FACULTAD DE BIOLOGÍA, CIENCIAS AMBIENTALES Y QUÍMICA

GRADO EN BIOLOGÍA

TRABAJO DE FIN DE GRADO

TRATAMIENTO DE DATOS QUÍMICO-FORENSES PARA LA DISCRIMINACIÓN DE FLUIDOS BIOLÓGICOS EN MATERIALES SUPERABSORBENTES

Autor: Ignacio Pachón Jiménez
Tutor/es: Carmen García Ruiz

Año 2016

FACULTAD DE BIOLOGÍA, CIENCIAS AMBIENTALES Y QUÍMICA
GRADO EN BIOLOGÍA
TRABAJO DE FIN DE GRADO

TRATAMIENTO DE DATOS QUÍMICO-FORENSES PARA LA DISCRIMINACIÓN DE FLUIDOS BIOLÓGICOS EN MATERIALES SUPERABSORBENTES

Tribunal de calificación:

(Firma)

Presidente: _____

(Firma)

Vocal 1º: _____

(Firma)

Vocal 2º: _____

Calificación: _____

Fecha: _____

Año 2016

Universidad de Alcalá

FACULTAD DE BIOLOGÍA, CIENCIAS AMBIENTALES Y QUÍMICA

INFORME PARA LA DEFENSA PÚBLICA DEL TRABAJO DE FIN DE GRADO

D/Dª CARMEN GARCIA RUIZ, profesor del Departamento de QUÍMICA ANALÍTICA, QUÍMICA FÍSICA E ING QUÍMICA de la UAH, como tutor del Trabajo de Fin de Grado en BIOLOGÍA de D/Dª IGNACIO PACHÓN JIMENEZ titulado TRATAMIENTO DE DATOS QUÍMICO-FORENSES PARA LA DISCRIMINACIÓN DE FLUIDOS BIOLÓGICOS EN MATERIALES SUPERABSORBENTES

INFORMA:

QUE ESTE TRABAJO HA SIDO REALIZADO Y REDACTADO POR EL MENCIONADO ESTUDIANTE BAJO MI DIRECCIÓN Y CON ESTA FECHA AUTORIZO A SU PRESENTACIÓN Y DEFENSA PÚBLICA.

Alcalá de Henares 7 de SEPTIEMBRE de 2016

Firma del tutor

Fdo.: CARMEN GARCIA RUIZ

Firma del cotutor (si lo hubiere)

Fdo.:

Agradecimientos:

A Carmen García Ruiz, Félix Zapata Arráez, Inés Gregorio Martins, Fernando Ortega Ojeda y todo el equipo de INQUIFOR;

A mis padres y hermanos, por todo el apoyo brindado.

To Bryan A. Hanson for his inestimable help with R code and ChemoSpec package.

A Alberto Jiménez-Valverde por su orientación y disposición.

A Eneritz Lamikiz Moreno, por todo.

Índice de contenidos

Resumen

En este trabajo fin de grado se persigue analizar los datos espectrales, obtenidos por espectroscopía infrarroja, de manchas de fluidos biológicos sobre materiales superabsorbentes, con el fin de discriminar la presencia de semen, fluido de interés en muchos delitos sexuales.

El trabajo consistió en el análisis estadístico de los espectros obtenidos mediante espectroscopía infrarroja con transformada de Fourier y reflexión total atenuada (ATR-FTIR) de muestras de distintos fluidos biológicos (fluido vaginal, orina y semen), sobre una serie de materiales superabsorbentes comerciales (compresas, salvaslips y pañales). El análisis de los espectros se ha realizado mediante: corrección de línea base, normalización y suavizado; y posteriormente por técnicas quimiométricas como el Análisis de Componentes Principales (PCA) y coeficiente de correlación de Pearson. Para validar el modelo se generaron curvas ROC. El procedimiento estadístico se ha ejecutado íntegramente con el software R.

Los resultados muestran que el PCA no permite la distinción entre semen y fluido vaginal. Sin embargo, la estadística bayesiana aplicada conlleva un alto rendimiento en la discriminación de estos fluidos. Por ello, se ha confirmado la posibilidad de llevar a cabo una metodología mejorada para la determinación de la presencia de semen en distintos materiales superabsorbentes en presencia de otros fluidos biológicos (fluido vaginal y orina) mediante espectroscopía infrarroja.

Abstract

In this thesis, the objective pursued is to analyze the spectral data obtained by infrared spectroscopy from biological fluid stains on super-absorbent materials, in order to discriminate the presence of semen, which is a fluid of interest in many sexual crimes.

The study consisted on the statistical analysis of the spectra obtained by infrared spectroscopy with Fourier transformed and attenuated total reflection (ATR-FTIR) of samples from several biological fluids (vaginal fluid, urine and semen), on a series of commercial super-absorbent materials (pads, salvaslips, and diapers). The spectra analysis has been performed by: baseline correction,

normalization, and smoothing followed by chemometric techniques like Principal Component Analysis (PCA) and Pearson's correlation coefficient. For the validation of the model, ROC curves were generated. The statistical procedure has been executed entirely in R software.

The results show that the PCA does not allow the differentiation between semen and vaginal fluid. However, the applied Bayesian statistic yields a high efficiency for the discrimination of these fluids. Finally, the possibility of carrying out an enhanced methodology for the determination of the presence of semen on different super-absorbent materials, mixed with other biological fluids (vaginal fluid and urine), by infrared spectroscopy, has been confirmed.

Palabras clave: Química forense, espectroscopía infrarroja, fluidos biológicos, semen, fluido vaginal, orina, materiales superabsorbentes, análisis cualitativo, Análisis de Componentes Principales (ACP), ROC.

Keywords: Forensic chemistry, infrared spectroscopy, biological fluids, semen, vaginal fluid, urine, super-absorbent materials, qualitative analysis, Principal Component Analysis (PCA), ROC.

1. Introducción

La ciencia forense se define como la aplicación de conocimientos científicos a problemas legales e investigaciones criminales (West's Encyclopedia of American Law, 2008). Engloba un amplio abanico de disciplinas científicas, siendo algunas la genética forense, antropología forense, medicina forense, química forense y entomología forense, entre otros. Todos estos campos componen la Criminalística, que persigue ayudar en la reconstrucción del hecho delictivo mediante las ciencias forenses. Existe legislación en cuanto al procedimiento que ha de seguirse desde la comisión de un delito hasta la aplicación del código penal por parte de un juez, de obligado cumplimiento por parte de todo el personal que pueda estar relacionado con la observación, la recogida y el transporte de evidencias, y la investigación y la comunicación de los resultados de esta.

A nivel judicial tiene una gran importancia el tratamiento estadístico de los datos porque permiten establecer el grado de certeza de un hecho.

La Real Academia Española define dato como 1. "la información sobre algo concreto que permite su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho", 2. "documento, testimonio, fundamento" y 3. "información dispuesta de manera adecuada para su tratamiento por un ordenador" (Real Academia Española, '*Diccionario de la Lengua Española*'). La estrategia para el análisis de datos obtenidos en el laboratorio consiste en la obtención de los datos, el procesamiento de estos, el análisis propiamente dicho y su posterior evaluación.

El análisis univariante permite realizar predicciones o explicaciones basadas en análisis de correlación y regresión. A diferencia de los análisis de datos univariantes, el análisis multivariante permite realizar predicciones o explicaciones basadas en análisis discriminantes y correlaciones canónicas y, como su propio nombre indica, estudiar la relación que existe entre más de una variable al mismo tiempo, es decir, su relación estructural. El ejemplo más claro de análisis multivariante es la predicción meteorológica, donde para predecir la temperatura, es necesario conocer la estación, la humedad, las condiciones atmosféricas… El análisis multivariante tiene aplicación en muchos sectores,

como las ciencias sociales (se fundó en el área de la psicología), ciencias económicas, ciencias físicas, etc. (Swarbrick, B., 2012).

"Se denomina modelo al conjunto de conocimientos sobre un sistema en el momento de la investigación. Se puede utilizar un buen modelo para predecir eventos futuros con confianza" (Swarbrick, B., 2012). Entendemos por sistema al conjunto de variables y la relación que tienen entre sí, de tal modo que la variación de una de ellas contribuye en mayor o menor medida a la variación de las demás. Debe tenerse en cuenta que, si las variables dependen totalmente del azar, esto no va a contribuir positivamente a la validez del modelo.

Un sistema de espectros IR se compone de longitudes de onda (transformadas a número de onda en cm^{-1}; observaciones), intensidad de la reflectancia (*R*) de cada espectro (transformadas a $Log\ (\frac{1}{R})$; variables), y una gran cantidad de espectros de diferente naturaleza (muestras). Además, conocemos la naturaleza de los espectros, es decir, qué fluido o fluidos contienen.

El análisis exploratorio multivariante por el método Análisis de Componentes Principales (PCA, por sus siglas en inglés), "es conocido como el caballo de batalla de los métodos de análisis multivariante". "El PCA aporta una de las herramientas gráficas más potentes para la comprensión de las relaciones entre muestras y variables." Este método es interesante también porque revela las tendencias básicas de los espectros y elimina el ruido que contienen, a la vez que reduce significativamente la cantidad de datos sin perder demasiada información (Swarbrick, B., 2012; Shlens, J., 2014).

Del PCA surgen tres conceptos, principalmente. Estos son: *scores*, *loadings* y Componentes Principales (PCs, por sus siglas en inglés). "El PCA transforma el sistema de coordenadas original", siendo los PCs los ejes, que apuntan hacia la varianza más alta (el PC1), y las varianzas subsiguientes (PC2, PC3…). Los componentes principales son las nuevas variables sin correlación lineal. Los *scores* son equivalentes a las bandas de los espectros iniciales, pero transformados al espacio tridimensional mediante el cálculo de varianzas (Figura 1). Posteriormente pueden ser representados tanto en dos dimensiones como en tres dimensiones. El PCA también aporta información sobre la contribución

de cada variable original a los *scores* mediante los *loadings*, es decir, el factor que multiplicado por la variable original da lugar a los *scores* (Bruker Daltonics, n.d.).

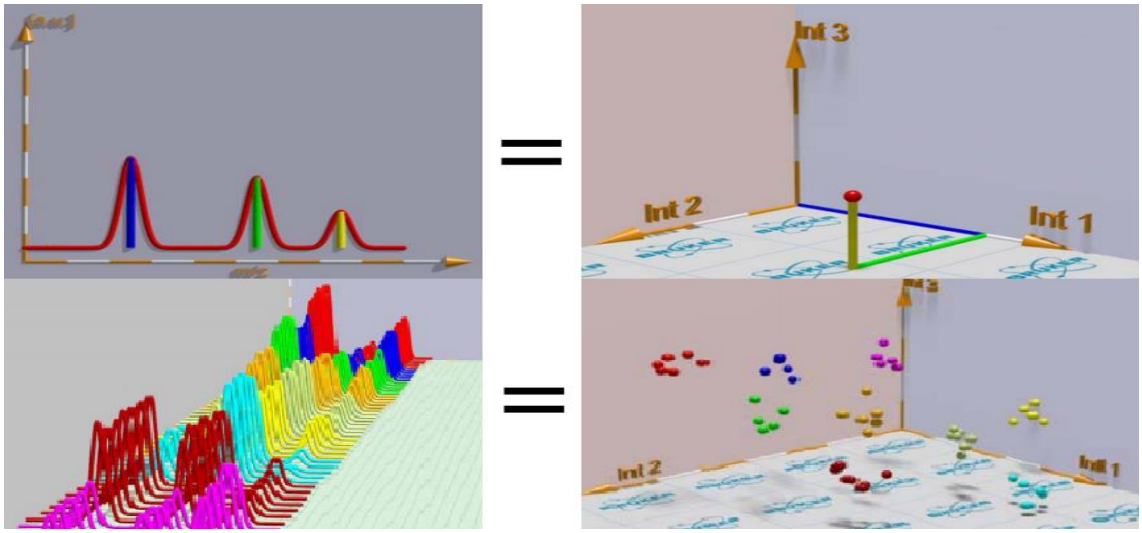

Figura 1. Transformación de los espectros a *scores* mediante PCA. Adaptada de *(Bruker Daltonics, n.d.)*.

En ocasiones interesa conocer el grado de similitud o covarianza entre una colección de datos conocidos. Para ello, uno de los métodos más utilizados es la correlación lineal de Pearson.

El método de correlación lineal de Pearson (r) (Camacho Martinez-Vara de Rey, C. G., n.d.), que se define como la media de los productos cruzados de las puntuaciones estandarizadas de dos valores *X* e *Y*, que corresponden a los valores de intensidad de los espectros. Consiste en la comparación de dos variables relacionadas linealmente mediante la expresión:

$$r_{xy} = \frac{\sum Z_x Z_y}{N}$$

donde r_{xy} es el coeficiente de correlación de Pearson entre dos muestras, *Z* es la puntuación estandarizada y *N* es el número de valores, dando lugar a un valor entre -1 y 1, siendo -1 una perfecta correlación negativa, 1 una perfecta correlación positiva y 0 una correlación nula. La expresión para obtener las puntuaciones estandarizadas:

$$Z_x = \frac{X - \overline{X}}{S_x};\ Z_y = \frac{Y - \overline{Y}}{S_y}$$

donde *X* e *Y* son los valores de intensidad de los espectros a correlacionar y *S* la desviación tipo, que se obtiene de la siguiente expresión:

$$S_x = \sqrt{\frac{\sum X^2}{N} - \overline{X^2}}\ ;\ S_y = \sqrt{\frac{\sum Y^2}{N} - \overline{Y^2}}$$

Los métodos anteriormente mostrados están basados en la estadística convencional ya que permiten obtener conclusiones basándose en la frecuencia estadística (estadística frecuentística), o aceptación o rechazo de una hipótesis, dentro del marco del estudio que se esté realizando. Cuando se posee información externa al estudio que se realiza y esta se incorpora al análisis, permitiendo el cálculo de probabilidades, hablamos de estadística bayesiana. Al incorporar información externa, por ejemplo, agrupando los datos según características comunes conocidas, se generan variables cualitativas discretas que pueden analizarse mediante un análisis cualitativo.

