es que

$$AA^T = U\Sigma V^T V\Sigma^T U^T = U\Sigma\Sigma^T U^T$$

Y, análogamente

$$A^T A = V\Sigma^T \Sigma V^T$$

Como AA^T y $A^T A$ son simétricas, pueden ser ortogonalmente diagonalizadas. De hecho los las matrices diagonalizadoras de AA^T y $A^T A$ son U y V respectivamente en tanto que las matrices diagonales (que contienen a los valores característicos) son $\Sigma\Sigma^T$ y $\Sigma^T\Sigma$ respectivamente.

Si ($m = k$) tendríamos

$$\Sigma = \Sigma^T$$

Y

$$\Sigma\Sigma^T = \Sigma^T\Sigma = \Sigma^2$$

Es por esto que

$$\langle xx^T \rangle = \langle Ass^T A^T \rangle$$

(aquí ‹› se refiere al proceso de calcular promedios) y tendremos que

$$\langle xx^T \rangle = \langle U\Sigma V^T ss^T V\Sigma^T U^T \rangle = U\Sigma V^T \langle ss^T \rangle V\Sigma^T U^T$$

Estas ecuaciones son exactas. En el análisis de componentes independientes se postula:

$$\langle ss^T \rangle = I$$

(donde I es una matriz identidad del tamaño adecuado) .
Por ello

$$\langle xx^T \rangle = U\Sigma V^T \langle ss^T \rangle V\Sigma^T U^T = U\Sigma\Sigma^T U^T$$

De hecho es más correcto tomar como postulado toda la ecuación anterior (es decir, que el promedio no depende de U ni de V). Esto se llama blanqueado y lo analizaremos posteriormente.

Pero la última ecuación tiene precisamente la forma de una ecuación de valor característico y vector característico. Esto significa que si conocemos $\langle xx^T \rangle$ podemos inferir U y Σ. Pero no podemos inferir V todavía.

Si usamos una rutina diagonalizadora para $\langle xx^T \rangle$

$$[E, D] = eig(\langle xx^T \rangle)$$

$$E = U$$
$$D = \Sigma\Sigma^T$$

y

$$\langle xx^T \rangle = EDE^T$$

Por todo esto

$$A = ED^{\frac{1}{2}}V^T$$

y (supondremos que A es invertible, si no, habrá que usar pseudoinversas)

$$W = VD^{\frac{-1}{2}}E^T$$

De la ecuación de valores propios es fácil ver que

$$\langle (E^T x)(Ex^T)^T \rangle = D$$

(salvo cambios en notación, esto es lo que hicimos en las componentes principales) .

Hasta aquí todo parece componentes principales, pero ahora se normaliza todo mediante D de modo que si definimos

$$x_w = (D^{\frac{-1}{2}} E^T)x$$

lograremos que

$$\langle x_w x_w^T \rangle = I$$

($a\ x_w$) se le suele denotar con el subíndice w por el blanqueo (que en inglés es white). x_w es la versión blanqueada de los datos de entrada. De la ecuación

$$W = VD^{\frac{-1}{2}}E^T$$

$$\hat{s} = W\ x\ =\ V\ x_w$$

y lo que falta es encontrar V.

4. Una implementación sencilla.

En la red hemos encontrado una rutina Matlab que enlistamos a continuación:

```
function[y, w] = ica(x)
 % programa simple por Damasceno
n = size(x, 1);
[E, D] = eig(cov(x' ));
 V = E ∗ D^(−0.5) ∗ E' ∗ x;
 z = repmat(sqrt(sum(V.^2 )), n, 1). ∗ V ;
 [EE, DD] = eig(cov(z'));
y = EE' ∗ V ;
w = EE' ;
 end
```

que hemos probado con el mismo ejemplo del autor:

```
clear
clc
n = 1 : 100;
s1 = sin(3 ∗ n);
s2 = randn(1, 100);
s = [s1; s2];
a = rand(2);
x = a ∗ s;
[y, w] = ica(x);
subplot(3, 2, 1); plot(s1);
subplot(3,2,2); plot(s2);
subplot(3,2,3); plot(x(1, :));
subplot(3,2,4); plot(x(2, :));
subplot(3,2,5); plot(y(1, :));
subplot(3,2,6); plot(−y(2, :));
```

Y los resultados se muestran en la siguiente figura:

