Universidad de Valladolid

FACULTAD DE CIENCIAS

DEPARTAMENTO DE ESTADÍSTICA
E INVESTIGACIÓN OPERATIVA

TESIS DOCTORAL:

MEJORAS EN REGLAS DE CLASIFICACIÓN MEDIANTE INCORPORACIÓN DE INFORMACIÓN ADICIONAL

Presentada por David Conde del Río para optar al
grado de doctor por la Universidad de Valladolid

Dirigida por:
Bonifacio Salvador, Miguel A. Fernández

Bonifacio Salvador González, Catedrático de Universidad, y Miguel A. Fernández Temprano, Profesor Titular de Universidad, certifican que la presente memoria ha sido realizada, bajo su dirección, por David Conde del Río en el Departamento de Estadística e Investigación Operativa de la Universidad de Valladolid.

Valladolid, 31 de enero de 2014

Mi más sincero agradecimiento a mis directores, Bonifacio Salvador y Miguel A. Fernández, y a Cristina Rueda, por su dedicación y tutela, por su comprensión, ideas y sugerencias, y, por extensión, a todos los miembros del departamento de Estadística e Investigación Operativa.

A Alma, a mi familia

Contenido

1 Introducción

Esta memoria se presenta en el formato denominado “compendio de publicaciones”, por lo que comenzamos describiendo el tópico en el que se enmarcan las contribuciones que se presentan y resumiendo los resultados más destacados existentes en el mismo hasta el momento. El objetivo de esta descripción no es otro que situar los resultados obtenidos en el contexto apropiado y justificar la unidad temática de los mismos. En las secciones subsiguientes se describirán las aportaciones realizadas y los trabajos en curso en este momento.

Las aportaciones contenidas en esta memoria se sitúan dentro de la inferencia estadística con restricciones, una rama de la estadística que nació en los años 50 del siglo pasado y que tiene por objeto, cuando se dispone de información a priori no muestral, diseñar procedimientos que aprovechen esta información y resulten más eficientes que los procedimientos que no la toman en cuenta.

En general, bajo la metodología de la estadística frecuentista, esta información adicional se incorpora al modelo por medio de restricciones (de orden, no negatividad, forma...) sobre los parámetros. En la práctica existen muchas situaciones en las que es razonable suponer restricciones de orden sobre los parámetros. Por ejemplo, en un ensayo clínico es habitual que los niveles medios de toxicidad de un medicamento crezcan con el nivel de la dosis, ver por ejemplo Gasparini y Eisele (2000). En otras ocasiones es el propio modelo el que impone dichas restricciones. Tal es el caso de la estimación no paramétrica de funciones de distribución, donde hay que tener en consideración la naturaleza monótona de dichas funciones, ver Lo (1987). Referencias fundamentales en el campo de la inferencia bajo restricciones son Barlow et al. (1972), Robertson et al. (1988) y Silvapulle y Sen (2004), libros que recogen las aportaciones más relevantes hasta su fecha de publicación.

Una alternativa que no consideraremos en esta memoria es la metodología bayesiana, donde la información adicional se incopora al modelo a través de las distribuciones a priori. Algunas referencias de interés en este campo son Marchand y Strawderman (2004), Taylor et al. (2007) y Hoitjink (2013).

Uno de los tópicos más desarrollados dentro de la Inferencia estadística con restricciones es el de los contrastes con restricciones para medias de poblaciones normales. Estos contrastes están motivados por la necesidad en muchas aplicaciones de contrastar hipótesis frente a alternativas restringidas, como por ejemplo homogeneidad o simetría frente a alternativas restringidas como son la monotonía o unimodalidad. Se han desarrollado muchos métodos estadísticos para detectar las citadas propiedades dentro del modelo normal, y también fuera del modelo normal utilizando métodos no paramétricos.

La metodología basada en la verosimilitud es la principal aproximación a los contrastes con restricciones. Bartholomew (1959) estudió el contraste de homogeneidad de medias de poblaciones normales univariantes con varianzas iguales contra la hipótesis alternativa de restricciones de orden entre las medias y determinó el estadístico de razón de verosimilitudes y su distribución bajo la hipótesis nula. Desde entonces, la bibliografía relativa a este problema ha sido abundante. Se ha obtenido la distribución del estadístico razón de verosimilitudes en distintas situaciones [Bartholomew (1961), Barlow et al. (1972), Sasabuchi et al. (1983), Kulatunga y Sasabuchi (1984), Shapiro (1988), Robertson et al. (1988), Cohen et al. (2000)], así como propiedades de la función de potencia de los mismos [Perlman (1969), Mukerjee et al. (1986), Tsai (1992), Praestgaard (2012)]. Se han detectado anomalías en estos tests de razón de verosimilitudes bajo restricciones, habiéndose estudiado y caracterizado las situaciones en que están dominados por otros tests que no tienen en cuenta la información dada por las restricciones: Menéndez et al. (1991, 1992a, 1992b), Menéndez y Salvador (1991, 1992), Hu y Wright (1994), Cohen et al. (2000), Perlman y Wu (2002), Cohen y Sackrowitz (2004). También se han considerado procedimientos ajenos a la razón de verosimilitudes con resultados positivos, como los tests lineales como el que proponen Abelson y Tukey (1963) y que después extienden, entre otros, Schaafsma y Smid (1966), Shi y Kudô (1987), Rueda (1989) y Tsai y Sen (1993). Muy recientemente, El Barmi y McKeague (2013) han propuesto un estadístico test basado en la verosimilitud empírica para contrastar la presencia de orden estocástico entre k distribuciones univariantes.

Otro de los grandes tópicos de la Inferencia bajo restricciones es el de la estimación de los parámetros. Ayer et al. (1955) estudiaron el modelo binomial bajo restricciones lineales de orden en las probabilidades de éxito y dieron por primera vez una formulación minimax para la solución. Van Eeden (1956, 1957a, 1957b) estudia condiciones para la existencia y propone algoritmos para encontrar el Estimador Máximo Verosímil (EMV) para el problema de varias poblaciones univariantes con desigualdades entre los parámetros y cotas sobre ellos. Brunk (1955, 1958) considera el problema de varias poblaciones de la familia exponencial uniparamétrica con restricciones entre los parámetros. Muchos de estos procedimientos se enmarcan dentro de la regresión isotónica o regresión isotónica generalizada, ver Brunk (1970) y Robertson et al. (1988), donde se trata el tema en detalle.

Cuando se considera globalmente la estimación del vector multidimensional de medias, el EMV restringido tiene menor error cuadrático medio que el no restringido, ver Robertson et al. (1988). Sin embargo, cuando el interés está en estimar una función lineal de las coordenadas, el resultado depende del tipo de restricciones y de la función lineal a estimar. Rueda y Salvador (1995) proporcionan condiciones bajo las que, para conos determinados por una o dos restricciones lineales, cualquier función lineal del estimador restringido se comporta mejor que la función lineal del estimador no restringido. En el caso de restricciones de orden, y para estimar las coordenadas del vector de medias, el resultado depende del tipo de restricción. En el caso de restricciones de orden simple, Kelly (1989) y Lee (1981) prueban que el EMV restringido domina al no restringido. Sin embargo, en el caso de restricciones tree order, Lee (1988) prueba que el estimador restringido no domina al no restringido para la componente raíz del árbol. Hwang y Peddada (1994) refuerzan y unifican los resultados de Lee (1981, 1988) y Kelly (1989) con restricciones de orden simple y tree order: proponen estimadores alternativos que son mejores que los no restringidos para la i-ésima coordenada en el sentido de que los intervalos de confianza centrados en dichos estimadores tienen una probabilidad de cubrimiento mayor. En este sentido, Rueda et al. (1997a) muestran que si el vector de medias está restringido a un semiespacio, la probabilidad de cubrimiento de los intervalos de confianza es mayor para el EMV restringido, tanto si la estimación es simultánea como si no. Fernández et al. (1998, 1999, 2000) caracterizan situaciones en las que el estimador restringido se comporta mejor que el no restringido en términos de la probabilidad de cubrimiento y del error cuadrático medio.

El problema de la estimación de parámetros de escala con varios tipos de restricciones de orden es estudiado también por Hwang y Peddada (1994), que obtienen resultados de dominación parcial: demuestran que, en un orden simple, el EMV restringido del parámetro más pequeño domina estocásticamente al no restringido, siendo dominado para el parámetro mayor. Fernández et al. (1997) también estudian este problema para poblaciones uniformes, y determinan funciones para las cuales el estimador restringido domina al no restringido y una clase de funciones para las que no.

Marchand y Strawderman (2004) y van Eeden (2006) repasan los métodos de estimación con restricciones en los parámetros presentes en la literatura y sus propiedades de admisibilidad y minimaximalidad, y cuándo estas buenas propiedades se mantienen al estimar funciones del parámetro, en particular la i-ésima coordenada. Estas propiedades dependen de la función de pérdida utilizada. La mayoría de los autores usan función de pérdida cuadrática, pero también se utilizan otras funciones de pérdida como, por ejemplo, las de Stein (Tsukuma y Kubokawa, 2011) o LINEX (Ma y Liu, 2013).

Mención aparte merece el modelo lineal. La literatura es abundante en lo referente a la estimación de los parámetros del modelo lineal en espacios paramétricos restringidos, es decir, cuando el vector de medias (de dimensión k) está restringido a un subconjunto convexo y cerrado de $\mathbb{R}^k$. Uno de los primeros trabajos se debe a Lovell y Prescott (1970). En Silvapulle y Sen (2004) se estudia en detalle este problema, con numerosas referencias. En van Eeden (2006) se repasa la bibliografía en términos de admisibilidad en tres situaciones: una restricción de desigualdad lineal, restricciones intervalo y cuadrante, y restricciones poliédricas y elipsoidales, centrándose en este último caso en los resultados de Moors y van Houwelingen (1993). Rueda et al. (1997b) tratan la estimación simultánea en un modelo lineal cuando el vector de parámetros está restringido a un cono poliédrico.

La literatura de la inferencia bajo restricciones en lo relativo a intervalos de confianza prácticamente se ha limitado a sugerir métodos para construir intervalos de confianza bajo restricciones de orden en los parámetros. Marcus y Peritz (1976) construyen cotas inferiores de confianza simultáneas bajo ciertos modelos lineales con restricciones y varianzas conocidas. Schoenfeld (1986) propone cotas de confianza para las medias ordenadas de poblaciones normales invirtiendo el test de razón de verosimilitudes. Hwang y Peddada (1994) construyen intervalos de confianza unidimensionales de amplitud constante para las coordenadas de la media de una variable simétrica con distribución elíptica bajo restricciones de orden en las coordenadas. Estos intervalos, centrados en algunos estimadores mejorados como los procedentes de la regresión isotónica, dominan a los intervalos de confianza estándar (centrados en el EMV) cuando la matriz de covarianzas es diagonal. Asimismo, presentan un nuevo esquema para construir mejores intervalos de confianza, centrados en el estimador propuesto, para otras restricciones de orden generales. Korn (1982) propone un procedimiento para construir bandas de confianza para curvas dosis-respuesta isotónicas para las que no asume ningún modelo paramétrico. Más recientemente, Sampson et al. (2008) han propuesto un nuevo prodecimiento que generaliza el de Korn (1982).

También se ha estudiado el bootstrap y otras metodologías de remuestreo en modelos con restricciones de orden. Peddada (1997) estudia la construcción de intervalos de confianza para medias de poblaciones con restricciones de orden utilizando metodologías jackknife y bootstrap. Rueda et al. (2002) proponen una metodología para estimar una combinación lineal de las coordenadas de la media basada en procedimientos bootstrap que produce estimadores con menor error cuadrático medio. Li et al. (2010) construyen intervalos y regiones de confianza para probabilidades binomiales ordenadas basados en distribuciones asintóticas y en diversas versiones del bootstrap. Sin embargo, hay que tener cuidado con la aplicación del bootstrap a modelos con restricciones: el estimador bootstrap es generalmente inconsistente en inferencia con restricciones de desigualdad (Shaw y Geyer, 1997). Andrews (2000) prueba que el estimador bootstrap es inconsistente cuando el parámetro está en la frontera del espacio paramétrico definido por restricciones lineales o no lineales siendo alguna de desigualdad.

Tras el estudio de los procedimientos fundamentales de estimación y contraste, los procedimientos de inferencia bajo restricciones se han aplicado también a técnicas multivariantes. Entre los métodos del análisis multivariante que han incluido restricciones en los parámetros figuran el análisis de correlación canónica, el análisis de correspondencias, el análisis clúster y el análisis discriminante.

Das y Sen (1994) introducen restricciones en los parámetros en el cálculo de correlaciones canónicas. En un principio, el análisis se centra en las restricciones de no negatividad, para después extenderlo a restricciones de desigualdad más generales, y también a restricciones en solo algunos de los coeficientes. Cuando las restricciones son de no negatividad, Das y Sen (1996) establecen resultados de normalidad asintótica para los estimadores (obtenidos a partir de la matriz de covarianzas muestral). Kuriki (2005) desarrolla contrastes para la independencia de tablas de contingencia de dos dimensiones ordenadas en un marco general de análisis de correlaciones canónicas con restricciones de desigualdad, y discute algunos métodos numéricos para el ajuste de modelos con restricciones de orden.

Varios son los autores que han incluido restricciones en el análisis de correspondencias. Bockenholt y Bockenholt (1990) y Takane et al. (1991) consideran diversas formas de incorporar restricciones lineales, Groenen y Poblone (2003) aplican el análisis de correspondencias con restricciones de orden al problema de la datación arqueológica, y van de Velden et al. (2009) estudian su comportamiento mediante un estudio de simulación.

En cuanto al análisis clúster, Peddada et al. (2003) proponen un algoritmo que utiliza la inferencia estadística bajo restricciones de orden para seleccionar genes y agruparlos en clústers de acuerdo con su evolución temporal. En la misma situación, Liu et al. (2009) proponen un algoritmo rápido y con buenas propiedades de precisión y robustez.

Las aportaciones contenidas en esta memoria se desarrollan en el contexto del análisis discriminante con información adicional, por lo que la incorporación de la información adicional al análisis discriminante a través de restricciones sobre los parámetros se repasará con mayor detenimiento.

Supongamos que tenemos un número finito de poblaciones $\Pi_1, \ldots, \Pi_k$, cuya existencia conocemos a priori. Una observación cualquiera pertenece a una, y solo una, de las k poblaciones. Sea Y la variable categórica que identifica la población a la que pertenece la observación, donde $Y = i$ significa que la observación pertenece a la población Π_i, $i = 1, ..., k$. Asimismo, sea $\boldsymbol{X} = (X_1, ..., X_p)$ un vector p-dimensional que contiene p características asociadas a la observación. Básicamente, el problema del análisis discriminante consiste en la predicción de Y a partir de $\boldsymbol{X}$, es decir, en la clasificación de la observación en una de las k

poblaciones a partir de la información contenida en $\boldsymbol{X}$. El análisis discriminante se conoce también como clasificación supervisada, dado que en general las funciones de distribución de las poblaciones son desconocidas y se dispone de una muestra de entrenamiento de tamaño n, $M_n = \{(\boldsymbol{X}_i, Y_i), i = 1, ..., n\}$, formada por observaciones de las que conocemos el vector p-dimensional de clasificadores, $\boldsymbol{X}_i$, y la población a la que pertenecen, Y_i, para $i = 1, ..., n$. El objetivo es encontrar una regla de clasificación que sea óptima en algún sentido.

El análisis discriminante tiene una cantidad muy numerosa de aplicaciones: clasificación biológica (Sneath y Sokal, 1973), diagnósticos médicos y reconocimiento de patrones (McLachlan, 2004), desarrollo de nuevos fármacos (Wang, 2008), reconocimiento óptico de caracteres (Bunke y Wang, 1997), reconocimiento facial (Li y Jain, 2011), reconocimiento de voz y caracteres escritos a mano (Hastie et al., 1995), escáners médicos (Sonka y Fitzpatrick, 2004), identificación biométrica (Boulgouris et al., 2009), calificaciones de crédito (Thomas et al., 2002), modelos de relaciones estructura-actividad cuantitativas (Benigni, 2013), geo-estadística (Fraley y Raftery, 2002), procesamiento del lenguaje natural (Kao y Poteet, 2007), clasificación de documentos (Berry, 2004), motores de búsqueda (Chang et al., 2001)...

Algunas de las ideas asociadas al análisis discriminante se remontan a la década de 1920, en concreto el coeficiente de concordancia racial de Pearson (1926) y el índice posicional y varias medidas de distancia entre grupos de Mahalanobis (1927). El primer trabajo sobre análisis discriminante se debe a Fisher (1936), y trata un problema taxonómico de dos poblaciones. Rao (1948) lo generaliza a más de dos poblaciones. Desde entonces, la bibliografía es extensa. Basten como ejemplos significativos: Lachenbruch y Goldstein (1979), Hand (1981), McLachlan (2004) y Huberty y Olejnik (2006).

En esta memoria se estudia el análisis discriminante desde una perspectiva paramétrica cuando se dispone de información adicional sobre los parámetros. No hay mucha bibliografía sobre este tema, si bien una referencia inicial es Long y Gupta (1998), que tratan el problema de dos poblaciones normales con la misma matriz de covarianzas. Obtienen resultados de aplicación muy limitada, pues se limitan al caso de matriz de covarianzas igual a la matriz identidad y con las restricciones (de orden) de que una de las medias es mayor o igual que la otra, componente a componente. Las reglas se construyen sustituyendo en la regla de clasificación de Fisher los estimadores usuales de las medias por otros que tienen en cuenta las restricciones. Demuestran que se logran algunas mejoras respecto de la regla de Fisher. Posteriormente, Fernández et al. (2006) proponen reglas para situaciones más generales, basadas en nuevos estimadores de la diferencia de medias que tienen en cuenta la información adicional de que dicha diferencia pertenece a un cono convexo, poliédrico y cerrado, y que muestran un mejor comportamiento cuando la matriz de covarianzas es la misma entre las poblaciones y conocida, y comprueban mediante simulación que esto también ocurre cuando la matriz de covarianzas es la misma pero desconocida. Salvador et al. (2008) estudian propiedades de robustez de las reglas de clasificación que incorporan la información adicional contra contaminación de la muestra de entrenamiento, y prueban que estas reglas no solo tienen menor probabilidad total de clasificación errónea sino que también previenen contra algunos tipos de contaminación.

Desde una perspectiva no paramétrica, en Dykstra et al. (1999) se presentan procedimientos de clasificación no paramétricos con restricciones de orden entre la variable respuesta y las variables explicativas. En este sentido han aparecido recientemente algunos trabajos sobre clasificación cuando la variable respuesta es de tipo ordinal y hay restricciones de orden entre las variables explicativas y la variable respuesta. Un ejemplo es el de Auh y Sampson (2006), quienes proponen un procedimiento de discriminación isotónica logística que generaliza la discriminación logística lineal, relajando la linealidad y exigiendo tan solo monotonía en los discriminadores. En Kotlowski y Slovinski (2013) se analizan los procedimientos de aprendizaje con restricciones de monotonía no paramétricos desde un punto de vista estadístico, a través de un análisis teórico confirmado por simulaciones.

Para finalizar esta introducción, terminamos con un breve resumen de los tres artículos que conforman esta memoria.

En Conde et al. (2005) consideramos una regla de clasificación para poblaciones exponenciales cuando se sabe que hay un orden entre los parámetros, y probamos que dicha regla se comporta mejor que la basada en la verosimilitud. Además, estudiamos su comportamiento en cada una de las poblaciones comparando las probabilidades de clasificación errónea, y consideramos la incorporación de datos con censura de tipo II. Por último, evaluamos la regla para más de dos poblaciones por medio de un estudio de simulación. El estudio de las poblaciones exponenciales muestra que se pueden mejorar las reglas de clasificación en otros contextos distintos al de la distribución normal.

En Conde et al. (2012) hacemos una generalización no trivial a más de dos poblaciones de las reglas de clasificación con restricciones que aparecen en Fernández et al. (2006), y ofrecemos evidencia empírica que demuestra que la metodología con restricciones propuesta se comporta mejor que la metodología sin restricciones existente. Aplicamos la metodología a una situación de diagnóstico y tratamiento del cáncer, donde se quiere clasificar a los pacientes en uno de varios grupos de diagnosis a partir de determinados marcadores biológicos y se sabe que algunas de las variables predictoras toman, en media, valores mayores o menores en los pacientes de algunos de los grupos que en los de otros grupos. Esta clase de datos es muy común en la práctica, por ejemplo al tratar con datos de proteínas o microarrays. Utilizamos un conjunto de datos de este tipo, en concreto de cáncer de vejiga, y comprobamos que las nuevas reglas restringidas se comportan mejor que las no restringidas. La mejora es tanto o más importante cuantas más restricciones no cumpla la muestra de entrenamiento.

En Conde et al. (2013) comparamos las reglas expuestas en Fernández et al. (2006) con las basadas en estimadores shrinkage para las medias propuestas por Tong et al. (2012) con respecto a varios de los criterios más utilizados en la práctica: probabilidad de clasificación errónea total, área bajo la curva ROC, calibrado y refinamiento. La comparación viene motivada por el hecho de que las reglas de Fernández et al. (2006) se pueden ver como reglas contractivas al estar basadas en proyecciones y ser la proyección un operador contractivo. Comprobamos que estas reglas compiten bien, mejorando los resultados en varios de los escenarios. Además, se prueban resultados acerca de la tasa de error aparente que muestran la necesidad de nuevos estimadores de la tasa de error verdadero para las reglas anteriores, y se proponen cuatro nuevos estimadores bootstrap. Comparamos el comportamiento de estos estimadores en un estudio de simulación y los aplicamos a un conjunto de datos de cáncer de vejiga.

A continuación, en las secciones 2, 3 y 4 se detallan los objetivos (especificando hipótesis de partida, antecedentes y objetivos concretos), la metodología, los resultados y las conclusiones más relevantes de estos tres artículos. Por último, en la sección 5 se describen los trabajos en desarrollo y futuros.

2 A classification rule for ordered exponential populations

David Conde, Miguel A. Fernández, Bonifacio Salvador.
Journal of Statistical Planning and Inference **135** (2), 339-356 (2005).

Factor impacto 2012: 0.713. Puesto en categoría Statistics and Probability: 72/117.

2.1 Objetivos

2.1.1 Hipótesis de partida

En muchos problemas reales tiene sentido considerar restricciones en los parámetros utilizando la información adicional de que se dispone, y así mejorar los resultados que se obtienen con métodos convencionales. Este es el caso del análisis discriminante. Hasta la fecha de este trabajo, no se había investigado en profundidad la incorporación de información adicional a las reglas de clasificación a través de restricciones en los parámetros.

Asimismo, mientras la bibliografía relativa al análisis discriminante con poblaciones normales es abundante, lo es menos en el caso no normal. La distribución exponencial es muy interesante dado que es frecuente encontrarla en contextos prácticos como análisis de fiabilidad o supervivencia. Ejemplos donde se utiliza este análisis discriminante en poblaciones exponenciales se pueden encontrar en Zelen (1966) en el contexto de la investigación sobre el cáncer, y en Basu y Gupta (1974) sobre fiabilidad.

En este trabajo abordamos el estudio del análisis discriminante con poblaciones exponenciales cuando hay restricciones de orden simple en sus parámetros, y el modo de mejorar las reglas en este contexto.

2.1.2 Antecedentes

Algunas referencias sobre clasificación en el caso exponencial con aplicaciones son Bhattacharya y Das Gupta (1964), Basu y Gupta (1974) y Adegboye (1993). Referencias interesantes relativas a la estimación de parámetros ordenados en el caso no normal son Kaur y Singh (1991) y Vijayasree y Singh (1991, 1993) para poblaciones exponenciales, y Chang y Shinozaki (2002) para poblaciones gamma. Aquí se combinan ambas situaciones, considerando la clasificación con poblaciones exponenciales con restricciones de orden simple sobre los parámetros.

2.1.3 Objetivos concretos

Los objetivos concretos que se persiguen son:

1. Definición de reglas de clasificación para el modelo de dos poblaciones exponenciales con restricciones de orden simple sobre los parámetros y estudio de su consistencia.

2. Estudio del comportamiento de las reglas anteriores en comparación con la basada en la verosimilitud, comparando la probabilidad de clasificación errónea a nivel global y en cada una de las dos poblaciones.

3. Incorporación de datos con censura de tipo II a las reglas anteriores: definición y comportamiento.

4. Definición de reglas de clasificación cuando hay más de dos poblaciones exponenciales con restricciones de orden simple sobre los parámetros y estudio de su consistencia.

5. Simulación de la ganancia de las reglas definidas en el apartado anterior con respecto a la regla usual basada en la razón de verosimilitudes.

2.2 Metodología

En este trabajo nos planteamos la introducción de restricciones de orden sobre los parámetros en el análisis discriminante de poblaciones exponenciales. Nuestro objetivo es mejorar las reglas de clasificación en este contexto.

La función de pérdida a utilizar es la 0-1, y por lo tanto la pérdida esperada viene dada por la probabilidad de clasificación errónea, y se persigue la mejora respecto de las reglas que no tienen en cuenta la información adicional. Para ello, se buscan resultados analíticos que nos permitan obtener estimadores y establecer propiedades de los mismos, y se trata de establecer las condiciones bajo las que las nuevas reglas con restricciones son consistentes y proporcionan mejores soluciones que las no restringidas.

Además, en todo momento se hace uso de computación intensiva, necesaria para la implementación y aplicación de los nuevos procedimientos, y a la vez imprescindible como apoyo de las hipótesis a contrastar analíticamente, siendo en algunos casos, debido a la complejidad de los cálculos (como cuando el número de poblaciones es elevado), el único modo de abordar el problema.

2.3 Resultados

2.3.1 Reglas para dos poblaciones

Sean Π_1 y Π_2 dos poblaciones con distribuciones exponenciales uniparamétricas cuya función de densidad es:

$$f_i(x) = \lambda_i \mathrm{e}^{-\lambda_i x}, x > 0, i = 1, 2.$$

Supongamos que las probabilidades a priori y los costes de clasificación errónea son iguales. La regla de clasificación óptima de Bayes es:

$$R : \text{Clasificar } z \text{ en } \Pi_1 \text{ sii } (z - x_0)(\lambda_1 - \lambda_2) < 0,$$

donde $x_0 = \frac{\ln\lambda_1 - \ln\lambda_2}{\lambda_1 - \lambda_2}$. Si λ_1 y λ_2 son desconocidos y se dispone de una muestra de entrenamiento de cada población de tamaños n_1 y n_2, respectivamente, podemos obtener la regla estimando los parámetros. La regla estimada es:

$$R_U : \text{Clasificar } z \text{ en } \Pi_1 \text{ sii } (z - \hat{x}_0)(\hat{\lambda}_1 - \hat{\lambda}_2) < 0,$$

donde $\hat{\lambda}_1$ y $\hat{\lambda}_2$ son los EMV de λ_1 y λ_2 y $\hat{x}_0 = \frac{\ln\hat{\lambda}_1 - \ln\hat{\lambda}_2}{\hat{\lambda}_1 - \hat{\lambda}_2}$.

Si conocemos que $\lambda_1 \geq \lambda_2 > 0$, es decir, el valor esperado en Π_1 es menor o igual que en Π_2, podemos incorporar esta información a la regla. La regla que proponemos, y a la que llamaremos regla ordenada, es:

$$R_O : \text{Clasificar } z \text{ en } \Pi_1 \text{ sii } (z - \hat{x}_0) < 0,$$

dado que al tener la información de que $\lambda_1 \geq \lambda_2 > 0$, el segundo término no es necesario.

Supongamos que tenemos m observaciones $z_1, \ldots, z_m$, de una de las poblaciones Π_1 o Π_2 (todas las observaciones provienen de la misma población), y queremos clasificarlas (obviamente, todas en la misma población). Sea $P_*(i/j)$ la probabilidad de que las m observaciones de la población Π_j se clasifiquen en la población Π_i cuando se usa la regla R_*, $i, j = 1, 2$, y sea $P_*(MC) = \frac{1}{2}(P_*(1/2) + P_*(2/1))$ la probabilidad total de clasificación errónea de R_*.

Teorema 1 *Sea $z_1, \ldots, z_m$ una muestra aleatoria de Π_1 o Π_2. Si $m, n_1, n_2 > N$, entonces:*

$$\lim_{N\to\infty} P_O(MC) \to 0.$$

Este resultado prueba que la regla ordenada es consistente, en el sentido de que la probabilidad de clasificación errónea tiende a 0 cuando los tamaños muestrales m, n_1 y n_2 crecen.

Veamos ahora un resultado que demuestra que la nueva regla ordenada se comporta mejor que la regla usual: la probabilidad de clasificación errónea es menor para la regla ordenada.

Teorema 2 *Si $\lambda_1 \geq \lambda_2 > 0$, entonces:*

$$P_O(MC) \leq P_U(MC).$$

Además, $P_O(MC) < P_U(MC)$ cuando $\lambda_1 > \lambda_2 > 0$.

Vamos a comparar ahora las probabilidades de clasificación errónea de las reglas R_O y R_U en cada una de las dos poblaciones.

Teorema 3 *Sean $n_1 = n_2 = n$ y $m = 1$.*

1. *Si $n > 1$, $P_O(1/1) \geq P_U(1/1)$, $\forall \lambda_1 \geq \lambda_2 > 0$.*
2. *Si $n = 1$, existe $\delta_0 \in (0,1)$ tal que*

$$P_O(1/1) \geq P_U(1/1),\ \forall \lambda_1, \lambda_2 > 0,\ 0 < \tfrac{\lambda_2}{\lambda_1} \leq \delta_0.$$

$$P_O(1/1) < P_U(1/1),\ \forall \lambda_1, \lambda_2 > 0,\ \delta_0 < \tfrac{\lambda_2}{\lambda_1} < 1.$$

Teorema 4 *Sean $n_1 = n_2 = n$ y $m = 1$.*

1. *Si $n = 1$, $P_O(2/2) \geq P_U(2/2)$, $\forall \lambda_1 \geq \lambda_2 > 0$.*
2. *Si $n > 1$, existen δ_0^* y δ_1^* en $(0,1)$ tales que*

$$P_O(2/2) \geq P_U(2/2),\ \forall \lambda_1, \lambda_2 > 0,\ 0 < \tfrac{\lambda_2}{\lambda_1} \leq \delta_0^*.$$

$$P_O(2/2) < P_U(2/2),\ \forall \lambda_1, \lambda_2 > 0,\ \delta_1^* < \tfrac{\lambda_2}{\lambda_1} < 1.$$

Cuando $m > 1$ los cálculos se complican. Sin embargo, sabemos por las simulaciones que, cuando m crece, $P_O(1/1) - P_U(1/1)$ decrece y $P_O(2/2) - P_U(2/2)$ crece, disminuyendo la diferencia global $P_O(MC) - P_U(MC)$. Por tanto, la regla R_O es todavía mejor que R_U cuando $m > 1$ que cuando $m = 1$.