La Real Academia Española define el análisis cualitativo como un "análisis que tiene por objeto identificar los componentes de una sustancia" (Real Academia Española, '*Diccionario de la Lengua Española*'). Hasta ahora, todos los métodos comentados han tenido por finalidad discriminar los componentes de una batería de datos.

Un tipo de análisis cualitativo es el que se lleva a cabo mediante las conocidas como curvas ROC (característica operativa del receptor). Estos análisis permiten llevar a cabo un entrenamiento del modelo, y obtener una valoración del rendimiento de la clasificación predictiva de este, es decir, averiguar la capacidad del modelo de predecir eventos favorables o desfavorables. La finalidad de este análisis es conocer la probabilidad de obtener resultados positivos o negativos y su tasa de acierto. Esto es: Verdaderos Positivos y Falsos Negativos (VP y FN; resultados positivos), y Verdaderos Negativos y Falsos Positivos (VN y FP; resultados negativos). A partir de estos valores se obtiene (Wen Zhu, Nancy Zeng, Ning Wang, 2010):

$Exactitud = \frac{VP+VN}{VP+VN+FP+FN}$ = Tasa de acierto en la clasificación de muestras

$Sensibilidad = \frac{VP}{VP+FN}$ = Tasa de acierto de resultados positivos

$Especificidad = \frac{VN}{VN+FP}$ = Tasa de acierto de resultados negativos

$Precisión = \frac{VP}{VP+FP}$ = Tasa de acierto de clasificaciones positivas

Además, se obtiene una calificación (razón de verosimilitudes positiva o *Diagnostic Likelihood Ratio* positivo; DLR+) para el test diagnóstico de los resultados positivos y negativos, es decir, la razón de Verdaderos Positivos respecto de la razón de Falsos Positivos. A mayor DLR+, mayor probabilidad de hallar un Verdadero Positivo que un Falso Positivo. A modo de ejemplo, un *DLR+=7* quiere decir que el hallazgo de datos que contienen la característica es siete veces más frecuente; en contraposición, un *DLR+=1* quiere decir que el hallazgo de datos que contienen la característica es tan frecuente como el de datos que no la contienen (Puebla Arredondo, C., 2010). Otra forma de calcular este valor es mediante la siguiente expresión, a partir de las tasas expuestas anteriormente:

$$DLR+= \frac{Sensibilidad}{(1-Especificidad)} \qquad DLR-= \frac{(1-Sensibilidad)}{Especificidad}$$

Los resultados de este valor califican la utilidad del método según la Tabla 1. Esto nos permite elegir entre varios test confirmativos o criterios cuál es el que aporta una fiabilidad más alta. Otro indicador muy utilizado para calificar el rendimiento es el área bajo la curva ROC (AUC, Tabla 2)

Tabla 1. Calificación del modelo a partir de su DLR+. Adaptada de *(Puebla Arredondo, C., 2010).*

DLR+	Calificación
DLR+ > 10	**Excelente**
6 < DLR+ < 10	**Bueno**
3 < DLR+ < 6	**Regular**
1 < DLR+ < 3	**Malo**
DLR+ ≤ 1	**Inútil**

Tabla 2. Calificación del Rendimiento por el Rango del área bajo la curva ROC. Adaptada de *(Hosmer, D. W., Lemeshow, S., 2000*; *Wen Zhu et al, 2010)*.

AUC	Calificación
0.9 < AUC < 1.0	**Excepcional**
0.8 < AUC < 0.9	**Excelente**
0.7 < AUC < 0.8	**Aceptable**
0.6 < AUC < 0.7	**Regular**
AUC = 0.5	**No hay discriminación**

2. Hipótesis y objetivos

Dado que la estadística bayesiana analiza la probabilidad de que un hecho sea cierto, desarrollamos una hipótesis para este trabajo fin de grado: es posible combinar la estadística clásica con la estadística bayesiana para aportar cierto grado de certeza al análisis de datos espectrales de manchas de fluidos biológicos como semen, fluido vaginal y orina en materiales superabsorbentes.

Para ello, se plantearon los siguientes objetivos concretos:

- Familiarización con conceptos básicos de espectroscopía infrarroja y la información que aporta (espectros).
- Conocimiento del funcionamiento del software R, un programa muy utilizado en ciencias de la vida para el análisis estadístico.
- Aplicación de las herramientas que aporta R para el tratamiento estadístico de datos espectrales.
- Evaluación de la capacidad del procedimiento estadístico desarrollado para conocer el grado de certeza de un hecho químico.

3. Materiales y Métodos

a. Datos utilizados: espectros infrarrojos de manchas de fluidos sobre materiales superabsorbentes.

En este trabajo fin de grado se han utilizado 236 espectros de manchas preparadas de semen (53 espectros), orina (58 espectros), fluido vaginal (45 espectros), y muestras correspondientes a mezclas (80 espectros). Las mezclas contenían orina, semen y fluido vaginal, excepto las preparadas para pañales en las que no se empleó fluido vaginal. Además, se incluyeron 170 espectros correspondientes a controles negativos en los soportes de material superabsorbente que carecían de muestra (que denominaremos “blancos”). En total, por tanto, la colección de espectros estudiados constó de 406 espectros.

Las manchas estudiadas estaban sobre compresas, pañales y salvaslips de nueve marcas comerciales distintas: Ausonia (Procter & Gamble, Ohio, EE.UU.),

Bonté (Ontex ID, Segovia, España), Carefree (Johnson & Johnson, New Jersey, USA), Deliplus (Mercadona, Valencia, España), Evax (Procter & Gamble, Ohio, EE.UU.), Chelino (Laboratorios INDAS, Madrid, España), Dodot (Procter & Gamble, Ohio, EE.UU.) y Renova (Renova, Torres Novas, Portugal). Los espectros estudiados correspondieron a la primera capa absorbente de estos materiales superabsorbentes.

Los espectros empleados se habían obtenido por la técnica de espectroscopía infrarroja con transformada de Fourier y Reflexión Total Atenuada Transformada de Fourier (ATR-FTIR por sus siglas en inglés), midiendo en diversos puntos de una mancha para cada preparación de fluido sobre material superabsorbente (una media de aproximadamente nueve puntos por muestra de fluidos y trece por blanco). Cada espectro fue una media de dieciséis medidas sobre el mismo punto. Se empleó el software OMNIC^{TM} Specta (ThermoScientific^{TM}) en el intervalo de números de onda entre 800 y 2000 cm^{-1}, y empleando la señal del logaritmo decimal de la inversa de la Reflexión (Log 1/R). En la bibliografía se ha encontrado que estas bandas corresponden a la vibración de los enlaces que contienen los aminoácidos constituyentes de las proteínas presentes en los fluidos (sobre todo amida I y amida II) (Pretsch, E., Bülhmann, P. Badertscher, M., 2009).

b. Tratamiento de datos (espectros) mediante el software R

El tratamiento estadístico de los espectros se realizó íntegramente con el software R (versión 3.2.3 "Wooden Christmas-Tree") (R Core Team, 2015). Para ello se utilizaron los siguientes paquetes de funciones: "baseline" (Kristian Hovde Liland and Bjørn-Helge Mevik, 2015), "ChemoSpec" (Bryan A. Hanson, 2016), "graphics" (R Core Team, 2015), "Hmisc" (Frank E. Harrel Jr. et al., 2015), "IDPmisc" (Rene Locher, Andreas Ruckstuhl et al., 2012), "OptimalCutpoints" (Mónica López-Ratón et al., 2014), "ROCR" (Sing, T. et al., 2005), "R.utils" (Henrik Bengtsson, 2015), "signal" (signal developers, 2013) y "stats" (R Core Team, 2015).

La selección de los paquetes necesarios para el estudio consistió en la búsqueda de las funciones adecuadas en las páginas web de ayuda de R (comando *help()*); repositorios CRAN (R-project, '*The Comprehensive R Archive*

Network') y GitHub (Github, 2016); páginas web R-bloggers (R-bloggers, 2016) e inside-R (Revolution Analytics, 2015); en los foros de ayuda Stack Exchange (stackoverflow y Cross Validated) (Stack Exchange, 2016); y por recomendaciones de los especialistas: Profesor Alberto Jiménez-Valverde (Universidad de Alcalá; material de sus clases magistrales del Máster Universitario de Investigación en Ciencias 2015-16) y Profesor Bryan A. Hanson (DePauw University; autor del paquete de funciones "ChemoSpec"). Este último, aportó las funciones personalizadas para corrección de línea base, suavizado y normalización de los espectros, además de consejos e indicaciones para el uso del paquete, por correspondencia vía correo electrónico.

Para el aprendizaje del uso del software se requirió el estudio de manuales como "R para principiantes" (Paradis, E., 2003), "Computación Estadística: Introducción a R" ('*Computación estadística: introducción a R*', n.d.), "*An introduction to the prospectr package*" (Stevens, A., Ramírez-López, L., 2014) y "*ChemoSpec: An R Package for Chemometric Analysis of Spectroscopic Data*" (Bryan A. Hanson, 2015).

i. Procesamiento de los espectros

Para el procesamiento de los espectros se necesitó un archivo de texto (de formato *.CSV*) de cada espectro con los datos de este en dos columnas, una correspondiente a las observaciones (eje de ordenadas) y otra a las variables (eje de abscisas), exportándolos desde OMNIC™ Specta. Posteriormente, se procedió a la lectura de los archivos mencionados con R mediante la función correspondiente de *ChemoSpec*, *files2SpectraObject*. Con esta función se designó un criterio de agrupamiento por tipo de muestra (blanco, semen, fluido vaginal, orina o mezcla), y se designó un color a cada una. De esta forma se comparó la tendencia de cada grupo observando las bandas características, determinando el intervalo que correspondía mayoritariamente a la composición de los materiales superabsorbentes (región soporte), y el intervalo en el que mejor se percibía una distinción entre el semen y los demás fluidos (región ID; véase la Figura A1 del anexo). Una vez seleccionada la región ID de interés, se realizó un pretratamiento de los espectros disponibles para recortar en un intervalo de número de onda comprendido entre 1500 y 1690 cm^{-1} (véase la Figura A2 del anexo) con la función *removeFreq* (*ChemoSpec*)

Para una mejor visualización de la tendencia general de las firmas espectrales de cada grupo, se obtuvo con la función *surveySpectra* la representación gráfica de las medias de los espectros junto a la desviación estándar de la media (argumento *method="sd"*).

Para corregir la contribución del ruido y escalar los espectros, se desarrolló un tratamiento de los datos que consistió en tres fases: corrección de línea base, suavizado y normalización, en este orden. Este tratamiento fue posible mediante la construcción de las funciones personalizadas que se detallan a continuación.

La corrección de línea base se llevó a cabo en dos pasos. Primero, se buscó el valor mínimo de cada espectro, y se restó a todos los puntos del espectro: $f(x) = x - \min(x)$. Posteriormente, se calculó la regresión lineal de los espectros respecto a las frecuencias (con la función *lm*) y se obtuvieron las predicciones (con la función *predict*), que fueron finalmente restadas a los espectros originales.

El suavizado consistió en la aplicación del filtro Savitzky-Golay a los espectros, aplicando la función *sgolayfilt* (paquete *signal*) con los siguientes parámetros: polinomio de orden dos (p=2), longitud del filtro o ventana once (cinco puntos a la izquierda y cinco a la derecha, además del central: n=11) y sin generar derivada (m=0). Fue necesario adaptar la función a la clase *Spectra* del paquete *ChemoSpec*.

Finalmente, los espectros se normalizaron restando el pico mínimo y dividiendo por el máximo todos los puntos de los espectros. Como consecuencia, quedaron todos escalados en una intensidad entre cero y la unidad, como se observa en las Figuras A3 y A4 del anexo.

ii. Análisis exploratorio multivariante de los espectros procesados

Se realizó un Análisis de Componentes Principales (PCA, por sus siglas en inglés) mediante la función *c_pcaSpectra* de *ChemoSpec*. Esta función se aplicó a los espectros no escalados. Se realizó el PCA de todos los datos (espectros) y, también, excluyendo los blancos para visualizar la capacidad discriminatoria del análisis respecto a los grupos de muestras de fluidos (semen, orina y fluido vaginal).

iii. Estudio de la variabilidad Inter e Intraespecífica de los espectros

La variabilidad de los espectros se estudió mediante la correlación de todos los espectros de la población semen consigo mismos, dando lugar a la llamada Intravaliabilidad; la correlación de los espectros de la población semen con los de la población no semen da lugar a la llamada Intervariabilidad. De esta forma se consiguió visualizar mediante un histograma (frecuencia vs coeficiente de correlación de Pearson) el grado de covarianza entre ambas poblaciones. Además, se han seleccionado dos espectros que contienen semen, fluido vaginal y orina para proponer dos escenarios hipotéticos y comprobar cómo se comportan sus correlaciones con espectros de semen. Todo esto permitió aproximar la frontera entre un resultado discriminativo positivo y negativo.

Lo primero que se hizo fue exportar como archivo de texto la transpuesta de la matriz de datos procesados, y leer este desde R. Posteriormente, se dividió la matriz general en submatrices (semen, no semen y mezclas). La submatriz de "no semen" contenía todos los espectros de orina y fluido vaginal. En el caso de las mezclas, se seleccionaron solamente dos espectros de todos los disponibles (concretamente, el espectro de menor intensidad y el de mayor intensidad, que se traduce en el caso más desfavorable y más favorable, respectivamente). Resultó más eficiente trabajar con estas matrices como marcos de datos (*as.data.frame*), dado que esta opción implicó no tener que eliminar las correlaciones diagonalmente simétricas que resultan de la correlación.