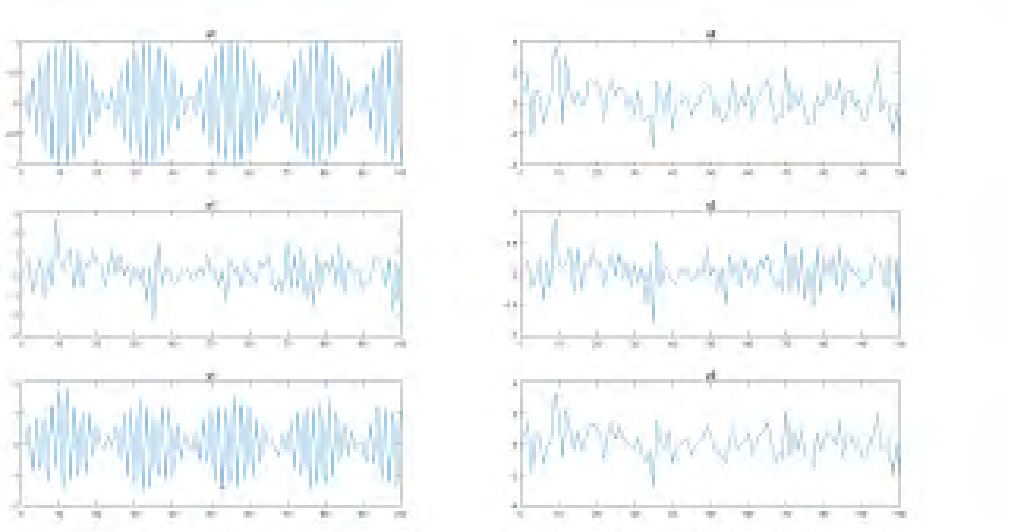

Fig. 5.10. Resultados del ejemplo encontrado en red.

Pero este ejemplo sigue líneas diferentes de las anteriormente expuestas. El artículo de Shlens discute la obtención de V usando información mutua.

5. Raman

El algoritmo anterior se aplicó a uno de nuestros espectros Raman con el siguiente resultado:

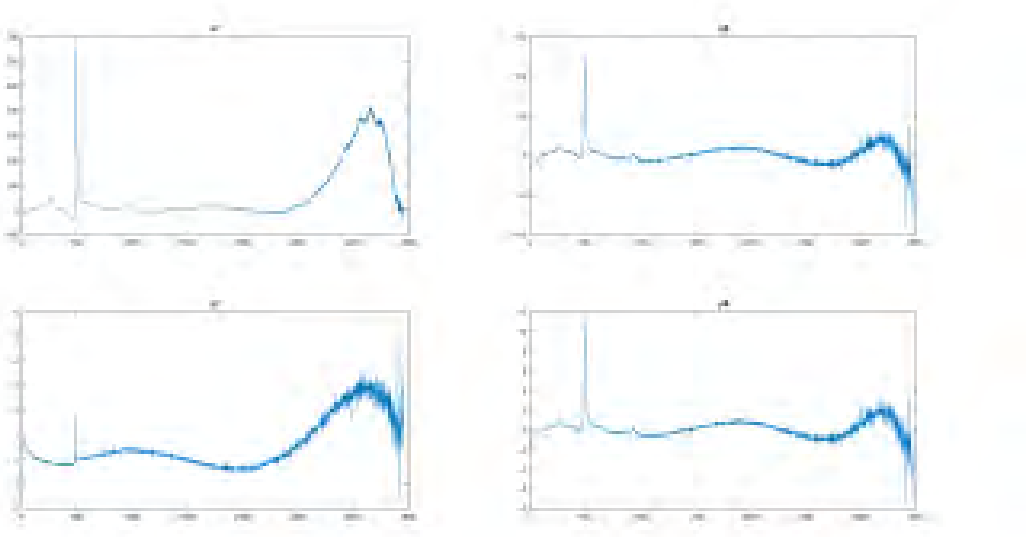

Fig. 5.11. Resultados de Espectros Raman mediante componentes independientes

Esto nos indica que el problema no es tan sencillo como con ejemplos prefabricados.

Aquí se ha pensado en que dos espectros juegan el papel de las señales. Pero nótese que los espectros no son registrados simultáneamente por lo que la analogía entre el problema del

cóctel y los espectros es cuestionable. Pero esto da mucha pauta para trabajo ulterior. Agradezco a un sinodal la sugerencia de usar componentes independientes para el tratamiento del ruido.