2.3.2 Reglas para dos poblaciones con datos censurados

La censura de tipo II aparece en experimentos que consisten en observar los tiempos de fallo de n unidades, finalizando cuando fallan un número predeterminado $r < n$ de unidades: las $n - r$ que no han fallado tienen censura por la derecha. El EMV del parámetro λ de una distribución exponencial a partir de una muestra aleatoria $X_1, \ldots, X_n$ es:

$$\lambda^* = \frac{r}{\Sigma_{i=1}^{r} X_{(i)} + (n-r)X_{(r)}}.$$

Es inmediato derivar las reglas anteriores bajo censura de tipo II:

$$R_{U^*} : \text{Clasificar } z \text{ en } \Pi_1 \text{ sii } (z - x_0^*)(\lambda_1^* - \lambda_2^*) < 0,$$

$$R_{O^*} : \text{Clasificar } z \text{ en } \Pi_1 \text{ sii } (z - x_0^*) < 0,$$

donde $x_0^* = \frac{\ln \lambda_1^* - \ln \lambda_2^*}{\lambda_1^* - \lambda_2^*}$.

En estas condiciones también demostramos que R_{O^*} es consistente, y que $P_{O^*}(MC) \leq P_{U^*}(MC)$ para $\lambda_1 \geq \lambda_2 > 0$.

2.3.3 Reglas para más de dos poblaciones

Cuando hay $k > 2$ poblaciones, las definiciones de las reglas son más complejas. Supongamos $\lambda_1 \geq ... \geq \lambda_k$, y sean $\hat{\lambda}_1,...,\hat{\lambda}_k$, los EMV de los parámetros en cada población y $\hat{\lambda}_{(1)} \leq ... \leq \hat{\lambda}_{(k)}$ los parámetros ordenados. Si $\hat{\lambda}_{(i)} = \hat{\lambda}_j$, entonces:

$$R_U : \text{Clasificar } z \text{ en } \Pi_j \text{ sii } \frac{\ln \hat{\lambda}_{(i)} - \ln \hat{\lambda}_{(i+1)}}{\hat{\lambda}_{(i)} - \hat{\lambda}_{(i+1)}} < z \leq \frac{\ln \hat{\lambda}_{(i-1)} - \ln \hat{\lambda}_{(i)}}{\hat{\lambda}_{(i-1)} - \hat{\lambda}_{(i)}},$$

$$R_O : \text{Clasificar } z \text{ en } \Pi_{k-i+1} \text{ sii } \frac{\ln \hat{\lambda}_{(i)} - \ln \hat{\lambda}_{(i+1)}}{\hat{\lambda}_{(i)} - \hat{\lambda}_{(i+1)}} < z \leq \frac{\ln \hat{\lambda}_{(i-1)} - \ln \hat{\lambda}_{(i)}}{\hat{\lambda}_{(i-1)} - \hat{\lambda}_{(i)}},$$

utilizando la desigualdad apropiada para las poblaciones Π_1 y Π_k.

Se puede probar la consistencia de la regla ordenada R_O. Las simulaciones indican que para $k > 2$ poblaciones, incluso para valores altos de k, R_O se comporta globalmente mejor que R_U. Para el caso particular de tres poblaciones y tamaños muestrales iguales $n_1 = n_2 = n_3 = n$, con $n \geq 1$, las simulaciones indican que $P_O(1/1) \geq P_U(1/1)$, $P_O(2/2) \leq P_U(2/2)$ y $P_O(3/3) \geq P_U(3/3)$.

2.4 Conclusiones

Proponemos una regla de clasificación R_O para dos poblaciones Π_1 y Π_2 con distribuciones exponenciales univariantes de parámetros λ_1 y λ_2, respectivamente, cuando se sabe que, en media, las observaciones de Π_1 son menores o iguales que las de Π_2, es decir, $\lambda_1 \geq \lambda_2$. Primero probamos que esta regla es consistente, es decir, la probabilidad de clasificación errónea de una muestra test cuyos m elementos pertenecen todos a una de las dos poblaciones tiende a 0 cuando los tamaños muestrales de las muestras test y entrenamiento tienden a infinito. Después probamos que la nueva regla ordenada se comporta mejor que la usual (basada en los EMV) R_U que no tiene en cuenta la información adicional contenida en las restricciones: la probabilidad de clasificación errónea es menor o igual para la regla ordenada para cualquier par de valores $\lambda_1 \geq \lambda_2 > 0$, y estrictamente menor para $\lambda_1 > \lambda_2 > 0$. Esto se cumple para una única observación a clasificar o para una muestra de observaciones provenientes todas de la misma población. Es interesante destacar que las probabilidades de clasificación errónea de las reglas R_O y R_U dependen de λ_1 y de λ_2 solo a través del cociente λ_2/λ_1, y también del tamaño muestral, ya sea el mismo $n_1 = n_2 = n$ o distinto.

Cuando se clasifica una muestra test de tamaño $m > 1$, hemos comprobado a partir de simulaciones que la probabilidad de clasificación correcta es mayor para la regla ordenada, creciendo con m la diferencia con respecto a la regla no restringida.

Hemos estudiado el comportamiento de esta regla R_O en cada una de las poblaciones, comparando las probabilidades de clasificación correcta, y hemos comprobado, para iguales tamaños muestrales $n_1 = n_2 = n$ y una única observación test a clasificar, que aunque la regla ordenada se comporta mejor globalmente, hay algunas configuraciones de los parámetros para las que esta propiedad no se verifica en una de las dos poblaciones. En el caso más habitual $n > 1$ la regla ordenada se comporta siempre mejor para la población Π_1, no así para Π_2, y lo contrario sucede cuando $n = 1$, caso de poco interés práctico. Los valores concretos que determinan estas configuraciones son interesantes desde un punto de vista práctico, y se pueden obtener numéricamente. Se proporciona una tabla con dichos valores para distintos tamaños muestrales.

En situaciones como los análisis de fiabilidad en que está presente la distribución exponencial, es habitual que los datos tengan censura de tipo II. En este caso se conoce la expresión del EMV para cada parámetro $\lambda_i, i = 1,2$. Las reglas anteriores se definen para datos censurados sustituyendo estos estimadores en lugar de los anteriores para datos no censurados. Es fácil comprobar que la regla ordenada también se comporta mejor que la no ordenada en el sentido de una menor probabilidad de clasificación errónea cuando $\lambda_1 \geq \lambda_2 > 0$.

Por último, definimos la regla cuando hay más de dos poblaciones y evaluamos su comportamiento por medio de un estudio de simulación, el cual indica que la regla ordenada se comporta globalmente mejor que la no ordenada sea cual sea el número de poblaciones.

3 Classification of samples into two or more ordered populations with application to a cancer trial

David Conde, Miguel A. Fernández, Cristina Rueda, Bonifacio Salvador.
Statistics in Medicine **31**, 3773-3786 (2012).

Factor impacto 2012: 2.044. Puesto en categoría Statistics and Probability: 11/117.

3.1 Objetivos

3.1.1 Hipótesis de partida

En muchas aplicaciones como las relacionadas con diagnóstico y tratamiento del cáncer, se quiere clasificar a los pacientes en uno de varios grupos de diagnosis a partir de determinados marcadores biológicos. Con frecuencia, se sabe que algunas de las variables predictoras toman, en media, valores mayores o menores en los pacientes de algunos de los grupos que en los de otros grupos. En muchas ocasiones se dispone de dicha información adicional, y en otras muchas esta información se puede obtener indagando en el problema. Las reglas de clasificación tradicionales que no tienen en cuenta la información adicional pueden presentar altas tasas de clasificación errónea, especialmente cuando el número de grupos es mayor que dos. Aquí nos proponemos diseñar reglas de clasificación para más de dos poblaciones normales con la misma matriz de covarianzas, en situaciones en las que existe información sobre el orden de las medias de algunos predictores, que tengan una menor tasa de clasificación errónea que las reglas sin restricciones. Desde un punto de vista teórico, lo que pretendemos es una extensión, que no es en absoluto trivial, de las reglas de clasificación con restricciones de orden para dos poblaciones.

Hay que destacar que la clase de datos que vamos a utilizar es muy común en la práctica, por ejemplo al tratar con datos de proteínas o microarrays, o, en general, cuando se dispone de un pequeño conjunto de datos a partir del cual hay que derivar una regla de clasificación y se dispone de información adicional.

3.1.2 Antecedentes

Los resultados previos más relevantes en el problema de discriminación se refieren al caso de dos poblaciones normales. Se puede destacar el trabajo de Fernández et al. (2006), donde se diseñan reglas de clasificación para dos poblaciones normales ordenadas, con matrices de covarianzas desconocidas pero iguales, que mejoran a la regla de Fisher incorporando nuevos estimadores. En Salvador et al. (2008) se estudian las propiedades de robustez de las reglas de Fernández et al. (2006).

Aquí se pretende generalizar a más de dos poblaciones la metodología existente para dos poblaciones en Fernández et al. (2006). Es importante señalar que la extensión a más de dos poblaciones no es directa: cuando hay solo dos poblaciones, no hay muchas posibilidades de ordenarlas, pero cuando hay más de dos poblaciones puede aparecer un tipo de orden diferente para cada una de las variables predictoras. Además, en Fernández et al. (2006) la definición de las reglas restringidas se basaba en la diferencia de las medias de las dos poblaciones mientras que en el caso de más de dos poblaciones existen múltiples diferencias por pares, por lo que la extensión no es en absoluto directa.

3.1.3 Objetivos concretos

Los objetivos concretos que se persiguen son:

1. Definición de estimadores para las medias de más de dos poblaciones normales con restricciones de orden entre los parámetros, apropiados para la discriminación.
2. Definición de reglas de clasificación con los estimadores definidos en el apartado anterior y demostración de que generalizan a las ya existentes para dos poblaciones.
3. Diseño de experimentos de simulación suficientemente completos para mostrar que las nuevas reglas mejoran a las reglas que no tienen en cuenta la información adicional.
4. Aplicación de las reglas de clasificación a problemas reales.

3.2 Metodología

En este trabajo nos planteamos la introducción de restricciones de orden en los parámetros en el análisis discriminante con más de dos poblaciones normales. El hecho de que el problema tenga implicaciones prácticas hace que el mayor interés de las reglas esté en su capacidad de clasificar correctamente las observaciones, por lo que los estimadores y las reglas definidas a partir de ellos deben tener en cuenta esta consideración.

La función de pérdida a utilizar es la 0-1, y por lo tanto la pérdida esperada viene dada por la probabilidad de clasificación errónea, y se compara con la de las reglas que no tienen en cuenta la información adicional. Para ello, se hace uso de computación intensiva, necesaria para la implementación y aplicación de los nuevos procedimientos, y al mismo tiempo imprescindible como apoyo de las hipótesis a contrastar, siendo en algunos casos, a causa de la complejidad de los cálculos cuando el número de poblaciones es elevado, la única manera de abordar el problema.

3.3 Resultados

3.3.1 Estimadores y reglas para más de dos poblaciones

Supongamos que tenemos k poblaciones $\Pi_1,...,\Pi_k$. Sea $\boldsymbol{X} = (X_1,...,X_p)$ un vector de variables predictoras que vamos a usar para clasificar las observaciones. Asumimos normalidad, es decir, que si la observación considerada proviene de la población Π_i, entonces $\boldsymbol{X} \sim N_p(\mu_i,\Sigma)$, $i = 1,...,k$. Suponemos diferentes medias en cada población pero matriz de covarianzas común Σ, función de pérdida 0-1 e iguales probabilidades a priori para las poblaciones (la extensión a probabilidades a priori distintas es trivial). En esta situación, la regla óptima de Bayes es:

$$\text{Clasificar } z \text{ en } \Pi_i \text{ sii } (z-\mu_i)'\Sigma^{-1}(z-\mu_i) \leq (z-\mu_j)'\Sigma^{-1}(z-\mu_j), j = 1,\ldots,k,$$

que, estimando los parámetros $\mu_1,...,\mu_k$ y Σ cuando son desconocidos por sus estimadores $\overline{X}_1,...,\overline{X}_k$ y S, se convierte en la regla habitual de Fisher:

$$\text{Clasificar } z \text{ en } \Pi_i \text{ sii } (z-\overline{X}_i)'S^{-1}(z-\overline{X}_i) \leq (z-\overline{X}_j)'S^{-1}(z-\overline{X}_j), j = 1,\ldots,k,$$

donde S es la matriz de covarianzas muestral pooled:

$$S = \frac{\sum\limits_{j=1}^{k}(n_j-1)S_j}{n-k},$$

siendo n_j y S_j el tamaño muestral y la matriz de covarianzas muestral de la muestra de la población Π_j, $j = 1,...,k$.

A continuación, se definen reglas de clasificación que tienen en cuenta la información adicional, que se incorpora a la regla restringiendo el espacio paramétrico. Proponemos trabajar en un espacio extendido $\mathbb{R}^{k\times p}$ y, con la información adicional disponible, definir un cono C de restricciones en $\mathbb{R}^{k\times p}$ al cual pertenezca el vector de medias: $(\mu_1',...,\mu_k')' \in C$. Si proyectamos el vector de medias muestral sobre el cono con la métrica adecuada nos aseguramos de que el estimador de las medias esté tan próximo al vector de medias muestral como sea posible al tiempo que verifica las restricciones que determinan la información adicional.

La métrica para la proyección que utilizaremos es la proporcionada por la matriz de dimensiones $(k\times p)\times(k\times p)$:

$$S_*^{-1} = diag\left(\frac{S}{n_1},...,\frac{S}{n_k}\right)^{-1}.$$

Es fácil expresar las restricciones de orden utilizando conos en $\mathbb{R}^{k\times p}$. En nuestro caso multivariante, si las restricciones afectan a un conjunto de variables $T \subseteq \{1,...,p\}$, los conos más habituales se pueden escribir:

- Octante positivo:

$$C_{O^+} = \left\{x \in \mathbb{R}^{k\times p} : x_{t+np} \geq 0, t \in T, n = 0,1,...,k-1\right\}$$

- Orden simple:

$$C_{SO} = \left\{x \in \mathbb{R}^{k\times p} : x_t \leq x_{t+p} \leq ... \leq x_{t+(k-1)p}, t \in T\right\} \tag{1}$$

- Árbol:

$$C_{TO} = \left\{x \in \mathbb{R}^{k\times p} : x_t \leq x_{t+np}, t \in T, n = 1,...,k-1\right\} \tag{2}$$

- Loop:

$$C_{LO} = \left\{x \in \mathbb{R}^{k\times p} : x_t \leq [x_{t+p},...,x_{t+(k-2)p}] \leq x_{t+(k-1)p}, t \in T\right\}$$

- Unimodalidad:

$$C_{UO} = \left\{x \in \mathbb{R}^{k\times p} : x_t \leq ... \leq x_{t+(r-1)p} \geq x_{t+rp} \geq ... \geq x_{t+(k-1)p}, t \in T\right\}$$

- Estrella (star shaped):

$$C_{SS} = \left\{x \in \mathbb{R}^{k\times p} : x_t \leq \frac{x_t + x_{t+p}}{2} \leq \frac{x_t + x_{t+p} + x_{t+2p}}{3} \leq ..., t \in T\right\}$$

Ahora definimos una familia de estimadores $\widehat{\mu}_i^{\gamma}$, $\gamma \in [0,1]$, $i = 1,...,k$, utilizando las proyecciones.

Definición 5 *Sea $\widehat{\mu}^{\gamma} \in \mathbb{R}^{k\times p}$ para $\gamma \in [0,1]$ el límite cuando $l \to \infty$ del siguiente procedimiento iterativo:*

$$\widehat{\mu}^{\gamma(l)} = P_{S_*^{-1}}(\widehat{\mu}^{\gamma(l-1)} \mid C) - \gamma P_{S_*^{-1}}(\widehat{\mu}^{\gamma(l-1)} \mid C^P), l = 1,2,...,$$

donde $\widehat{\mu}^{\gamma(0)} = \left(\overline{X}_1',...,\overline{X}_g'\right)' \in \mathbb{R}^{k\times p}$, $P_{S_*^{-1}}(Y \mid C)$ es la proyección de Y sobre el cono C utilizando la métrica dada por S_*^{-1} y $C^P = \{y \in \mathbb{R}^{k\times p} : y' S_*^{-1} x \leq 0, \forall x \in C\}$ es el cono polar de C.

Teorema 6 *El procedimiento anterior converge y, además, se cumple:*

$$\widehat{\mu}^{\gamma} = \lim_{l\to\infty} \mu^{\gamma(l)} \in C,\ \forall \gamma \in [0,1].$$

La Figura 1 muestra la necesidad del procedimiento iterativo para algunos tipos de conos de restricciones, ya que en general un único paso del procedimiento no garantiza la pertenencia al cono.

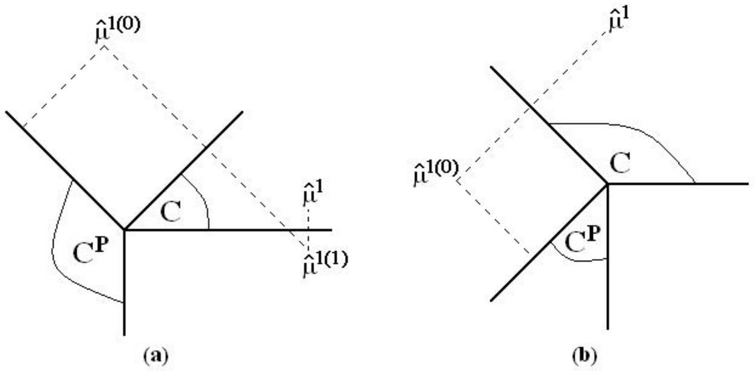

Figura 1: Ejemplos del procedimiento iterativo para la estimación de un vector de medias con un cono agudo (a) y no agudo (b).

Sea $\widehat{\mu}_i^\gamma = \left((\widehat{\mu}^\gamma)_{(i-1)p+1}, (\widehat{\mu}^\gamma)_{(i-1)p+2}, ..., (\widehat{\mu}^\gamma)_{ip}\right)', i = 1, ..., k$. Las nuevas reglas, que denotamos $R_\mu(\gamma)$, se definen:

$$\text{Clasificar } z \text{ en } \Pi_i \text{ sii } (z-\widehat{\mu}_i^\gamma)'S^{-1}(z-\widehat{\mu}_i^\gamma) \leq (z-\widehat{\mu}_j^\gamma)'S^{-1}(z-\widehat{\mu}_j^\gamma), j = 1, \ldots, k.$$

Si $\gamma = 0$, $\widehat{\mu}^0$ es el EMV estándar de la inferencia con restricciones de orden: es la proyección sobre el cono de restricciones. Cuando $\gamma > 0$, los estimadores pertenecen al interior del cono (ver Figura 1). Se comprueba que las reglas $R_\mu(\gamma)$ con $\gamma > 0$ se comportan mejor que las reglas $R_\mu(0)$.

El siguiente resultado prueba que estas reglas generalizan a las definidas para dos poblaciones en Fernández et al. (2006), denominadas $R_\delta(\gamma)$:

Teorema 7 *Si $k = 2$, entonces $R_\delta(\gamma)$ y $R_\mu(\gamma)$ son equivalentes para $\gamma \in [0, 1]$.*

3.3.2 Comportamiento de las reglas en simulaciones

Para estudiar el comportamiento de las reglas $R_\mu(\gamma)$ hemos realizado un estudio de simulación. Por simplicidad, nos concentramos en el caso $k = 3$ poblaciones con distribución $N_3(\mu_i, \Sigma)$, $i = 1, 2, 3$, bajo las dos restricciones de orden más comunes en la práctica: el orden simple (1) y el orden en árbol (2). Para cada una de ellas, se consideran los escenarios correspondientes a varias configuraciones de los vectores de medias y varias matrices de covarianzas distintas que tratan de cubrir los valores de los coeficientes de correlación más comunes en la práctica. Una primera conclusión es que las reglas $R_\mu(\gamma)$ se comportan mejor que la no restringida en todos los escenarios y restricciones de orden. Una segunda conclusión es la conveniencia de utilizar valores $\gamma > 0$: las simulaciones muestran que $R_\mu(1)$ se comporta mejor que $R_\mu(0)$ en prácticamente todas las situaciones. Sin embargo, como cabe esperar, las diferencias no son muy grandes, puesto que cuando las muestras de entrenamiento verifican las restricciones, las reglas coinciden con la de Fisher. Hemos estudiado los resultados según el número de restricciones que verifican las muestras de entrenamiento, y hemos comprobado que las mejoras son mayores cuando las muestras de entrenamiento no cumplen varias de las restricciones.

Las restricciones de orden simple son más restrictivas que las de orden en árbol, y los resultados de las simulaciones son mejores para las restricciones de orden simple, independientemente de la matriz de covarianzas. Podemos concluir que cuanto más precisa sea la información adicional disponible, mejores serán las reglas.

3.3.3 Aplicación

Por último, tras obtener la autorización correspondiente, hemos aplicado las reglas a un conjunto de datos de cáncer de vejiga proporcionados por Proteomika S.L. y Laboratorios SALVAT, S.A. y tomados con el objetivo de desarrollar un test in vitro no invasivo para el diagnóstico de la recurrencia del cáncer de vejiga y así complementar y reducir el número de cistoscopias. Se clasifican los pacientes en cinco niveles según los resultados de la cistoscopia. El primer nivel es el control (ausencia de cáncer de vejiga), siendo los demás T_a, T_1G_1, T_1G_3 y T_2, niveles progresivamente más avanzados del cáncer: la etapa T describe el tamaño del tumor y su extensión, y el grado G está relacionado con la apariencia de las células en el microscopio.

Hemos considerado dos conjuntos de datos, ambos obtenidos con la tecnología xMAP® de Luminex para medir los valores de las proteínas. El primer conjunto de datos que recibimos, D_1, contenía información de 141 pacientes y 11 proteínas además del nivel real al que pertenecían los pacientes. Este conjunto inicial de datos es con el que construimos las reglas: la muestra de entrenamiento. El segundo conjunto de datos, D_2, es la muestra test, que utilizamos para medir el comportamiento de las reglas de clasificación construidas con el primero. Fue recibido con posterioridad y contiene información relativa a otros 149 pacientes y las mismas 11 proteínas, así como el nivel real de la enfermedad. Tanto en una muestra como en la otra, tomamos logaritmos para que las variables tengan una distribucón aproximadamente normal.

A partir de la información adicional proporcionada, utilizamos para el análisis solo tres proteínas que eran consideradas relevantes, y que denotamos por P_1, P_2 y P_3: los valores medios de estas tres proteínas aumentan con el nivel de la enfermedad. Además, estudios sencillos utilizando t-tests nos indicaron que no todas las proteínas eran significativas. A causa de que los tamaños muestrales en algunos de los niveles son muy pequeños en comparación con los otros y porque clínicamente parecía razonable, redujimos los niveles a tres: el grupo control y los grupos $T_a + T_1 G_1$ y $T_1 G_3 + T_2$.

Hemos comprobado que no solo las nuevas reglas $R_\mu(\gamma)$ mejoran a la de Fisher, sino a otros muchos métodos de construcción de reglas de clasificación como support vector machines, k vecinos más próximos, clasificadores de Bayes naive o redes neuronales, e incluso clasificación ordinal, y hemos comprobado que las nuevas reglas se comportan mucho mejor que las otras tanto para $\gamma = 0$ como, y sobre todo, para $\gamma = 1$, como puede verse en el artículo.

3.4 Conclusiones

Los resultados de las simulaciones y el ejemplo muestran que las nuevas reglas restringidas definidas en este artículo se comportan mejor que las no restringidas, y que la mejora puede ser muy significativa cuando la muestra de entrenamiento no cumple algunas de las restricciones.

El hecho de que las reglas restringidas coincidan con las no restringidas cuando la muestra de entrenamiento verifica las restricciones permite recomendar el uso de las reglas propuestas en la práctica, ya que nunca se pierde con respecto a la regla usual no restringida. Para facilitar su uso, hemos compilado una librería de R para análisis discriminante con información adicional, de nombre `dawai`, que comentaremos más adelante.

Hemos probado que cuanto más precisa sea la información adicional, más potentes serán las reglas que la incorporen. Por tanto, recomendamos incorporar cuanta información tengamos a nuestra disposición. Esta clase de información es frecuente, por ejemplo, en problemas de diagnóstico médico, donde las variables predictoras suelen estar relacionadas isotónicamente con la variable respuesta.

Una cuestión importante es la estimación de la probabilidad de clasificación errónea. En este artículo queda abierta esta cuestión, si bien se intuye que el comportamiento de la tasa de error aparente de estas reglas es diferente al de la tasa de error aparente de las reglas usuales, por lo que puede ser necesaria la definición de nuevos procedimientos correctores del sesgo de la tasa de error aparente de las reglas restringidas.

4 Performance and estimation of the true error rate of classification rules built with additional information. An application to a cancer trial

David Conde, Bonifacio Salvador, Cristina Rueda, Miguel A. Fernández.
Statistical Applications in Genetics and Molecular Biology **12** (5), 583-602 (2013).

Factor impacto 2012: 1.717. Puesto en categoría Statistics and Probability: 18/117.

4.1 Objetivos

4.1.1 Hipótesis de partida

Como ya hemos comentado, las reglas de clasificación que incorporan la información adicional disponible en el problema a través de restricciones sobre los parámetros se comportan mejor que las reglas que no tienen en cuenta esta información adicional, dado que tienen menor probabilidad de clasificación errónea total que la regla de Fisher (Fernández et al., 2006). En este artículo nos planteamos comparar estas reglas con las basadas en estimadores shrinkage para las medias propuestas por Tong et al. (2012) en escenarios de alta dimensionalidad con respecto a varios de los criterios más utilizados en la práctica. La comparación viene motivada por el hecho de que estas reglas pueden verse como reglas contractivas al estar basadas en proyecciones.

Un segundo objetivo de este artículo es completar el estudio de las reglas con restricciones, en el caso de dos poblaciones, mediante la evaluación del rendimiento de las mismas para una muestra de entrenamiento dada, lo que es más útil en la práctica que la probabilidad (incondicional) de clasificación errónea. La tasa de error verdadero es la probabilidad de clasificación errónea condicionada a la muestra de entrenamiento disponible. La mejor forma de estimar el error verdadero es con una muestra test, lo cual habitualmente no es posible en la práctica. Por lo tanto, nos proponemos encontrar estimadores de la tasa de error verdadero basados en la muestra de entrenamiento y la información adicional disponible.

4.1.2 Antecedentes

Como ya hemos mencionado anteriormente, en las aplicaciones es habitual que haya información adicional, y hemos visto que la incorporación de esta clase de información en las reglas de clasificación ha demostrado que mejora el rendimiento de la regla.

Dado que las reglas de Fernández et al. (2006) se pueden ver como reglas contractivas al estar basadas en proyecciones y ser la proyección un operador contractivo, comenzamos comparando estas reglas con las reglas propuestas por Tong et al. (2012) con respecto a varios de los criterios más utilizados: probabilidad de clasificación errónea total, área bajo la curva ROC (ver Pepe et al., 2006) y calibrado y refinamiento (introducidos por Kim y Simon, 2011).

La otra cuestión de interés considerada en este artículo, la estimación de la tasa de error verdadero, es una cuestión ampliamente estudiada en la literatura en reglas como la de Fisher, la discriminante cuadrática, las de los vecinos más próximos o las de random forests. Se han propuestos estimadores paramétricos y no paramétricos del error verdadero, y los estimadores no paramétricos basados en remuestreo se comportan bien para las reglas mencionadas. Schiavo y Hand (2000) resumen el trabajo realizado hasta la fecha. Referencias más recientes son Fu et al. (2005) o Kim (2009).

4.1.3 Objetivos concretos

Los objetivos concretos que se persiguen son:

1. Comparación de las reglas restringidas con las contractivas de Tong et al. (2012) en escenarios de alta dimensionalidad.

2. Obtención de propiedades de la tasa de error aparente de las reglas restringidas.

3. Definición de nuevos estimadores de la tasa de error verdadero basados en la información adicional.

4. Estudio del comportamiento de los nuevos estimadores de la tasa de error verdadero respecto de estimadores clásicos que no tienen en cuenta la información adicional.

5. Aplicación de las reglas de clasificación y los nuevos estimadores de la tasa de error verdadero a datos reales.

4.2 Metodología

El primer objetivo se lleva a cabo utilizando varios de los criterios más utilizados en la actualidad, como son la probabilidad de clasificación errónea total, el área bajo la curva ROC, el calibrado y el refinamiento, haciendo uso de simulaciones, dado que, a causa de la alta dimensionalidad, es la única manera de abordar el problema.

- El área bajo la curva ROC se utiliza con frecuencia (aunque no en exclusiva) en contextos médicos. Un diagnóstico es positivo o negativo dependiendo de si el correspondiente clasificador probabilístico $\tilde{p}(u)$ es mayor que o menor o igual que un determinado valor umbral perteneciente al intervalo $(0,1)$. Para cada valor umbral, la *sensibilidad* se define como la probabilidad de un verdadero positivo y la *especifidad* como la probabilidad de un verdadero negativo. La curva ROC es la gráfica *sensibilidad* contra $1-$*especifidad* para cada umbral. El área bajo la curva es obviamente la comprendida entre la curva y el eje $1-$*especifidad*. Cuando se dispone de una muestra de tamaños n_1 y n_2 en cada población, el área bajo la curva ROC (AUC) se puede estimar a través del estadístico U de Mann-Whitney (ver Pepe et al., 2006):

$$\widehat{AUC} = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left(I_{[\tilde{p}(u_i)>\tilde{p}(u_j)]} + \frac{1}{2}I_{[\tilde{p}(u_i)=\tilde{p}(u_j)]}\right).$$

- Se dice que un clasificador probabilístico $\tilde{p}$ está bien calibrado si se cumple $P(Y=1|\tilde{p}(u)=w)=w$ para cada probabilidad de predicción w, con $0<w<1$. Es decir, si la proporción de los individuos que el clasificador clasifica en la población $Y=1$ con probabilidad w es efectivamente w.