La intervariabilidad se estudió correlacionando todos los espectros excepto los blancos (espectros de semen contra espectros de no semen) y la intravariabilidad se estudió correlacionando todos los espectros de semen consigo mismos, mediante el coeficiente de correlación de Pearson empleando la función *cor.test*, del paquete *stats*. Para generar los histogramas de coeficientes de correlación se utilizó la función *hist* del paquete *graphics*.

iv. Análisis cualitativo mediante curvas ROC

Para el análisis ROC del modelo cualitativo, se obtuvo un objeto de clase *prediction* (paquete *ROCR*) compuesto por dos vectores (coeficientes de correlación de Pearson de Inter e Intravariabilidad por un lado y valores indicadores de positivo o negativo a semen por otro lado) a partir del cual se

generó la curva ROC con la función *performance* del paquete *ROCR* (argumentos *"tpr"* y *"fpr"* en el mismo comando). Con la misma función, se generaron las curvas de Verdaderos y Falsos Positivos y Verdaderos y Falsos Negativos (VP, FP, VN y FN, respectivamente; argumentos "tpr", "fpr", "tnr" y fnr" por separado). Para obtener los mejores valores para la selección de un umbral, se obtuvo un objeto de clase *optimal.cutpoints* (del paquete *OptimalCutpoints*), que aportaba una amplia variedad de criterios para ello.

En el anexo se muestra una síntesis del código R desarrollado para el procesamiento de los espectros, así como de todos los análisis estadísticos.

4. Resultados y discusión

En primer lugar, se procedió a una visualización de espectros considerados (Fig. A1 y A2 del anexo). Una vez determinado el intervalo de número de onda de interés (1500 a 1690 cm^{-1}) se procesaron los espectros para obtener su media y desviación estándar. La Figura 2 muestra la media y desviación estándar de los 170 espectros considerados para los blancos, 58 espectros considerados para las manchas de orina, 53 espectros considerados para las manchas de semen, 45 espectros considerados para el fluido vaginal y 80 espectros considerados para las manchas de mezclas de fluidos. Destacar que, en mayor o menor medida, la desviación estándar se debe a la variación de concentración de los fluidos en los materiales superabsorbentes. Los espectros que resultaron del pretratamiento se muestran en las Figuras A3 y A4 del anexo.

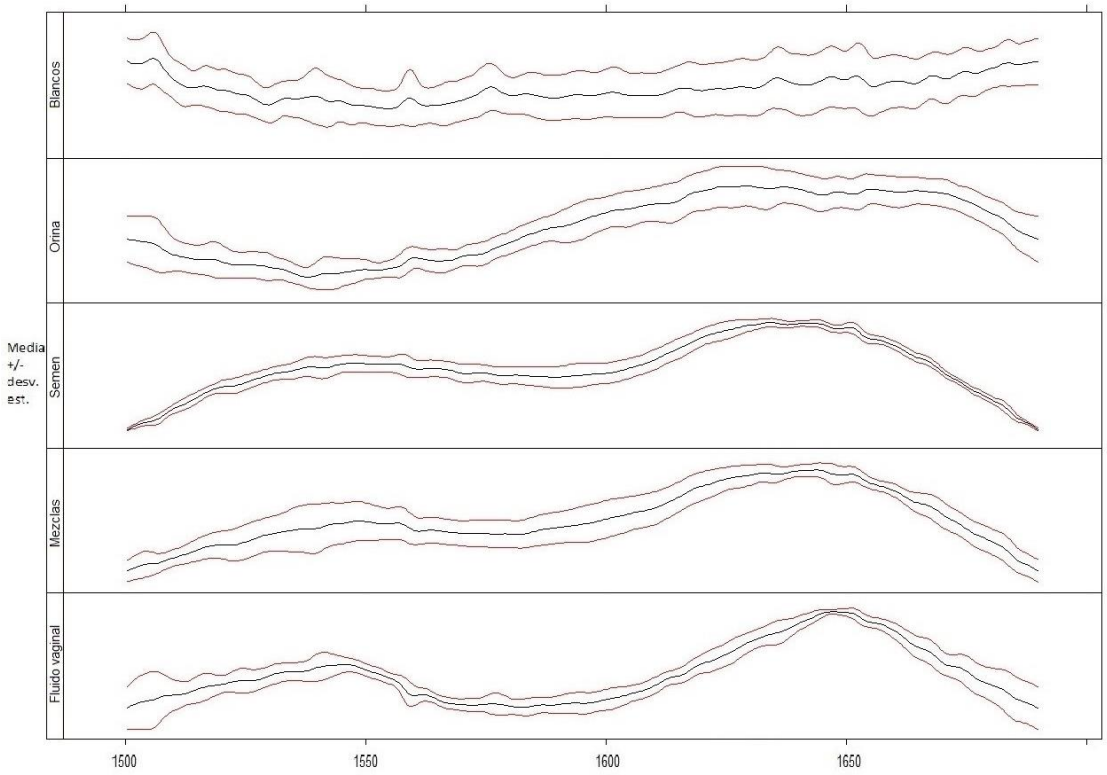

Figura 2. Medias de los espectros y sus desviaciones estándar, posteriores al tratamiento de corrección de línea base, suavizado y normalización. De arriba a abajo: Blancos, Orina, Semen, Mezclas y Fluido vaginal. Eje horizontal, números de onda en cm^{-1}. Eje vertical, escala de intensidad.

a. Análisis de Componentes Principales

Con el fin de estudiar si existía una diferenciación entre grupos de espectros de manchas de semen, fluido vaginal, orina, mezclas y blancos, se realizó un PCA. En la Figura 3, donde se muestran los scores respecto al PC1 (primer componente principal) y al PC2 (segundo componente principal). Se observa claramente la diferenciación de las manchas de orina (morado). Sin embargo, en cuanto al semen (verde) y al fluido vaginal (marrón), como se esperaba por su composición similar, la distribución de los scores no permite su discriminación. Las mezclas (azul), se muestran más cerca del fluido vaginal y del semen, aunque se aprecia una gran dispersión. La gran dispersión de los scores de blancos puede deberse a la heterogeneidad de los soportes además de la complejidad de sus materiales.

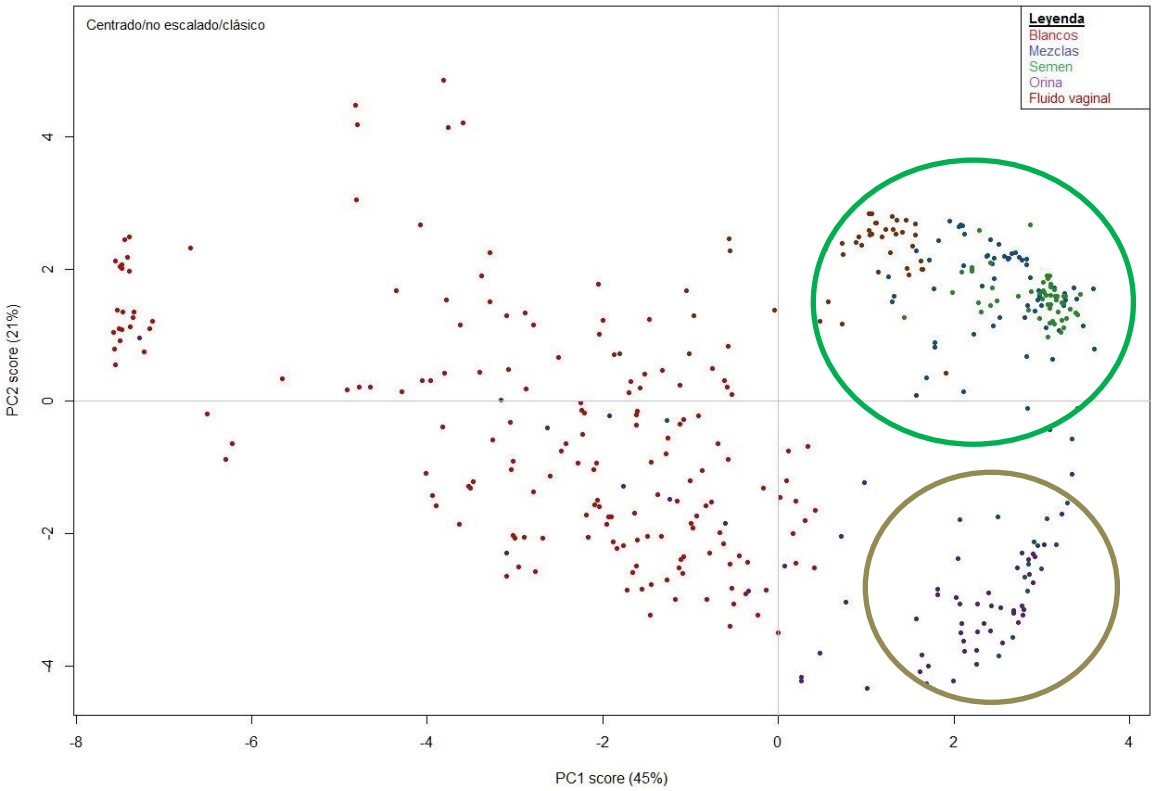

Figura 3. Representación de los *scores* (PCA) en dos dimensiones del PC1 (45%; eje horizontal) vs el PC2 (21%; eje vertical).

En la Figura 4 se muestran los *loadings* correspondientes al PC1 y al PC2 del *score* realizado con la matriz de espectros de las manchas de fluido vaginal. Semen, orina, mezclas y blancos.

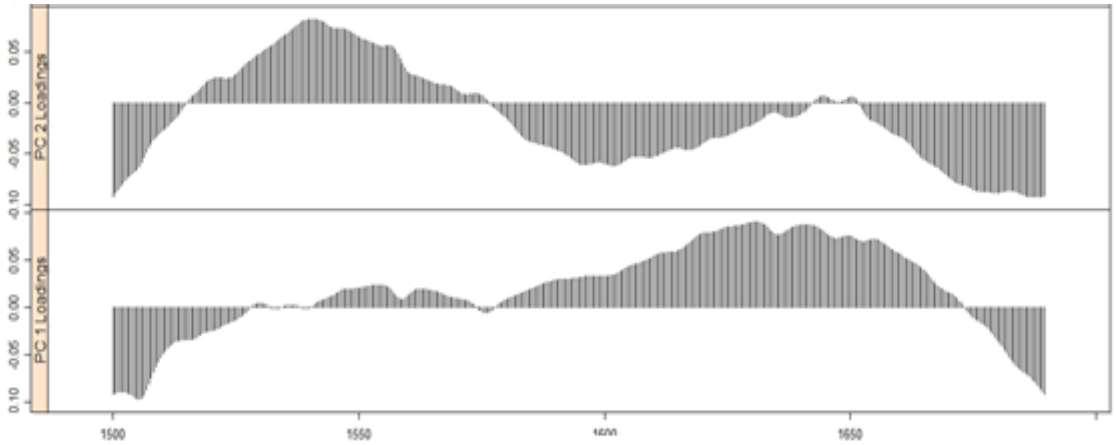

Figura 4. Representación de los *loadings* (PCA) de los PC1 y PC2 (de abajo a arriba). Editada a partir de la imagen que genera el software R. Eje vertical valor de los *loadings* y eje horizontal número de onda (cm^{-1}).

En la Figura 4, el primer componente principal (PC1) contribuye positivamente a la distribución de los scores el rango que comprende

aproximadamente entre 1530 y 1675 cm^{-1}, destacando alrededor de 1630 cm^{-1}. El segundo componente principal (PC2) contribuye positivamente a la distribución de los scores las bandas del intervalo entre 1515 y 1575 cm^{-1}, destacando aproximadamente a 1540 cm^{-1}.

Una vez comprobada la aportación de este análisis al estudio, se decidió analizar de nuevo los mismos grupos de espectros excluyendo los blancos, para eliminar la contribución de estos al análisis multivariante. Se comprobó que, de esta forma, los grupos de manchas de fluidos y mezclas se distinguían bien. La Figura 5 muestra la distribución de los *scores* de los dos primeros componentes principales (PC1 y PC2), donde se confirma la discriminación entre semen y fluido vaginal (principal interés del estudio). También se puede observar la discriminación de la orina frente al semen y al fluido vaginal. Las mezclas muestran una dispersión más o menos homogénea entre las nubes de *scores* de los tres fluidos que las componen. El hecho de que la tendencia de las mezclas sea tan próxima a la del semen aporta una información positiva hacia el objetivo del estudio, ya que incita a creer que una mezcla que contenga semen se clasificará como positivo en semen en un gran número de ocasiones (aunque esto se discutirá en el próximo apartado).