CONCLUSIONES

Hemos revisado la técnica de componentes principales y escrito un programa Matlab para analizar espectros Raman obtenidos en el laboratorio de materiales avanzados del Instituto de Física, UNAM.

Se escogió Matlab por ser el más usado entre los físicos. Después de todo deseamos que ellos puedan modificar nuestro programa de ser necesario.

Aunque tanto Matlab como R tienen rutinas que lo hacen todo, hemos elegido programar nuestra propia rutina como único camino para entender cabalmente el proceso.

A diferencia de lo hecho en la referencia (4) en la que se usaron espectros simulados, nosotros hemos utilizado espectros reales.

Con estos espectros se encuentra que la técnica da resultados similares para espectros similares, pero permite discriminar cuando los espectros presentan diferencias o anomalías. Por ejemplo, puede ocurrir que una muestra, que nominalmente tenga sólo una substancia, tenga contaminantes accidentales.

El trabajo a futuro, en el laboratorio, incluye el diseño de una base de datos de tamaño real teniendo no sólo muchos espectros diferentes para una misma substancia, sino datos para muy diversas muestras.

Desde la perspectiva matemática se ha optado por usar en Matlab un proceso de diagonalización en vez de una descomposición en valores singulares. En parte se debió la elección a que la diagonalización es conceptualmente más simple, pero sobre todo a que es más rápida en nuestras máquinas.

Hemos explorado también el uso de la técnica de componentes independientes y esto abre nuevas posibilidades a explorar.

APENDICE

1. Descomposición en valores singulares.

Sea A una matriz real de tamaño $n \times k$ (no tiene que ser cuadrada). Entonces existen matrices reales $U\Sigma$ y V tales que

$$A = U\Sigma V^T$$

Y donde

- U es de $m \times m$ y es ortogonal, es decir, $UU^T = U^T U = I$ donde la T indica transposición e I es la matriz identidad de $m \times m$
- V es de $k \times k$ y es ortogonal, es decir $VV^T = V^T V = I$ donde ahora I es la matriz identidad de$k \times k$.
- Σ es una matriz de $m \times k$, es diagonal (sus elementos σ_{ij} son cero si $i \neq j$), y sus elementos diagonales aparecen en orden decreciente.
- el número de elementos $\sigma_{ii} \neq 0$ es el rango de A.

Los elementos $\sigma_i = \sigma_{ii}$ reciben el nombre de valores singulares. Las columnas de U reciben el nombre de vectores singulares izquierdos en tanto que las columnas de V se llaman vectores singulares derechos.

La factorización $A = U\Sigma V^T$ recibe el nombre de descomposición en valores singulares.

En matlab la diagonalización se calcula con el comando *eig* en tanto que la descomposición en valores singulares está dada por *svd*.

2. La inversa de A

Si conociéramos V podríamos tener completamente a A. En caso de que A sea invertible podemos calcular w. Pero procedamos con cautela.

2.1. La pseudoinversa de Moore-Penrose. Para una matriz A siempre existirá otra, única, A^t llamada pseudoinversa de Moore-Penrose, y que satisface:

$$AA^tA = A$$
$$A^tAA^t = A^t$$
$$(AA^t)^T = (AA^t)$$
$$(A^tA)^T = (A^tA)$$

En matlab el comando es *pinv*.

Cuando la matriz A tiene inversa, A^t nos dará dicha inversa. Si no, A^t será una mejor aproximación a la inversa, en un sentido que el lector mejor hallará en la literatura. En breve, la pseudoinversa es la opción idónea, en todo caso.

BIBLIOGRAFÍA

[1] Carles M. Cuadras “Nuevos Métodos de Análisis Multivariante” CMC Editions, Barcelona, 2007.

[2] Daniel Peña “Análisis de Datos Multivariantes” McGraw Hill, 2002

[3] José Luis Vicente Villardón “Análisis de Componentes Principales” Departamento de Estadística

[4] Juan José González Vidal “Identificación Automática de Espectros Raman de pigmentos mediante Análisis por Componentes Principales” Universidad Politécnica de Cataluña.

[5] Ferraro J. R., Nakamoto K. Introductory Raman Sprectoscopy. Primera edición. Academic Press. 1994.

[6] José Luis Pérez, Rogelio Murillo, Raúl Gómez “Espectroscopias infrarroja y Raman”. Sistemas.fciencias.unam.mx/fam/EsRaman.pdf

[7] Jonathon Shlens “A tutorial on Independent Component Analysis”. Mountain View, CA 94043