Se dice que un clasificador probabilístico $\tilde{p}$ es refinado si las probabilidades de predicción w tienden a 0 o 1. Se define el refinamiento como:

$$E_w\left[P(Y=1|\tilde{p}(u)=w)P(Y=2|\tilde{p}(u)=w)\right].$$

Para evaluar el calibrado y el refinamiento, Kim y Simon (2011) definen dos medidas: *CS (calibration score)* y *RS (refinement score)*. Se divide el intervalo unidad en m subintervalos iguales $B_k = ((k-1)/m, k/m]$, $k=1,\ldots,m$; para cada B_k, sea q_k la proporción de predicciones w que caen en B_k, r_k la frecuencia relativa de predicciones en B_k para la población 1, y u_k el punto central de B_k, $k=1,\ldots,m$. En estas condiciones:

$$CS = \sum_{k=1}^{m}(r_k-u_k)^2 q_k, \quad RS = \sum_{k=1}^{m} r_k(1-r_k)q_k.$$

Los restantes objetivos están relacionados con la estimación de la tasa de error verdadero. Primero, se buscan resultados analíticos de la tasa de error aparente que proporcionen información del comportamiento del sesgo del error aparente en las reglas con restricciones. Después, se definen nuevos estimadores de la tasa de error verdadero para las reglas con restricciones que corrijan el sesgo del error aparente. Estos estimadores están basados en técnicas de remuestreo, en concreto bootstrap y validación cruzada, y su definición parte de la idea de que el mundo bootstrap debería reflejar el mundo real. Por último, se evalúa, mediante un estudio de simulación y en un conjunto de datos reales, el comportamiento de los estimadores propuestos a partir de la denominada *deviation distribution* (Braga-Neto y Dougherty, 2004): si $\hat{E}$ es un estimador del error verdadero E_n, la distribución de la variable aleatoria $(\hat{E}-E_n)$ se conoce como *deviation distribution*, y una medida global del comportamiento del estimador viene dada por $E[(\hat{E}-E_n)^2]$.

4.3 Resultados

4.3.1 Comparación con reglas contractivas basadas en shrinkage

Sean Π_1 y Π_2 dos poblaciones con distribuciones $N_p(\mu_1,\Sigma)$ y $N_p(\mu_2,\Sigma)$. Suponemos diferentes medias en cada población pero matriz de covarianzas común, función de pérdida 0-1 e iguales probabilidades a priori para las poblaciones. Si $z = (z_1,...,z_p)'$ es una nueva observación a clasificar, la regla óptima (teórica) de Bayes se puede escribir, como muestran Fernández et al. (2006):

$$\text{Clasificar } z \text{ en } \Pi_1 \text{ sii } (z-(c_1\mu_1+c_2\mu_2)+c\delta)'\Sigma^{-1}\delta \geq 0,$$

con $c_i = n_i/(n_1+n_2), i = 1,2$, $c = c_1-c_2$ y $\delta = \mu_1-\mu_2$. Si $\overline{X}_1$ y $\overline{X}_2$ son los vectores de medias muestrales para las observaciones de las poblaciones Π_1 y Π_2, respectivamente, entonces la regla de Fisher se obtiene sustituyendo los parámetros desconocidos por sus estimadores:

$$\text{Clasificar } z \text{ en } \Pi_1 \text{ sii } (z-(c_1\overline{X}_1+c_2\overline{X}_2)+c\overline{\delta})'S^{-1}\overline{\delta} \geq 0,$$

siendo $\overline{\delta} = \overline{X}_1-\overline{X}_2$ y S la matriz de covarianzas muestral pooled. Asumimos que la información adicional se incorpora a través de la restricción $\delta \in C$, con C un cono poliédrico, cerrado y convexo de $\mathbb{R}^p$: $C = \{x \in \mathbb{R}^p : a_j'x \geq 0, j = 1,...,m\}$.

Las reglas de clasificación definidas en Fernández et al. (2006), denominadas $R_\delta(\gamma)$, se obtienen sustituyendo en la regla de Bayes Σ por S, $c_1\mu_1+c_2\mu_2$ por $c_1\overline{X}_1+c_2\overline{X}_2$ y δ por un miembro de la familia δ_γ^*, $\gamma \in [0,1]$, definida como el límite de un procedimiento iterativo cuya convergencia se demuestra en dicho artículo:

$$\widehat{\delta}_\gamma^{(0)} = \overline{X}_1-\overline{X}_2, \widehat{\delta}_\gamma^{(i)} = P_{S^{-1}}(\widehat{\delta}_\gamma^{(i-1)}/C)-\gamma P_{S^{-1}}(\widehat{\delta}_\gamma^{(i-1)}/C^P), i = 1,2,...$$

Así, las nuevas reglas son:

$$R_\delta(\gamma): \text{Clasificar } z \text{ en } \Pi_i \text{ sii } (z-(c_1\overline{X}_1+c_2\overline{X}_2)+c\delta_\gamma^*)'S^{-1}\overline{\delta}_\gamma^* \geq 0.$$

Si denotamos $\mu_{\gamma 1}^* = c_1\overline{X}_1+c_2(\overline{X}_2+\delta_\gamma^*)$ y $\mu_{\gamma 2}^* = c_1(\overline{X}_1-\delta_\gamma^*)+c_2\overline{X}_2$, entonces las reglas $R_\delta(\gamma)$ se pueden obtener sustituyendo en la regla de Bayes μ_1, μ_2, δ y Σ por $\mu_{\gamma 1}^*$, $\mu_{\gamma 2}^*$, $\mu_{\gamma 1}^*-\mu_{\gamma 2}^*$ y S.

Dado que la proyección es un operador contractivo, los estimadores $\mu_{\gamma 1}^*$ y $\mu_{\gamma 2}^*$ anteriores se pueden considerar como estimadores contractivos de las medias. Tong et al. (2012) proponen un estimador shrinkage del tipo James-Stein para la media bajo función de pérdida cuadrática, para tamaño muestral fijo n_i, $i = 1,2$, y dimensión p elevada, y proponen la regla que denominan *SmDLDA*, reemplazando las medias poblacionales por los estimadores shrinkage de las medias, y considerando una matriz de covarianzas diagonal para evitar problemas de singularidad si p es mayor que n_i, $i = 1,2$. Ver Tong et al. (2012) para una descripción detallada de su regla y el buen funcionamiento de la misma. A través de un estudio de simulación en escenarios similares a los que aparecen en Tong et al. (2012) (p grande y no tanto el tamaño muestral), comparamos el comportamiento de las reglas de *Fisher*, *SmDLDA* y $R_\delta(\gamma)$ para $\gamma = 0,1$, es decir, $R_\delta(0)$ y $R_\delta(1)$, con respecto a varios de los criterios más utilizados en la actualidad: probabilidad de clasificación errónea total (también conocida como tasa de clasificación errónea o error de predicción), área bajo la curva ROC, y calibrado y refinamiento (ver Pepe et al., 2006, y Kim y Simon, 2011). Como resultado de las simulaciones, podemos decir que, incluso con datos de altas dimensiones en problemas de tamaño muestral pequeño, $R_\delta(0)$ y $R_\delta(1)$ compiten bien con *SmDLDA*, superándola en varias configuraciones.

4.3.2 Estimación de la tasa de error verdadero de las reglas restringidas

En la práctica, es necesario poder evaluar el comportamiento de una regla de clasificación con una muestra de entrenamiento dada, y dado que no es habitual disponer de una muestra test independiente con la que poder estimar el error verdadero (probabilidad de clasificación errónea condicionada a la muestra de entrenamiento disponible), nos proponemos encontrar estimadores de la tasa de error verdadero. La tasa de error aparente (*APP*) estima la tasa de error verdadero como la proporción de observaciones de la muestra de entrenamiento mal clasificadas por la regla. Se sabe que, en general, subestima la tasa de error verdadero pues la muestra de entrenamiento se utiliza a la vez para crear la regla y para evaluarla.

Para las reglas con restricciones obtenemos el siguiente resultado en lo que se refiere a *APP*:

Teorema 8 *Si $n_1 = n_2$, para matriz de covarianzas conocida Σ y $\gamma \in [0,1]$, se cumple:*

$$E(APP(R_\delta(\gamma))) \geq E(APP(R_\delta(0))) \geq E(APP(Fisher))).$$

En Fernández et al. (2006) se prueba que, si $n_1 = n_2$, la tasa de error verdadero de las reglas $R_\delta(\gamma)$, $\gamma \in [0,1]$, es menor que la de la regla de Fisher. Además, se prueba en las simulaciones que la tasa de error verdadero es mayor que *APP*. En virtud del teorema anterior, si $n_1 = n_2$ el sesgo del error aparente de las reglas $R_\delta(\gamma)$, $\gamma \in [0,1]$, es menor que el de la regla de Fisher. Por lo tanto, no se espera que los procedimientos habituales basados en bootstrap para corregir el sesgo del error aparente funcionen bien en las reglas $R_\delta(\gamma)$, lo que nos lleva a la necesidad de proponer nuevos estimadores para la tasa de error verdadero.

Proponemos cuatro métodos basados en técnicas de remuestreo, que modifican *LOOBT* (Efron, 1983) y *BCV* (Fu et al., 2005) para que tengan en cuenta la información adicional. Para obtener *LOOBT* se toman B muestras bootstrap, con cada una se obtiene la correspondiente versión bootstrap de la regla de clasificación y con esta regla se clasifican las observaciones de la muestra original que no están en la muestra bootstrap, siendo el estimador *LOOBT* la proporción de observaciones mal clasificadas. Efron (1983) observa que *LOOBT* sobreestima la tasa de error verdadero y propone $BT632 = 0.368 \cdot APP + 0.632 \cdot LOOBT$. *BCV* es el promedio de los errores de validación cruzada (*CV*, Lachenbruck y Mickey, 1968) de B muestras bootstrap. Para cada muestra, se deja fuera cada una de las observaciones, se crea la regla con el resto y se clasifica la que se ha dejado fuera, siendo el error de validación cruzada la proporción de observaciones mal clasificadas.

El problema de la inconsistencia del bootstrap en modelos con restricciones proviene del hecho de que el mundo bootstrap no representa bien el mundo real porque en el mundo real los parámetros de la población están en el cono, $\delta \in C$, mientras que en el mundo bootstrap la muestra es la población y, en general, $\overline{\delta} \notin C$. La definición de los nuevos estimadores del error verdadero para las reglas con restricciones se basa en la idea de que el mundo bootstrap debería reflejar el mundo real. Para ello hacemos dos propuestas. En la primera se modifica el cono de restricciones de acuerdo a lo observado y en la segunda se modifica la muestra para que verifique las restricciones.

Sea $\overline{C}$ el cono aleatorio definido a partir de C y $\overline{\delta} = \overline{X}_1 - \overline{X}_2$ de la siguiente manera:

$$\overline{C} = \left\{ x \in \mathbb{R}^p : \begin{array}{l} a_j'x \geq 0 \text{ si } a_j'\overline{\delta} \geq 0 \\ a_j'x \leq 0 \text{ si } a_j'\overline{\delta} < 0 \end{array}, j = 1, \ldots, m \right\}.$$

Definimos $BT2$ como *LOOBT* con la particularidad de que las reglas de clasificación a partir de cada muestra bootstrap se obtienen proyectando no sobre C sino sobre $\overline{C}$. De igual manera, definimos $BT2CV$ como *BCV* pero proyectando sobre $\overline{C}$ y no sobre C, observar que $\overline{\delta} = \overline{X}_1 - \overline{X}_2 \in \overline{C}$.

Como ya hemos dicho, nuestra segunda propuesta consiste en modificar la muestra de entrenamiento de forma que las nuevas medias muestrales verifiquen las restricciones. Para ello, transformamos la muestra de entrenamiento original $\{(X_i, Y_i), i = 1, \ldots, n\}$ de la forma siguiente:

$$W_i = X_i - \overline{X}_j + \mu^*_{\gamma j}, \text{ si } Y_i = j, j = 1, 2.$$

De esta forma, la muestra de entrenamiento transformada $\{(W_i, Y_i), i = 1, \ldots, n\}$

verifica las restricciones:

$$\overline{W}_1 - \overline{W}_2 = \mu^*_{\gamma 1} - \mu^*_{\gamma 2} \in C,$$

siendo $\overline{W}_1$ y $\overline{W}_2$ las nuevas medias muestrales. Definimos $BT3$ y $BT3CV$ como $LOOBT$ y BCV, respectivamente, una vez reemplazada la muestra de entrenamiento original por la transformada.

Llevamos a cabo un estudio de simulación para comparar los estimadores del error verdadero APP, CV, $LOOBT$, $BT632$, BCV, $BT2$, $BT2CV$, $BT3$ y $BT3CV$ de las reglas $R_\delta(\gamma)$. El comportamiento de un estimador $\hat{E}$ del error verdadero E_n se analiza a partir de la distribución de la variable aleatoria $(\hat{E} - E_n)$ (*deviation distribution*, Braga-Neto y Dougherty, 2004). Como medida global del comportamiento de $\hat{E}$ utilizamos $E[(\hat{E} - E_n)^2]$, que se puede descomponer:

$$E[(\hat{E} - E_n)^2] = Var(\hat{E} - E_n) + [E(\hat{E} - E_n)]^2.$$

Denotamos $A(\hat{E}) = E[(\hat{E} - E_n)^2]^{\frac{1}{2}}$ y $B(\hat{E}) = E(\hat{E} - E_n)$. Como en otros estudios de simulación, comprobamos que APP tiene el mayor sesgo $(B(\hat{E}))$ negativo, CV el menor sesgo pero el mayor $A(\hat{E})$, $LOOBT$ sesgo positivo, y BCV, $BT2CV$, $BT3CV$ y $BT632$, sesgo negativo. $BT2$ y $BT3$ tienen un comportamiento similar, lo que sorprende al estar motivados por ideas muy diferentes, siendo los mejores estimadores de la tasa de error verdadero para los valores más pequeños de $\|\delta\|^2$, precisamente la situación más interesante en la práctica, cuando la clasificación es más difícil y cuando la información adicional puede jugar un papel importante.

4.3.3 Aplicación

Aplicamos la metodología presentada a un conjunto de datos reales de cáncer de vejiga, descritos anteriormente en el apartado 3.3.3. Basándonos en el conocimiento previo y en nuestros resultados, decidimos utilizar $p = 4$ proteínas, denominadas P_1, P_2, P_3 y P_4, para discriminar entre dos poblaciones: Π_1, formada por el grupo control, y Π_2, formada por los grupos $T_1G_3 + T_2$. Para cada una de estas cuatro proteínas se esperaba que, en media, tomaran valores más altos cuanto más avanzado fuese el nivel de la enfermedad, es decir, $\delta = \mu_2 - \mu_1 \in O_4^+$ (octante positivo), restricciones que no se verifican en la muestra de entrenamiento, por lo que las reglas $R_\delta(\gamma)$ son relevantes en este problema. Se consideran cuatro reglas: $Fisher$, $SmDLDA$, $R_\delta(0)$ y $R_\delta(1)$, siendo $R_\delta(1)$ la que mejor comportamiento presenta. También comparamos las estimaciones de la tasa de error verdadero para los nuevos estimadores, y comprobamos el buen comportamiento de $BT2$ para $R_\delta(0)$ y $R_\delta(1)$.

4.4 Conclusiones

Las reglas $R_\delta(\gamma)$, $\gamma \in [0,1]$, se obtienen sustituyendo los parámetros desconocidos de la regla de Bayes por estimadores que tienen en cuenta la información adicional. Estos estimadores se definen a partir de proyecciones de las medias muestrales sobre el cono definido por la información adicional y su cono polar. Como la proyección es un operador contractivo, estas reglas restringidas se pueden ver como reglas contractivas. Tong et al. (2012) proponen estimadores shrinkage de las medias para definir reglas de clasificación conocidas como reglas *SmDLDA*. Hemos comparado el comportamiento de las reglas $R_\delta(\gamma)$, *SmDLDA* y *Fisher* en escenarios similares a los de Tong et al. (2012) y usando varios de los más comunes criterios utilizados en la literatura, comprobando que las reglas $R_\delta(\gamma)$ compiten bien bajo dichos criterios, aun en situaciones de alta dimensionalidad, mejorando los resultados en muchos de los escenarios. Entendemos que el hecho de que los estimadores utilizados en las reglas $R_\delta(\gamma)$ incluyan la información adicional disponible en el problema es una ventaja conceptual respecto a la contracción efectuada en *SmDLDA*, al no estar esta motivada por información asociada al problema.

Una cuestión importante para cualquier regla de clasificación es la estimación de la tasa de error verdadero. Hemos comprobado que el error aparente de las reglas restringidas tiene un sesgo menor que el de la regla de Fisher. Una posible explicación es que las reglas $R_\delta(\gamma)$, $\gamma \in [0,1]$, son menos dependientes de la muestra de entrenamiento. Por tanto, los procedimientos habituales para reducir el sesgo de la tasa de error aparente para estimar la tasa de error verdadero no funcionan bien en este contexto, y nos proponemos encontrar nuevos estimadores para la tasa de error verdadero, específicos para las reglas restringidas. Consideramos dos métodos basados en procedimientos bootstrap diferentes que toman en consideración la información adicional disponible. El primero, $BT2$, ajusta el cono de restricciones a la muestra de entrenamiento, mientras el segundo, $BT3$, ajusta la muestra de entrenamiento al cono de restricciones. $BT2CV$ y $BT3CV$ son las correspondientes validaciones cruzadas después del bootstrap de estos procedimientos. A partir de un estudio de simulación, comprobamos que $BT2$ y $BT3$ son los mejores estimadores de la tasa de error verdadero para los valores más pequeños de $||\delta||^2$, situaciones donde las poblaciones no están muy separadas, el caso más probable para las muestras de entrenamiento que no verifiquen las restricciones, mientras que $BT2CV$ y $BT3CV$ son los mejores para mayores valores de $||\delta||^2$. En la librería de R que hemos compilado, de nombre `dawai` y que describimos en la siguiente sección, se incluyen estos métodos.

Se ha aplicado la metodología presentada a un conjunto de datos reales de cáncer de vejiga. Se han considerado cuatro reglas: *Fisher*, *SmDLDA*, $R_\delta(0)$ y $R_\delta(1)$, siendo $R_\delta(1)$ la que mejor comportamiento presenta. También comparamos las estimaciones de la tasa de error verdadero para los nuevos estimadores y comprobamos el buen comportamiento de $BT2$ para $R_{\delta}(0)$ y $R_{\delta}(1)$. Podemos finalizar señalando que la combinación de la regla $R_{\delta}(1)$ con el estimador $BT2$ proporciona buenos resultados en este caso.

5 Trabajos en desarrollo y futuros

Como complemento de los trabajos expuestos en las secciones anteriores, hemos programado una librería en el entorno R, de nombre `dawai`, acrónimo de análisis discriminante con información adicional (d*iscriminant* a*nalysis* w*ith* a*dditional* i*nformation*), que se puede descargar en http://cran.r-project.org/web/packages/dawai/. Esta librería contiene todas las funciones que permiten definir las reglas de clasificación lineal $R_\mu(\gamma)$ definidas en Conde et al. (2012) que tienen en cuenta la información adicional en forma de restricciones sobre las medias, y clasificar muestras, así como evaluar la precisión de los resultados a través de los estimadores del error verdadero $BT2$, $BT3$, $BT2CV$ y $BT3CV$ propuestos en Conde et al. (2013).

En un artículo que se encuentra en proceso de referee, presentamos la librería `dawai`, extendemos los resultados y las definiciones que aparecen en Fernández et al. (2006), Conde et al. (2012) y Conde et al. (2013) para el caso de matrices de covarianzas iguales en las poblaciones al caso de matrices de covarianzas diferentes en las poblaciones, y por tanto definimos las correspondientes reglas de clasificación cuadrática y sus estimadores del error verdadero, y también extendemos la definición de estimadores del error verdadero al caso general de más de dos poblaciones, incluyéndose todas estas funcionalidades en la librería `dawai`. El software se puede aplicar a una amplia variedad de contextos, y su uso se ilustra aplicándolo a dos conjuntos de datos de campos diferentes como la biología y el reconocimiento de patrones.

Veamos una breve descripción de las principales funciones contenidas en la librería `dawai`:

- `rlda()`: Construye reglas de clasificación para poblaciones normales con matrices de covarianzas iguales que tienen en cuenta la información adicional en forma de restricciones sobre las medias. Además, proporciona la tasa de error aparente de las reglas.

- `predict.rlda()`: Permite clasificar observaciones multivariantes con las reglas de clasificación con restricciones construidas con `rlda()`.

- `err.est.rlda()`: Proporciona los estimadores del error verdadero $BT2$, $BT3$, $BT2CV$ y $BT3CV$ de las reglas de clasificación con restricciones construidas con `rlda()`.

- `rqda()`: Es el equivalente de `rlda()` para poblaciones normales con matrices de covarianzas distintas.

- `predict.rqda()`: Permite clasificar observaciones multivariantes con las reglas de clasificación con restricciones construidas con `rqda()`.

- `err.est.rqda()`: Proporciona los estimadores del error verdadero $BT2$, $BT3$, $BT2CV$ y $BT3CV$ de las reglas de clasificación con restricciones construidas con `rqda()`.

En un futuro pretendemos incorporar esta metodología a otras técnicas de clasificación. En algunos trabajos se han modificado varios algoritmos de machine learning para garantizar la monotonía, incluyendo, por ejemplo, inducción de reglas (Dembezynski et al., 2001), árboles de decisión (Potharst y Feelders, 2002), k vecinos más próximos (Duivesteijn y Feelders, 2008) y aprendizaje de reglas (Kotlowski y Slovinski, 2009). En Tian et al. (2012) se recogen los últimos avances al respecto de support vector machines. En otros trabajos, en lugar de modificar algoritmos para garantizar la monotonía, se han desarrollado métodos de preprocesamiento de los datos, tales como técnicas de reetiquetado, que reparan las posibles inconsistencias en la muestra de entrenamiento para que los clasificadores sean monótonos, ver Feelders (2010). Creemos que existen bastantes posibilidades aun de mejora en este tipo de procedimientos mediante la incorporación de información adicional a través de procedimientos como los desarrollados en la presente memoria.

Referencias

[1] Abelson, R. P., Tukey, J. W. (1963). Efficient utilization of non-numerical information in quantitative analysis: general theory and the case of simple order. *The Annals of Mathematical Statistics* **34**, 1347-1369.

[2] Adegboye, O. S. (1993). The optimal classification rule for exponential populations. *Australian Journal of Statistics* **35** (2), 185-194.

[3] Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* **68**, 399-405.

[4] Auh, S., Sampson, A. R. (2006). Isotonic logistic discrimination. *Biometrika* **93** (4), 961-972.

[5] Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics* **26** (4), 641-647.

[6] Barlow, R. E., Bartholomew, D. J., Bremner, J. M., Brunk, H. D. (1972). *Statistical inference under order restrictions*. Wiley. New York.

[7] Bartholomew, D. J. (1959). A test of homogeneity for ordered alternatives. *Biometrika* **46**, 36-48.

[8] Bartholomew, D. J. (1961). A test of homogeneity of means under restricted alternatives. *Journal of the Royal Statistical Society, Series B* **23** (2), 239-281.

[9] Basu, A. P., Gupta, A. K. (1974). Classification rules for exponential populations. *Proc. Conference on Reliability and Biometry*. SIAM Philadelphia, 637-650.

[10] Benigni, R. (2013). *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*. CRC Press.

[11] Berry, M. W. (2004). *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer.

[12] Bhattacharya, P. K., Das Gupta, S. (1964). Classification between univariate exponential distributions. *Sankhyā: The Indian Journal of Statistics, Series A* **26**, 17-24.

[13] Bockenholt, U., Bockenholt, I. (1990). Canonical analysis of contingency tables with linear constraints. *Psychometrika* **55**, 633-639.

[14] Boulgouris, N. V., Plataniotis, K. N., Micheli-Tzanakou, E. (2009). *Biometrics: Theory, Methods, and Applications*. Wiley.

[15] Braga-Neto, U. M., Dougherty, E. R. (2004). Is croos-validation valid for small-sample microarray classification? *Bioinformatics* **20**, 374-380.

[16] Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics* **26** (4), 607-616.

[17] Brunk, H. D. (1958). On the estimation of parameters restricted by inequalities. *The Annals of Mathematical Statistics* **29** (2), 437-454.

[18] Brunk, H. D. (1970). Estimation of isotonic regression. Nonparametric Techniques in Statistical Inference. *Cambridge University Press*, 177-195.

[19] Bunke, H., Wang, P. S. P. (1997). *Handbook of Character Recognition and Document Image Analysis*. World Scientific.

[20] Chang, G., Healey, M. J., McHugh, J. A. M., Wang, J. T. L. (2001). *Mining the World Wide Web: An Information Search Approach*. Kluwer Academic Publishers.

[21] Chang, Y. T., Shinozaki, N. (2002). A comparison of restricted and unrestricted estimators in estimating linear functions of ordered scale parameters of two gamma distributions. *Annals of the Institute of Statistical Mathematics* **54** (4), 848-860.

[22] Cohen, A., Kemperman, J. H. B., Sackrowitz, H. B. (2000). Properties of likelihood inference for order restricted models. *Journal of Multivariate Analysis* **72**, 50-77.

[23] Cohen, A., Sackrowitz, H. B. (2004). A discussion of some inference issues in order restricted models. *The Canadian Journal of Statistics* **32**, 199-205.

[24] Conde, D., Fernández, M. A., Salvador, B. (2005). A classification rule for ordered exponential populations. *Journal of Statistical Planning and Inference* **135** (2), 339-356.

[25] Conde, D., Fernández, M. A., Rueda, C., Salvador, B. (2012). Classification of samples into two or more ordered populations with application to a cancer trial. *Statistics in Medicine* **31**, 3773-3786.

[26] Conde, D., Salvador, B., Rueda, C., Fernández, M. A. (2013). Performance and estimation of the true error rate of classification rules built with additional information. An application to a cancer trial. *Statistical Applications in Genetics and Molecular Biology* **12** (5), 583-602.

[27] Das S., Sen, P. K. (1994). Restricted canonical correlations. *Linear Algebra and its Applications* **210**, 29-47.

[28] Das S., Sen, P. K. (1996). Asymptotic distribution of restricted canonical correlations and relevant resampling methods. *Journal of Multivariate Analysis* **56** (1), 1-19.

[29] Dembezynski, K., Kotlowski, W., Slowinski, R. (2001). Learning rule ensembles for ordinal classification with monotonicity constraints. *Fundamenta Informaticae* **XXI**, 1001-1016.

[30] Duivesteijn, W., Feelders, A. (2008). Nearest neighbour classification with monotonicity constraints. *Lecture Notes in Computer Science* **5211**, 301-316.

[31] Dykstra, R., Hewett, J., Robertson, T. (1999). Nonparametric, isotonic discriminant procedures. *Biometrika* **86**, 429-438.

[32] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78**, 316-331.

[33] El Barmi, H., McKeague, I. W. (2013). Empirical likelihood-based tests for stochastic ordering. *Bernoulli* **19** (1), 295-367.

[34] Feelders, A. (2010). Monotone relabelling in ordinal classification. *ICDM '10 Proceedings of the 2010 IEEE International Conference on Data Mining*, 803-808.

[35] Fernández, M. A., Rueda, C., Salvador, B. (1997). On the maximum likelihood estimator under order restrictions in uniform probability models. *Communications in Statistics - Theory and Methods* **26** (8), 1971-1980.

[36] Fernández, M. A., Rueda, C., Salvador, B. (1998). Simultaneous estimation by isotonic regression. *Journal of Statistical Planning and Inference* **70**, 111-119.

[37] Fernández, M. A., Rueda, C., Salvador, B. (1999). The loss of efficiency estimating contrast under restrictions. *Scandinavian Journal of Statistics* **26** (4), 579-592.

[38] Fernández, M. A., Rueda, C., Salvador, B. (2000). Parameter estimation under orthant restrictions. *The Canadian Journal of Statistics* **28**, 171-181.

[39] Fernández, M. A., Rueda, C., Salvador, B. (2006). Incorporating additional information to normal linear discriminant rules. *Journal of the American Statistical Association* **101** (474), 569-577.

[40] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7** (2), 179-188.

[41] Fraley, C., Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association* **97** (458), 611-631.

[42] Fu, W. J., Carroll, R. J., Wang, S. (2005). Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics* **21**, 1979-1986.

[43] Gasparini, M., Eisele, J. (2000). A curve-free method for phase I clinical trials. *Biometrics* **56**, 609-615.

[44] Groenen, P. J. F., Poblome, J. (2003). Constrained correspondence analysis for seriation in archaeology applied to Sagalassos ceramic tablewares. *Exploratory Data Analysis in Empirical Research*, 90-97.

[45] Hand, D. J. 1981. *Discrimination and Classification*. New York. Wiley.

[46] Hastie, T., Buja, A., Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics*, 73-102.

[47] Hoitjink, H. (2013). Objective Bayes factors for inequality constrained hypotheses. *International Statistical Review* **81** (2), 207-229.

[48] Hu, X., Wright, F. T. (1994). Likelihood ratio tests for a class of non-oblique hypotheses. *Annals of the Institute of Statistical Mathematics* **46**, 137-145.

[49] Huberty, C. J., Olejnik, S. (2006). *Applied MANOVA and Discriminant Analysis*. John Wiley & Sons.

[50] Hwang, J. T. G., Peddada, S. D. (1994). Confidence interval estimation subject to order restrictions. *The Annals of Statistics* **22**, 67-93.

[51] Kao, A., Poteet, S. R. (2007). *Natural Language Processing and Text Mining*. Springer.

[52] Kaur, A., Singh, H. (1991). On the estimation of ordered means of two exponential populations. *Annals of the Institute of Statistical Mathematics* **43** (2), 347-356.

[53] Kelly, R. (1989). Stochastic reduction of loss in estimating normal means by isotonic regression. *The Annals of Statistics* **17** (2), 937-910.

[54] Kim, J. H. (2009). Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis* **53** (11), 3735-3745.

[55] Kim, K. I., Simon, R. (2011). Probabilistic classifiers with high-dimensional data. *Biostatistics* **12** (3), 399-412.

[56] Korn, E. L. (1982). Confidence bands for isotonic dose-response curves. *Applied Statistics* **31** (1), 59-63.

[57] Kotlowski, W., Slowinski, R. (2009). Rule learning with monotonicity constraints. *ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning*, 537-544.

[58] Kotlowski, W., Slowinski, R. (2013). On nonparametric ordinal classification with monotonicity constraints. *IEEE Transactions on Knowledge and Data Engineering* **25** (11), 2576-2589.

[59] Kulatunga, D. D. S., Sasabuchi, S. (1984). A test of homogeneity of mean vectors against multivariate isotonic alternatives. *Mem. Fac. Sci. Kyushu Univ. Ser. A Math.* **38**, 151-161.

[60] Kuriki, S. (2005). Asymptotic distribution of inequality-restricted canonical correlation with application to tests for independence in ordered contingency tables. *Journal of Multivariate Analysis* **94** (2), 420-449.