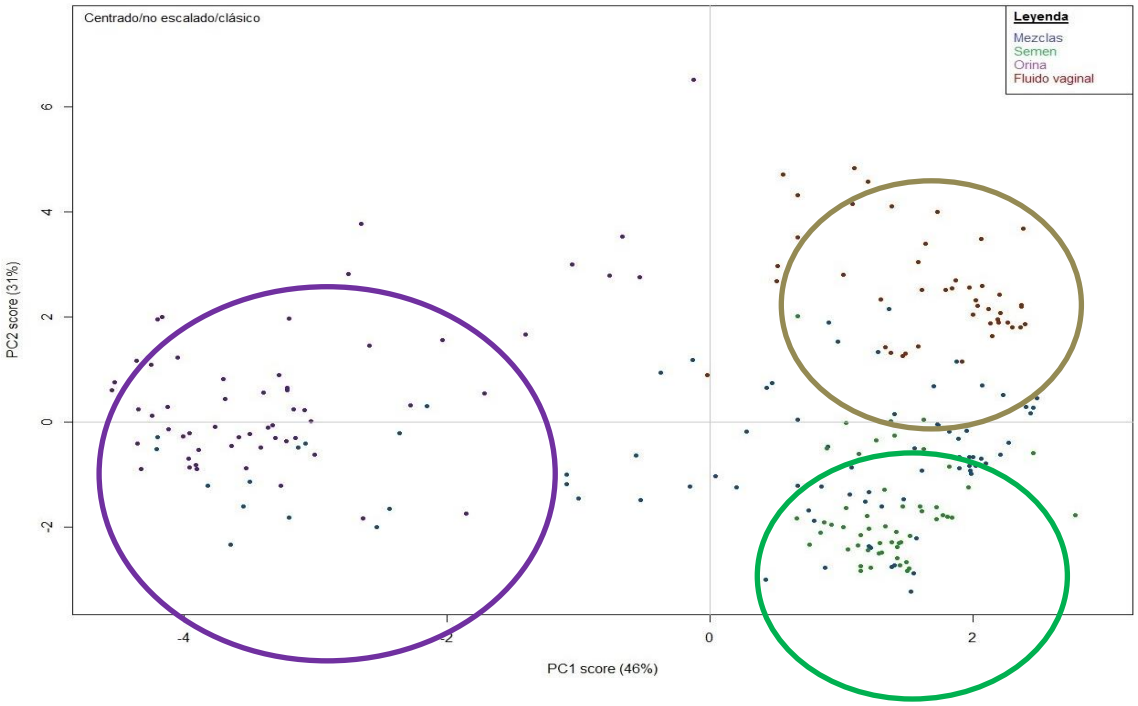

Figura 5. Representación en dos dimensiones de los *scores* (PCA) de los PC1 (46%; eje horizontal) vs. PC2 (31%; eje vertical), excluyendo los blancos.

En cuanto a los *loadings* de este segundo PCA, se observa en la Figura 6 cómo contribuyen los dos primeros componentes principales a la discriminación de los tres fluidos y de las mezclas. El primer componente principal (PC1) contribuye a la distribución de los scores principalmente en los rangos entre 1510 y 1575 cm^{-1} y entre 1640 y 1655 cm^{-1}, aproximadamente, dando la mayor contribución alrededor de 1540 cm^{-1}. El segundo componente principal (PC2) contribuye en los rangos entre 1500 y 1520 cm^{-1} y entre 1650 y 1690 cm^{-1}, aproximadamente, siendo igual de contribuyente en ambos rangos.

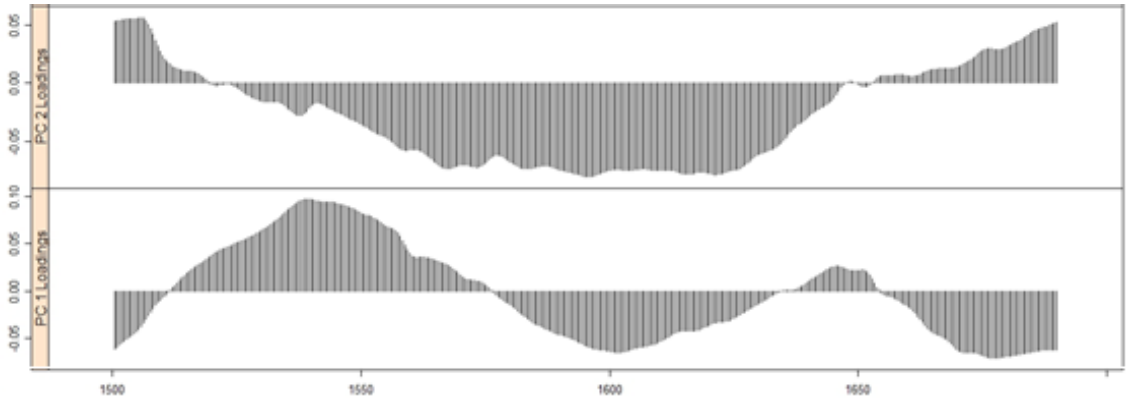

Figura 6. Representación de los *loadings* (PCA) de los PC1, PC2 (de abajo a arriba). Eje vertical valor de los *loadings* y eje horizontal número de onda (cm^{-1}).

b. Variabilidad Inter e Intraespecífica respecto al semen

Interesa conocer el grado de similitud o covarianza de los espectros, pero en este caso, debido al interés del estudio, se estudió la correlación de los espectros dividiéndolos en dos poblaciones o grupos: Semen y No Semen. La finalidad es conocer el grado de covarianza (y, consecuentemente, la capacidad de diferenciación) entre una colección de espectros conocidos que contienen únicamente semen y otra que contiene fluidos corporales distintos (fluido vaginal y orina).

En este apartado, se relacionan los espectros de semen consigo mismos (intravariabilidad), y con los espectros de semen con los que no contienen semen (intervariabilidad). Se consideraron espectros que carecen de semen aquellos que poseen muestra de un fluido distinto al semen, que en el caso que se estudia son: espectros de manchas de fluido vaginal y de orina. Además, se correlacionaron los espectros de manchas de semen con los espectros de mezclas, para proponer escenarios que permitieran validar el modelo.

La correlación de los espectros fue posible tras confeccionar las matrices correspondientes a los grupos mencionados (semen, no semen y mezclas) y se realizó mediante el coeficiente de Pearson. La Figura 7 muestra el resultado de las correlaciones Semen vs Semen y Semen vs No semen, donde se puede ver la frecuencia relativa de las correlaciones, y como consecuencia, el grado de relación entre dichos grupos. Se puede ver que a partir del 88,5 % (r=0,885) la frecuencia de coeficientes de correlación del grupo de semen vs semen (intravariabilidad) es mayor que la del grupo de semen vs no semen (intervariabilidad), es decir, que si la correlación entre un espectro de semen conocido y uno desconocido es mayor de 0,885 es probable que en el último haya semen. Sin embargo, esto no es suficiente para determinar un valor umbral.

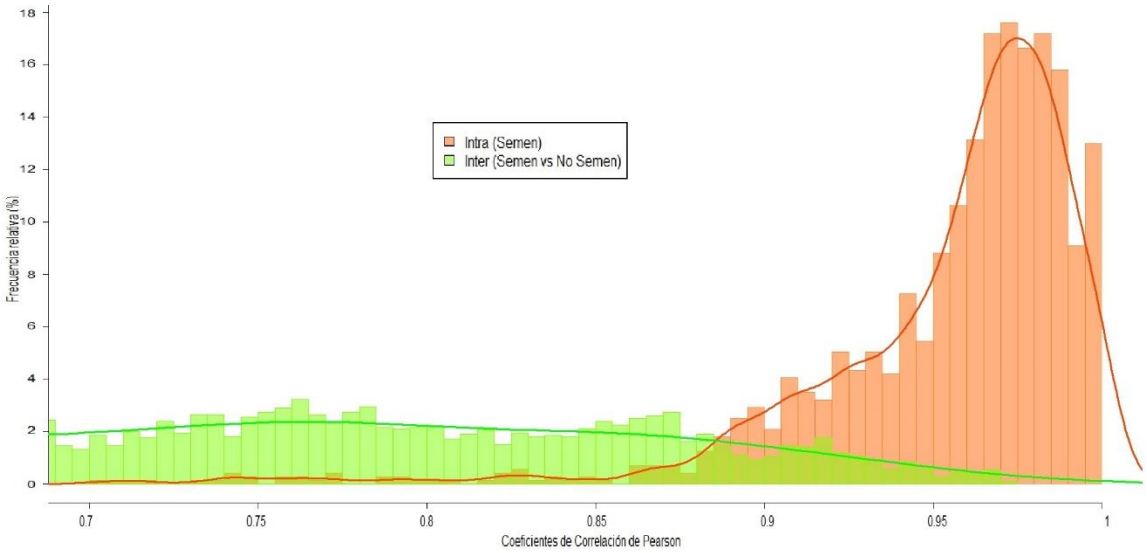

Figura 7. Representación de los coeficientes de correlación de Pearson (eje de ordenadas) frente a su frecuencia relativa (%; eje de abscisas). Se muestran las curvas de densidad de los grupos (naranja: intravariabilidad; verde: intervariabilidad).

Para conocer el comportamiento del modelo con manchas que contengan mezclas de fluidos biológicos, se plantearon dos estrategias: el estudio de la mezcla cuyo espectro tuviera la intensidad más alta (escenario más favorable; que llamaremos Escenario 1) y la mezcla cuya intensidad fuera la más baja (escenario menos favorable; Escenario 2). La Figura 8 muestra que la presencia de otros fluidos en la mancha apenas influye en la discriminación mediante

correlación, pero sí la cantidad de muestra, la cual determina la intensidad de la señal del espectro. Sin embargo, la correlación más baja en el Escenario 2 es 0,745 con una frecuencia relativa del 4 %. La frecuencia relativa más alta de este Escenario ocurre a un coeficiente de relación de 0,890 siendo 19 %, mientras que en el Escenario 1 ocurre a 0,925, siendo 34 %, en ambos casos por encima del valor en que la frecuencia relativa de la intravariabilidad supera la intervariabilidad (r = 0,885).

c. Evaluación mediante curvas ROC

Además de averiguar si el análisis multivariante mediante PCA permitía la distinción del semen en una mancha, se pretendió demostrar si este modelo era viable para ser aplicado por las Fuerzas y Cuerpos de Seguridad del Estado para estudiar un delito donde intervengan manchas de fluidos biológicos, como ocurre en los casos de agresiones sexuales. Además, se pretendía emplear una herramienta que se entendiera fácilmente por los componentes de los tribunales y especialmente por los jueces, quienes finalmente deben dictar sentencia. Para ello, se consideró necesario evaluar qué tasa de acierto y error implicaba la aproximación cualitativa empleada.

Como tasa de acierto entendimos qué proporción de resultados positivos son realmente manchas que contienen semen (Verdaderos Positivos). La tasa de error la interpretamos como la proporción de resultados negativos que son realmente negativos (Verdaderos Negativos, es decir, no contienen semen). Complementarios a estos, también calculamos los Falsos Positivos y los Falsos Negativos. De hecho, lo más importante es reducir estos últimos, ya que no debe aceptarse que el modelo estadístico interprete como negativo un Verdadero Positivo, ya que esto conlleva el descarte de una evidencia.

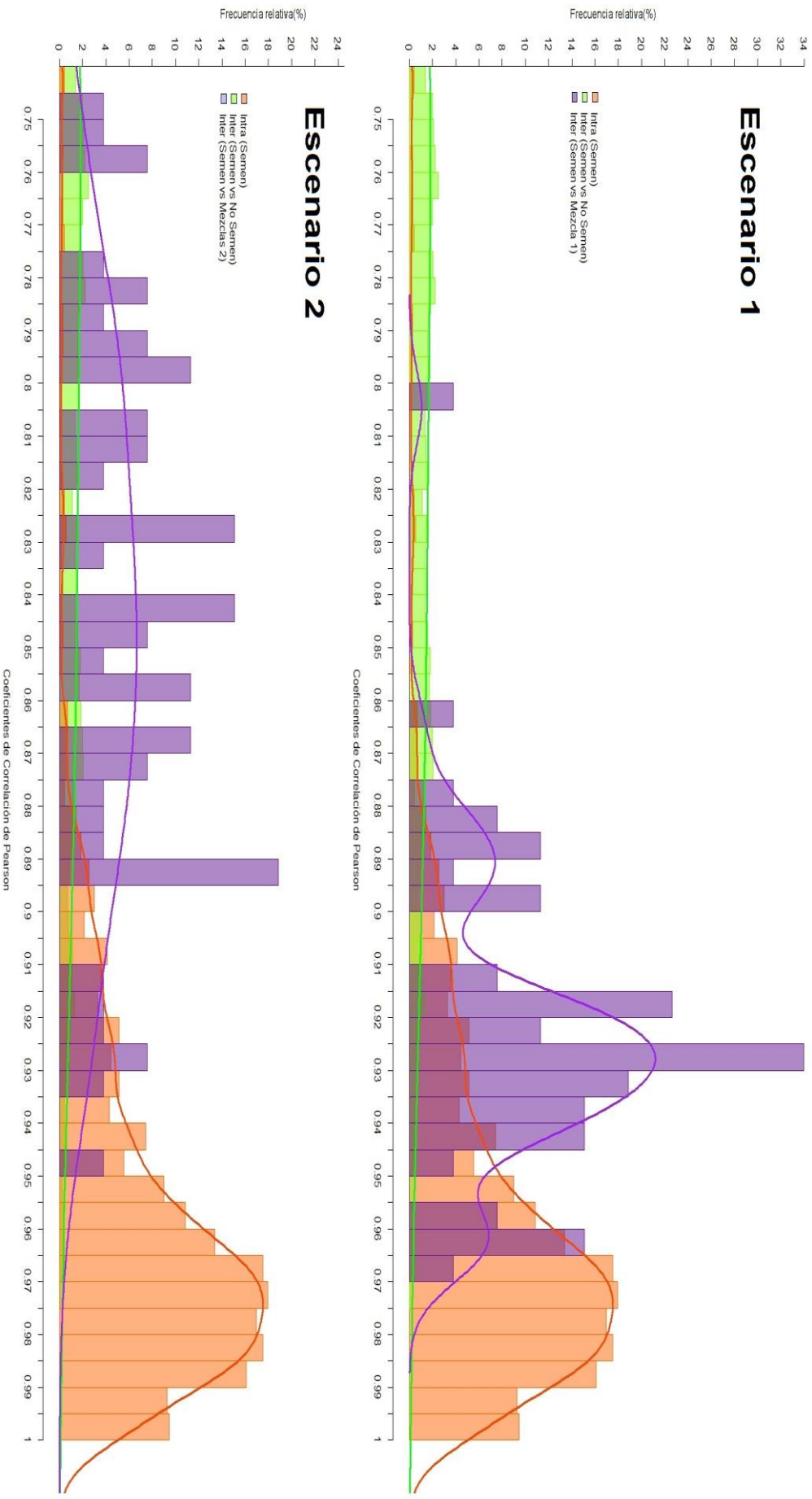

Figura 8. Histogramas de las correlaciones de Pearson de la Intravariabilidad, Intervariabilidad en A) Escenario más favorable y B) Escenario menos favorable, con líneas de densidad naranja, verde y morada, respectivamente.