[61] Lachenbruch, P. A., Goldstein, M. (1979). Discriminant analysis. *Biometrics* **35**, 69-85.

[62] Lachenbruch, P., Mickey, M.(1968). Estimation of error rates in discriminant analysis. *Technometrics* **10**, 167-178.

[63] Lee, C. I. C. (1981). The quadratic loss of isotonic regression under normality. *The Annals of Statistics* **9** (3), 686-688.

[64] Lee, C. I. C. (1988). The quadratic loss of order restricted estimators for several treatment means and a control mean. *The Annals of Statistics* **16**, 751-758.

[65] Li, Z., Taylor, J. M. G., Nan, B. (2010). Construction of confidence intervals and regions for ordered binomial probabilities. *The American Statistician* **64** (4), 291-298.

[66] Li, S. Z., Jain, A. K. (2011). *Handbook of Face Recognition*. Springer.

[67] Liu, T., Lin, N., Shi, N., Zhang, B. (2009). Information criterion-based clustering with order-restricted candidate profiles in short time-course microarray experiments. *Bioinformatics* **10**, 146.

[68] Lo, S. H.(1987). Estimation of distribution functions under order restrictions. *Statistics and decisions* **5**, 251-262.

[69] Long, T., Gupta, R. D. (1998). Alternative linear classification rules under order restrictions. *Communications in Statistics - Theory and Methods* **27** (3), 559-575.

[70] Lovell, M. C., Prescott, E. (1970). Multiple regression with inequality constraints: pretesting bias, hypothesis testing and efficiency. *Journal of the American Statistical Association* **65** (330), 913-925.

[71] Ma, T., Liu, S. (2013). Estimation of order-restricted means of two normal populations under the LINEX loss function. *Metrika* **76** (3), 409-425.

[72] Mahalanobis P. C. (1927). Analysis of race mixture in Bengal. *Journal and Proceedings of the Asiatic Society of Bengal* **23**, 301-333.

[73] Marchand, E., Strawderman, W. E. (2004). Estimation in restricted parameter spaces: A review. *A Festschrift for Herman Rubin. Institute of Mathematical Statistics Lecture Notes - Monograph Series* **45**, 21-44.

[74] Marcus, R., Peritz, E. (1976). Some simultaneous confidence bounds in normal models with restricted alternatives. *Journal of the Royal Statistical Society, Series B (Methodological)* **38** (2), 157-165.

[75] McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley-Interscience.

[76] Menéndez, J. A., Rueda, C., Salvador, B. (1991). Conditional test for testing a face of the tree order cone. *Communications in Statistics - Simulation and Computation* **20**, 751-762.

[77] Menéndez, J. A., Rueda, C., Salvador, B. (1992a). Dominance of likelihood ratio tests under order constraints. *The Annals of Statistics* **20**, 2087-2099.

[78] Menéndez, J. A., Rueda, C., Salvador, B. (1992b). Testing non-oblique hypotheses. *Communications in Statistics - Theory and Methods* **21**, 471-484.

[79] Menéndez, J. A., Salvador, B. (1991). Anomalies of the likelihood ratio test for testing restricted hypothesis. *The Annals of Statistics* **19**, 889-898.

[80] Menéndez, J. A., Salvador, B. (1992). Equivalence of likelihood ratio tests and obliquity. *Statistics & Probability Letters* **14**, 223-229.

[81] Moors, J. J. A., van Houwelingen, J. C. (1993). Estimation of linear models with inequality restrictions. *Statististica Neerlandica* **47**, 185-198.

[82] Mukerjee, H., Robertson, T., Wright, F. T. (1986). A probability inequality for elliptically contoured densities with applications in order restricted inference. *The Annals of Statistics* **14**, 1544-1554.

[83] Pearson, K. (1926). On the coefficient of racial likeness. *Biometrika* **18** (1/2), 105-117.

[84] Peddada, S. D. (1997). Confidence interval estimation of population means subject to order restrictions using resampling procedures. *Statistics & Probability Letters* **31**, 255-265.

[85] Peddada, S., Lobenhofer, E., Li, L., Afshari, C., Weinberg, C., Umbach, D. (2003). Gene Selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Bioinformatics* **19**, 834-841.

[86] Pepe, M. S., Cai, T., Lognton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* **62** (1), 221-229.

[87] Perlman, M. D. (1969). One-sided testing problems in multivariate analysis. *The Annals of Mathematical Statistics* **40** (2), 549-567.

[88] Perlman, M. D., Wu, L. (2002). A defense of the likelihood ratio test for one-sided and order-restricted alternatives. *Journal of Statistical Planning and Inference* **107**, 173-186.

[89] Potharst, R., Feelders, A. (2002). Classification trees for problems with monotonicity constraints. *ACM SIGKDD Explorations Newsletter* **4** (1), 1-10.

[90] Praestgaard, J. (2012). A note on the power superiority of the restricted likelihood ratio test. *Journal of Multivariate Analysis* **104**, 1-15.

[91] Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)* **10** (2), 159-203.

[92] Robertson, T., Wright, F. T., Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

[93] Rueda, C. (1989). *Contrates de Hipótesis con Restricciones bajo Condiciones de Oblicuidad*. Tesis Doctoral. Universidad de Valladolid.

[94] Rueda, C. Salvador, B. (1995). Reduction of risk using restricted estimators. *Communications in Statistics - Theory and Methods* **24**, 1011-1022.

[95] Rueda, C., Salvador, B., Fernández, M. A. (1997a). A good property of the maximum likelihood estimator in a restricted normal model. *Test* **6** (1), 127-135.

[96] Rueda, C., Salvador, B, Fernández, M. A. (1997b). Simultaneous estimation in a restricted linear model. *Journal of Multivariate Analysis* **61**, 61-66.

[97] Rueda, C., Menéndez, J. A., Salvador, B. (2002). Bootstrap adjusted estimators in a restricted setting. *Journal of Statistical Planning and Inference* **107**, 123-131.

[98] Sampson, A. R., Singh, H., Whitaker, L. R. (2008). Simultaneous confidence bands for isotonic functions. *Journal of Statistical Planning and Inference* **139**, 828-842.

[99] Salvador, B., Fernández, M. A., Martín, I., Rueda, C. (2008). Robustness of classification rules that incorporate additional information. *Computational Statistics & Data Analysis* **52**, 2489-2495.

[100] Sasabuchi, S., Inutsuka, M., Kulatunga, D. D. S. (1983). A multivariate version of isotonic regression. *Biometrika* **70** (2), 465-472.

[101] Schaafsma, W., Smid, L. J. (1966). Most stringent somehere most powerful tests against alternatives restricted by a number of linear inequalities. *The Annals of Mathematical Statistics* **37**, 1161-1172.

[102] Schiavo, R. A., Hand, D. J. (2000). Ten more years of error rate research. *International Statistical Review* **68**, 295-310.

[103] Schoenfeld, D. A. (1986). Confidence bounds for normal means under order restrictions, with application to dose-response curves, toxicology experiments, and low-dose extrapolation. *Journal of the American Statistical Association* **81**, 186-195.

[104] Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review* **56** (1), 49-62.

[105] Shaw, F. H., Geyer, C. J. (1997). Estimation and testing in constrained covariance component models. *Biometrika* **84** (1), 95-102.

[106] Shi, N. Z., Kudô, A. (1987). The most stringent somehere most powerful one-sided test of the multivariate normal mean. *Mem. Fac. Sci. Kyushu Univ.* **29**, 303-328.

[107] Silvapulle, M. J., Sen, P. K. (2004). *Constrained Statistical Inference*. John Wiley & Sons, New Jersey.

[108] Simmons, S., Peddada, S. (2007). Order-restricted inference for ordered gene expression (ORIOGEN) data under heteroscedastic variances. *Bioinformation* **1**, 414-419.

[109] Sneath, P. H. A., Sokal, R. R. (1973). *Numerical taxonomy. The principles and practice of numerical classification*. W. H. Freeman Limited.

[110] Sonka, M., Fitzpatrick, J. (2004). *Handbook of Medical Imaging: Medical image processing and analysis*. The Society of Photo-Optical Instrumentation Engineers.

[111] Takane, Y., Yanai, H., Mayekawa, S. (1991). Relationships among several methods of linearly constrained correspondence analysis. *Psychometrika* **56**, 667-684.

[112] Taylor, J. M. G., Wang, L., Li, Z. (2007). Analysis on binary responses with ordered covariates and missing data. *Statistics in Medicine* **26**, 3443-3458.

[113] Thomas, L. C., Edelman, D. B., Crook, J. N. (2002). *Credit Scoring and Its Applications*. SIAM.

[114] Tian, Y., Shi, Y., Liu, X. (2012). Recent advances in support vector machines research. *Technological and Economic Development of Economy* **18**, 5-33.

[115] Tong, T., Chen, L., Zhao, H. (2012). Improved mean estimation and its application to diagonal discriminant analysis. *Bioinformatics* **28** (4), 531-537.

[116] Touissant, G. T. (1974). Bibliography on estimation of misclassification. *IEEE Transactions on Information Theory* **20** (4), 472-479.

[117] Tsai, M. (1992). On the power superiority of likelihood ratio tests for restricted alternatives. *Journal of Multivariate Annalysis* **42**, 102-109.

[118] Tsai, M., Sen, P. K. (1993). On the local optimality of optimal linear tests for restricted alternatives. *Statistica Sinica* **3**, 103-115.

[119] Tsukuma, H., Kubokawa, T. (2011). Modifying estimators of ordered positive parameters under the Stein loss. *Journal of Multivariate Analysis* **102** (1), 164-181.

[120] Van de Velden, M., Groenen, P. J. F., Poblome, J. (2009). Seriation by constrained correspondence analysis: A simulation study. *Computational Statistics and Data Analysis* **53**, 3129-3138.

[121] Van Eeden, C. (1956). Maximum likelihood estimation of ordered probabilities. *Proc. Kon. Nederl. Akad. Wetensch. Ser. A* **59**, 444-455.

[122] Van Eeden, C. (1957a). Maximum likelihood estimation of partially or completely ordered parameters. *Proc. Kon. Nederl. Akad. Wetensch. Ser. A* **60**, 128-136.

[123] Van Eeden, C. (1957b). Note on two methods for estimating ordered parameters of probability distributions. *Proc. Kon. Nederl. Akad. Wetensch. Ser. A* **60**, 506-512.

[124] Van Eeden, C. (2006). *Restricted parameter space estimation problems: admissibility and minimaxity properties*. Springer.

[125] Vijayasree, G., Singh, H. (1991). Simultaneous estimation of two ordered exponential parameters. *Communications in Statistics - Theory and Methods* **20** (8), 2559-2576.

[126] Vijayasree, G., Singh, H. (1993). Mixed estimators of two ordered exponential means. *Journal of Statistical Planning and Inference* **35**, 47-56.

[127] Wang, F. (2008). *Biomarker Methods in Drug Discovery and Development*. Humana Press.

[128] Zelen, M. (1966). Application of exponential models to problems in cancer research with discussion. *Journal of the Royal Statistical Society: Series A* **129**, 368-398.

ANEXOS: Artículos publicados

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Statistical Planning and
Inference 135 (2005) 339 – 356

journal of
statistical planning
and inference

www.elsevier.com/locate/jspi

A classification rule for ordered exponential populations[☆]

David Conde, Miguel A. Fernández[*], Bonifacio Salvador

Departamento de Estadística e Investigación Operativa, C/Prado de la Magdalena s/n, Universidad de Valladolid, 47005 Valladolid, Spain

Received 30 April 2003; accepted 1 May 2004
Available online 21 July 2004

Abstract

In this paper, we consider classification procedures for exponential populations when an order on the populations parameters is known. We define and study the behavior of a classification rule which takes into account the additional information and outperforms the likelihood-ratio-based rule when two populations are considered. Moreover, we study the behavior of this rule in each of the two populations and compare the misclassification probabilities with the classical ones. Type II censorship, which is usual in practice, is considered and results obtained. The performance for more than two populations is evaluated by simulation.

MSC: primary 62H30; 62F30

Keywords: Classification rules; Exponential populations; Restricted parameter spaces; Misclassification probabilities

1. Introduction

There is an extensive literature on classification in normal populations. However not much work has been done for the non-normal case. The exponential distribution is perhaps one of the most interesting ones to be considered since it is quite frequent to find it in practical

[☆] Research partially supported by Spanish DGES Grant PB 97-0475.

[*] Corresponding author. Fax: +34-9834-23013.
E-mail address: miguelaf@eio.uva.es (M.A. Fernández).

0378-3758/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2004.05.004

contexts such as reliability or survival analysis. Some papers dealing with classification for the exponential case and its applications are Basu and Gupta (1974), Bhattacharya and Das Gupta (1964) and Adegboye (1993). Interesting papers dealing with estimation of non-normal ordered parameters are Vijayasree and Singh (1991, 1993) for exponential populations and Chang and Shinozaki (2002) for gamma populations.

However, the study of classification rules when additional information in the form of parameter ordering is available and how the rules can be improved in this context has not been thoroughly investigated. To our best knowledge the only paper in this line is Long and Gupta (1998) where an order restriction on the means of two normal populations is assumed and some improvements over Anderson's classification rule are achieved.

In this paper, we combine both situations and consider classification on exponential populations whose parameters are known to follow a simple order. Examples where this sort of scheme can be used may be found in Zelen (1966) in the cancer research context and in Basu and Gupta (1974) in the reliability one.

The layout of the paper is the following. In Section 2, we study the two populations case. First, we define a rule which takes into account the information given by the order restrictions and compare it with the usual and unrestricted likelihood-ratio based one. We find that the proposed rule is consistent and that it performs globally better than the usual rule. The second part of Section 2 is related to the question of misclassification in each of the two populations for the new rule and results comparing these probabilities with the ones for the usual rule are obtained.

In Section 3, we consider type II censorship and how this scheme affects the rule. We obtain results for this situation quite usual in reliability and survival analysis applications.

Section 4 is devoted to the case of more than two populations. The proposed rule is extended to this case and shown to be consistent. We also compare it with the likelihood-ratio based-one and using simulation we evaluate its performance in terms of misclassification probabilities both globally and in each of the populations involved. An appendix contains the more involved proofs that have been delayed to improve the readability of the paper.

2. Two populations case

2.1. Global rule behavior

Let π_1 and π_2 be two one-parameter exponential populations with probability density functions

$$f_i(x) = \lambda_i \mathrm{e}^{-\lambda_i x}, \;\; x > 0, \quad \lambda_i > 0, \quad i = 1, 2,$$

such that it is known that $\lambda_1 \geqslant \lambda_2$. In other words, the expected value in the first population is not bigger than in the second one.

We also assume that we have a training sample from each of the two populations $X = (X_1, \ldots, X_{n_1})$, $Y = (Y_1, \ldots, Y_{n_2})$. Now, we want to classify a new observation z coming from one of those two populations but whose exact origin is unknown.

In what follows, we will suppose that the populations are equally likely and that the costs of misclassifications are equal.

The usual likelihood classification rule appearing in Basu and Gupta (1974) can be written in the following way:

$$R: \text{Classify } z \text{ into } \begin{cases} \pi_1 & \text{iff } (z-x_0)(\lambda_1-\lambda_2)<0, \\ \pi_2 & \text{otherwise,} \end{cases}$$

where $x_0 = \frac{\ln\ \lambda_1 - \ln\ \lambda_2}{\lambda_1-\lambda_2}$. If we consider the usual case where the parameters are unknown the rule turns into

$$R_{\mathrm{U}}: \text{ Classify } z \text{ into } \begin{cases} \pi_1 & \text{iff } (z-\widehat{x}_0)(\widehat{\lambda}_1-\widehat{\lambda}_2)<0, \\ \pi_2 & \text{otherwise,} \end{cases} \tag{1}$$

where $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ are the MLEs of λ_1 and λ_2 and

$$\widehat{x}_0 = \frac{\ln\ \widehat{\lambda}_1 - \ln\ \widehat{\lambda}_2}{\widehat{\lambda}_1 - \widehat{\lambda}_2}.$$

This rule is shown to be consistent in Theorem 3 of the same aforementioned paper and has been compared with the one based on Fisher's Discriminant Function for normal populations in Adegboye (1993) where the superiority of R_{U} over the normal translated one is exhibited.

The rule we propose, which we will refer to as the ordered rule, takes into account that we know that $\lambda_1 \geqslant \lambda_2$ and is defined as follows:

$$R_{\mathrm{O}}: \text{ Classify } z \text{ into } \begin{cases} \pi_1 & \text{iff } (z-\widehat{x}_0)<0, \\ \pi_2 & \text{otherwise.} \end{cases} \tag{2}$$

This rule is based on the estimatior given in Vijayasree and Singh (1991) with mixing parameter $\alpha = 0$.

First, we prove the consistency of this rule. Let us denote as $P_*(i/j)$ the probability that an observation coming from population j is classified in population i when the rule R_* is being used and let $P_*(\mathrm{MC}) = \frac{1}{2}(P_*(\frac{1}{2}) + P_*(\frac{2}{1}))$ be the global misclassification probability of R_*.

Theorem 1. *Let $z_1,\ldots,z_m$ be a random sample from π_0 where $\pi_0=\pi_i$ for exactly one i ($i=1$ or 2) then*

$$P_{\mathrm{O}}(\mathrm{MC}) \longrightarrow 0 \ for\ m, n_1, n_2 > N\ when\ N \to \infty.$$

Proof. The proof of this result follows the same lines of Theorems 2 and 3 in Basu and Gupta (1974) and therefore is not developed here. □

Next, we prove that the new ordered rule outperforms the usual one which does not take into account the additional information given by the restrictions. Notice that this is not obvious at all since restricted procedures do not always perform better that the unrestricted ones. This has been observed as much in likelihood ratio tests under order restrictions (see Menéndez and Salvador, 1991) as in restricted parameter estimation (see Lee, 1988; Fernández et al., 1999).

Theorem 2. *Let* R_{O} *and* R_{U} *be the classification rules defined in* (2) *and* (1), *respectively, then*

$$P_{\mathrm{O}}(\mathrm{MC}) \leqslant P_{\mathrm{U}}(\mathrm{MC})$$

for any $\lambda_1 \geqslant \lambda_2 > 0$.

Proof. Let us work with the correct classification probabilities

$$\begin{aligned} 1 - P_{\mathrm{O}}(\mathrm{MC}) &= \tfrac{1}{2}\left(P_{\mathrm{O}}\left(\tfrac{1}{1}\right) + P_{\mathrm{O}}\left(\tfrac{2}{2}\right)\right) \\ &= \tfrac{1}{2}[P_{\lambda_1}(Z \leqslant \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) + P_{\lambda_1}(Z \leqslant \widehat{x}_0, \widehat{\lambda}_1 > \widehat{\lambda}_2)] \\ &\quad + \tfrac{1}{2}[P_{\lambda_2}(Z > \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) + P_{\lambda_2}(Z > \widehat{x}_0, \widehat{\lambda}_1 > \widehat{\lambda}_2)] \end{aligned} \tag{3}$$

and for the usual rule

$$\begin{aligned} 1 - P_{\mathrm{U}}(\mathrm{MC}) &= \tfrac{1}{2}\left(P_{\mathrm{U}}\left(\tfrac{1}{1}\right) + P_{\mathrm{U}}\left(\tfrac{2}{2}\right)\right) \\ &= \tfrac{1}{2}[P_{\lambda_1}(Z > \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) + P_{\lambda_1}(Z \leqslant \widehat{x}_0, \widehat{\lambda}_1 > \widehat{\lambda}_2)] \\ &\quad + \tfrac{1}{2}[P_{\lambda_2}(Z \leqslant \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) + P_{\lambda_2}(Z > \widehat{x}_0, \widehat{\lambda}_1 > \widehat{\lambda}_2)]. \end{aligned} \tag{4}$$

Obviously, it is enough to prove

$$\begin{aligned} &P_{\lambda_1}(Z \leqslant \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) + P_{\lambda_2}(Z > \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) \\ &\quad \geqslant P_{\lambda_1}(Z > \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) + P_{\lambda_2}(Z \leqslant \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2). \end{aligned}$$

Taking into account that $\widehat{\lambda}_1 = \frac{n_1}{\sum_{i=1}^{n_1} X_i}$, $\widehat{\lambda}_2 = \frac{n_2}{\sum_{i=1}^{n_2} Y_i}$ and that $X_{\mathrm{T}} = \sum_{i=1}^{n_1} X_i \rightsquigarrow \gamma(\lambda_1, n_1)$ and $Y_{\mathrm{T}} = \sum_{i=1}^{n_2} Y_i \rightsquigarrow \gamma(\lambda_2, n_2)$ if we denote

$$f(x, \lambda, n) = \frac{\lambda^n}{(n-1)!} \mathrm{e}^{-\lambda x} x^{n-1}, \tag{5}$$

we can write

$$\begin{aligned} P_\lambda(Z \leqslant \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) &= P_\lambda\left(Z \leqslant \widehat{x}_0, X_{\mathrm{T}} \geqslant \frac{n_1}{n_2} Y_{\mathrm{T}}\right) \\ &= \int\int_{x \geqslant \frac{n_1}{n_2} y} P_\lambda\left(Z \leqslant \widehat{x}_0, X_{\mathrm{T}} \geqslant \frac{n_1}{n_2} Y_{\mathrm{T}} / X_{\mathrm{T}} = x, Y_{\mathrm{T}} = y\right) \\ &\quad \times \mathrm{d}P_{X_{\mathrm{T}}}(x)\, \mathrm{d}P_{Y_{\mathrm{T}}}(y) \\ &= \int\int_{x \geqslant \frac{n_1}{n_2} y} P_\lambda(Z \leqslant \widehat{x}_0)\, \mathrm{d}P_{X_{\mathrm{T}}}(x)\, \mathrm{d}P_{Y_{\mathrm{T}}}(y) \\ &= \int_0^\infty \int_{\frac{n_1}{n_2} y}^\infty (1 - \mathrm{e}^{-\lambda \widehat{x}_0}) f(x, \lambda_1, n_1) f(y, \lambda_2, n_2)\, \mathrm{d}x\, \mathrm{d}y. \end{aligned} \tag{6}$$

From this expression it is clear that the probability, we are calculating is increasing in λ so that as $\lambda_1 \geqslant \lambda_2$

$$P_{\lambda_1}(Z \leqslant \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) \geqslant P_{\lambda_2}(Z \leqslant \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2). \tag{7}$$

Moreover, as

$$\begin{aligned} &P_{\lambda_1}(Z \leqslant \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) + P_{\lambda_1}(Z > \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) \\ &\quad = P_{\lambda_1}(\widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) = P_{\lambda_2}(\widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) \\ &\quad = P_{\lambda_2}(Z \leqslant \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) + P_{\lambda_2}(Z > \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2), \end{aligned}$$

we have

$$P_{\lambda_2}(Z > \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) \geqslant P_{\lambda_1}(Z > \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) \tag{8}$$

and the theorem is done. □

Remark 3. Notice that from the proof it is clear that $P_{\mathrm{O}}(\mathrm{MC}) > P_{\mathrm{U}}(\mathrm{MC})$ whenever $\lambda_1 > \lambda_2 > 0$, that is the equality only holds when the parameters of the two populations are equal.

Corollary 4. *Theorem 2 is also true when we have to classify a sample of size $m > 1$ coming from the same unknown population.*

Proof. In this case we use the sufficient statistic $\overline{Z} = \frac{1}{m}(Z_1 + \cdots + Z_m)$ instead of Z for the classification. Following the same lines as the previous theorem it is enough to replace $P_\lambda(Z < \widehat{x}_0) = 1 - \mathrm{e}^{-\lambda \widehat{x}_0}$ in Eq. (6) with the probability corresponding to the gamma distribution $P_\lambda(\overline{Z} < \widehat{x}_0) = 1 - \sum_{i=0}^{m-1} \mathrm{e}^{-\lambda \widehat{x}_0 m} \frac{(\lambda \widehat{x}_0 m)^i}{i!}$ which is also increasing in λ when all other parameters are fixed. □

These results can be applied to situations interesting in practice as the air-conditioning equipment in two aircraft example appearing in Basu and Gupta (1974), if it is known that one of the two aircraft has a mean time to failure lower than the other. In that case, the use of rule R_{O} instead R_{U} will decrease the misclassification probabilities as shown in the previous results.

It is clearly interesting to know more precisely how much we gain. It will become apparent in the proof of Theorems 5 and 6 in next subsection that the misclassification probabilities of R_{O} and R_{U} depend on λ_1 and λ_2 only through λ_1/λ_2. This fact has been already noticed by Adegboye (1993) and we use it in the graphics appearing in the paper. The horizontal axe represents $\lambda_2/\lambda_1 \in (0, 1]$ and the vertical axe the correct classification probability differences between R_{O} and R_{U}. In Fig. 1, we show how the correct classification probabilities change when the samples sizes are equal ($n_1 = n_2 = n$). When samples sizes are different we have observed that the value depends much more strongly on n_2 than on n_1. The maximum value is obtained when $n_2 = 1$ and $n_1 \to \infty$ and is about 0.0747.

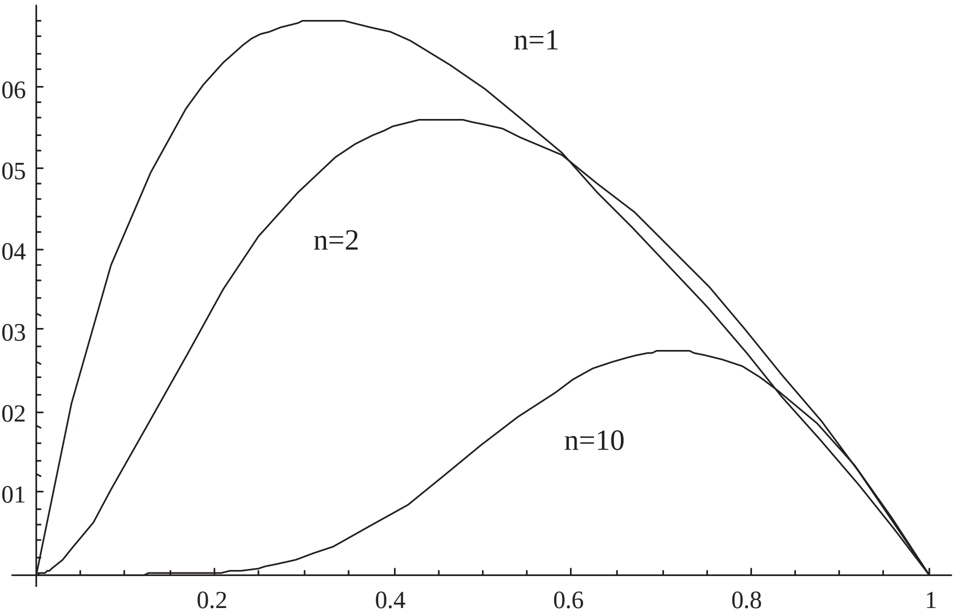

Fig. 1. Global correct probability difference between R_{O} and R_{U} for several sample sizes.

2.2. *Misclassification probabilities in each population*

The next step in the evaluation of the rule is to compare the misclassification probabilities of R_{O} and R_{U} in each of the two populations. Notice that the proof of Theorem 2 cannot be used for this purpose as the inequalities (7) and (8) appearing there do not involve the appropriate sets to be compared.

The comparison of individual populations probabilities is interesting since in Long and Gupta (1998) some ordered rules for the restricted normal classification problem are defined and one of them is shown to improve the misclassification probabilities for both populations. Unfortunately, this property is not true for all possible parameter values when dealing with exponential populations.

In this subsection of the paper, we will assume that both training samples sizes are equal, that is $n_1 = n_2 = n$ and that $m = 1$. The proof of the results is delayed to the appendix in order to improve the readability of the paper. The results are written in terms of the correct classification probabilities.

We have the following result for the first population

Theorem 5. *Assume* $n_1 = n_2 = n$.

1. *If* $n > 1$, $P_{\mathrm{O}}(\frac{1}{1}) > P_{\mathrm{U}}(\frac{1}{1})$ $\forall \lambda_1 \geqslant \lambda_2 > 0$.
2. *If* $n = 1$ *there is* $\delta_0 \in (0, 1)$ *such that*

$$P_{\mathrm{O}}\left(\tfrac{1}{1}\right) \geqslant P_{\mathrm{U}}\left(\tfrac{1}{1}\right), \quad \forall \lambda_1, \lambda_2 > 0, \quad 0 < \frac{\lambda_2}{\lambda_1} \leqslant \delta_0,$$

$$P_{\mathrm{O}}\left(\tfrac{1}{1}\right) < P_{\mathrm{U}}\left(\tfrac{1}{1}\right), \quad \forall \lambda_1, \lambda_2 > 0, \quad \delta_0 < \frac{\lambda_2}{\lambda_1} < 1.$$

And we have the following theorem for the second one

Table 1
Values of δ_0, δ_0^* and δ_1^* for different sample sizes

Sample size (n)	1	2	5	10	25	50
δ_0	0.56108	—	—	—	—	—
$\delta_0^* = \delta_1^*$	—	0.84678	0.63983	0.58354	0.55164	0.54110

Theorem 6. *Assume $n_1 = n_2 = n$.*

1. *If $n = 1$, $P_{\mathrm{O}}(\frac{2}{2}) \geqslant P_{\mathrm{U}}(\frac{2}{2})\ \forall \lambda_1 \geqslant \lambda_2 > 0$.*
2. *If $n > 1$ there are δ_0^* and δ_1^* in $(0, 1)$ such that*

$$P_{\mathrm{O}}\left(\tfrac{2}{2}\right) \geqslant P_{\mathrm{U}}\left(\tfrac{2}{2}\right) \quad \forall \lambda_1, \lambda_2 > 0, \quad 0 < \frac{\lambda_2}{\lambda_1} \leqslant \delta_0^*,$$
$$P_{\mathrm{O}}\left(\tfrac{2}{2}\right) < P_{\mathrm{U}}\left(\tfrac{2}{2}\right) \quad \forall \lambda_1, \lambda_2 > 0, \quad \delta_1^* < \frac{\lambda_2}{\lambda_1} < 1.$$

From these results it is apparent that although the rule performs globally better there are certain configurations of the parameters for which that property does not hold for each of the populations. For the most usual, $n > 1$, case performance is always better for the first population but not for the second one while the opposite is true when the training samples are of size 1. These results are obviously interesting when interest is focused on correct classification in one of the two populations.

In Fig. 2 these results can be checked graphically for different sample sizes. The figure also shows how big the differences are.

The values of δ_0, δ_0^* and δ_1^* are interesting from a practical point of view. From the proof of the theorems it is clear that they cannot be written in terms of elementary functions but these values can be computed numerically and from Fig. 2 it is also clear that $\delta_0^* = \delta_1^*$. Table 1 gives the values of these parameters for different sample sizes.