Como se puede comprobar en la Figura 9, a r = 1 (perfecta correlación positiva) la relación (ratio) de Falsos Positivos es igual a cero, como cabría esperar, ya que el extremo es el punto más restrictivo; sin embargo, por la misma razón, el ratio de Verdaderos positivos también es igual a cero, ya que tan solo la comparación de un espectro de semen consigo mismo daría lugar a una correlación positiva perfecta. De la misma forma, las curvas opuestas a las anteriores (Falsos negativos y Verdaderos negativos), por la máxima restricción del coeficiente de correlación de Pearson, dan lugar a ratios de 1. A medida que desciende la restricción (disminuye *r*) la pendiente de la curva de Verdaderos Positivos (VP) es mucho mayor que la de Falsos Positivos (FP), y esta es la clave para distinguir una buena discriminación de una mala. Por ejemplo, cuando *r = 0.7*, el Ratio de FP es igual a 0 y el de VP es igual a 1. Cuando *r > 0.7* la precisión comienza a disminuir.

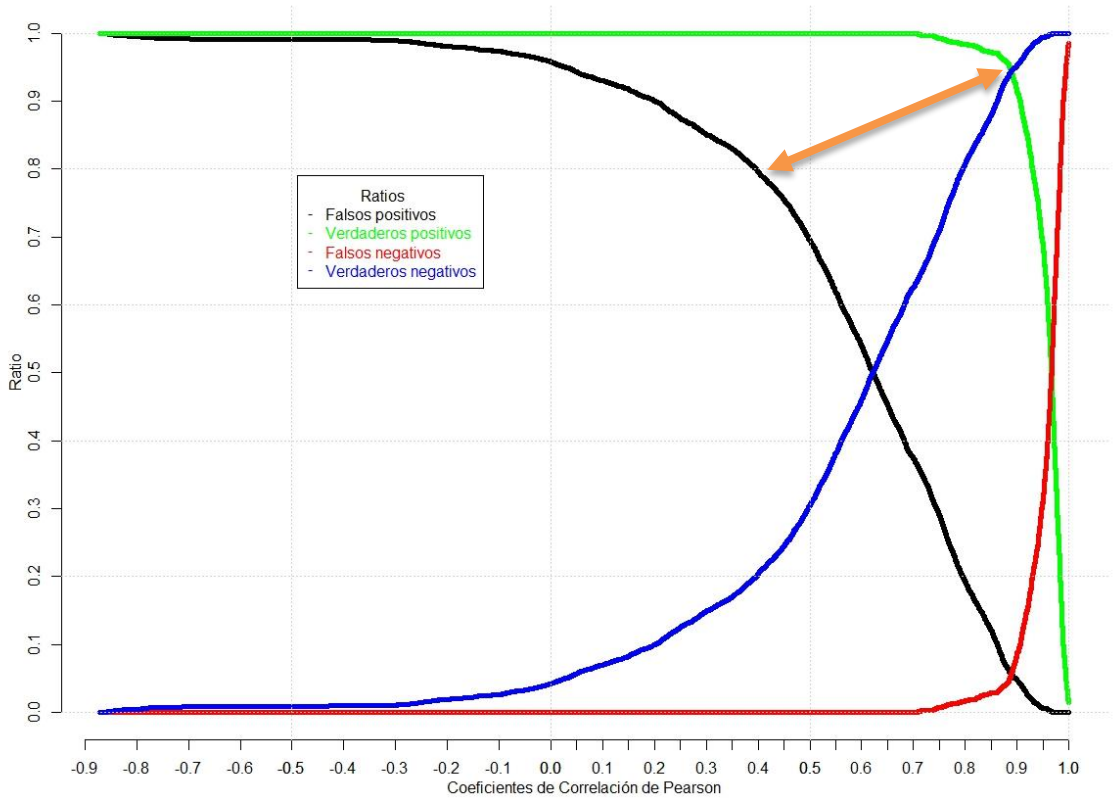

Figura 9. Representación de la relación de los coeficientes de correlación de Pearson con los ratios de Falsos Positivos (curva negra), Verdaderos Positivos (curva verde), Falsos Negativos (curva roja) y Verdaderos Negativos (curva azul). La flecha resalta la distancia entre los ratios de VP y FP.

Como herramienta para evaluar la capacidad de rendimiento diagnóstico, se llevó a cabo el análisis ROC, generando la curva (razón de Verdaderos

Positivos frente a razón de Falsos Positivos) representada en la Figura 10. Esta se muestra notablemente lejana de la diagonal, lo que a priori indica un buen rendimiento del modelo. Para conocer el rendimiento se calculó el área bajo la curva (AUC, por sus siglas en inglés), siendo esta AUC = 0.984, es decir, un rendimiento excepcional según la bibliografía (ver Tabla 2 de la Introducción; Hosmer D. W. et al., 2000).

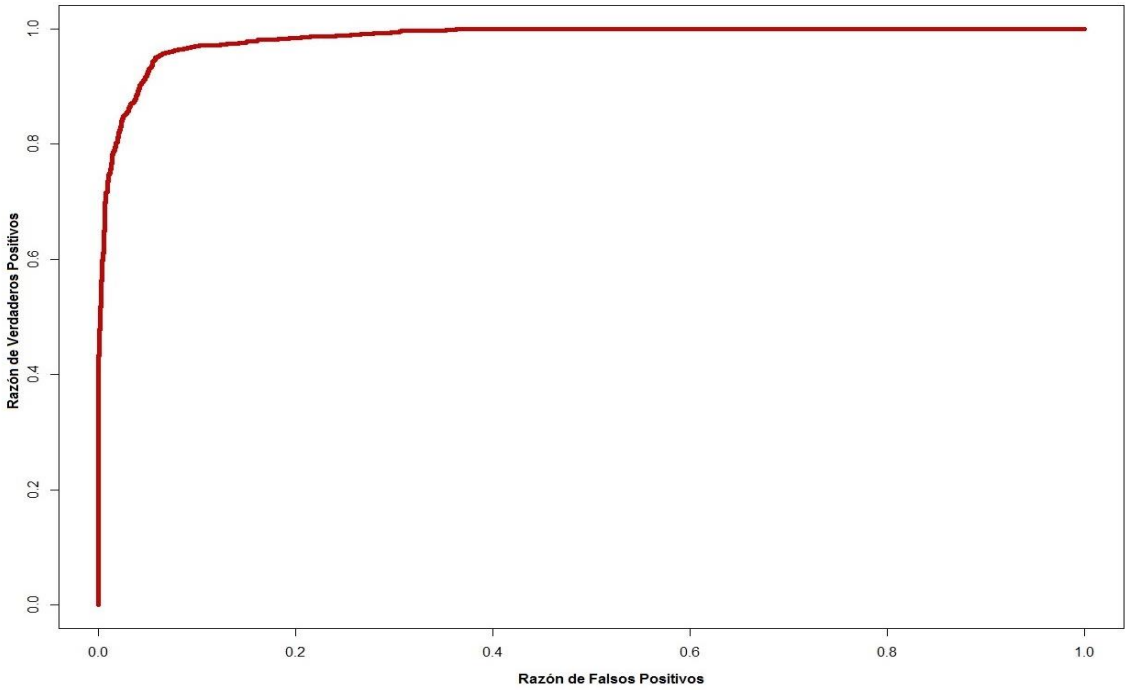

Figura 10. Curva ROC. Representada en base a la razón de Falsos Positivos frente la razón de Verdaderos Positivos.

Por último, se estudiaron cuatro criterios distintos para seleccionar un coeficiente de correlación de Pearson como umbral óptimo. En la Tabla 3 se muestran los valores obtenidos de Verdaderos Positivos, Falsos Positivos, Verdaderos Negativos y Falsos Negativos en sus correspondientes matrices de confusión, y en la Tabla 4 se muestran los valores que se calcularon de rendimiento (sensibilidad, especificidad, precisión, razón de verosimilitudes o DLR, y exactitud) del modelo correspondientes a cada hipotético umbral a partir de los valores obtenidos de la matriz de confusión generada con cada criterio. Esto permite comprender que el umbral deberá conllevar un buen equilibrio entre estos valores.

Tabla 3. Matrices de confusión.

Tabla 3-1. Matriz de confusión del criterio: Máxima Sensibilidad.

Criterio: MaxSe			
Clase estimada	Clase real		
	Positivo	Negativo	
Positivo	2809	2036	4845
Negativo	0	3423	3423
	2809	5459	

Tabla 3-2. Matriz de confusión del criterio: Índice de Youden.

Criterio: Youden			
Clase estimada	Clase real		
	Positivo	Negativo	
Positivo	2669	317	2986
Negativo	140	5142	5282
	2809	5459	

Tabla 3-3. Matriz de confusión del criterio: Máxima Eficiencia.

Criterio: MaxEfficiency			
Clase estimada	Clase real		
	Positivo	Negativo	
Positivo	2671	314	2985
Negativo	142	5145	6287
	2813	5459	

Tabla 3-4. Matriz de confusión del criterio: Máxima Especificidad y Sensibilidad.

Criterio: MaxSpSe			
Clase estimada	Clase real		
	Positivo	Negativo	
Positivo	2653	305	2958
Negativo	156	5154	5310
	2809	5459	

Tabla 3-5. Matriz de confusión del criterio: Máxima Especificidad.

Criterio: MaxSp			
Clase estimada	Clase real		
	Positivo	Negativo	
Positivo	1147	0	1147
Negativo	1662	5459	7121
	2809	5459	

Tabla 4. Valores obtenidos de exactitud, sensibilidad, especificidad y precisión para cuatro criterios de elección de hipotéticos umbrales: Sensibilidad Máxima (MaxSe), Especificidad Máxima (MaxSp), Sensibilidad y Especificidad Máximas (MaxSpSe), Exactitud Máxima (MaxEfficiency), Índice de Youden (Youden). Se muestran los valores de Exactitud calculados.

	MaxSe	**Youden**	**MaxEfficiency**	**MaxSpSe**	**MaxSp**
Umbral óptimo	0.701	0.887	0.887	0.889	0.972
Sensibilidad	1	0.95	0.949	0.944	0.408
Especificidad	0.627	0.942	0.942	0.944	1
Precisión (PPV)	0.58	0.894	0.895	0.897	1
Precisión (NPV)	1	0.973	0.973	0.971	0.767
DLR+	2.68	16.4	16.5	16.9	∞
DLR-	0	0.053	0.054	0.059	0.592
Exactitud	0.754	0.945	0.945	0.945	0.799

Descubrimos que el único criterio que conlleva la total ausencia de Falsos Negativos es el que maximiza la sensibilidad. Este criterio sugiere como umbral óptimo el r = 0.701, y es el más estricto. Sin embargo, el ahorro de tiempo y recursos es menor, ya que conlleva gran cantidad de Falsos Positivos. Por contraposición, el umbral más restrictivo es el que se obtiene de maximizar la especificidad, siendo r = 0.972, pero este es el menos aceptable ya que desecha gran cantidad de pruebas. El criterio basado en maximizar tanto especificidad como sensibilidad simultáneamente, así como el basado en el índice de Youden, son los que conllevan la mayor exactitud, siendo equilibradas la sensibilidad, la especificidad y la precisión. Estos últimos sugieren como umbral óptimo valores de r = 0.887 y 0.889. Dado que estos conllevan un DLR+ > 16, estos criterios obtienen un valor diagnóstico muy bueno.

5. Conclusiones

Los hallazgos más innovadores de este modelo son:

- Mediante estadística clásica se ha demostrado que el PCA supone un rápido análisis presuntivo de una mancha de composición desconocida, gracias a la determinación de un rango de identificación de la firma espectral infrarroja de tres fluidos biológicos (semen, fluido vaginal y orina), cuando están sobre materiales superabsorbentes.

- La evaluación de la eficacia y el rendimiento del modelo estadístico se puede hacer mediante estadística bayesiana más fácilmente interpretable por la justicia. Se propone un valor umbral para ser implementado en los procedimientos que llevan a cabo los laboratorios forenses (entre r=0.887 y r=0.889). Dada la necesidad de mejorar dichos procedimientos, esta investigación aporta una buena demostración de la posibilidad de determinar la presencia de semen en una mancha de composición desconocida de una forma no destructiva, y su posterior análisis genético.

Como trabajo futuro, se sugiere aumentar el abanico de soportes posibles, pero, sobre todo, aumentar la base de datos de espectros de estos fluidos biológicos de una mayor cantidad de donantes. En el futuro también se debe realizar el estudio estadístico de variabilidad y la posterior estadística bayesiana incluyendo espectros de manchas que contengan otros fluidos además de semen en el grupo semen. De esta forma, se llevará a cabo un entrenamiento del método mucho más robusto que el realizado en el presente trabajo, aumentando la certeza con la que se realiza la predicción. Además, se pretende explorar la idea de implementar un *script* a un sistema portátil de espectroscopía infrarroja, de forma que incluso en la escena del delito pueda aportar información acerca de las manchas que pudieran ser halladas, y la selección de aquellas que puedan arrojar pistas acerca del agresor. Para todo esto, es necesaria la colaboración de los cuerpos de seguridad del Estado, que podrán validar y aplicar una nueva metodología en la resolución de casos de agresiones sexuales.