Remark 7. When $m > 1$ the calculations are much more involved so that no analytic results can be easily obtained. However some comments can be given from what we have observed in the simulations made for this situation. For $m > 1$ the difference of probabilities of correct classification for the first population $P_{\mathrm{O}}(\frac{1}{1}) - P_{\mathrm{U}}(\frac{1}{1})$ behaves worse when m increases while the difference for the second population $P_{\mathrm{O}}(\frac{2}{2}) - P_{\mathrm{U}}(\frac{2}{2})$ increases when $m > 1$. The final result is that the global difference $P_{\mathrm{O}}(\mathrm{MC}) - P_{\mathrm{U}}(\mathrm{MC})$ behaves better for $m > 1$ than for $m = 1$. Therefore, in this situation our rule R_0 is even better than in the usual $m = 1$ case when compared with R_{U}.

3. Classification under type II censorship

Up to this point we have assumed that complete samples are available for classification. In this section, we will study how the proposed rule performs when data are censored. It is well known that in many physical situations where the exponential distribution appears

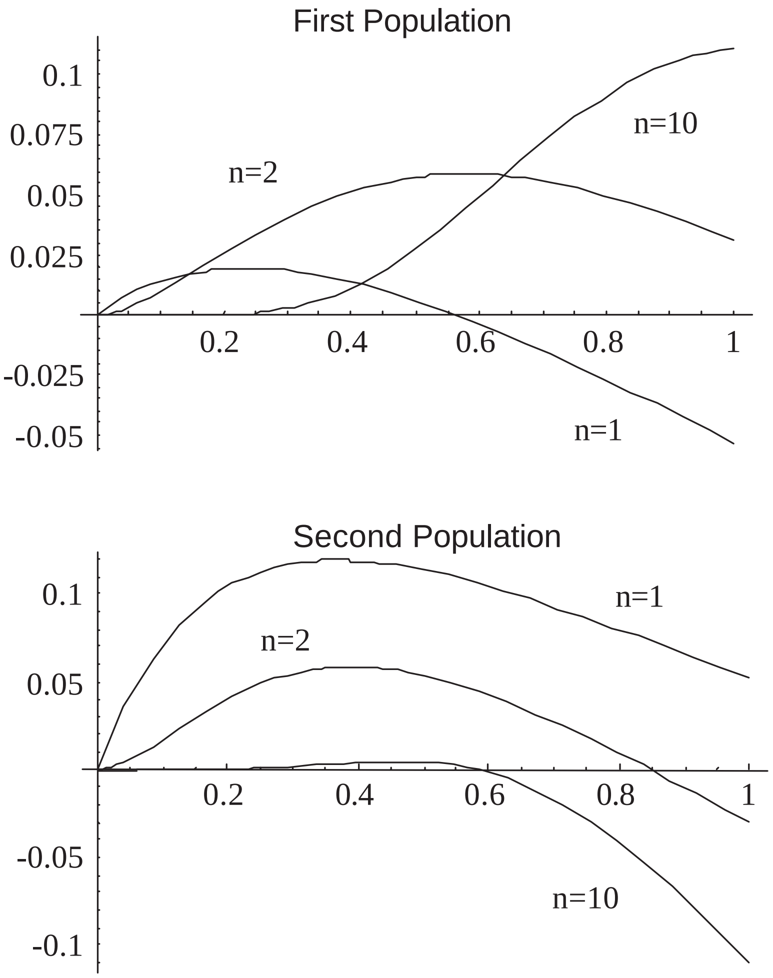

Fig. 2. Difference between the correct classification probabilities for R_{O} and R_{U} for each of the two populations for several sample sizes.

such as life-testing problems or reliability analysis the data are often censored. We will consider here one of the most frequent types of censorship, the so-called type II censorship. In the reliability context, this kind of censorship appears, for example, when *n* units are put to test and the study stops when the first *r* units have failed. In this type of censorship, the number of data we collect is known in advance but the time needed to complete the study is random. In type I censorship the opposite happens, number of data is random and time to end is fixed.

Under type II censorship, the MLE of the exponential parameter λ from a size *n* random sample $X_1, \ldots, X_n$ is known to be

$$\lambda^* = \frac{r}{\sum_{i=1}^{r} X_{(i)} + (n-r)X_{(r)}}$$

using the common notation for the ordered statistic.

The usual and ordered rules under censorship R_{U^*} and R_{O^*} can be easily defined replacing $\widehat{\lambda}_i$ by λ_i^* in the uncensored rules R_U and R_O defined in (1) and (2).

$$R_{U^*}\ :\ \text{Classify } z \text{ into } \begin{cases} \pi_1 & \text{iff } (z-x_0^*)(\lambda_1^*-\lambda_2^*)<0,\\ \pi_2 & \text{otherwise,}\end{cases}$$

where $x_0^* = \frac{\ln\lambda_1^* - \ln\lambda_2^*}{\lambda_1^*-\lambda_2^*}$ and

$$R_{O^*}\ :\ \text{Classify } z \text{ into } \begin{cases} \pi_1 & \text{iff } (z-x_0^*)<0,\\ \pi_2 & \text{otherwise.}\end{cases} .$$

Now taking into account that $\sum_{i=1}^{r} X_{(i)} + (n-r)X_{(r)}$ follows a $\gamma(r,\lambda)$ distribution (cf. Basu, 1965) and the proofs of the theorems in the previous section we can obtain the following results in this case:

From Theorem 1 we can prove that the rule R_{O^*} is consistent. Using the proof of Theorem 2 it is easy to check that this rule outperforms R_{U^*}, i.e. $P_{O^*}(\mathrm{MC}) \leqslant P_{U^*}(\mathrm{MC})$ for any $\lambda_1 \geqslant \lambda_2 > 0$. Finally, from the proof of Theorems 5 and 6 we can study the behavior in each population provided the same number of observations is fixed in each population ($r_1 = r_2$).

4. More than two populations case

When $k>2$ populations are present the rules definitions are more complex. Assume $\lambda_1 \geqslant \cdots \geqslant \lambda_k$. Let $\widehat{\lambda}_1, \widehat{\lambda}_2, \cdots, \widehat{\lambda}_k$ be the (unrestricted) MLEs of the parameters in each population, $\widehat{\lambda}_{(1)} \leqslant \widehat{\lambda}_{(2)} \leqslant \cdots \leqslant \widehat{\lambda}_{(k)}$ the ordered parameters and suppose $\widehat{\lambda}_{(i)} = \widehat{\lambda}_j$. Then the usual rule can be written in the following way:

$$z \in P_j \quad \Leftrightarrow \quad \frac{\ln \widehat{\lambda}_{(i)} - \ln \widehat{\lambda}_{(i+1)}}{\widehat{\lambda}_{(i)} - \widehat{\lambda}_{(i+1)}} < z \leqslant \frac{\ln \widehat{\lambda}_{(i-1)} - \ln \widehat{\lambda}_{(i)}}{\widehat{\lambda}_{(i-1)} - \widehat{\lambda}_{(i)}},$$

while the ordered rule is

$$z \in P_{k-i+1} \quad \Leftrightarrow \quad \frac{\ln \widehat{\lambda}_{(i)} - \ln \widehat{\lambda}_{(i+1)}}{\widehat{\lambda}_{(i)} - \widehat{\lambda}_{(i+1)}} < z \leqslant \frac{\ln \widehat{\lambda}_{(i-1)} - \ln \widehat{\lambda}_{(i)}}{\widehat{\lambda}_{(i-1)} - \widehat{\lambda}_{(i)}},$$

where in both cases only the appropriate inequality is considered for the extreme populations.

The consistency of the ordered rule can be proved using the same techniques appearing in Basu and Gupta (1974). We have used simulations to evaluate the global behavior of the ordered rule. The simulations considered for these cases indicate that the ordered rule globally outperforms the usual one no matter how many populations are considered.

Table 2 deals with the three populations case taking into account that the classification probabilities only depend on $\lambda_1, \lambda_2, \lambda_3$ through λ_2/λ_1 and λ_3/λ_2. In this table we have assumed $n_1 = \cdots = n_k = n = 10$. In each cell of the table we give the following values: left column contains values for R_U, first the probability of correct classification and then $P_U(\frac{1}{1})$, $P_U(\frac{2}{2})$ and $P_U(\frac{3}{3})$. Right column of each cell contains same values for the ordered rule R_O. Notice that in all cases $P_O(\frac{1}{1}) \geqslant P_U(\frac{1}{1})$, $P_O(\frac{2}{2}) \leqslant P_U(\frac{2}{2})$ and $P_O(\frac{3}{3}) \geqslant P_O(\frac{3}{3})$. Same configuration holds for other values of n including $n=1$.

Table 2
Global and individual correct classification probabilities for $k = 3$

$\lambda_3/\lambda_2 \backslash \lambda_2/\lambda_1$	0.1		0.25		0.5		0.75		0.9		1	
0.1	0.77	0.77	0.69	0.69	0.62	0.62	0.59	0.59	0.57	0.57	0.56	0.56
	0.98	0.98	0.96	0.96	0.94	0.94	0.91	0.91	0.89	0.89	0.88	0.88
	0.54	0.54	0.32	0.32	0.14	0.14	0.07	0.07	0.06	0.06	0.05	0.05
	0.78	0.78	0.78	0.78	0.78	0.78	0.77	0.77	0.76	0.76	0.75	0.75
0.25	0.71	0.71	0.63	0.63	0.56	0.56	0.52	0.52	0.50	0.50	0.49	0.49
	0.96	0.96	0.93	0.93	0.89	0.89	0.84	0.84	0.81	0.81	0.78	0.78
	0.52	0.52	0.30	0.30	0.15	0.15	0.08	0.08	0.07	0.07	0.07	0.07
	0.64	0.64	0.65	0.65	0.64	0.64	0.63	0.63	0.63	0.63	0.62	0.62
0.5	0.61	0.64	0.55	0.56	0.47	0.49	0.44	0.45	0.42	0.43	0.40	0.41
	0.86	0.94	0.84	0.89	0.72	0.81	0.70	0.76	0.69	0.73	0.62	0.69
	0.46	0.46	0.30	0.28	0.22	0.15	0.14	0.08	0.10	0.06	0.16	0.07
	0.50	0.52	0.51	0.52	0.49	0.52	0.47	0.50	0.47	0.49	0.43	0.49
0.75	0.51	0.60	0.47	0.52	0.42	0.45	0.38	0.40	0.36	0.38	0.35	0.37
	0.64	0.92	0.63	0.87	0.58	0.78	0.50	0.71	0.49	0.67	0.45	0.65
	0.46	0.43	0.38	0.26	0.29	0.13	0.27	0.07	0.25	0.06	0.27	0.06
	0.43	0.44	0.39	0.44	0.37	0.44	0.35	0.43	0.34	0.41	0.33	0.40
0.9	0.48	0.57	0.43	0.50	0.38	0.43	0.36	0.38	0.34	0.36	0.33	0.35
	0.56	0.92	0.51	0.83	0.45	0.76	0.43	0.68	0.38	0.64	0.36	0.62
	0.47	0.39	0.43	0.25	0.39	0.12	0.31	0.06	0.30	0.06	0.32	0.05
	0.40	0.41	0.35	0.41	0.31	0.40	0.32	0.40	0.33	0.37	0.32	0.36
1	0.46	0.56	0.41	0.49	0.37	0.42	0.34	0.37	0.33	0.35	0.33	0.33
	0.49	0.91	0.41	0.84	0.39	0.76	0.39	0.68	0.35	0.64	0.32	0.61
	0.47	0.38	0.50	0.23	0.42	0.11	0.35	0.06	0.30	0.05	0.35	0.06
	0.40	0.40	0.31	0.39	0.29	0.38	0.30	0.37	0.35	0.35	0.33	0.34

Table 3
Global correct classification probabilities for different values of k

$\delta \backslash k$	4		5		7		10	
0.1	0.70	0.70	0.64	0.64	0.55	0.55	0.44	0.44
0.25	0.56	0.56	0.50	0.50	0.40	0.41	0.31	0.31
0.5	0.41	0.42	0.37	0.37	0.30	0.31	0.24	0.24
0.75	0.30	0.33	0.26	0.28	0.22	0.23	0.17	0.18
0.9	0.26	0.28	0.21	0.23	0.16	0.18	0.12	0.13
0.99	0.25	0.25	0.20	0.20	0.14	0.15	0.10	0.10

In Table 3, we have considered several values of k from 4 to 10. The simulations have also been done for $n_i = 10, i = 1, \dots, k$ and now we have further assumed $\lambda_{i+1}/\lambda_i = \delta \in (0, 1]$ for $i = 1, \dots, k-1$. For this case only the global classification probabilities are given. The first value in each cell corresponds to R_{U} and the second one to R_{O}. We can observe that the ordered rule seems to perform globally better even for high values of k.

Acknowledgements

The authors thank the referees for their suggestions and comments which led to this improved version of the paper.

Appendix A.

Here, we develop the proofs of Theorems 5 and 6. First, we state a pair of lemmas that will be used in the proof of the theorems.

Lemma A.1. *For any $n > 0$ and $a, b \in \Re^+$*

$$\frac{1}{(1+a)^2} \geqslant \frac{(1+a+\frac{1}{n}b)^{2n}}{(1+a+\frac{1}{n+1}b)^{2n+2}}.$$

Let us denote

$$g(t,n,\delta) = \frac{1}{(1+\delta t)^{2n}} - \frac{2}{(1+\delta t+\frac{1}{n}\frac{t\ \ln(t)}{t-1})^{2n}}. \tag{A.1}$$

Lemma A.2. *For each $\delta \in (0,1]$ and $n > 0$ there is a single point $t_0(\delta,n) \in (0,1)$ such that $g(t,n,\delta) < 0$ for $0 < t < t_0(\delta,n)$ and $g(t,n,\delta) > 0$ for $t_0(\delta,n) < t < 1$.*

Proof of Lemma A.1. Define

$$f(b) = \frac{(1+a+\frac{1}{n}b)^{2n}}{(1+a+\frac{1}{n+1}b)^{2n+2}}.$$

It is straightforward to check that $f'(b) < 0$ as a and b are positive numbers. Therefore, $f(b)$ is decreasing in b, and we have $f(b) \leqslant f(0), \forall b \geqslant 0$, so that

$$\frac{(1+a+\frac{1}{n}b)^{2n}}{(1+a+\frac{1}{n+1}b)^{2n+2}} \leqslant \frac{(1+a)^{2n}}{(1+a)^{2n+2}} = \frac{1}{(1+a)^2}. \qquad \square$$

Proof of Lemma A.2. Define

$$g^*(t) = \left(1+\delta t+\frac{1}{n}\frac{t\ \ln(t)}{t-1}\right)^{2n} - 2(1+\delta t)^{2n}$$

and assume that δ and n are fixed numbers. It is clear that to prove the lemma it is enough to show that for these fixed values of δ and n there is a single value $t_0 \in (0,1)$ such that $g^*(t_0) = 0$. It is easy to check that

$$\lim_{t\to 0^+} g^*(t) = -1 < 0.$$

Next, we prove that

$$\lim_{t\to 1^-} g^*(t) = \left(1+\delta+\frac{1}{n}\right)^{2n} - 2(1+\delta)^{2n} > 0.$$

This is equivalent to

$$\frac{1}{n}(2^{1/2n}-1)^{-1} > 1+\delta$$

but this is true since $\frac{1}{n}(2^{1/2n}-1)^{-1}$ is increasing in n and we have

$$\frac{\frac{1}{n}}{2^{1/2n}-1} \geqslant \frac{1}{\sqrt{2}-1} > 2 \geqslant 1+\delta, \text{ for any } \delta \in (0,1].$$

We have proved that there is a point $t_0 \in (0,1)$ such that $g^*(t_0)=0$. To conclude the result we just have to check that this point is unique.

$$g^*(t)=0 \quad \Leftrightarrow \quad \left(1+\frac{1}{n}\frac{t\ \ln(t)}{(t-1)(1+\delta t)}\right)^{2n} - 2 = 0.$$

Then to prove the result it is enough to prove that

$$f(t) = \frac{t\ \ln(t)}{(t-1)(1+\delta t)}$$

is strictly increasing in $t \in (0,1)$. The first derivative is

$$f'(t) = \frac{(t-1)(1+t\delta)-(1+t^2\delta)\ln(t)}{(t-1)^2(1+\delta t)^2}.$$

Using Taylor's expansion of logarithm

$$-\ln(t) = \sum_{i=1}^{\infty}\frac{(1-t)^i}{i} > (1-t)+\frac{(1-t)^2}{2} \quad \forall t \in (0,1),$$

we have

$$\begin{aligned}&(t-1)(1+t\delta)-(1+t^2\delta)\ln(t)\\ &\quad > (t-1)(1+t\delta)+(1+t^2\delta)\left[(1-t)+\frac{(1-t)^2}{2}\right]\\ &\quad = (1-t)^2\left[\frac{1+t^2\delta}{2}-t\delta\right]\end{aligned}$$

and $\frac{1+t^2\delta}{2} - t\delta > 0$ in $(0,1)$ so that $f'(t) > 0\ \forall t \in (0,1)$, $f(t)$ is strictly increasing $\forall t \in (0,1)$ and the result is done. □

Now we prove the theorems.

Proof of Theorem 5. *First part*: Taking into account (3) and (4) we just have to prove

$$P_{\lambda_1}(Z \leqslant \widehat{x_0}, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) \geqslant P_{\lambda_1}(Z > \widehat{x_0}, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2)$$

or

$$P_{\lambda_1}\left(Z \leqslant \widehat{x_0}, \sum_{i=1}^{n} X_i \geqslant \sum_{i=1}^{n} Y_i\right) \geqslant P_{\lambda_1}\left(Z > \widehat{x_0}, \sum_{i=1}^{n} X_i \geqslant \sum_{i=1}^{n} Y_i\right).$$

If we use (5) as in Theorem 2 we have

$$\begin{aligned} &P_{\lambda_1}\left(Z > \widehat{x_0}, \sum_{i=1}^{n} X_i \geqslant \sum_{i=1}^{n} Y_i\right) \\ &\quad = \int_0^\infty \int_0^x \int_{\widehat{x_0}}^\infty \lambda_1 \mathrm{e}^{-\lambda_1 z} f(x, \lambda_1, n) f(y, \lambda_2, n)\, \mathrm{d}z\, \mathrm{d}y\, \mathrm{d}x \\ &\quad = \int_0^\infty \int_0^x \exp\left(\frac{\lambda_1}{n} \frac{\ln \frac{y}{x}}{1 - \frac{y}{x}} y\right) f(x, \lambda_1, n) f(y, \lambda_2, n)\, \mathrm{d}y\, \mathrm{d}x. \end{aligned}$$

Now, if we consider the change of variable $t = \frac{y}{x}$, $u = x$ the integral can be written as

$$\begin{aligned} &\int_0^\infty \int_0^1 u \exp\left(-\frac{\lambda_1}{n} \frac{ut}{t-1} \ln(t)\right) f(u, \lambda_1, n) f(ut, \lambda_2, n)\, \mathrm{d}t\, \mathrm{d}u \\ &\quad = \int_0^1 \int_0^\infty \frac{(\lambda_1 \lambda_2)^n t^{n-1}}{((n-1)!)^2} u^{2n-1} \exp\left(-\left(\lambda_1 + \lambda_2 t + \frac{\lambda_1 t \ln(t)}{n(t-1)}\right) u\right) \mathrm{d}u\, \mathrm{d}t \\ &\quad = \int_0^1 \frac{(\lambda_1 \lambda_2)^n}{((n-1)!)^2} \frac{(2n-1)!}{(\lambda_1 + \lambda_2 t + \frac{\lambda_1}{n} \frac{t}{t-1} \ln(t))^{2n}} t^{n-1}\, \mathrm{d}t \\ &\quad = K \int_0^1 h_1(t, n, \delta)\, \mathrm{d}t, \end{aligned}$$

where

$$\delta = \frac{\lambda_2}{\lambda_1} \in (0, 1], \qquad K = \frac{\delta^n (2n-1)!}{((n-1)!)^2}$$

and

$$h_1(t, n, \delta) = \frac{t^{n-1}}{(1 + \delta t + \frac{1}{n} \frac{t}{t-1} \ln(t))^{2n}}. \tag{A.2}$$

With similar arguments

$$\begin{aligned}P(\widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) &= P\left(\sum_{i=1}^{n} X_i \geqslant \sum_{i=1}^{n} Y_i\right)\\ &= \frac{(\lambda_1\lambda_2)^n(2n-1)!}{((n-1)!)^2}\int_0^1 \frac{t^{n-1}}{(\lambda_1+\lambda_2 t)^{2n}}\,\mathrm{d}t\\ &= K\int_0^1 h_0(t,n,\delta)\,\mathrm{d}t\end{aligned}$$

where

$$h_0(t,n,\delta) = \frac{t^{n-1}}{(1+\delta t)^{2n}}. \tag{A.3}$$

Now from (A.2) and (A.3)

$$\begin{aligned}P_{\lambda_1}(Z \leqslant \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) &= P(\widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) - P_{\lambda_1}(Z > \widehat{x}_0, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2)\\ &= K\int_0^1 (h_0(t,n,\delta) - h_1(t,n,\delta))\,\mathrm{d}t\end{aligned}$$

and it is enough to prove

$$P_{\lambda_1}(Z \leqslant \widehat{x_0}, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) - P_{\lambda_1}(Z > \widehat{x_0}, \widehat{\lambda}_1 \leqslant \widehat{\lambda}_2) > 0$$

or

$$K\int_0^1 (h_0(t,n,\delta) - 2h_1(t,n,\delta))\,\mathrm{d}t > 0.$$

As $K > 0$ we will just check that the integral is positive. To prove this we use induction on n. Denote

$$H(n,\delta) = \int_0^1 (h_0(t,n,\delta) - 2h_1(t,n,\delta))\,\mathrm{d}t. \tag{A.4}$$

We start with $n = 2$. In this case

$$H(2,\delta) = \int_0^1 t\left(\frac{1}{(1+\delta t)^4} - \frac{2}{(1+\delta t + \frac{1}{2}\frac{t\ \ln(t)}{t-1})^4}\right)\mathrm{d}t$$

since $t\ \ln(t)/(t-1)$ is a concave function it is not difficult to check

$$\frac{t\ \ln(t)}{t-1} \geqslant \begin{cases} 2t, & t \in [0, 0.2),\\ \frac{3}{4}t + \frac{1}{4}, & t \in [0.2, 1],\end{cases}$$

so that

$$\begin{aligned}H(2,\delta) &\geqslant \int_0^1 \frac{t}{(1+\delta t)^4}\,\mathrm{d}t - \int_0^{0.2} \frac{2t}{(1+\delta t + t)^4}\,\mathrm{d}t - \int_{0.2}^1 \frac{2t}{(1+\delta t + \frac{3}{8}t + \frac{1}{8})^4}\,\mathrm{d}t\\ &= \frac{7128 + 10350\delta - 1008\delta^2 - 8347\delta^3 - 5041\delta^4 - 1338\delta^5 - 172\delta^6 - 8\delta^7}{6(1+\delta)^3(3+2\delta)^2(6+\delta)^3}\end{aligned}$$

and the quotient is positive because the denominator is positive in $(0, 1]$ and the numerator is a concave function with positive values in 0 and 1.

As we are using induction on n now we assume the result is true for n and we check it for $n+1$. From (A.2) and (A.3)

$$\begin{aligned} H(n+1,\delta) &= \int_0^1 (h_0(t,n+1,\delta) - 2h_1(t,n+1,\delta))\,\mathrm{d}t \\ &= \int_0^1 t\left(\frac{h_0(t,n,\delta)}{(1+\delta t)^2} - 2h_1(t,n,\delta)\frac{(1+\delta t+\frac{1}{n}\frac{t\ \ln(t)}{t-1})^{2n}}{(1+\delta t+\frac{1}{n+1}\frac{t\ \ln(t)}{t-1})^{2n+2}}\right)\mathrm{d}t. \end{aligned}$$

Now using Lemma A.1

$$\begin{aligned} H(n+1,\delta) &\geqslant \int_0^1 \frac{t}{(1+\delta t)^2}(h_0(t,n,\delta) - 2h_1(t,n,\delta))\,\mathrm{d}t \\ &= \int_0^1 \frac{t}{(1+\delta t)^2} t^{n-1} g(t,n,\delta)\,\mathrm{d}t, \end{aligned}$$

where $g(t,n,\delta)$ is the function defined in (A.1) and used in Lemma A.2.

It is straightforward to check that $r(t) = \frac{t}{(1+\delta t)^2}$ is positive and increasing in $t \in [0,1]$ so if we consider t_0 defined in Lemma A.2

$$\begin{aligned} H(n+1,\delta) &\geqslant \int_0^{t_0} r(t)t^{n-1}g(t,n,\delta)\,\mathrm{d}t + \int_{t_0}^1 r(t)t^{n-1}g(t,n,\delta)\,\mathrm{d}t \\ &\geqslant \int_0^{t_0} r(t_0)t^{n-1}g(t,n,\delta)\,\mathrm{d}t + \int_{t_0}^1 r(t_0)t^{n-1}g(t,n,\delta)\,\mathrm{d}t \\ &= r(t_0)\int_0^1 t^{n-1}g(t,n,\delta)\,\mathrm{d}t \\ &= r(t_0)H(n,\delta) > 0 \end{aligned}$$

for each $\delta \in (0,1]$ since in Lemma A.2 we proved, $g(t,n,\delta) < 0$ for $t < t_0$ and positive for $t > t_0$. Then the first part of the theorem is proved.

Second part:

Now $n = 1$. For this part it is enough to prove $\exists \delta_0 \in (0,1]$ such that

$$P_{\lambda_1}(Z \leqslant \widehat{x_0}, \widehat{\lambda_1} \leqslant \widehat{\lambda_2}) \geqslant P_{\lambda_1}(Z > \widehat{x_0}, \widehat{\lambda_1} \leqslant \widehat{\lambda_2})$$

for any $\lambda_1, \lambda_2 > 0$ satisfying $0 < \frac{\lambda_2}{\lambda_1} \leqslant \delta_0$ and the opposite inequality for $\delta_0 < \frac{\lambda_2}{\lambda_1} \leqslant 1$. Using (A.4), we can write

$$P_{\lambda_1}(Z \leqslant \widehat{x_0}, \widehat{\lambda_1} \leqslant \widehat{\lambda_2}) - P_{\lambda_1}(Z > \widehat{x_0}, \widehat{\lambda_1} \leqslant \widehat{\lambda_2}) = \delta \cdot H(1,\delta).$$

$H(1,\delta)$ is strictly decreasing in $\delta \in (0,1)$ because

$$\begin{aligned}
H'(1,\delta)&=(-2)\int_0^1 t\left(\frac{1}{(1+\delta t)^3}-\frac{2}{(1+\delta t+t\frac{\ln(t)}{t-1})^3}\right)\mathrm{d}t\\
&\leqslant(-2)\int_0^1 t\left(\frac{1}{(1+\delta t)^3}-\frac{2}{(1+(\delta+1)t)^3}\right)\mathrm{d}t\\
&=\frac{\delta^2-2}{(1+\delta)^2(2+\delta)^2}<0 \qquad \text{(A.5)}
\end{aligned}$$

where in (A.5) we have used the left side of the inequality

$$t<t\,\frac{\ln(t)}{t-1}\leqslant t+0.22 \quad \forall t\in(0,1). \qquad \text{(A.6)}$$

Using (A.6) again

$$H(0,1)=\int_0^1\left(1-\frac{2}{(1+t\frac{\ln(t)}{t-1})^2}\right)\mathrm{d}t>\int_0^1\left(1-\frac{2}{(1+t)^2}\right)\mathrm{d}t=0.$$

$$\begin{aligned}
H(1,1)&=\int_0^1\left(\frac{1}{(1+t)^2}-\frac{2}{(1+t+t\,\frac{\ln(t)}{t-1})^2}\right)\mathrm{d}t\\
&\leqslant\int_0^1\left(\frac{1}{(1+t)^2}-\frac{2}{(1+t+t+0.22)^2}\right)\mathrm{d}t=-0.009<0,
\end{aligned}$$

and the result is done. □

Proof of Theorem 6. Using arguments similar to those in the proof of Theorem 5 we can write

$$\begin{aligned}
P_{\mathrm{O}}\left(\tfrac{2}{2}\right)-P_{\mathrm{U}}\left(\tfrac{2}{2}\right)&=P_{\lambda_2}(Z>\widehat{x}_0,\widehat{\lambda}_1\leqslant\widehat{\lambda}_2)-P_{\lambda_2}(Z\leqslant\widehat{x}_0,\widehat{\lambda}_1\leqslant\widehat{\lambda}_2)\\
&=\frac{\delta^n(2n-1)!}{((n-1)!)^2}\,H_2(n,\delta),
\end{aligned}$$

where

$$H_2(n,\delta)=\int_0^1 t^{n-1}\left(\frac{2}{(1+\delta t+\frac{\delta}{n}\,\frac{t\ln(t)}{t-1})^{2n}}-\frac{1}{(1+\delta t)^{2n}}\right)\mathrm{d}t. \qquad \text{(A.7)}$$

First part: Now, we fix $n=1$. From (A.6) and (A.7)

$$H_2(1,0)=\int_0^1(2-1)\,\mathrm{d}t=1$$

$$\begin{aligned}
H_2(1,1)&=\int_0^1\left(\frac{2}{(1+t+t\frac{\ln(t)}{t-1})^2}-\frac{1}{(1+t)^2}\right)\mathrm{d}t\\
&\geqslant\int_0^1\left(\frac{2}{(1+t+t+0.22)^2}-\frac{1}{(1+t)^2}\right)\mathrm{d}t=0.009>0.
\end{aligned}$$

$$
\begin{aligned}
H_2'(1,\delta) &= (-2)\int_0^1 \left[\frac{2(t + t\frac{\ln(t)}{t-1})}{(1+\delta t + \delta t\,\frac{\ln(t)}{t-1})^3} - \frac{t}{(1+\delta t)^3} \right] \mathrm{d}t \\
&\leqslant (-2)\int_0^1 \left[\frac{4t}{(1+\delta t + \delta(t+0.22))^3} - \frac{t}{(1+\delta t)^3} \right] \mathrm{d}t \\
&= \frac{135531\delta^3 + 238150\delta^2 - 417500\delta - 375000}{(540-111\delta)^2(1+\delta)^2(50+11\delta)} < 0,
\end{aligned}
$$

the last fraction is negative in $(0, 1]$ since the denominator is positive and the numerator is a convex function whose values in 0 and 1 are negative. Therefore, $H_2(1,\delta)$ is decreasing in δ and the first part of the theorem is proved.

Second part: From Theorem 2

$$P_{\mathrm{O}}\left(\tfrac{1}{1}\right) + P_{\mathrm{O}}\left(\tfrac{2}{2}\right) \geqslant P_{\mathrm{U}}\left(\tfrac{1}{1}\right) + P_{\mathrm{U}}\left(\tfrac{2}{2}\right)$$

and in Remark 3 we have noticed that the equality only holds when $\lambda_1 = \lambda_2$, i. e. when $\delta = 1$. Moreover from Theorem 5 when $n > 1$

$$P_{\mathrm{O}}\left(\tfrac{1}{1}\right) > P_{\mathrm{U}}\left(\tfrac{1}{1}\right)$$

for any $\lambda_1 \geqslant \lambda_2 > 0$. Then we have

$$P_{\mathrm{O}}\left(\tfrac{2}{2}\right) < P_{\mathrm{U}}\left(\tfrac{2}{2}\right) \qquad \text{for } \delta = 1,$$

and as the probabilities are continuous functions in δ that is true in a neighborhood of $\delta = 1$.

On the other hand

$$H_2(n,0) = \int_0^1 t^{n-1}\,\mathrm{d}t = \frac{1}{n} > 0$$

and as $\frac{\delta^n(2n-1)!}{((n-1)!)^2} > 0$ when $\delta > 0$ using continuity in δ again the difference is positive in a neighborhood of $\delta = 0$ and the result is done. □

References

Adegboye, O.S., 1993. The optimal classification rule for exponential populations. Austral. J. Statist. 35 (2), 185–194.

Basu, A.P., 1965. On some tests of hypotheses relating to the exponential distribution when some outliers are present. J. Amer. Statist. Assoc. 60, 548–559.

Basu, A.P., Gupta, A.K., 1974. Classification rules for exponential populations. In: Proschan, F., Serfling, R.J. (Eds.), Reliability and Biometry. SIAM, Philadelphia, pp. 637–650.

Bhattacharya, P.K., Das Gupta, S., 1964. Classification between univariate exponential distributions. Sankhyā Ser. A 26, 17–24.

Chang, Y.-T., Shinozaki, N., 2002. A comparison of restricted and unrestricted estimators in estimating linear functions of ordered scale parameters of two gamma distributions. Ann. Inst. Statist. Math. 54 (4), 848–860.

Fernández, M.A., Rueda, C., Salvador, B., 1999. The loss of efficiency estimating contrast under restrictions. Scand. J. Statist. 26, 79–92.

Lee, C.C., 1988. The quadratic loss of order restricted estimators for treatment means with a control. Ann. Statist. 16, 751–758.

Long, T., Gupta, R.D., 1998. Alternative linear classification rules under order restrictions. Comm. Statist.—Theory Methods 27 (3), 559–575.

Menéndez, J.A., Salvador, B., 1991. Anomalies of the likelihood ratio test for testing restricted hypotheses. Ann. Statist. 19, 889–898.

Vijayasree, G., Singh, H., 1991. Simultaneous estimation of two ordered exponential parameters. Comm. Statist.—Theory Methods 20 (8), 2559–2576.

Vijayasree, G., Singh, H., 1993. Mixed estimators of two ordered exponential means. J. Statist. Plann. Inference 35, 47–56.

Zelen, M., 1966. Application of exponential models to problems in cancer research with discussion. J. Roy. Statist. Soc. Ser. A 129, 368–398.

Research Article

Statistics in Medicine

Received 11 August 2011, Accepted 14 May 2012 Published online 19 July 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5476

Classification of samples into two or more ordered populations with application to a cancer trial

D. Conde, M. A. Fernández,*[†] C. Rueda and B. Salvador

In many applications, especially in cancer treatment and diagnosis, investigators are interested in classifying patients into various diagnosis groups on the basis of molecular data such as gene expression or proteomic data. Often, some of the diagnosis groups are known to be related to higher or lower values of some of the predictors. The standard methods of classifying patients into various groups do not take into account the underlying order. This could potentially result in high misclasiffication rates, especially when the number of groups is larger than two.

In this article, we develop classification procedures that exploit the underlying order among the mean values of the predictor variables and the diagnostic groups by using ideas from order-restricted inference. We generalize the existing methodology on discrimination under restrictions and provide empirical evidence to demonstrate that the proposed methodology improves over the existing unrestricted methodology. The proposed methodology is applied to a bladder cancer data set where the researchers are interested in classifying patients into various groups. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: classification rules; order-restricted inference; cancer diagnostic test research

1. Introduction

An important component of proper patient care, especially in cancer treatment, is correct classification of the patient into one of the disease stages. Such a classification problem in bladder cancer motivated this work. We aim to investigate new methodologies to select the best classifiers in the context of an in vitro diagnostic tool for the disease. Our industrial and pharmaceutical partners in this research are Proteomika S.L. and Laboratorios SALVAT, S.A.

The final aim of the research is to develop a non-invasive in vitro test for the diagnosis of bladder cancer recurrence (transitional cell carcinoma). Because of its high recurrence rates, the disease is perceived as chronic (as many as 80% of patients have at least one recurrence). Therefore, high rates of recurrence and progression make careful long-term follow-up a clinical priority. An ideal diagnostic test must show elevated levels of sensitivity and specificity and could be used on patients with a history of transitional cell carcinoma and on a clinical follow-up. Development of new non-invasive methods to monitor patients could be very useful to complement and reduce the number of cystoscopes, which up to date remains the gold standard for diagnosis. The institutional review board of the clinical center approved the study, and all patients signed an informed consent.

We classified patients in five levels on the basis of cytoscopy. The first level is the control level (i.e., negative result of cytoscopy, therefore considered as absence of bladder cancer), and the other four levels are denoted as Ta, T1G1, T1G3, and T2, each of them corresponding to increasingly advanced levels of cancer. This combines the TNM staging, which uses the size and extension of the primary Tumor, its lymphatic Nodes involvement, and the presence of Metastases to classify the progression of cancer [1, 2] and the grading. Grading is also important as there is interobserver variability in classifying

Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Spain

Correspondence to: M. A. Fernández, Departamento de Estadística e Investigación Operativa. C/Prado de la Magdalena s/n. Universidad de Valladolid, 47005 Valladolid, Spain.

[†] *E-mail: miguelaf@eio.uva.es*

due not only to staging but also to grading [3]. To be more precise, stage T describes the size of the tumor and whether it has spread, and grade G refers to the appearance of the cells under the microscope. In Ta stage, the tumor is only in the innermost lining of the bladder, whereas in T1, it has started to grow into the connective tissue just under the bladder lining. Ta and T1 are non-invasive tumors, but T2 is an invasive tumor because in that stage, the tumor has grown through the connective tissue into the muscle. The grade also gives an idea about how rapidly the cancer may develop [4]. G1, low grade, means that the cancer looks much like normal bladder cells, whereas in the G3 grade, the cells look very abnormal and are likely to grow more quickly and are more likely to spread.

As usual, in this kind of research, an initial database with a moderate number of patients was provided. The purpose of this pilot study was to confirm or discard the associations among the proteins and the illness (i.e., to check if the protein levels are useful to predict the level of the illness) in order to establish a larger multicenter study.

On the basis of a priori information provided by our pharmaceutical and industrial partners, we use three proteins for our analysis. As, for intellectual property reasons, the names of the proteins used in this study are not disclosed in the paper, we denote these proteins as P_1, P_2, and P_3. We expect the mean values of these three proteins to increase with the severity of the illness. Because sample sizes in some of the illness groups are relatively very small compared with those of the others and also because clinically it seemed reasonable, we merged the five initial levels of the illness in the following three groups: the control group, G_1; the Ta + T1G1 group, G_2; and the T1G3 + T2 group, G_3. As usual, the values of the protein levels have been transformed logarithmically so that the variables are approximately normally distributed.

When the database was provided, the classification results based on standard statistical methodology were surprisingly poor. Some protein levels were expected to have a good classification power for discriminating among the levels of the illness as those protein levels were expected to increase with the level of the sickness. However, the percentages of correct classification obtained with these protein levels as predictors were quite low. A careful exploration of the database exposed that the levels of those proteins known to be associated with the disease did not conform to the level of the sickness. In situations such as these, the classical classification rules, which are designed to discriminate between groups, do not exploit the underlying information regarding the order relations among the mean values of the predictor values and the diagnostic groups. Consequently, as will be seen in this paper, they are not expected to perform as well as rules that honor the underlying ‘inequality’ structure.

In this work, we present discriminant rules designed to deal with that sort of information. From the theoretical point of view, this method is an extension and non-trivial generalization of the previous efforts made in trying to define classification rules under order restrictions for the two population cases [5, 6]. However, the extension to more than two populations is not straightforward. When only two populations are present, there are not many ordering chances among them; but when there are more than two populations, a different ordering for each of the predictors may appear. Notice that, although the populations may be ordered, our focus is different from those considering an ordinal response variable as those ordinal rules only rely on the natural ordering existing among the populations.

From the practical point of view, we would like to remark the fact that the kind of data set used in this work is quite usual in the statistical practice when dealing with protein or microarray data or, in general, when a small data set is provided to derive a classification rule and there is some additional knowledge on the problem. This work shows the power of order-restricted inference in the solution of real statistical problems. Some recent papers in this line are [7–11].

The layout of the work is as follows. In Section 2, we develop the methodology underlying the proposed solution and its properties. We devote Section 3 to the simulation study exposing the good behavior of the new rules. We present in Section 4 the results for the bladder cancer problem; finally, we give the discussion and future developments in Section 5.

2. Isotonized rules

In this section, we describe the new classification rules we propose. These rules are specially interesting for the analysis of problems such as the one described in the introduction. Initially, our problem is the classical statistical problem of classifying observations in one of k populations. The main difference is that we assume additional information on the parameters of the populations. We describe Fisher’s linear discriminant rule, which is the classical solution for this situation and is available in any statistical package dealing with discrimination problems. Then, we detail how the additional information is mathematically taken into account to improve the rules and define the new rules incorporating that additional information. The new isotonized rules have a very similar structure to Fisher's rule. The main difference is that we used alternative estimators of the parameters that incorporate the additional information available on the problem.

Let us denote population i as G_i and as $X = \left(X_1, \ldots, X_p\right)'$ the vector of predictors, that is, the variables to be used for classifying the observations. We assume normality, so that if the observation considered comes from G_i, then $X \sim N_p\left(\mu_i, \Sigma\right)$, where $\mu_i = \left(\mu_{i1}, \ldots, \mu_{ip}\right)'$ is the mean of vector X in population i and Σ is the covariance matrix of vector X. We are assuming different means in each population but a common covariance matrix. In the rest of this work, we will also assume equal a priori probabilities for the groups. The case of different a priori probabilities is straightforward from what follows.

Under these assumptions, it is well known that the best way of classifying a new observation $z = \left(z_1, \ldots, z_p\right)'$ is the following linear classification rule

$$\text{classify } z \text{ in } G_i \text{ iff } i = \arg_j \min \left\{ \left(z - \mu_j\right)' \Sigma^{-1} \left(z - \mu_j\right), j = 1, \ldots, k \right\}.$$

As the mean vectors and the covariance matrix are usually unknown, they are estimated so that the usual classification rule (Fisher's rule) for this case is

$$\text{classify } z \text{ in } G_i \text{ iff } i = \arg_j \min \left\{ \left(z - \overline{x}_j\right)' S^{-1} \left(z - \overline{x}_j\right), j = 1, \ldots, k \right\}. \tag{1}$$

Next, we define rules that take into account the additional information. This information is incorporated into the rule restricting the parameter space. For our bladder cancer case study, we are assuming that the levels of the proteins increase with the stage of the illness. In the usual statistical terminology, we can say that there is a simple order among the three population means in each of the three variables. Mathematically, the additional information is formulated using cones. In our case, we can write that

$$\left(\mu_{1s}, \mu_{2s}, \mu_{3s}\right) \in C^*_{\text{BC}} = \left\{ x \in \mathbb{R}^3 : x_1 \leqslant x_2 \leqslant x_3 \right\} \text{ for } s = 1, 2, 3.$$

Fernández et al [6] dealt with case $k = 2$. In that case, the problem can be simplified rewriting the rule and the additional information on the parameters in terms of the difference of means $\delta = \mu_1 - \mu_2$. Then, the usual estimator of the difference of means is projected onto the cone containing the information. A family of estimators of the difference of means, δ^*_γ, with $\gamma \in [0, 1]$, that fulfill the additional information is obtained by means of an iterative procedure that is shown to converge. These estimators are plugged into the original rule to obtain the following isotonized rules

$$R_\delta(\gamma)\text{: classify } z \text{ in } G_1 \text{ iff } \left(z - \left[\frac{n_1 \overline{x}_1 + n_2 \overline{x}_2}{n_1 + n_2} - \frac{n_1 - n_2}{2(n_1 + n_2)} \delta^*_\gamma \right] \right)' S^{-1} \delta^*_\gamma \geqslant 0, \tag{2}$$

where n_1 and n_2 are the sample sizes of the populations.

For full details on these rules, the reader is referred to [6], where results and simulations showing that this rule performs better than the two populations version of Fisher's rule are given.

However, these rules do not have a direct generalization to the case of $k > 2$ populations as in that case, the reduction of the problem to a single parameter $\delta \in \mathbb{R}^p$ is not possible. Our proposal in this paper is to work in an extended space $\mathbb{R}^{kp}$ and to define a cone of restrictions in $\mathbb{R}^{kp}$ representing the additional information available on the problem. For the bladder cancer case study, we have that

$$\left(\mu'_1, \mu'_2, \mu'_3\right) \in C_{\text{BC}} = \left\{ x \in \mathbb{R}^9 : x_i \leqslant x_{i+3} \leqslant x_{i+6}, i = 1, 2, 3 \right\}.$$

Projecting the sample means onto the cone with the appropriate metric ensures that we find the estimator of the means that is as close as possible to the sample means and also verifies the additional information.

It is easy to write the most usual order restrictions using cones. For example, the simple order cone appears, as in our case study, when there is an increasing relation between the level of the illness and the predictors. The cone can be expressed as

$$C_{\text{SO}} = \left\{ x \in \mathbb{R}^{kp} : x_t \leqslant x_{t+p} \leqslant \ldots \leqslant x_{t+(k-1)p}, t \in T \right\},$$

where $T \subseteq \{1, \ldots, p\}$ is the set of predictors for which the simple order relation is known to hold. It is obvious that C_{BC} is a particular case of C_{SO} with $k = p = 3$ and $T = \{1, 2, 3\}$.

Other usual cone of restrictions among the mean values of variables is the so-called tree order. This cone usually appears when comparing $k-1$ treatments with a control. In this case, it is common to know that, if G_1 is the control group, $\mu_{1j} \leqslant \mu_{ij}$ for $i = 2, \ldots, k$ and $j \in T$. This cone can be written as

$$C_{\mathrm{TO}} = \left\{x \in \mathbb{R}^{kp} : x_t \leqslant x_{t+np}, t \in T, n = 1, \ldots, k-1\right\}.$$

Because the projections are made in a kp dimensional space, we also need a metric for the projection. The appropriate metric is the one given by the kp square matrix

$$S_*^{-1} = \operatorname{diag}\left(\frac{S}{n_1}, \frac{S}{n_2}, \ldots, \frac{S}{n_k}\right)^{-1}, \tag{3}$$

where $n_1, \ldots, n_k$ are the sample sizes in each of the populations $i = 1, 2, \ldots, k$.

For details on cones of restrictions, their geometry, and projections on them, the reader may consult [12, 13] which are books containing the main results in restricted inference.

Now, a family of estimators indexed by $\gamma \in [0, 1]$, $\widehat{\mu}_i^{\gamma}$, $i = 1, \ldots k$, is defined using projections as follows.

Definition 1
Let $\widehat{\mu}^{\gamma} \in \mathbb{R}^{kp}$ for $\gamma \in [0, 1]$ be the limit value, when $m \to \infty$, of the following iterative procedure

$$\widehat{\mu}^{\gamma(m)} = P_{S_*^{-1}}\left(\widehat{\mu}^{\gamma(m-1)}\,|C\right) - \gamma P_{S_*^{-1}}\left(\widehat{\mu}^{\gamma(m-1)}\,|C^P\right), m = 1, 2, \ldots,$$

where $\widehat{\mu}^{\gamma(0)} = \left(\overline{x}_1', \ldots, \overline{x}_k'\right)' \in \mathbb{R}^{kp}$, $P_{S_*^{-1}}(Y\,|C)$ is the projection of Y onto cone C using the metric given by S_*^{-1}, and $C^P = \left\{y \in \mathbb{R}^{kp} : y'x \leqslant 0, \forall x \in C\right\}$.

Remark 2
From this $\widehat{\mu}_i^{\gamma} = \left((\widehat{\mu}^{\gamma})_{(i-1)p+1}, \ldots, (\widehat{\mu}^{\gamma})_{ip}\right)'$ for $i = 1, \ldots, k$. Notice that the iterative procedure is the same one appearing in [6]. In that paper, the procedure was used to estimate $\delta = \mu_1 - \mu_2$, whereas here, we estimate μ_1, μ_2,…,μ_p. The proof of the convergence of this iterative procedure is similar to the one in [6] and therefore is not given here.

When $\gamma = 0$, $\widehat{\mu}_i^0$ is a standard estimator in order-restricted inference. It is the projection of the sample means onto the cone of restrictions and also the restricted maximum likelihood estimator. When the value of γ increases, we obtain estimators that are more deeply inside the cone. We will see in what follows that rules based in estimators with $\gamma > 0$ perform better than the one with $\gamma = 0$. The main reason for this, which also justifies the choice of this family of estimators, may be the fact that restricted estimators belonging to the frontier of the cone of restrictions (i.e., the points of the cone for which at least one of the inequalities in the definition of the cone is verified as an equality) are not admissible under the squared error loss [14]. The question of the choice of γ will be further considered in the final Discussion section.

Now, the new rules, that we denote as $R_\mu(\gamma)$, can be defined as

$$R_\mu(\gamma)\text{: classify } z \text{ in } G_i \text{ iff } i = \arg_j \min\left\{(z - \widehat{\mu}_j^{\gamma})' S^{-1}(z - \widehat{\mu}_j^{\gamma}), j = 1, \ldots, k\right\}.$$

The following result states that these rules generalize the ones defined for two populations. Therefore, the isotonized rules share the good properties already proved for the $k = 2$ rules defined in the aforementioned paper [6] (i.e., the probability of correct classification is higher than that of Fisher's rule). The proof of the result is deferred to the appendix to improve the readability of the paper.

Theorem 3
If $k = 2$, rules $R_\delta(\gamma)$ and $R_\mu(\gamma)$ are equivalent for $\gamma \in [0, 1]$.

Therefore, we have managed to extend the rules defined for two populations to the situation of more than two populations. Notice that the case of three or more populations is more interesting from the additional information point of view, because more different ordering relations can be established in this case. For example, in the case of two populations, the simple order C_{SO} and tree order cones C_{TO} coincide, whereas this does not happen if there are more populations. All these different situations can be treated in a homogeneous way with the procedure we have just developed.

3. Simulation study

3.1. Study design

In order to check the good performance of the $R_\mu(\gamma)$ rules, we have designed and developed the following simulation study.

For simplicity, we concentrate on the case $k = 3$. We consider three 3-dimensional populations with distribution $N_3(\mu_i, \Sigma)$ under two different order restrictions. Namely, we consider the simple order and tree order restrictions that are the most common in practice. The simple order cone for this three populations and three variables case is again the cone C_{BC} appearing in our bladder cancer example, and the tree order cone is the C_{TO} cone defined earlier for $k = 3$ and $p = 3$.

We generate training samples of size $n_1 = n_2 = n_3 = 5$. It can be objected that these training sample sizes are too small for what is usual in practice. However, notice that also variability has to be taken into account. We have also run simulations with higher sample sizes ($n_1 = n_2 = n_3 = 50$) rescaling the covariance matrices accordingly, obtaining similar results.

Sixty four different simulations are conducted to show cases where the means and/or Σ are different for each of the two order cones. The parameter configurations are different for each cone and are generated by 16 mean vectors and 4 covariance matrices. The values chosen for the covariance matrix are intended to cover usual values in practice for the correlation coefficients. As for the means, they are chosen to be able to compare the performance of the different restricted rules for different parameter configurations. In all cases, the mean μ_1 of the first population has been set equal to $(0, 0, 0)$. Then, the 64 configurations are generated as $\mu_{2_i}\mu_{3_j}\Sigma_k$ with $i, j, k = 1, \ldots, 4$, where full details of the configurations are given in Table I. For each scenario, we generated 10, 000 training samples for which the rules are determined. For each of these training samples, 100 test observations coming from each of the three populations have been classified.

3.2. Simulation results

The results of the simulation study appear in Tables II and III. Each cell of those tables contains the total probability of correct classification $\widehat{p}$ for the corresponding rule. We obtained this probability from the estimated probabilities of correct classification in each population as $\widehat{p} = \frac{1}{3}\left(\widehat{p}_{11} + \widehat{p}_{22} + \widehat{p}_{33}\right)$ where

Table I. Mean vectors and covariance matrices.

Simple order cone means	
$\mu_{2_1} = \mu_1 + (0.1, 0.1, 0.1)$	$\mu_{3_1} = \mu_2 + (0.1, 0.1, 0.1)$
$\mu_{2_2} = \mu_1 + (0.1, 0.5, 0.5)$	$\mu_{3_2} = \mu_2 + (0.1, 0.5, 0.5)$
$\mu_{2_3} = \mu_1 + (0.5, 0.5, 1)$	$\mu_{3_3} = \mu_2 + (0.5, 0.5, 1)$
$\mu_{2_4} = \mu_1 + (1, 1, 1)$	$\mu_{3_4} = \mu_2 + (1, 1, 1)$
Tree order cone means	
$\mu_{2_1} = \mu_1 + (0.1, 0.1, 0.1)$	$\mu_{3_1} = \mu_1 + (0.2, 0.2, 0.2)$
$\mu_{2_2} = \mu_1 + (0.1, 0.5, 0.5)$	$\mu_{3_2} = \mu_1 + (0.2, 1, 1)$
$\mu_{2_3} = \mu_1 + (0.5, 0.5, 1)$	$\mu_{3_3} = \mu_1 + (1, 1, 2)$
$\mu_{2_4} = \mu_1 + (1, 1, 1)$	$\mu_{3_4} = \mu_1 + (2, 2, 2)$
Covariance matrices	
$\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\Sigma_2 = \begin{bmatrix} 1 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 1 \end{bmatrix}$
$\Sigma_3 = \begin{bmatrix} 1 & 0.3 & -0.3 \\ 0.3 & 1 & -0.3 \\ -0.3 & -0.3 & 1 \end{bmatrix}$	$\Sigma_4 = \begin{bmatrix} 1 & 0.7 & 0.7 \\ 0.7 & 1 & 0.7 \\ 0.7 & 0.7 & 1 \end{bmatrix}$

Table II. Correct classification probabilities under C_{SO}.

	CovarianceΣ_1			CovarianceΣ_2		
Means	Fisher	$R_\mu(0)$	$R_\mu(1)$	Fisher	$R_\mu(0)$	$R_\mu(1)$
$\mu_{2_1}\mu_{3_1}$	0.342	0.359	0.368	0.339	0.352	0.360
μ_{3_2}	0.388	0.410	0.419	0.376	0.394	0.401
μ_{3_3}	0.450	0.466	0.473	0.423	0.439	0.445
μ_{3_4}	0.507	0.516	0.519	0.464	0.474	0.477
$\mu_{2_2}\mu_{3_1}$	0.388	0.409	0.419	0.377	0.395	0.402
μ_{3_2}	0.443	0.465	0.473	0.425	0.444	0.451
μ_{3_3}	0.500	0.517	0.525	0.469	0.485	0.492
μ_{3_4}	0.554	0.567	0.571	0.507	0.518	0.522
$\mu_{2_3}\mu_{3_1}$	0.449	0.466	0.472	0.424	0.440	0.446
μ_{3_2}	0.501	0.519	0.526	0.469	0.486	0.492
μ_{3_3}	0.559	0.574	0.581	0.517	0.531	0.538
μ_{3_4}	0.612	0.622	0.627	0.554	0.563	0.568
$\mu_{2_4}\mu_{3_1}$	0.506	0.516	0.518	0.464	0.474	0.477
μ_{3_2}	0.554	0.566	0.571	0.507	0.518	0.522
μ_{3_3}	0.611	0.622	0.626	0.554	0.563	0.568
μ_{3_4}	0.666	0.671	0.675	0.592	0.598	0.602
	Covariance Σ_3			Covariance Σ_4		
Means	Fisher	$R_\mu(0)$	$R_\mu(1)$	Fisher	$R_\mu(0)$	$R_\mu(1)$
$\mu_{2_1}\mu_{3_1}$	0.345	0.364	0.375	0.337	0.346	0.352
μ_{3_2}	0.406	0.428	0.438	0.383	0.396	0.400
μ_{3_3}	0.480	0.494	0.500	0.422	0.434	0.440
μ_{3_4}	0.534	0.544	0.547	0.432	0.442	0.445
$\mu_{2_2}\mu_{3_1}$	0.406	0.428	0.437	0.383	0.396	0.400
μ_{3_2}	0.475	0.496	0.505	0.445	0.457	0.460
μ_{3_3}	0.543	0.560	0.566	0.478	0.490	0.493
μ_{3_4}	0.594	0.607	0.611	0.484	0.492	0.494
$\mu_{2_3}\mu_{3_1}$	0.480	0.494	0.500	0.424	0.436	0.440
μ_{3_2}	0.544	0.560	0.567	0.478	0.489	0.493
μ_{3_3}	0.611	0.624	0.630	0.521	0.531	0.536
μ_{3_4}	0.663	0.672	0.676	0.525	0.533	0.537
$\mu_{2_4}\mu_{3_1}$	0.534	0.544	0.547	0.432	0.441	0.444
μ_{3_2}	0.595	0.608	0.613	0.483	0.492	0.495
μ_{3_3}	0.664	0.672	0.676	0.524	0.532	0.536
μ_{3_4}	0.716	0.720	0.723	0.535	0.541	0.545

$\widehat{p}_{ii}$ is the estimated probability that an observation coming from population G_i is correctly classified in that population for $i = 1, 2, 3$. We performed all these simulations by using R.

It is worth commenting the results appearing in Tables II and III. The first obvious conclusion is that the new rules perform better than the one not taking into account the additional information for all the scenarios and order restrictions considered. Another interesting conclusion is the convenience of values of γ strictly bigger than 0, as in all situations but one, the rule with $\gamma = 1$ outperforms the one with $\gamma = 0$ based on the restricted maximum likelihood estimator.

The differences between the correct classification probabilities of the rules are not too big. The main reason is that the isotonized rules are equal to Fisher's rule when the training sample verifies the restrictions imposed by the additional information. The isotonized rules show their power when the training set does not verify all the restrictions which is obviously when these rules should be used in statistical practice. In fact, the results improve more when the number of restrictions that are not verified is higher. For example, consider the scenario given by C_{SO}, covariance Σ_1, and means μ_{2_3} and μ_{3_3}. Table IV details the results obtained for this case depending on the number of restrictions fulfilled. The improvements in the correct classification probabilities with respect to the unrestricted Fisher rule are bigger when there are several restrictions not verified by the training set. The scenario appearing in our case study is one of these.

Table III. Correct classification probabilities under C_{TO}.

	Covariance Σ_1			Covariance Σ_2		
Means	Fisher	$R_\mu(0)$	$R_\mu(1)$	Fisher	$R_\mu(0)$	$R_\mu(1)$
$\mu_{2_1}\mu_{3_1}$	0.342	0.356	0.361	0.339	0.349	0.354
μ_{3_2}	0.456	0.466	0.466	0.437	0.448	0.448
μ_{3_3}	0.556	0.563	0.564	0.525	0.531	0.531
μ_{3_4}	0.616	0.623	0.626	0.577	0.582	0.582
$\mu_{2_2}\mu_{3_1}$	0.370	0.388	0.394	0.366	0.380	0.385
μ_{3_2}	0.442	0.457	0.460	0.425	0.437	0.439
μ_{3_3}	0.562	0.573	0.577	0.527	0.536	0.538
μ_{3_4}	0.638	0.650	0.654	0.591	0.600	0.601
$\mu_{2_3}\mu_{3_1}$	0.425	0.440	0.444	0.406	0.420	0.423
μ_{3_2}	0.471	0.483	0.485	0.455	0.467	0.471
μ_{3_3}	0.560	0.567	0.570	0.517	0.525	0.528
μ_{3_4}	0.664	0.672	0.675	0.605	0.612	0.615
$\mu_{2_4}\mu_{3_1}$	0.482	0.493	0.495	0.442	0.452	0.452
μ_{3_2}	0.518	0.525	0.525	0.492	0.500	0.501
μ_{3_3}	0.583	0.586	0.589	0.551	0.556	0.559
μ_{3_4}	0.665	0.668	0.670	0.591	0.594	0.597
	Covariance Σ_3			Covariance Σ_4		
Means	Fisher	$R_\mu(0)$	$R_\mu(1)$	Fisher	$R_\mu(0)$	$R_\mu(1)$
$\mu_{2_1}\mu_{3_1}$	0.345	0.360	0.366	0.337	0.344	0.349
μ_{3_2}	0.485	0.496	0.497	0.463	0.471	0.473
μ_{3_3}	0.589	0.597	0.599	0.530	0.534	0.536
μ_{3_4}	0.637	0.645	0.649	0.538	0.543	0.542
$\mu_{2_2}\mu_{3_1}$	0.383	0.402	0.409	0.376	0.386	0.391
μ_{3_2}	0.474	0.488	0.490	0.445	0.453	0.453
μ_{3_3}	0.609	0.621	0.625	0.544	0.551	0.552
μ_{3_4}	0.674	0.687	0.692	0.564	0.571	0.571
$\mu_{2_3}\mu_{3_1}$	0.451	0.466	0.470	0.411	0.421	0.426
μ_{3_2}	0.510	0.519	0.521	0.494	0.504	0.506
μ_{3_3}	0.611	0.618	0.620	0.521	0.526	0.529
μ_{3_4}	0.708	0.715	0.718	0.577	0.583	0.586
$\mu_{2_4}\mu_{3_1}$	0.511	0.523	0.527	0.414	0.422	0.423
μ_{3_2}	0.554	0.559	0.558	0.513	0.522	0.524
μ_{3_3}	0.617	0.622	0.623	0.570	0.574	0.578
μ_{3_4}	0.714	0.717	0.718	0.537	0.539	0.542

Table IV. Correct classification probabilities under C_{SO} depending on the number of restrictions verified by the training set for covariance Σ_1 and means μ_{2_3} and μ_{3_3}.