6. Bibliografía más relevante

- Bruker Daltonics, n.d., '*Principal Component Analysis (PCA): Basics'*, material didáctico. http://research.med.helsinki.fi/corefacilities/proteinchem/pca_introduction_basics.pdf
- Bryan A. Hanson (2016) '*ChemoSpec: Exploratory Chemometrics for Spectroscopy*'. R package version 4.2.8.[1] https://github.com/bryanhanson/ChemoSpec https://cran.r-project.org/web/packages/ChemoSpec/vignettes/ChemoSpec.pdf
- Camacho Martinez-Vara de Rey, C. G., n.d., '*Coeficiente de Correlación Lineal de Pearson*', material didáctico de Análisis De Datos En Psicología, Universidad de Sevilla. http://personales.us.es//vararey/adatos2/correlacion.pdf
- '*Computación estadística: introducción a R*', n.d., http://mmeixide.pbworks.com/f/diapositivas_r.pdf
- Frank E. Harrell Jr., with contibutions from Charles Dupont and many others. (2015). '*Hmisc: Harrell miscellaneous*'. R package versión 3.17-1.[1] https://CRAN.R-project.org/package=Hmisc
- GitHub, 2016. https://github.com
- Henrik Bengtsson (2015). '*R.utils: Various Programming Utilities*', R package version 2.2.0.[1] https://CRAN.R-project.org/package=R.utils
- Hosmer, D. W., Lemeshow. S., 2000, '*Applied Logistic Regression*', 2ª ed., Editorial Wiley-Interscience, Nueva York.
- Kristian Hovde Liland and Bjørn-Helge Mevik (2015). '*baseline: Baseline Correction of Spectra*'. R package version 1.2-1.[1] https://CRAN.R-project.org/package=baseline

[1] Se ha mantenido fidelidad con el formato de citación de paquetes del software R por requerimiento del propio software.

- Mónica López-Ratón, María Xose Rodríguez-Álvarez, Carmen Cadarso Suárez, Francisco Gude Sampedro (2014). '*OptimalCutpoints: An R package for Selecting Optimal Cutpoints in Diagnostic Tests*'. Journal of Statistical Software, 61(8), 1-36.[1] http://www.jstatsoft.org/v61/i08/
- Paradis E., 2003. '*R para principiantes*', Institut des Sciences de l'Évolution, Universit Montpellier II, France.
- Pretsch, E., Bühlmann, P., Badertscher, M., 2009, '*Structure Determination of Organic Compounds: Tables of Spectral Data'*, 4ª ed., Editorial Springer.
- Puebla Arredondo, C. 2010, '*Likelihood ratios y teorema de Bayes*', material didáctico de Medicina Basada en la Evidencia, Universidad de Valparaíso, Chile. https://mbeuv.files.wordpress.com/2010/09/12-likelihood-ratios-y-teorema-de-bayes.pdf
- R Core Team (2015). '*R: A language and environment for statistical computing*'. R Foundation for Statistical Computing, Vienna, Austria.[1] https://www.R-project.org/
- R-bloggers, 2016. https://www.r-bloggers.com/
- R-project, '*The Comprehensive R Archive Network'*. https://cran.r-project.org/
- Real Academia Española, n.d., '*Diccionario de la Lengua Española*', 23ª ed., Consultado el día 8 de septiembre de 2016. http://dle.rae.es/
- Rene Locher and Andreas Ruckstuhl et al. (2012). '*IDPmisc: Utilities of Institute of Data Analyses and Process Design (www.idp.zhaw.ch)*'. R package version 1.1.17.[1] https://CRAN.R-project.org/package=IDPmisc
- Revolution Analytics, 2015. http://www.inside-r.org/

[1] Se ha mantenido fidelidad con el formato de citación de paquetes del software R por requerimiento del propio software.

- Shlens, J., April 2014, '*A tutorial on Principal Component Analysis*', Versión 3.02, Cornell University Library, Nueva York. https://arxiv.org/pdf/1404.1100.pdf
- signal developers (2013). '*signal: Signal processing*'.[1] http://r-forge.r-project.org/projects/signal/
- Sing T., Sander O., Beerenwinkel N. and Lengauer T. (2005). '*ROCR: visualizing classifier performance in R*', Bioinformatics, 21(20), pp. 7881.[1] http://rocr.bioinf.mpi-sb.mpg.de
- Stack Exchange, 2016. Stackoverflow y Cross Validated http://stackoverflow.com/ http://stats.stackexchange.com/
- Stevens, A., Ramirez-Lopez, L., 2014, '*An introduction to the prospectr package*', vignette del paquete *prospectr*. https://cran.r-project.org/web/packages/prospectr/vignettes/prospectr-intro.pdf
- Swarbrick, B., 2012, '*Multivariate Data Analysis For Dummies: CAMO Software Special Edition'*, Editorial Wiley, Inglaterra.
- West's Encyclopedia of American Law, 2008, 2ª ed. Consultado el 8 de septiembre de 2016. http://legal-dictionary.thefreedictionary.com/Forensic+Science
- Zhu, W., Zeng, N., Wang, N., 2010, '*Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations*', Health Care and Life Sciences.

Otra bibliografía consultada:

- Antunes, J., Silva, D.S.B.S., Balamurugan, K., Duncan, G., Alho, C.S. & McCord, B., 2016, "*High-resolution melt analysis of DNA methylation to discriminate semen in biological stains*", Analytical Biochemistry, vol. 494, pp. 40-45.
- Cerdas, L., Herrera, F., Arrieta, G., Morelli, C., Alvarez, K. & Gomez, A. ,2016, "*Menstrual cycle phase at the time of rape does not affect*

[1] Se ha mantenido fidelidad con el formato de citación de paquetes del software R por requerimiento del propio software.

recovery of semen or amplification of STR profiles of a suspect in vaginal swabs", Forensic science international, vol. 259, pp. 36-40.

- De Moors, A., Georgalis, T., Armstrong, G., Modler, J. & Fregeau, C.J., 2013, "*Sperm Hy-Liter: An effective tool for the detection of spermatozoa in sexual assault exhibits*", Forensic Science International: Genetics, vol. 7, no. 3, pp. 367-379.
- Elkins, K.M., 2011, "*Rapid Presumptive "*Fingerprinting*" of Body Fluids and Materials by ATR FT-IR Spectroscopy*", Journal of forensic sciences, vol. 56, no. 6, pp. 1580-1587.
- Elvik, S.L., 1990, "*Vaginal discharge in the prepubertal girl*", Journal of pediatric health care : official publication of National Association of Pediatric Nurse Associates & Practitioners, vol. 4, no. 4, pp. 181-5.
- Escola García, M. A., 2014, '*Perfilado de cocaína por Cromatografía de Gases y Espectrometría de Masas*', Trabajo Fin de Máster, Máster en Ciencias Policiales (IUICP), Universidad de Alcalá, Madrid.
- Ferragina, A., de los Campos, G., Vazquez, A.I., Cecchinato, A. & Bittante, G., 2015, "*Bayesian regression models outperform partial least squares methods for predicting milk components and technological properties using infrared spectral data*", Journal of dairy science, vol. 98, no. 11, pp. 8133-8151.
- Fu, X.D., Wu, J., Wang, J., Huang, Y., Hou, Y.P. & Yan, J., 2015, "*Identification of body fluid using tissue-specific DNA methylation markers*", Forensic Science International Genetics Supplement Series, vol. 5, pp. E151-E153.
- Lee, W.C., Khoo, B.E. & Lim bin Abdullah, A.F., 2012, "*A simple, low-cost and portable LED-based multi-wavelength light source for forensic application*", Proceedings of SPIE, vol. 8560, no. LED and Display Technologies II, pp. 856005/1-856005/9.
- Morelato, M., Beavis, A., Tahtouh, M., Ribaux, O., Kirkbride, P., Roux, C., 2013, '*The use of forensic case data in intelligence-led policing: The example of drug profiling*', Forensic Science International, vol. 226, pp. 1-9.

- Morelato, M., Beavis, A., Tahtouh, M., Ribaux, O., Kirkbride, P., Roux, C., 2014, '*The use of organic and inorganic impurities found in MDMA police seizures in a drug intelligence perspective*', Science and Justice, vol. 54, pp. 32-41.
- Morelato, M., Beavis, A., Tahtouh, M., Ribaux, O., Kirkbride, P., Roux, C., 2015, '*The use of methylamphetamine chemical profiling in an intelligence-led perspective and the observation of inhomogeneity within seizures*', Forensic Science International.
- Morillas, P. P. et al., 2016, '*Guía Eurochem: La adecuación al uso de los métodos analíticos. Una guía de laboratorio para la validación de métodos y temas relacionados*', 1ª ed., Eurolab España.
- Muehlethaler, C., Massonnet, G., Hicks, T., 2015, '*Evaluation of infrared spectra analyses using a likehood ratio approach: A practical example of spray paint examination*', Science and Justice.
- Olek, K. & Forat, S., 2014, *Method for detection and distinction of body fluids from forensic material*, Patent Application Country: Application: WO; WO; Priority Application Country: DE.
- Orphanou, C., 2015, "*The detection and discrimination of human body fluids using ATR FT-IR spectroscopy*", Forensic science international, vol. 252, pp. E10-E16.
- Sikirzhytski, V., Sikirzhytskaya, A. & Lednev, I.K., 2012, "*Advanced statistical analysis of Raman spectroscopic data for the identification of body fluid traces: Semen and blood mixtures*", Forensic science international, vol. 222, no. 1-3, pp. 259-265.
- Simard, A., Des Groseillers, L. & Sarafian, V., 2012, "*Assessment of RNA stability for age determination of body fluid stains*", Journal - Canadian Society of Forensic Science, vol. 45, no. 4, pp. 179-194.
- Szymanska, E., Gerreten, J., Engel, J., Geurts, B., Blanchet, L., Buydens, M. C. L., 2015, '*Chemometrics and qualitative analysis have a vibrant relationship*', Trends in Analytical Chemistry, vol. 69, pp. 34-51.
- Zapata, F., 2013, '*Identificación forense de fluidos biológicos mediante herramientas espectroscópicas y quimiométricas*', Trabajo Fin de

Máster, Máster en Ciencias Policiales (IUICP), Universidad de Alcalá, Madrid.

- Zapata, F., Angeles Fernandez de la Ossa, M. A. & Garcia-Ruiz, C. 2015, "*Emerging spectrometric techniques for the forensic analysis of body fluids*", Trends in Analytical Chemistry, vol. 64, pp. 53-63.

7. Anexo: Información complementaria

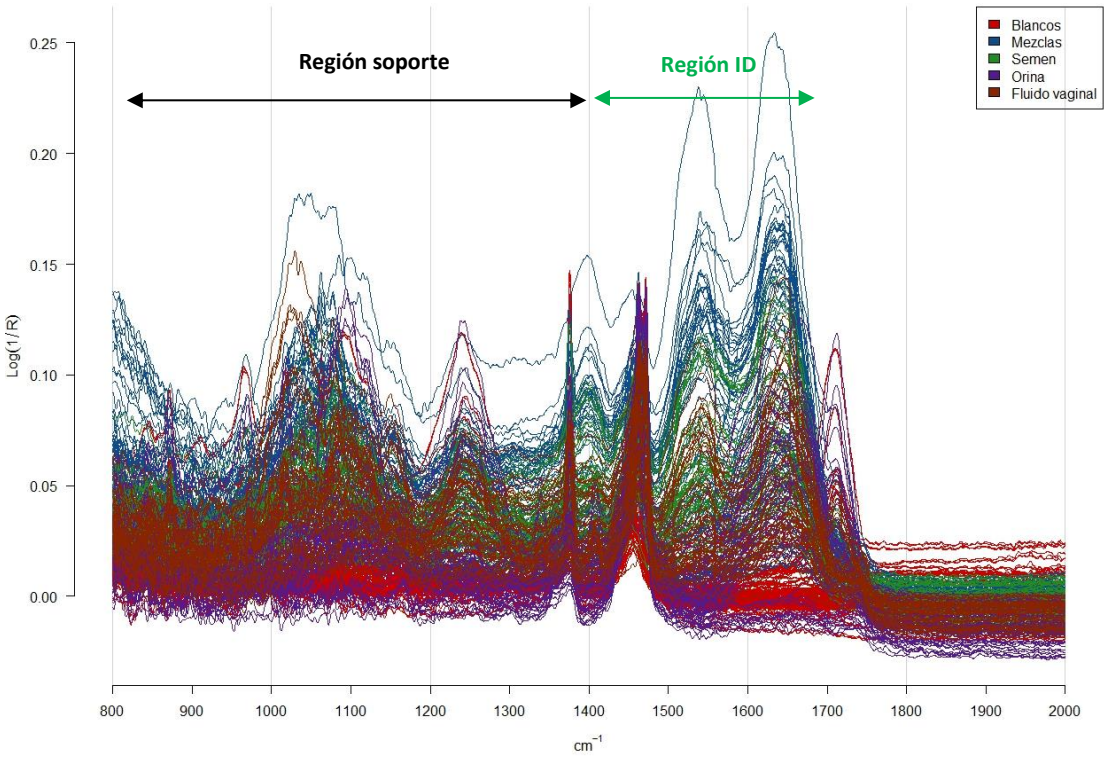

Figura A1. Todos los espectros obtenidos en el rango inicial (de 800 a 2000 cm^{-1}). Eje vertical $Log\frac{1}{R}$ y eje horizontal en cm^{-1}.