Restrictions verified	Fisher	$R_\mu(0)$	R_μ (1)
6	0.583	0.583	0.583
5	0.560	0.574	0.581
4	0.535	0.565	0.580
3	0.495	0.543	0.576
2	0.431	0.512	0.573
1	0.354	0.501	0.606

Next, we compare the results obtained using different cones. Notice that the pairs of means $\mu_{2_i}\mu_{3_i}$ for $i = 1, \ldots, 4$ are the same for the two cones we are considering (C_{SO} and C_{TO}). Figure 1 shows the results obtained for the first two pairs of means ($i = 1, 2$) for the unrestricted rule and for the new rules when γ equals 0 or 1. It is easy to check that, independently from the covariance matrix, better classification results are obtained if the information given by C_{SO} is included instead of the one given by C_{TO}. Similar results are obtained for the other two pairs of means ($i = 3, 4$). Therefore, as $C_{\text{SO}} \subset C_{\text{TO}}$, we conclude that better classification results are obtained if the additional information available is more precise.

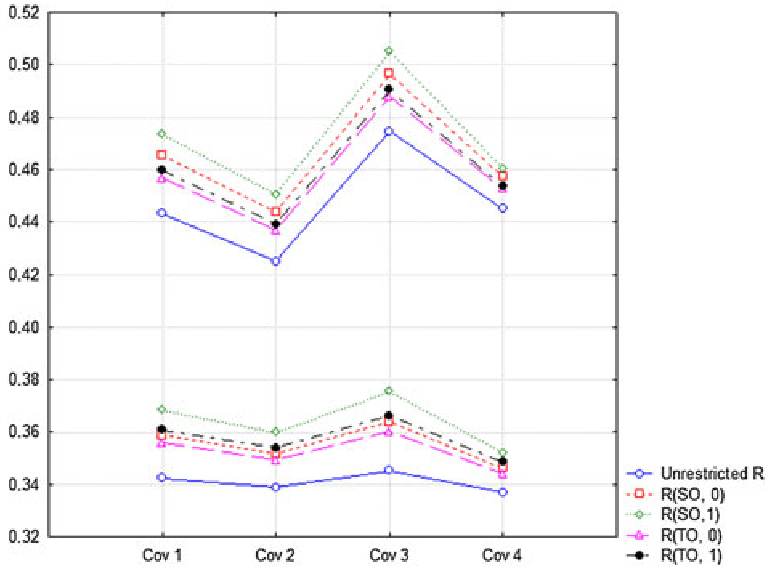

Figure 1. Correct classification probabilities for means $\mu_{2_1}\mu_{3_1}$ (lower set of lines) and μ_{2_2} μ_{3_2} (upper set of lines).

4. Case study

We will consider two data sets in this work. Both data sets are obtained using the xMAP® technology from Luminex for measuring the levels of the proteins. This technology is widely used in many medical and biological research areas such as gene expression or cancer markers. An example of its recent use in bladder cancer research is in [15].

The first data set we received, D_1, contained information on 141 patients and 11 proteins together with the real stage of the illness the patients belonged to. The proteins were expected to be useful for correctly classifying the patients in one of the five levels we have defined. This is the initial data set and it is the one we will use to build the rules. In the usual statistical terminology, this is the training set.

The second data set, D_2, is the test set, that is, the one we will use for checking the goodness of the classification rule built with the first one. This data set was received in a later stage during the research and contains information on a different set of 149 patients for whom we have measures on the same 11 proteins and we know their real illness stage.

As stated in the Introduction, using the judgment provided by our pharmaceutical and industrial partners, we consider three proteins, P_1, P_2, and P_3, for illustrating the methodology. However, we also analyzed the data using all 11 available proteins. As expected by our partners, the 11 proteins when used together for classifying subjects were not as informative as the three specific proteins they suggested us to focus on, and the estimated probability of misclassification increased significantly when all proteins were considered.

The mean values in each of the groups and the pooled covariance matrix for P_1, P_2 and P_3, obtained from data set D_1, appear in Table V.

Table V. Mean for each group and pooled covariance matrix from D_1.

	Means			
	N	$\log(P_1)$	$\log(P_2)$	$\log(P_3)$
G_1	41	2.935	3.879	1.416
G_2	68	2.670	3.944	1.029
G_3	32	3.245	4.348	1.578
		$S = \begin{pmatrix} 1.018 & 0.469 & 0.410 \\ 0.469 & 0.985 & 0.284 \\ 0.410 & 0.284 & 0.575 \end{pmatrix}$		

Table VI. Classification of test data using Fisher's rule.

	Predicted		
Observed	G_1	G_2	G_3
G_1	12	66	7
G_2	5	18	8
G_3	2	11	20
Correct		33.56%	

Table VII. Classification results using several discrimination procedures.

	Correct classification	
Procedure	Training set (%)	Test set (%)
Fisher	46.10	33.56
SVM type 1	48.23	20.80
SVM type 2	45.39	26.85
1-NN	–	28.85
2-NN	–	29.53
NBC	48.22	21.47
ANN	51.77	32.88
Ordinal	43.97	35.57

The results of the application of the Fisher rule to our test data set, D_2, appear in Table VI. We can check that the overall percentage of correct classification is 33.56%, which is quite a poor result. The overall percentage of correct classification in the training sample was 46.10%, showing that the bias in correct classification appearing in this rule is quite high.

In Table VII, we summarize the results obtained considering other more 'state-of-the-art' methods of building classification rules such as Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), Naive Bayes Classifiers (NBC), or Artificial Neural Networks (ANN). Because this situation could also be viewed as an ordinal classification problem, in order to compare this ordinal focus with the one proposed here, we also include ordinal classification in the comparison list. There are many recent references where these techniques are used in cancer research [16–19].

For a type 1 SVM with capacity 10 and a radial basis function as kernel ($\gamma = 0.333$), we obtained a classification accuracy of 48.23% for the training set and 20.80% for the test set. When we considered a type 2 SVM with $nu = 0.1$ and a radial basis kernel with the same value for gamma, the accuracy improved for the test set as we obtain 45.39% for the training set and 26.85% for the test set. Other methods such as K-NN, NBC, ANN, or Ordinal classification did not reach 36% of observations correctly classified in the test data set. Therefore, we can conclude that the usual methods do not perform well in this case. The main reason is that, as it can be observed in Figure 2, the training set does not verify the expected ordering among the levels of the illness and those of the proteins while the test data set does follow that ordering. Consequently, rules built using D_1 do not perform as expected.

The usual solution would be to discard the training data set. That would be time wasting and expensive. Now, we show the results obtained applying the isotonized rules. In this particular case, there are several restrictions that are not fulfilled (check Table V) so that an improvement when the new rules are used is expected. For this case, the new rules defined in the previous section are

$$R_\mu(\gamma)\text{: classify } z \text{ in } G_i \text{ iff } i = \arg_j \min \left\{(z - \widehat{\mu}_j^\gamma)' S^{-1} (z - \widehat{\mu}_j^\gamma),\ j = 1, 2, 3\right\},$$

where the $\widehat{\mu}_j^\gamma$ values for $j = 1, 2, 3$ and $\gamma = 0, 1$ appear in Table VIII.

The classification matrices obtained when these rules are used for classifying the observations in the test set D_2 appear in Table IX.

In that Table, we can check that the new rules yield very good results for our case study. In fact, using the new rules, we can see that the global probabilities of correct classification in the test set are 53.02%

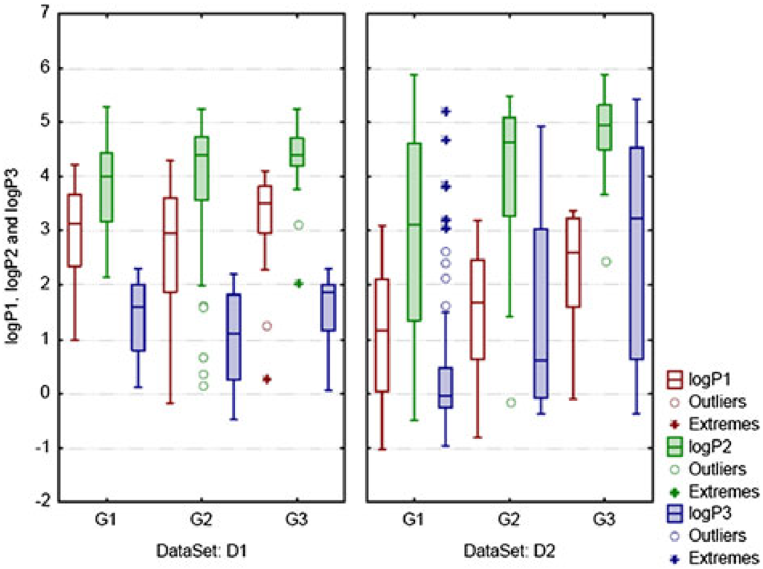

Figure 2. Box plot of variables $\log(P_1)$, $\log(P_2)$, and $\log(P_3)$ for each data set and illness category.

Table VIII. Restricted estimators of the means for $\gamma = 0, 1$.

	$\widehat{\mu}_j^0$	$\widehat{\mu}_j^1$
G_1	(2.762, 3.760, 1.175)	(2.589, 3.640, 0.933)
G_2	(2.774, 4.016, 1.175)	(2.878, 4.088, 1.321)
G_3	(3.245, 4.348, 1.578)	(3.245, 4.348, 1.578)

Table IX. Classification of test data using rules $R_\mu(\gamma)$.

	Predicted					
	$\gamma = 0$			$\gamma = 1$		
Observed	G_1	G_2	G_3	G_1	G_2	G_3
G_1	42	34	9	67	11	7
G_2	5	16	10	15	8	8
G_3	1	11	21	5	7	21
Correct		53.02%			64.43%	

for $\gamma = 0$ and 64.43% for $\gamma = 1$. It is certainly an improvement over the values under 36% obtained with the non-restricted classification procedures in Table VII.

5. Discussion

The case study and the simulations results in Sections 3 and 4 show that the new restricted rules defined in this paper outperform the unrestricted ones and that, as in Tables IV and IX, the improvement can be quite significative when several restrictions are not fulfilled by the training sample.

These results are also very interesting from the industrial and pharmaceutical points of view because they confirm the results from previous works that suggested the orderings used to define the rules. Moreover, these new rules allow research to go on without dropping the initial sample, which saves time and money.

The fact that there is no loss in classification when the training set verifies the restrictions allows us to recommend the use of the restricted rules proposed in this paper in the general practice. In order to make them easy to use, an R library for restricted classification is being compiled. For the moment, R programs for performing classification under additional information can be obtained from the authors on request.

Table X. Error rates for the discrimination procedures.

	Fisher	$\gamma = 0$	$\gamma = 1$
Actual error estimate (%)	66.44	46.98	35.57
Apparent error rate (%)	53.90	58.15	67.37

We have also shown that more precise additional information translates into more powerful rules. Therefore, we recommend researchers to incorporate as much information as the application at hand provides. This kind of information is often available in applications as, for example, in medical diagnosis problems, where the predictors are usually isotonically related to the response.

Another interesting point is the estimation of the actual probability of misclassification. In our case study, a test sample is available to evaluate the misclassification probability so that we have a good estimator of the actual error rate. However, in practice, it is quite usual that there is no second sample to do this. The usual procedure in practice is to evaluate the rule on the same sample that has been used to build it. In this way, the so-called apparent error rate is obtained. This procedure is obviously biased downward, and correction procedures have to be used in order to obtain a good estimator of the actual error rate. There is much literature on this subject. For example, the research of Jiang and Simon [20] contains a good study of the most usual estimators of the actual error rate. However, it is clear from Table X that, for our case study, the bias of the apparent error rate for the restricted rules does not behave like that of Fisher's rule. Therefore, new procedures specifically designed to correct the apparent error rate in restricted procedures have to be developed. These procedures will also help the user make the optimal choice of parameter γ. Although we have seen in this work that $\gamma = 1$ is a good choice, the value of γ for which the estimator of the actual error rate is lowest will obviously be the natural one. This is part of our present research and will hopefully appear in future works.

Finally, we would like to mention that, as pointed by one of the reviewers, the methodology developed in this work opens the possibility of modifying other classification procedures to make them able to cope with the sort of additional information considered here.

Appendix: Proofs

In order to prove Theorem 3, we will first prove the following lemma.

We denote as K and K_* the cones $K = \left\{ z \in \mathbb{R}^p : a'_j z \geqslant 0, j = 1, \ldots, r \right\}$ and $K_* = \left\{ z \in \mathbb{R}^{2p} : b'_j z \geqslant 0, j = 1, \ldots, r \text{ with } b_j = \begin{pmatrix} a_j \\ -a_j \end{pmatrix} \right\}$. Let S be a p dimensional positive definite matrix and denote as $S_*^{-1} = \begin{pmatrix} d_1 S^{-1} & 0 \\ 0 & d_2 S^{-1} \end{pmatrix}$ where $d_1, d_2 > 0$.

Lemma 4
If $x, y \in \Re^p$ and $v = P_{S^{-1}}(x - y|K)$, then

$$P_{S_*^{-1}}\left(\begin{pmatrix} x \\ y \end{pmatrix} |K_* \right) = \begin{pmatrix} d_1 x + d_2(y + v) \\ d_1(x - v) + d_2 y \end{pmatrix}.$$

Proof
First, let us check that we can assume $d_1 + d_2 = 1$ without loss of generality. If that condition is not fulfilled, as in (3), we can define $d_1^* = d_1/(d_1 + d_2)$, $d_2^* = d_2/(d_1 + d_2)$ and $S_{**}^{-1} = \begin{pmatrix} d_1^* S^{-1} & 0 \\ 0 & d_2^* S^{-1} \end{pmatrix}$. The new values d_1^*, d_2^* verify the condition and, as these changes of scale in the projection matrix do not change the projection, we have that $P_{S_*^{-1}}\left(\begin{pmatrix} x \\ y \end{pmatrix} |K_* \right) = P_{S_{**}^{-1}}\left(\begin{pmatrix} x \\ y \end{pmatrix} |K_* \right)$. Therefore, in the rest of the proof, we assume $d_1 + d_2 = 1$.

Now, denote $w_1 = d_1x + d_2(y + v)$ and $w_2 = d_1(x - v) + d_2y$. To prove the result, we have to check (cf. [12] Theorem 1.3.2) that

$$\left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}\right)' S_*^{-1} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = 0$$
$$\left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}\right)' S_*^{-1} h \leqslant 0, \forall h \in K_*.$$

As $v = P_{S^{-1}}(x - y\,|K)$, we know that

$$(x - y - v)'S^{-1}v = 0$$
$$(x - y - v)'S^{-1}f \leqslant 0, \forall f \in K.$$

Now,

$$\begin{aligned}\left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}\right)' S_*^{-1} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} &= d_1d_2(x - y - v)'S^{-1}\left(d_1x + d_2(y + v)\right) \\ &\quad - d_1d_2(x - y - v)'S^{-1}\left(d_1(x - v) + d_2y\right) \\ &= d_1d_2(x - y - v)'S^{-1}v = 0.\end{aligned}$$

Let $h = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \in K_*$. Notice that $h \in K_*$ if and only if $h_1 - h_2 \in K$. Then,

$$\left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}\right)' S_*^{-1} h = d_1d_2(x - y - v)'S^{-1}(h_1 - h_2) \leqslant 0$$

and the proof is done. □

Corollary 5
Under the conditions of Lemma 4 and assuming $d_1 + d_2 = 1$, $v = w_1 - w_2$ and $d_1w_1 + d_2w_2 = d_1x + d_2y$.

Now we can prove the pending result.

Proof of Theorem 3
After some easy calculations, we can write the rule $R_\mu(\gamma)$ for two populations as

$$\text{classify} z \text{ in } G_1 \text{ iff } \left(z - \frac{\widehat{\mu}_1^\gamma + \widehat{\mu}_2^\gamma}{2}\right) S^{-1}\left(\widehat{\mu}_1^\gamma - \widehat{\mu}_2^\gamma\right) \geqslant 0$$

so that, taking into account (2), to prove the result, we have to check that

$$\delta_\gamma^* = \widehat{\mu}_1^\gamma - \widehat{\mu}_2^\gamma$$
$$c_1\overline{x}_1 + c_2\overline{x}_2 - c\delta_\gamma^* = \frac{\widehat{\mu}_1^\gamma + \widehat{\mu}_2^\gamma}{2}.$$

These estimators are obtained as the limit of the iterative procedure appearing in Definition 1 so that

$$\begin{pmatrix} \widehat{\mu}_1^{\gamma(i)} \\ \widehat{\mu}_2^{\gamma(i)} \end{pmatrix} = (1 + \gamma)\begin{pmatrix} \omega_1^{\gamma(i)} \\ \omega_2^{\gamma(i)} \end{pmatrix} - \gamma\begin{pmatrix} \widehat{\mu}_1^{\gamma(i-1)} \\ \widehat{\mu}_2^{\gamma(i-1)} \end{pmatrix}, \text{ where}$$

$$\begin{pmatrix} \omega_1^{\gamma(i)} \\ \omega_2^{\gamma(i)} \end{pmatrix} = P_{S_*^{-1}}\left(\begin{pmatrix} \widehat{\mu}_1^{\gamma(i-1)} \\ \widehat{\mu}_2^{\gamma(i-1)} \end{pmatrix} |K\right) = \begin{pmatrix} c_1\widehat{\mu}_1^{\gamma(i-1)} + c_2(\widehat{\mu}_2^{\gamma(i-1)} + \nu^{\gamma(i)}) \\ c_1(\widehat{\mu}_1^{\gamma(i-1)} - \nu^{\gamma(i)}) + c_2\widehat{\mu}_2^{\gamma(i-1)} \end{pmatrix},$$
$$\nu^{\gamma(i)} = P_{S^{-1}}\left((\widehat{\mu}_1^{\gamma(i-1)} - \widehat{\mu}_2^{\gamma(i-1)})\,|K\right) \text{ and } \begin{pmatrix} \widehat{\mu}_1^{\gamma(0)} \\ \widehat{\mu}_2^{\gamma(0)} \end{pmatrix} = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \end{pmatrix}.$$

It is easy to prove that, for any $i \geqslant 1$,

$$\begin{aligned} c_1\widehat{\mu}_1^{\gamma(i)} + c_2\widehat{\mu}_2^{\gamma(i)} &= (1+\gamma)(c_1\omega_1^{\gamma(i)} + c_2\omega_2^{\gamma(i)}) - \gamma(c_1\widehat{\mu}_1^{\gamma(i-1)} + c_2\widehat{\mu}_2^{\gamma(i-1)}) \\ &= (1+\gamma)(c_1\widehat{\mu}_1^{\gamma(i-1)} + c_2\widehat{\mu}_2^{\gamma(i-1)}) - \gamma(c_1\widehat{\mu}_1^{\gamma(i-1)} + c_2\widehat{\mu}_2^{\gamma(i-1)}) \\ &= c_1\widehat{\mu}_1^{\gamma(i-1)} + c_2\widehat{\mu}_2^{\gamma(i-1)} = c_1\widehat{\mu}_1^{\gamma(0)} + c_2\widehat{\mu}_2^{\gamma(0)} \\ &= c_1\overline{x}_1 + c_2\overline{x}_2. \end{aligned} \tag{4}$$

Now we use an induction argument to obtain the proof. First, we assume $m = 1$. Then from Corollary 5,

$$\begin{aligned} \widehat{\mu}_1^{\gamma(1)} - \widehat{\mu}_2^{\gamma(1)} &= (1+\gamma)(\omega_1^{\gamma(1)} - \omega_2^{\gamma(1)}) - \gamma\,(\overline{x}_1 - \overline{x}_2) \\ &= (1+\gamma)\nu^{\gamma(1)} - \gamma\,(\overline{x}_1 - \overline{x}_2) = \delta_\gamma^{*(1)}. \end{aligned}$$

$$\begin{aligned} \frac{\widehat{\mu}_1^{\gamma(1)} + \widehat{\mu}_2^{\gamma(1)}}{2} &= (1+\gamma)\frac{\omega_1^{\gamma(1)} + \omega_2^{\gamma(1)}}{2} - \gamma\frac{\overline{x}_1 + \overline{x}_2}{2} \\ &= (1+\gamma)\left(c_1\overline{x}_1 + c_2\overline{x}_2 - c\nu^{\gamma(1)}\right) - \gamma\,(c_1\overline{x}_1 + c_2\overline{x}_2 - c(\overline{x}_1 - \overline{x}_2)) \\ &= c_1\overline{x}_1 + c_2\overline{x}_2 - c\left((1+\gamma)\nu^{\gamma(1)} - \gamma(\overline{x}_1 - \overline{x}_2)\right) \\ &= c_1\overline{x}_1 + c_2\overline{x}_2 - c\delta_\gamma^{*(1)}. \end{aligned}$$

Now we assume that the equalities hold for $m = i - 1$ and we prove them for $m = i$. Taking into account Corollary 5 and (4), we have

$$\begin{aligned} \widehat{\mu}_1^{\gamma(i)} - \widehat{\mu}_2^{\gamma(i)} &= (1+\gamma)(\omega_1^{\gamma(i)} - \omega_2^{\gamma(i)}) - \gamma(\widehat{\mu}_1^{\gamma(i-1)} - \widehat{\mu}_2^{\gamma(i-1)}) \\ &= (1+\gamma)\nu^{\gamma(i)} - \gamma\delta_\gamma^{*(i-1)} = \delta_\gamma^{*(i)}. \end{aligned}$$

$$\begin{aligned} \frac{\widehat{\mu}_1^{\gamma(i)} + \widehat{\mu}_2^{\gamma(i)}}{2} &= (1+\gamma)\frac{\omega_1^{\gamma(i)} + \omega_2^{\gamma(i)}}{2} - \gamma\frac{\widehat{\mu}_1^{\gamma(i-1)} + \widehat{\mu}_2^{\gamma(i-1)}}{2} \\ &= (1+\gamma)(c_1\widehat{\mu}_1^{\gamma(i-1)} + c_2\widehat{\mu}_2^{\gamma(i-1)} - c\nu^{\gamma(i)}) - \gamma(c_1\overline{x}_1 + c_2\overline{x}_2 - c\delta_\gamma^{*(i-1)}) \\ &= (1+\gamma)\left(c_1\overline{x}_1 + c_2\overline{x}_2 - c\nu^{\gamma(i)}\right) - \gamma(c_1\overline{x}_1 + c_2\overline{x}_2 - c\delta_\gamma^{*(i-1)}) \\ &= c_1\overline{x}_1 + c_2\overline{x}_2 - c\delta_\gamma^{*(i)}, \end{aligned}$$

and the proof is finished. □

Acknowledgements

The authors would like to thank the editor and two anonymous reviewers for their detailed reading that resulted in this improved version of the paper. This research is partially supported by Spanish DGES grant MTM2009-11161.

References

1. UICC. *TNM Classification of Malignant Tumours, 7th edition*. Wiley-Blackwell: New Jersey, 2009.
2. Oosterlinck W, Bernard L, Jakse G, Malmström P, Stöckle M, Sternberg C. Guidelines on bladder cancer. *European Urology* 2002; **41**:105–112.
3. Tosoni I, Wagner U, Sauter G, Egloff M, Knönagel H, Alund G, Bannwart F, Mihatsch MJ, Gasser TC, Maurer R. Clinical significance of interobserver differences in the staging and grading of superficial bladder cancer. *British Journal of Urology International* 2000; **85**:48–53.
4. Lokershwar VB, Young MJ, Gourdazi G, Iida N, Yudin AI, Cherr GN, Selzer MG. Identification of bladder tumor-derived hyaluronidase: its similarity to HYAL1. *Cancer Research* 1999; **59**:4464–4470.
5. Long T, Gupta RD. Alternative linear classification rules under order restrictions. *Communications in Statistics - Theory and Methods* 1998; **27**:559–575.
6. Fernández MA, Rueda C, Salvador B. Incorporating additional information to normal linear discriminant rules. *Journal of the American Statistical Association* 2006; **101**:569–577.
7. Peddada SD, Lobenhofer E, Li LP, Afshari CA, Weinberg CR, Umbach D. Gene selection and clustering for time course and dose-response microarray experiments using order-restricted inference. *Bioinformatics* 2003; **19**:834–841.

8. Peddada SD, Dinse G, Haseman J. A survival-adjusted quantal response test for comparing tumor incidence rates. *Journal of the Royal Statistical Society, Series C* 2005; **54**:51–61.
9. Cai B, Dunson DB. Bayesian multivariate isotonic regression splines: applications to carcinogenicity studies. *Journal of the American Statistical Association* 2007; **102**:1158–1171.
10. Simmons S, Peddada SD. Order-restricted inference for ordered gene expression (ORIOGEN) data under heteroscedastic variances. *Bioinformation* 2007; **1**:414–419.
11. Ghosh D, Banerjee M, Biswas P. Inference for constrained estimation of tumor size distributions. *Biometrics* 2008; **64**:1009–1017.
12. Robertson T, Wright FT, Dykstra RL. *Order restricted statistical inference*. John Wiley & Sons: New York, 1988.
13. Silvapulle MJ, Sen PK. *Constrained Statistical Inference*. John Wiley & Sons: New Jersey, 2005.
14. Van Eeden C. *Restricted Parameter Space Estimation Problems*, Lecture notes in Statistics 188. Springer: New York, 2006.
15. Wang M, Wang M, Zhang W, Yuan L, Fu G, Wei Q, Zhang Z. Common genetic variants on 8q24 contribute to susceptibility to bladder cancer in a Chinese population. *Carcinogenesis* 2009; **30**:991–996.
16. Rapaport F, Barillot E, Vert J-P. Classification of arrayCGH data using fused SVM. *Bioinformatics* 2008; **24**:375–382.
17. Dulewicz A, Jaszezak P, Pietka BD. Pattern Recognition Techniques in Recognition of Neoplastic Changes in Images of Cell Nuclei. In *IFMBE proceedings. Vol.25/5. Information and Communication in Medicine, Telemedicine and e-health*, Dössel O, Schlegel WC (eds). Springer: New York, 2010; 105–108.
18. Bengtsson S, Krogh M, Al-Khalili C, Uhlen M, Schedvins K, Silfverswärd C, Linder S, Auer G, Alaiya A, James P. Large-scale proteomics analysis of human ovarian cancer for biomarkers. *Journal of Proteome Research* 2007; **6**:1440–1450.
19. Margulis V, Lotan Y, Montorsi F, Shariat SF. Predicting survival after radical cystectomy for bladder cancer. *British Journal of Urology International* 2008; **102**:15–22.
20. Jiang W, Simon R. A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. *Statistics in Medicine* 2007; **26**:5320–5334.

David Conde, Bonifacio Salvador, Cristina Rueda and Miguel A. Fernández*

Performance and estimation of the true error rate of classification rules built with additional information. An application to a cancer trial

Abstract: Classification rules that incorporate additional information usually present in discrimination problems are receiving certain attention during the last years as they perform better than the usual rules. Fernández, M. A., C. Rueda and B. Salvador (2006): "Incorporating additional information to normal linear discriminant rules," J. Am. Stat. Assoc., 101, 569–577, proved that these rules have lower total misclassification probability than the usual Fisher's rule. In this paper we consider two issues; on the one hand, we compare these rules with those based on shrinkage estimators of the mean proposed by Tong, T., L. Chen and H. Zhao (2012): "Improved mean estimation and its application to diagonal discriminant analysis," Bioinformatics, 28(4): 531–537. with regard to four criteria: total misclassification probability, area under ROC curve, well-calibratedness and refinement; on the other hand, we consider the estimation of the true error rate, which is a very interesting parameter in applications. We prove results on the apparent error rate of the rules that expose the need of new estimators of their true error rate. We propose four such new estimators. Two of them are defined incorporating the additional information into the leave-one-out-bootstrap. The other two are the corresponding cross-validation after bootstrap versions. We compare these estimators with the usual ones in a simulation study and in a cancer trial application, showing the good behavior of the rules that incorporate additional information and of the new leave-one-out bootstrap estimators of their true error rate.

Keywords: area under ROC curve; bootstrap; cancer diagnostic test research; discriminant analysis; order restrictions; true error rate.

***Corresponding author: Miguel A. Fernández,** Departamento de Estadística e I.O.,Universidad de Valladolid, 47011 Valladolid, Spain, e-mail: miguelaf@eio.uva.es
David Conde, Bonifacio Salvador and Cristina Rueda: Departamento de Estadística e I.O.,Universidad de Valladolid, 47011 Valladolid, Spain

1 Introduction

Consider the classical discrimination problem with two populations Π_1 and Π_2. Denote the training sample from which the rule is built as $M_n=\{(X_i, Y_i), i=1, \dots, n\}$, where X is the p-dimensional vector of classifiers and Y=1, 2 is the binary variable identifying the population. Denote also as P_{XY} the joint distribution of the vector (X, Y). With this scheme, a classification rule is an application $R_n:\{\mathbb{R}^p\times\{1,2\}\}^n\times\mathbb{R}^p\to\{1,2\}$ that classifies a new observation $u\in\mathbb{R}^p$ into one of the two available populations: $R_n(M_n, u)\in\{1, 2\}$.

In applications it is usual that some additional information is available. Recent papers considering additional information issues are, for example, Lin et al. (2007), Simmons and Peddada (2007), Beran and Dümbgen (2010) and Oh and Shin (2011). It is frequent that this information tells us that the observations from one of the populations, for example Π_1, take higher (or lower) values than those coming from the other, i.e., Π_2. The incorporation of this kind of information into the classification rule has been shown to improve the performance of the rule. To our best knowledge, the first paper in this line was Long and Gupta (1998). More recently, Fernández et al. (2006) generalized and improved the results in that paper and proposed rules that take into account this additional information and have lower total misclassification probabilities (TMP) than the classical rules that do not consider this information. A good example of this situation appears in Section 5 where bladder cancer patients are known to usually take higher values in some variables (and lower in others) than healthy people. This information is then used to build a classification rule that outperforms Fisher's rule and that, as studied in Salvador et al. (2008), has good robustness properties with respect to its theoretical assumptions. From now on, we will refer to these rules as restricted rules, as they come from order restriction information on the populations.

There are more rules proposed in the literature that may be compared with those in Fernández et al. (2006). In this paper we compare the latter with those based on shrinkage estimators of the mean proposed by Tong et al. (2012). The comparison is motivated by the fact that the restricted rules can also be viewed as "shrinkage" rules since they are based on projections, and projection is a contractive operator.