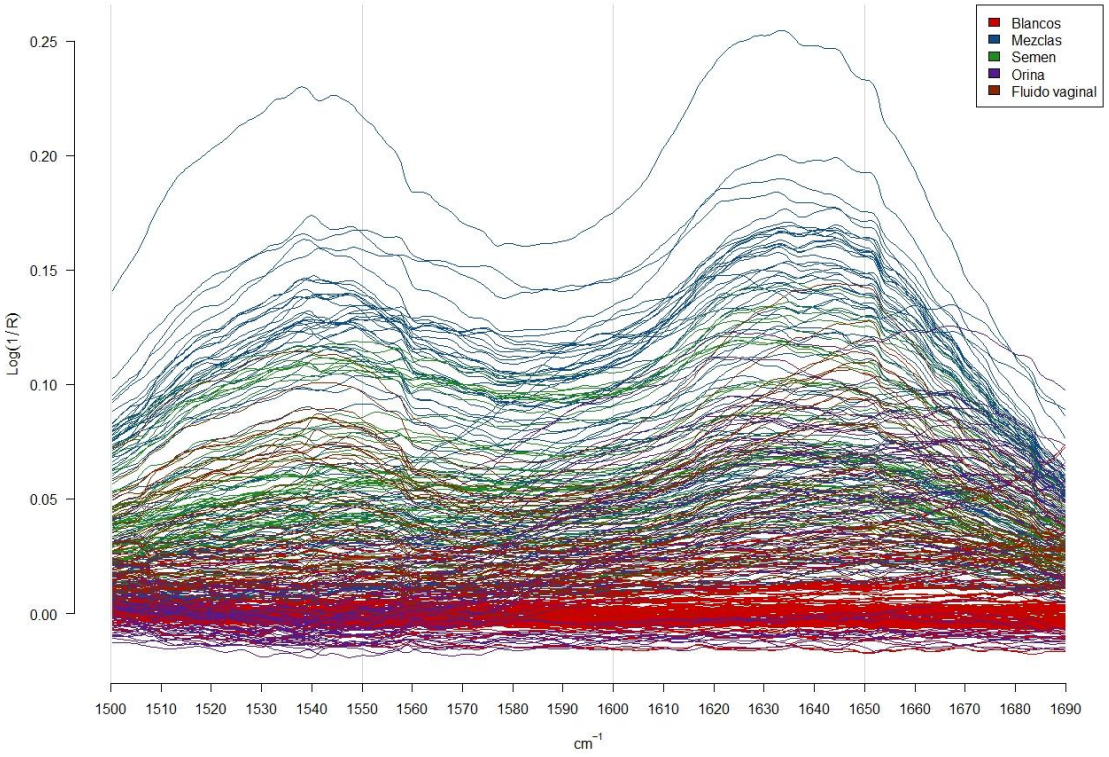

Figura A2. Todos los espectros en el rango de estudio (desde 1500 hasta 1690 cm^{-1}: región ID). Eje vertical $Log\frac{1}{R}$ y eje horizontal en cm^{-1}.

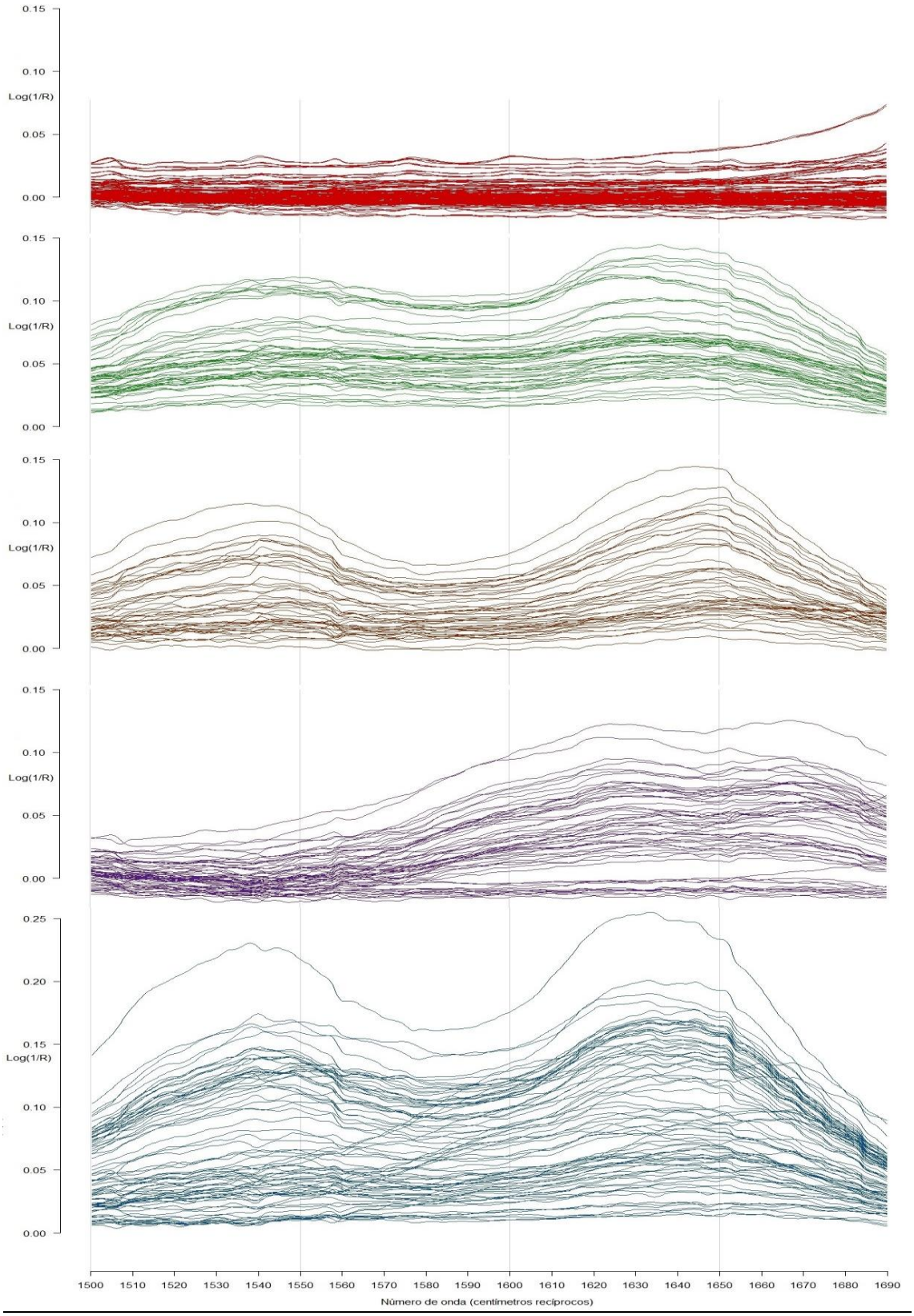

Figura A3. Espectros por grupos en el rango de estudio (desde 1500 hasta 1690 cm-1). En color rojo los espectros de los blancos; en color verde los espectros de semen; en color marrón los espectros de fluido vaginal; en color morado los espectros de orina; y en color azul los espectros de las mezclas. Eje vertical $Log\frac{1}{R}$ y eje horizontal en cm^{-1}.

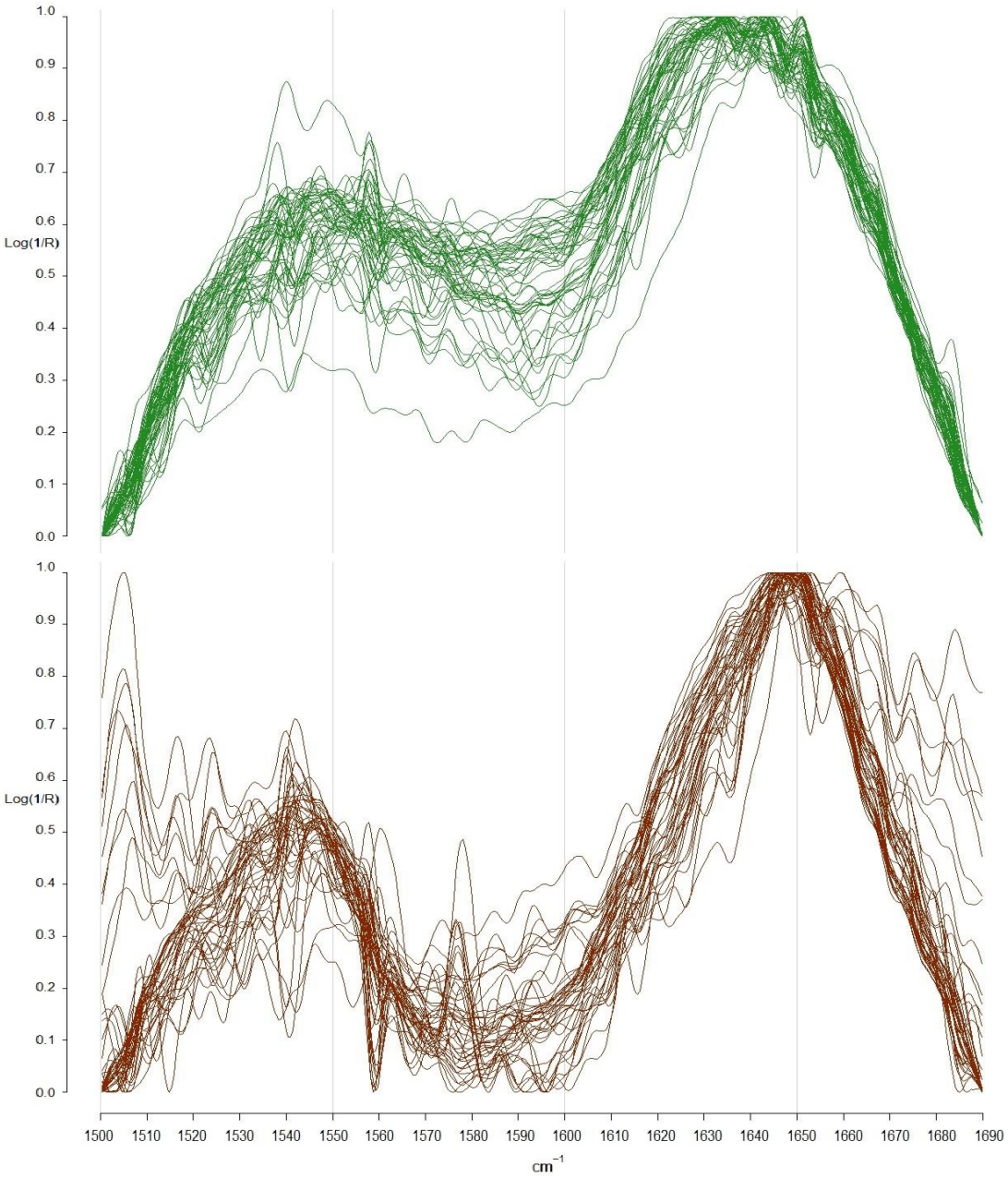

Figura A4. Espectros de semen (verde) y fluido vaginal (marrón) después del tratamiento con corrección de línea base, suavizado y normalización, en la región de interés (Rango ID). Eje vertical $Log\frac{1}{R}$ y eje horizontal en cm^{-1}.

Script simplificado:

```
########################################################################

#                     TRABAJO DE FIN DE GRADO

#      Tratamiento de datos químico-forenses para la discriminación

#      de fluidos biológicos en materiales superabsorbentes

########################################################################

#             Autor: Ignacio Pachón Jiménez                    #


#CARGAR PAQUETES USADOS#

library("ChemoSpec")

library("R.utils")

library("baseline")

library("IDPmisc")

library("signal")

library("stats")

library("Hmisc")

library("graphics")

library("ROCR")

library("OptimalCutpoints")


#CARGAR LOS DATOS ESPECTRALES#


#Read the Dataset

files2SpectraObject(gr.crit=c("Blank","Mixture","Semen","Urine","Vaginal fluid"),

gr.cols=c("red3", "dodgerblue4", "forestgreen", "purple4",
"orangered4"),freq.unit="",

int.unit="",descrip="Fluidos biológicos en materiales absorbentes",out.file="1
TFGSpectra")


xaxis<-expression(cm^-1)

yaxis<-expression(Log (1/R))


#Load de Dataset
```

```
All <- loadObject("1   TFGSpectra.RData")

########################


#CARGAR FUNCIONES PARA EL TRATAMIENTO DE LOS ESPECTROS#


#      Cargar la función de normalizacion personalizada (normNacho)

normNacho <- function(spectra) {


# Function to normalize a Spectra object so that each spectrum

# is on a [0...1] scale

# Bryan Hanson, DePauw University, Feb 2016


       if (missing(spectra)) stop("No spectral data provided")

       chkSpectra(spectra)


       for (i in 1:length(spectra$names)) {

              rMin <- min(spectra$data[i,])

              spectra$data[i,] <- spectra$data[i,] - rMin

              rMax <- max(spectra$data[i,])

              spectra$data[i,] <- spectra$data[i,]/rMax

              }


       chkSpectra(spectra)

       return(spectra)

       }


#      Cargar la función personalizada para el suavizado (Smoothing) Savitzky-Golay

sgfSpectra <- function(spectra, m = 0) {


# Function to filter a Spectra object

# Bryan Hanson, DePauw University, Feb 2016
```

```
	if (!requireNamespace("signal", quietly = TRUE)) {
		stop("You need to install package signal to use this function")
		}

	if (missing(spectra)) stop("No spectral data provided")
	chkSpectra(spectra)


	for (i in 1:length(spectra$names)) {
		spectra$data[i,] <- sgolayfilt(spectra$data[i,],p=2,n=11,m=0)
		}

	chkSpectra(spectra)
	return(spectra)
	}


#	Cargar la función de baseline correction personalizada (baselineNacho)
baselineNacho <- function(spectra) {

# Bryan Hanson, DePauw University, Feb 2016

	if (missing(spectra)) stop("No spectral data provided")
	chkSpectra(spectra)
	np <- length(spectra$freq)
	for (i in 1:length(spectra$names)) {
		rMin <- min(spectra$data[i,])
		spectra$data[i,] <- spectra$data[i,] - rMin
		# Do an lm from end to the other
		DF <- data.frame(
			x = c(spectra$freq[1], spectra$freq[np]),
			y = c(spectra$data[i,1], spectra$data[i,np]))
```

```
                fit <- lm(y ~ x, DF)
                spectra$data[i,] <- spectra$data[i,]- predict(fit,
                        newdata = data.frame(x = spectra$freq))
                }

        chkSpectra(spectra)
        return(spectra)
        }

#PRELIMINARY INSPECTION OF DATA#
################################
sumSpectra(All)

#DATA PRE-PROCESSING#

##      Remove frequencies. Selecting research's Range
Ranged<-removeFreq(All,rem.freq=All$freq>1690|All$freq<1500)

meanRAllsd<-surveySpectra(Ranged,method="sd",main="Media de espectros +/- desviación
estandar")

#BASELINE CORRECTION#
#Baseline offset f(x)=x-min(X)--->baselineNacho
#Linear Baseline Correction.
BsRanged<-baselineNacho(Ranged)

#SMOOTHING#
SmBsRanged<-sgfSpectra(BsRanged)

#NORMALIZATION#
NSBRanged<-normNacho(SmBsRanged)
```

```
######################

meanNSBRAllsd<-surveySpectra(NSBRanged,method="sd",main="Media de espectros +/-
desviación estandar")


#      PCA (Análisis de Componentes Principales)         #

PCA_NSBR <- c_pcaSpectra(NSBRanged, choice = "noscale")

plotScores(NSBRanged,main="Scores PCA con Blancos",PCA_NSBR,pcs=c(1,2))

diagnosticsOD <- pcaDiag(NSBRanged, PCA_NSBR, pcs = 10, plot = "OD")

diagnosticsSD <- pcaDiag(NSBRanged, PCA_NSBR, pcs = 5, plot = "SD")

plotScoresRGL(NSBRanged, PCA_NSBR,leg.pos = "A",t.pos = "B")

plotScores3D(NSBRanged, PCA_NSBR, main = title, ellipse = T)

plotLoadings(NSBRanged, PCA_NSBR, main = title,loads = c(1,2,3),ref=1)

#########################

NSBRPuros<-removeGroup(NSBRanged,"Blank")

Puros_PCA_NSBR <- c_pcaSpectra(NSBRPuros, choice = "noscale")

plotScores(NSBRPuros,main="Scores sin Blancos",Puros_PCA_NSBR,pcs=c(1,2))

diagnosticsOD <- pcaDiag(NSBRPuros, Puros_PCA_NSBR, pcs = 10, plot = "OD")

diagnosticsSD <- pcaDiag(NSBRPuros, Puros_PCA_NSBR, pcs = 5, plot = "SD")

plotScoresRGL(NSBRPuros, Puros_PCA_NSBR,leg.pos = "A",t.pos = "B")

plotScores3D(NSBRPuros, Puros_PCA_NSBR, main = title, ellipse = T)

plotLoadings(NSBRPuros, Puros_PCA_NSBR, main = title,loads = c(1,2,3),ref=1)


######  WARNING!!!! Set dir to "Pearson"  #################

#      PEARSON (r)          #

#      Cargar funciones para los Coef Corr Inter e Intra

#cor.test {stats}

cor.testInter <- function(x,y){

    FUN <- function(x, y) cor.test(x, y)[["estimate"]]

    z <- outer(

      colnames(x),

      colnames(y),

      Vectorize(function(i,j) FUN(x[,i], y[,j]))
```

```
    )
    dimnames(z) <- list(colnames(x), colnames(y))
    z
}

cor.testIntra <- function(x){
    FUN <- function(x, y) cor.test(x, y)[["estimate"]]
    z <- outer(
      colnames(x),
      colnames(x),
      Vectorize(function(i,j) FUN(x[,i], x[,j]))
    )
    dimnames(z) <- list(colnames(x), colnames(x))
    z
}

#      Exportar los espectros procesados
#Los exportamos de tal manera que los espectros son COLUMNAS
#Blank== 1:170
#Mix== 171:250
#Sem== 251:303
#Uri== 304:361
#Vag== 362:406

write.table(t(NSBRanged$data),file="PearsonMatrix.csv",
quote=F,sep=";",dec=",",row.names=F,col.names=F)

#Una vez tenemos la matriz creada y revisada en Excel, importamos los datos.
#Los espectros siguen siendo COLUMNAS en R

pearsonMatrix<-read.csv("PearsonMatrix.csv",header=F,sep=";",dec=",")
```

```
#Comprobamos que los gráficos son iguales

plot(pearsonMatrix$V1,type="l")

plotSpectra(NSBRanged,which=c(1))


##                   AQUÍ HAY QUE ELEGIR UNAS POCAS MUESTAS DE MEZCLAS PARA GRAFICAR

#      Definir submatrices que vamos a correlacionar


#      All

dfAll<-as.data.frame(pearsonMatrix)

#      Semen

dfSemen<-as.data.frame(dfAll[,251:303])

#      No Semen (Fluido vaginal y Orina)

dfNoSemen<-as.data.frame(dfAll[,304:406])

#      Mezclas

#227 +++

#217 ---

#192 -- (Escenario 0, o Escenario 2 Alternativo)

dfMezclasEscenario1<-as.data.frame(dfAll[,227])

dfMezclasEscenario2<-as.data.frame(dfAll[,217])

dfMezclas<-cbind(dfMezclasEscenario1,dfMezclasEscenario2)

#dfMezclasEscenario0<-as.data.frame(dfAll[,192])


#227 +++Intensity    (Scenario 1)

#217 ---Intensity (Scenario 2)

#192 --Intensity  (Scenario 0, or Alternative Scenario 2 )

dfAll<-as.data.frame(pearsonMatrix)

dfSemen<-as.data.frame(dfAll[,251:303])

dfNoSemen<-as.data.frame(dfAll[,304:406])

dfMezclasEscenario1<-as.data.frame(dfAll[,227])

dfMezclasEscenario2<-as.data.frame(dfAll[,217])

dfMezclas<-cbind(dfMezclasEscenario1,dfMezclasEscenario2)

#dfMezclasEscenario0<-as.data.frame(dfAll[,192])
```

```
rIntraSemen<-cor.testIntra(dfSemen)

rInter<-cor.testInter(dfSemen,dfNoSemen)

rInterM1<-cor.testInter(dfSemen,dfMezclasEscenario1)

rInterM2<-cor.testInter(dfSemen,dfMezclasEscenario2)


###################

range(rInter)

range(rInterM)

range(rIntraSemen)


#PLOT HISTOGRAMS#

histIntraSemen<-hist(rIntraSemen,freq=F,col="green",main="Intravariabilidad Semen vs
Intervariabilidad",border="green",breaks=90,xlim=c(-0.86,1),ylim=c(0,35),add=F)

histInterM1<-
hist(rInterM1,freq=F,col="purple",border="purple",main="Intervariabilidad Escenario
1",breaks=50)

histInterM2<-
hist(rInterM2,freq=F,col="purple",border="purple",main="Intervariabilidad Escenario
2",breaks=50,ylim=c(0,35),xlim=c(0.73,1))

histInter<-hist(rInter,freq=F,col="red",border="red",main="Intervariabilidad Semen
vs. No Semen",breaks=555)

#       1.Inter vs Intra

        plot(histIntraSemen,col=rgb(1,0.4,0,1/2),axes=F,border=rgb(1,0.4,0,1/2),freq=F
,xlab="Coeficientes de Correlación de Pearson",ylab="Frecuencia relativa
(%)",main="Inter vs. Intra")

        plot(histInter,col=rgb(0.5,1,0,1/2),axes=F,border=rgb(0.5,1,0,1),freq=F,add=T,
xlab="Coeficientes de Correlación de Pearson",ylab="Frecuencia relativa (%)")

        legend(0.8,14,bty="n",legend=c("Intra (Semen)","Inter (Semen vs No Semen)"),

        text.col="black",fill=c(rgb(1,0.4,0,1/2),rgb(0.5,1,0,1/2)))

        axis(1,at=seq(0.5,1,by=0.5),labels=seq(0.5,1,by=0.5))

        axis(1,at=seq(0.5,1,by=0.05),labels=seq(0.5,1,by=0.05))

        axis(2,at=seq(0,24,by=2),las=1)

        lines(density(rIntraSemen),col="orangered",lwd=3)

        lines(density(rInter),col="green",lwd=3)

#       2.Scenario 1
```

```
	plot(histIntraSemen,col=rgb(1,0.4,0,1/2),border=rgb(1,0.4,0,1),freq=F,axes=F,a
dd=F,xlab="Coeficientes de Correlación de Pearson",ylab="Frecuencia
relativa(%)",xlim=c(0.75,1),ylim=c(0,35),main="Escenario 1")

	plot(histInter,col=rgb(0.5,1,0,1/2),border=rgb(0.5,1,0,1),freq=F,add=T,axes=F,
xlab="Coeficientes de Correlación de Pearson",ylab="Frecuencia relativa(%)")

	plot(histInterM1,col=rgb(0.37,0.07,0.56,1/2),border=rgb(0.37,0.07,0.56,1),axes
=F,freq=F,add=T,xlab="Coeficientes de Correlación de Pearson",ylab="Frecuencia
relativa(%)")

	legend(0.85,26,bty="n",legend=c("Intra (Semen)","Inter (Semen vs No
Semen)","Inter (Semen vs Mezcla
1)"),fill=c(rgb(1,0.4,0,1/2),rgb(0.5,1,0,1/2),rgb(0.37,0.07,0.56,1/2)))

	lines(density(rIntraSemen),col="orangered",lwd=3)

	lines(density(rInter),col="green",lwd=3)

	lines(density(rInterM1),col="purple",lwd=3)

	axis(1,at=seq(0.75,1,by=0.005),labels=seq(0.75,1,by=0.005))

	axis(2,at=seq(0,35,by=2),las=1)

#	3.Scenario 2

	plot(histInter,axes=F,col=rgb(0.5,1,0,1/2),border=rgb(0.5,1,0,1),freq=F,add=F,
xlab="Coeficientes de Correlación de Pearson",ylab="Frecuencia
relativa(%)",main="Escenario 2",xlim=c(0.75,1),ylim=c(0,35))

	plot(histIntraSemen,axes=F,col=rgb(1,0.4,0,1/2),border=rgb(1,0.4,0,1),freq=F,a
dd=T,xlab="Coeficientes de Correlación de Pearson",ylab="Frecuencia relativa(%)")

	plot(histInterM2,axes=F,add=T,col=rgb(0.37,0.07,0.56,1/2),border=rgb(0.37,0.07
,0.56,1),freq=F,xlab="Coeficientes de Correlación de Pearson",ylab="Frecuencia
relativa(%)")

	legend(0.75,18.5,bty="n",legend=c("Intra (Semen)","Inter (Semen vs No
Semen)","Inter (Semen vs Mezclas
2)"),fill=c(rgb(1,0.4,0,1/2),rgb(0.5,1,0,1/2),rgb(0.5,0.5,1,1/2)))

	axis(1,at=seq(0.75,1,by=0.005),labels=seq(0.75,1,by=0.005))

	axis(2,at=seq(0,35,by=2),las=1)

	lines(density(rIntraSemen),col="orangered",lwd=3)

	lines(density(rInter),col="green",lwd=3)

	lines(density(rInterM2),col="purple",lwd=3)


# ROC & roll #

labIntra<-seq(1,1,length=length(rIntraSemen))

labInter<-seq(0,0,length=length(rInter))

labels<-c(labIntra,labInter)
```

```
preds<-c(rIntraSemen,rInter)

pred.obj<-prediction(preds,labels)

tpr<-performance(pred.obj,"tpr")

fpr<-performance(pred.obj,"fpr")

fnr<-performance(pred.obj,"fnr")

tnr<-performance(pred.obj,"tnr")

TP<-as.data.frame(tpr@"y.values")

FP<-as.data.frame(fpr@"y.values")


#PLOT CURVES#

plot(fpr,col="black",ylab="",xlab="",box.lty=0,lwd=5)

plot(tpr,col="green",ylab="",xlab="",add=T,lwd=5)

plot(fnr,col="red",ylab="",xlab="",add=T,lwd=5)

plot(tnr,col="blue",ylab="",xlab="",add=T,lwd=5)

mtext("Ratio",side=2,line=2)

axis(1,at=seq(0,1,by=0.05),labels=F)

axis(1,at=seq(-0.9,1,by=0.1),labels=T)

axis(2,at=seq(0.1,0.9,by=0.2),labels=T)

mtext("Coeficientes de Correlación de Pearson",side=1,line=2)

grid()

legend(-0.49,0.79,bty="",legend=c("         Ratios","Falsos positivos","Verdaderos
positivos","Falsos negativos","Verdaderos negativos"),

       text.col=c("black","black","green","red","blue"),pch=c("","--","--","--","--
"),col=c("black","black","green","red","blue"))


ROCcurve<-performance(pred.obj,"tpr","fpr")

ROCcurve

plot(ROCcurve,col="red3",lwd=5,main="Curva ROC")

ROCauc<-performance(pred.obj,"auc")

ROCauc@"y.values"


# Otros cálculos ROC
```

```
#       AUC

ROCauc<-performance(pred.obj,"auc")

ROCauc@"y.values"


cutpoints.obj<-data.frame(preds,labels)

data<-cutpoints.obj

MaxSpSe<-optimal.cutpoints(preds~labels,tag.healthy=0,"MaxSpSe",cutpoints.obj)

MaxSp<-optimal.cutpoints(preds~labels,tag.healthy=0,"MaxSp",cutpoints.obj)

MaxSe<-optimal.cutpoints(preds~labels,tag.healthy=0,"MaxSe",cutpoints.obj)

Youden<-optimal.cutpoints(preds~labels,tag.healthy=0,"Youden",cutpoints.obj)

MaxEffi<-optimal.cutpoints(preds~labels,tag.healthy=0,"MaxEfficiency",cutpoints.obj)

str(MaxSpSe)

str(MaxSp)

str(MaxSe)

str(Youden)

str(MaxEffi)

##############################################################################################
############################
```