In this paper we will not only compare the behavior of the discrimination procedures through the TMP (which was the only criterion considered in Fernández et al.). It is well known that there are some other ways of comparing the behavior of the procedures. For example, in medical classification problems, probabilistic classifiers are more frequently used than classification rules since they allow making complex clinical decisions. For probabilistic classifiers, it is frequent to use other measures such as the area under the ROC (receiving operating characteristic) curve, usually known as AUC (see, for example, Faraggi and Reiser, 2002; Pepe et al., 2004, and references therein), or well-calibratedness and refinement, introduced and developed by Kim and Simon (2011). All these measures will be considered in the comparison of the procedures.

Another key issue for a discrimination procedure is the evaluation of its performance for a given training sample. When TMP is considered, this is usually done by estimating the true error rate E_n of the rule R_n, which is the misclassification probability of the rule conditioned to the available training sample. Namely, $E_n=Error(R_n)=P_{XY}(R_n(M_n, X)\neq Y/M_n)$. The evaluation of the behavior of the rule using TMP, which is the expected, or unconditional, true error rate $E(E_n)$, allows the study of global properties of the rule but not the evaluation of E_n for a given sample M_n.

The best way of estimating E_n for a classification rule is to use an independent sample, usually called test sample. However, in practice it is common that the sample size is not large enough to split it into a training and a test sample as that would decrease the efficiency of the rule. For this reason, the estimation of E_n for the usual rules such as Fisher's linear rule, the quadratic discriminant rule, the nearest neighbors rules or random forest rules, is a widely studied topic in the literature. Parametric and non-parametric estimators of E_n have been proposed, and non-parametric estimators based on resampling have shown a good performance for the above mentioned rules. Schiavo and Hand (2000) summarizes the work made on this topic until that date. More recent references are, for example, Steele and Patterson (2000), Wehberg and Schumacher (2004), Fu et al. (2005), Molinaro et al. (2005), Kim and Cha (2006) or Kim (2009). In this paper, we check that the usual estimators are not expected to work well for the restricted rules and tackle the problem of proper estimation of E_n for the restricted rules defining new estimators that will also be useful when other performance measures, such as AUC, are considered.

The layout of the paper is as follows. In Section 2 we start reviewing the restricted rules defined in Fernández et al. and those in Tong et al. (2012). In Section 3 we consider different performance measures and compare the behavior of the rules described in Section 2 with respect to those measures. Section 4 is devoted to the practical estimation of E_n. A real data case dealing with bladder cancer is presented in Section 5. Finally, in Section 6 we discuss the results and summarize the conclusions.

2 Classification rules

In this Section we present the rules that will be compared throughout the paper. The rules that incorporate additional information defined in Fernández et al. are summarized in Subsection 2.1 and related to Fisher's discriminant rule and to those based on shrinkage defined in Tong et al., which are described in Subsection 2.2.

From now on, we assume two p-dimensional normal populations Π_1 and Π_2 with means μ_1 and μ_2 respectively, and common covariance matrix Σ. Using the notation given in the Introduction, we have that

X/Y=j~$N_p(\mu_j, \Sigma)$, j=1, 2. Let us denote as $\overline{X}_j$ the sample mean vector of the observations coming from population j (i.e., $\overline{X}_j=(\Sigma_{i=1}^{n} X_i I_{(Y_i=j)})/(\Sigma_{i=1}^{n} I_{(Y_i=j)})$), for j=1, 2, and S the pooled sample covariance matrix.

If we denote as $u\in\mathbb{R}^p$ a new observation to be classified, the optimal (theoretical) Bayes rule is:

$$\text{Classify } u \text{ in } \Pi_1 \text{ iff } \left(u-\frac{\mu_1+\mu_2}{2}\right)' \Sigma^{-1}\delta\geq\log\frac{\pi_2}{\pi_1}, \tag{1}$$

where $\delta=\mu_1-\mu_2$.

Let us further assume equal a priori probabilities π_j, j=1, 2, so, from now on, $\log\frac{\pi_2}{\pi_1}=0$.

The usual linear classification rule, also known as Fisher's discriminant rule, is obtained replacing the unknown parameters μ_1, μ_2 and Σ by their estimators $\overline{X}_1$, $\overline{X}_2$ and S:

$$\text{Classify } u \text{ in } \Pi_1 \text{ iff } \left(u-\frac{\overline{X}_1+\overline{X}_2}{2}\right)' S^{-1}\overline{\delta}\geq 0,$$

where $\overline{\delta}=\overline{X}_1-\overline{X}_2$.

As already mentioned, we will also assume that there is some additional information available on the problem being considered. For example, in medical contexts it is usual that patients take higher values in some variables (and lower in others) than healthy people. Mathematically, this information on the mean vectors can be incorporated using cones that restrict the parameter space. Namely, we assume that the difference between the mean vectors δ belongs to C, where C is a closed, convex, polyhedral cone in $\mathbb{R}^p$,

$$C=\{z\in\mathbb{R}^p : a_j'z\geq 0, j=1,\ldots,m\}.$$

For the two populations case, one of the most interesting cones is the positive orthant cone $O^+=\{x\in\mathbb{R}^p : x_i\geq 0, i\in T\}$, where T is the subset of predictors for which the means of the two populations are known to be higher in the first of the two populations.

Generally speaking, polyhedral cones are widely used in restricted inference, because they cover the most interesting cases from a practical standpoint. Among these cones are the simple order cone $C_{SO}=\{z\in\mathbb{R}^q : z_1\leq\ldots\leq z_q\}$ (where q is the number of illness levels with 1 denoting absence of illness or the control level), used when there is an increasing relation between the level of the illness and the predictors; the tree order cone $C_{TO}=\{z\in\mathbb{R}^q : z_1\leq z_i, i=2,\ldots,q\}$, used when it is known that the level of the predictors increase when the illness is present but it is not sure that they increase when the severity of the illness increases; and the already mentioned positive orthant cone.

For more details on cones of restrictions and their geometry, the reader may consult Robertson et al. (1988) or Silvapulle and Sen (2005).

2.1 Classification rules with additional information

In order to obtain a classification rule that incorporates the additional information available for the problem, Fernández et al. (2006) start rewriting rule (1) as

$$\text{Classify } u \text{ in } \Pi_1 \text{ iff } (u-(c_1\mu_1+c_2\mu_2)+c\delta)'\Sigma^{-1}\delta\geq 0,$$

where $c_i=n_i/(n_1+n_2)$, i=1, 2 and $c=(c_1-c_2)/2$.

The new classification rule is then obtained replacing Σ by S, $c_1\mu_1+c_2\mu_2$ by $c_1\overline{X}_1+c_2\overline{X}_2$ and the restricted parameter δ by an estimator that incorporates the additional information. To be more precise, δ is replaced by a member of the family δ_γ^*, with $\gamma\in[0, 1]$, defined as the limit of the following iterative procedure that Fernández et al. show to be convergent. Let $\hat{\delta}_\gamma^{(0)}=\overline{X}_1-\overline{X}_2$, and $\hat{\delta}_\gamma^{(i)}=p_{S^{-1}}(\hat{\delta}_\gamma^{(i-1)}/C)-\gamma p_{S^{-1}}(\hat{\delta}_\gamma^{(i-1)}/C^P)$ for i=1, 2, …, where

$p_{S^{-1}}(Z/C)$ is the projection of Z onto cone C with the metric given by S^{-1} and $C^P=\{z\in\mathbb{R}^p:z'x\le 0,x\in C\}$ is the so called polar cone of C. In this way the following family of new classification rules $R_n(\gamma)$ with $\gamma\in[0, 1]$ is obtained:

$$\text{Classify } u \text{ in } \Pi_1 \text{ iff } (u-(c_1\bar{X}_1+c_2\bar{X}_2)+c\delta^*_\gamma)'S^{-1}\delta^*_\gamma\ge 0. \tag{2}$$

If we denote as $\mu^*_{\gamma 1}=c_1\bar{X}_1+c_2(\bar{X}_2+\delta^*_\gamma)$ and $\mu^*_{\gamma 2}=c_1(\bar{X}_1-\delta^*_\gamma)+c_2\bar{X}_2$, then rule (2) can also be obtained replacing μ_1, μ_2 and Σ in (1) by their estimators $\mu^*_{\gamma 1}$, $\mu^*_{\gamma 2}$ and S.

The computation of the projection of a vector onto a cone can been carried out using the “lsConstrain.fit” method contained in “ibdreg” R package. Moreover, the R programs with which results presented throughout the paper have been calculated will be included in an R library that is being compiled. For the moment, these programs, which include the computation of the estimators, can be obtained from the authors on request.

For more details on these rules and their properties the reader is referred to Fernández et al. (2006). These rules can also be extended to deal with more than two populations. Full details on how this is done can be found in Conde et al. (2012).

2.2 Rules based on shrinkage estimators

Since projection is a contractive operator, the estimators $\mu^*_{\gamma 1}$ and $\mu^*_{\gamma 2}$ proposed in the previous subsection may be seen as “shrinkage estimators” for the mean. Other shrinkage estimators have been proposed for discrimination rules. Tong et al. describe an analytical James-Stein type shrinkage estimator for the mean under the cuadratic loss function. The estimator is proposed for fixed sampling size n_j, j=1, 2, for each population, and large dimension p. They construct a shrinkage-based discriminant diagonal rule replacing the population means by the proposed shrinkage means and considering a diagonal covariance matrix. The reason for considering a diagonal covariance matrix is the fact that, if p is greater than n_j, j=1, 2, singularity problems appear. To overcome these problems, Dudoit et al. (2002) introduced the diagonal linear discriminant analysis (*DLDA*), which, when the sample size is small, performed very well compared with more sophisticated classifiers in terms of accuracy and stability (Dettling, 2005; Lee et al., 2005).

The shrinkage mean-based diagonal linear discriminant analysis (*SmDLDA*) proposed by Tong et al. is based on the following shrinkage estimators for μ_j, j=1, 2:

$$\tilde{\mu}_j(\hat{r}_j^{opt})=\bar{\bar{X}}^j+\left(1-\frac{\hat{r}_j^{opt}}{||\bar{X}_j-\bar{\bar{X}}^j||^2_{S_j}}\right)(\bar{X}_j-\bar{\bar{X}}^j),$$
$$\hat{r}_j^{opt}\approx\frac{(n_j-1)(p-2)}{n_j(n_j-3)},$$

where $\hat{r}_j^{opt}$ is the optimal shrinkage parameter estimation, S_j the diagonal of the sample covariance matrix, $\bar{X}_j$ the sample mean vector and $\bar{\bar{X}}^j$ the grand sample mean across all variables in population Π_j for j=1, 2:

$$\bar{\bar{X}}^j=(\bar{X}_{j\cdot},\ldots,\bar{X}_{j\cdot}),$$
$$\bar{X}_{j\cdot}=\frac{1}{p}\sum_{k=1}^{p}\bar{X}_{jk}$$

Tong et al. show that *SmDLDA* outperforms *DLDA* in a wide range of situations when TMP criterion is considered.

3 Performance measures

In this section we compare the performance of *Fisher, SmDLDA*, $R_n(0)$ and $R_n(1)$ rules with regard to some of the most usual performance measures considered in practice, in scenarios similar to those in Tong et al. (with large p and not so large n), via a simulation study.

As already mentioned, we consider two populations Π_1 and Π_2 with multivariate normal distributions $N_p(\mu_1, \Sigma)$ and $N_p(\mu_2, \Sigma)$.

In the first scenario we assume an identity covariance matrix, i.e., $\Sigma=I_p$. We consider $\mu_1=(0, \ldots, 0, \mu_{1, d+1}, \ldots, \mu_{1, p})$ and $\mu_2=-\mu_1$, where $(\mu_{1, d+1}, \ldots, \mu_{1, p})$ is a random sample of size $p-d$ from the uniform distribution $U(0, 0.5)$. With these vector means, it is clear that we must focus on the positive orthant restrictions case, that is, $\delta=\mu_1-\mu_2\in O^+=\{x\in\mathbb{R}^p:x_i\geq 0, i=1,\ldots,p\}$. We consider two values of p (50 and 200) and two values of $n=n_1=n_2$ (10 and 20). For each p, we consider six different values of d: 0, $0.1\times p$, …, $0.5\times p$. For each simulation, we generate a training set of size n and a test set of size $5n$ for each population Π_j, $j=1, 2$. We repeat the procedure 1000 times. To overcome the singularity of the pooled sample covariance matrix, all rules use as estimated covariance matrix the diagonal of the pooled sample covariance matrix, *diag*(S). The results for $p=50$ are similar to those for $p=200$, so we only include the results for $p=50$.

The second scenario considers the case where the observations are correlated. Let the following block diagonal structure be the true covariance matrix:

$$\Sigma=\begin{pmatrix} \Sigma_\rho & 0 & 0 & 0 & \ldots \\ 0 & \Sigma_{-\rho} & 0 & 0 & \ldots \\ 0 & 0 & \Sigma_\rho & 0 & \ldots \\ 0 & 0 & 0 & \Sigma_{-\rho} & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{pmatrix}_{p\times p},$$

where Σ_ρ has the auto-regresive structure:

$$\Sigma_\rho=\begin{pmatrix} 1 & \rho & \ldots & \rho^9 \\ \rho & 1 & \ldots & \rho^8 \\ \ldots & \ldots & \ldots & \ldots \\ \rho^9 & \rho^8 & \ldots & 1 \end{pmatrix}_{10\times 10}.$$

We use different values of the correlation coefficient: $\rho=0$, 0.2, 0.4, 0.6, 0.8. Except for $d=0.1\times p$, all other settings remain the same.

3.1 TMP

The total misclassification probability (TMP), also known as misclassification rate or prediction error, is the probability that a new observation is misclassified. To approximate this value we use the proportion of misclassified observations over the total number of observations in the test sets.

The TMP values obtained for *Fisher, SmDLDA*, $R_n(0)$ and $R_n(1)$ rules are summarized in Figure 1.

As we can see in the figure, $R_n(0)$ takes the lowest TMP values for high values of d followed by $R_n(1)$, while for low values of d *SmDLDA* is the one that yields the lowest TMP values. These values are not far from those of $R_n(0)$ and $R_n(1)$, which take very similar TMP values.

For the second scenario, we can see that, except for $\rho=0$, $R_n(0)$ takes lowest values with regard to the TMP, followed very closely by $R_n(1)$.

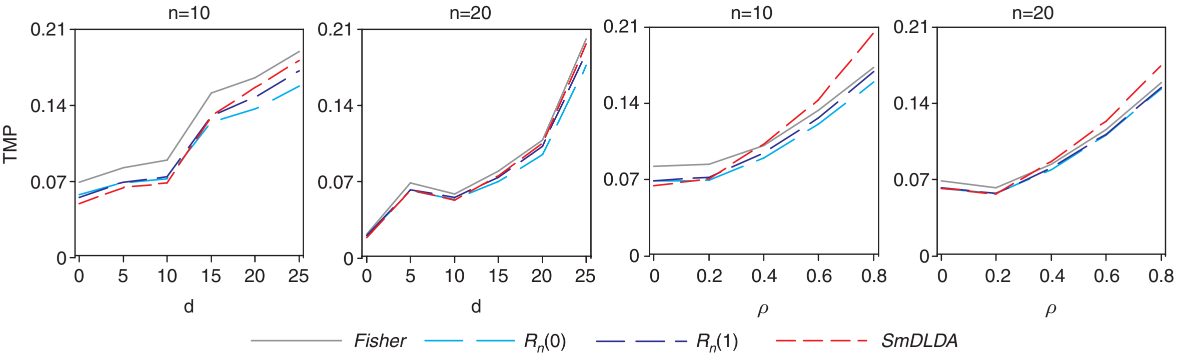

Figure 1 TMP values for Σ=*I* (two left graphs) and auto-regresive Σ (two right graphs).

3.2 Other performance measures

As already commented in the Introduction, probabilistic classifiers that provide an estimate of the probability of membership in each class for new cases are often more useful than classification rules that just assign cases to a class. For this reason, in the construction of medical probabilistic classifiers, other performance measures than just the misclassification rates are used. In fact, one-dimensional summary measures such as the misclassification rate are rarely used in practice (Pepe et al., 2004). Specifically, the most commonly used global index of diagnosis accuracy is the area under the ROC curve (AUC). Additionally, other measures such as well-calibratedness and refinement, introduced and developed by Kim and Simon, can be used.

In order to assess performance measures such as AUC, well-calibratedness and refinement, we need to transform the classification rules [*Fisher, SmDLDA* and $R_n(\gamma)$, γ=0, 1] into probabilistic classifiers. Let f_1 and f_2 be the normal probability density functions of populations Π_1 and Π_2. The optimal (theoretical) Bayes rule (1) is equivalent to

$$\text{Classify } u \text{ in } \Pi_1 \text{ iff } p(u;\mu_1,\mu_2,\Sigma)=\frac{\pi_1 f_1(u;\mu_1,\Sigma)}{\pi_1 f_1(u;\mu_1,\Sigma)+\pi_2 f_2(u;\mu_2,\Sigma)}\geq\frac{1}{2},$$

where $p(u;\mu_1,\mu_2,\Sigma)$ is the predictive probability of class 1 for a new case u. The function $p(.;\mu_1,\mu_2,\Sigma):\mathbb{R}^p\to\left[0,1\right]$ is the corresponding probabilistic classifier.

As we did in the previous section, in the rest of the paper we will continue assuming that $\pi_1=\pi_2=\frac{1}{2}$. If we replace in $p(u;\mu_1,\mu_2,\Sigma)$ the unknown parameters μ_1,μ_2 and Σ by each of the estimators considered in Section 2, we have different predictive probabilities of class 1 for a new case u, and, therefore, different probabilistic classifiers. If the estimators are $\tilde{\mu}_1(\hat{r}_1^{opt})$, $\tilde{\mu}_2(\hat{r}_2^{opt})$ and $diag(S)$, we obtain *SmDLDA* probabilistic classifier, that we will denote as p^*_{SmDLDA}. If the estimators are $\bar{X}_1$, $\bar{X}_2$ and S, we obtain Fisher probabilistic classifier, that we will denote as p^*_{Fisher}. Finally, if the estimators are $\mu^*_{\gamma1}$, $\mu^*_{\gamma2}$ and S, we obtain the restricted $R_n(\gamma)$ probabilistic classifiers, that we will denote as $p^*_{R_n(\gamma)}$. For the simulations, as the dimension p is too high and it is not possible to estimate all values in S, we will replace S by $diag(S)$.

3.2.1 AUC

The AUC is very frequently used in medical contexts as a measure for the effectiveness of diagnostic markers. In these contexts, a patient is assessed as positive or negative depending on whether the corresponding probabilistic classifier value $\tilde{p}(u)$ is greater than or less than or equal to a given threshold value. For each threshold value, there is a corresponding probability of a true positive (sensitivity) and a corresponding probability of a true negative (specifity). The ROC curve is a plot of sensitivity versus 1–specifity for all threshold values.

The AUC of the empirical ROC curve for a probabilistic classifier $\tilde{p}$ is the Mann-Whitney U statistic (cf. Pepe et al., 2006):

$$\widehat{AUC}=\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left(I_{[\tilde{p}(u_i)>\tilde{p}(u_j)]}+\frac{1}{2}I_{[\tilde{p}(u_i)=\tilde{p}(u_j)]}\right).$$

The results of the simulations conducted to compare p^*_{Fisher}, p^*_{SmDLDA}, $p^*_{R_n(0)}$ and $p^*_{R_n(1)}$ with regard to AUC appear in Figure 2.

From this figure, it is clear that $p^*_{R_n(0)}$ takes the highest mean AUC values with the lowest standard deviations for high values of d, followed by $p^*_{R_n(1)}$, while p^*_{SmDLDA} yields the highest mean AUC values with lowest standard deviations for low values of d.

For the second scenario, $p^*_{R_n(0)}$ and $p^*_{R_n(1)}$ take the highest mean AUC values ($p^*_{R_n(1)}$ for n=10, $p^*_{R_n(0)}$ for n=20) with the lowest standard deviations for ρ>0, while p^*_{SmDLDA} yields highest mean AUC values with lowest standard deviations for ρ=0.

3.2.2 Well-calibratedness and refinement

A probabilistic classifier $\tilde{p}$ is well calibrated if for any predictive probability w, 0<w<1, $P(C=1|\tilde{p}(u)=w)=w$. That is, the proportion of the cases that the probabilistic classifier predicts with probability w are actually class 1, is w.

A probabilistic classifier is refined if the predictive probabilities w tend to be close to 0 or 1. Refinement is defined as the expected value of

$$P(C=1|\tilde{p}(u)=w)(1-P(C=1|\tilde{p}(u)=w))$$

with respect to the predictive probability w.

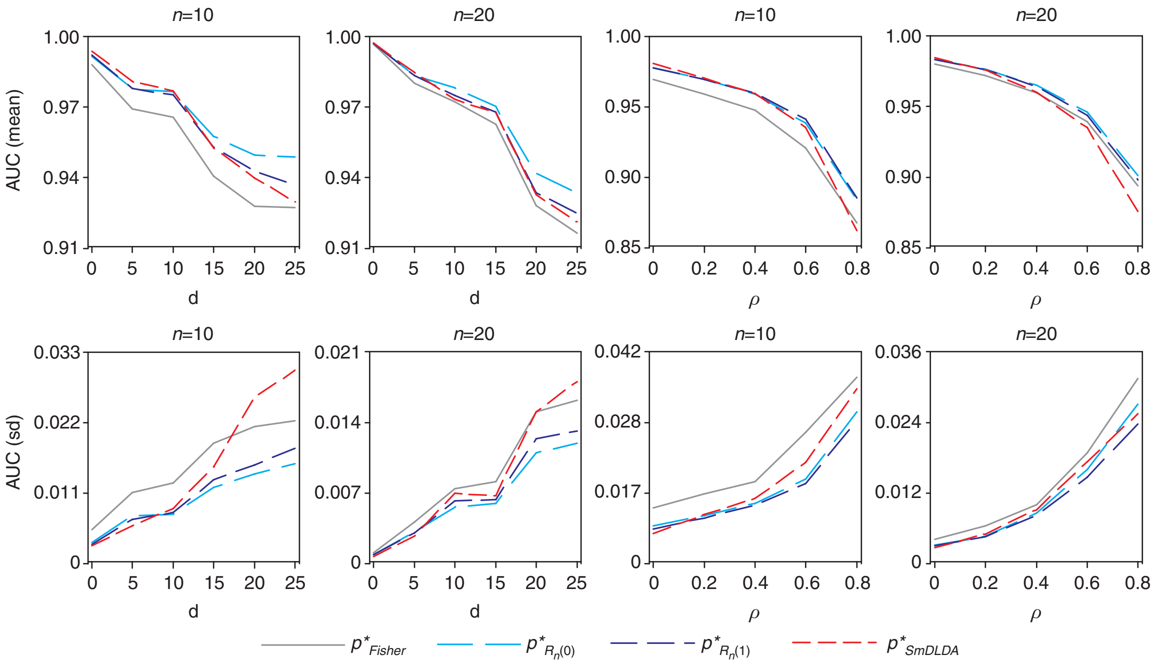

Figure 2 AUC values for Σ=I (four left graphs) and auto-regresive Σ (four right graphs).

To assess the calibration and refinement, Kim and Simon define two measures: calibration score (CS) and refinement score (RS). First, let us first partition the unit interval into m equal subintervals or bins $B_k=[(k-1)/m, k/m]$, $k=1, \ldots, m$. For each bin B_k, let q_k be the proportion of predictions w that fall into B_k, r_k the relative frequency of predictions in B_k for class 1, and u_k the center point of B_k. Then

$$CS=\sum_{k=1}^{m}(r_k-u_k)^2 q_k, \quad RS=\sum_{k=1}^{m} r_k(1-r_k)q_k.$$

The results of the simulations conducted to compare p^*_{Fisher}, p^*_{SmDLDA}, $p^*_{R_n(0)}$ and $p^*_{R_n(1)}$ with regard to well-calibratedness and refinement appear in Figure 3.

From this figure, it is clear that, again, $p^*_{R_n(0)}$ and $p^*_{R_n(1)}$ take very close CS and RS values. We can also see that $p^*_{R_n(0)}$ is the one with lowest calibration scores for high values of ρ (ρ=0.6, 0.8), while p^*_{SmDLDA} is the one with lowest calibration scores for all values of d and for low values of ρ.

With regard to refinement, it can be noticed that $p^*_{R_n(0)}$ has lowest refinement scores for high values of d and ρ, while p^*_{SmDLDA} takes lowest refinement scores for low values of d and ρ.

As a final conclusion for this Section, we can say that, even in high dimensional data with small sample size problems, $R_n(0)$ and $R_n(1)$ compete well with *SmDLDA* outperforming it in a number of configurations. We also think that the fact that the estimators used in restricted rules are motivated by the additional information available on the problem to be considered can be seen as a conceptual advantage over the “blind” shrinkage estimators considered by *SmDLDA*.

4 Practical estimation of true error rate

In the previous Section we considered the evaluation of the global (unconditional) performance of the discrimination procedures. As mentioned in the Introduction, in practice it is necessary to evaluate the behavior of the discrimination procedure for a given training data set. The best way of performing this conditional

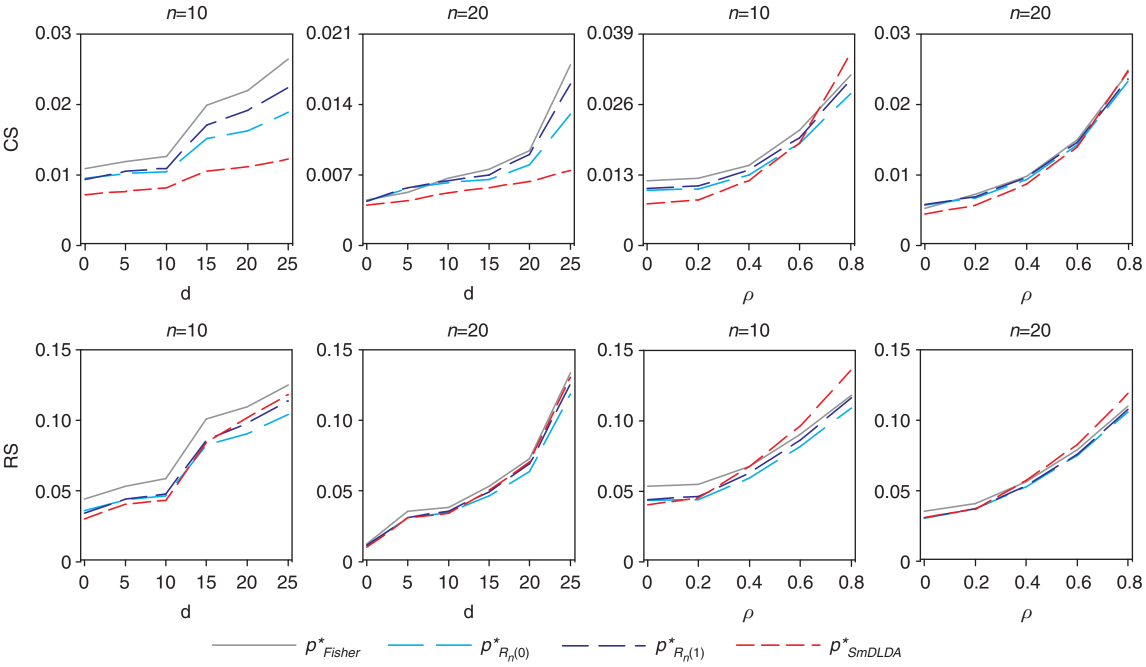

Figure 3 CS and RS values for $\Sigma=I$ (four left graphs) and auto-regresive Σ (four right graphs).

evaluation would be to have an independent test data set. However, in practice it is not common to have this second data set, so estimators of performance measures not based on the existence of the test set have been developed. Many of these procedures are, in one way or another, based on the resubstitution error also called apparent error rate.

The resubstitution estimator or apparent error rate, *APP*, estimates the true error rate E_n as the proportion of observations in the training sample that are wrongly classified by the rule. It is well known (see, for example, McLachlan, 1976, or Efron, 1983) that *APP* is a biased estimator that underestimates the true error rate because the training sample data are used twice, both to build the rule and to check its accuracy.

We start this Section by describing, in Subsection 4.1, the most usual procedures designed to correct the bias of *APP*. Then, in Subsection 4.2 we prove a property of the *APP* of the restricted rules that expose that the *APP* of these rules does not behave in the same way than that of Fisher's rule. For this reason, the usual correction procedures cannot be expected to perform well. Consequently, in Subsection 4.3 we give new proposals for the estimation of the true error rate for the restricted rules. Finally, in Subsection 4.4 we show how these estimators perform via a simulation study.

Notice that these estimators can also be useful for other performance measures such as AUC since the estimation of these measures can be based on the computation of misclassification probabilities (1–specificity and 1–sensibility).

4.1 Usual resampling based estimators

There are many non parametric estimators for the true error rate of a classification rule based on resampling techniques. In this subsection we describe the most usual ones.

The cross-validation, or leave-one-out, method was proposed in Lachenbruch and Mickey (1968). With this method one of the observations in the training sample is left out, then the classification rule is computed and the excluded observation is classified. This is repeated with each of the observations in the training sample. Then the cross-validation error, *CV*, is just the proportion of observations misclassified using this procedure. It is well known that this estimator has lower bias than *APP*.

Efron shows that *CV* has a low bias but a not so low variability and proposes estimators based on the bootstrap methodology. Let us denote as $M_n=\{(X_i, Y_i), i=1, \ldots, n\}$ the original training sample. A bootstrap training sample $M_n^*=\{(X_i^*,Y_i^*), i=1,\ldots,n\}$ is a size *n* randomly obtained (with replacement) sample from the original training sample (i.e., $P((X_i^*,Y_i^*)=(X_s,Y_s))=\frac{1}{n}$ with $s, i\in\{1, \ldots, n\}$). The probability that an observation from the original training sample is not included in the bootstrap training sample depends on *n* and is approximately 0.368. The bootstrap version of the classification rule is the rule based on the bootstrap training sample. From this methodology Efron proposes several ways of estimating the classification error. We consider two of them: the leave-one-out bootstrap (*LOOBT*) and the bootstrap 632 (*BT*632).

For the *LOOBT* estimator, *B* bootstrap training samples are considered and *B* bootstrap versions of the classification rule are obtained. Then each of these rules is used for classifying the original observations not belonging to the corresponding bootstrap training sample. Finally, *LOOBT* is the proportion of observations not correctly classified using this procedure. Efron notices that *LOOBT* tends to overestimate the true error rate and then proposes $BT632=0.368\cdot APP+0.632\cdot LOOBT$. In certain cases the value of *APP* is close to 0 (overfitting) so that *BT*632 is close to 0.632·*LOOBT* and the true error is underestimated. For these situations with high overfitting, Efron and Tibshirani (1997) propose the bootstrap 632+, defined as $BT632+=(1-\alpha)APP+\alpha LOOBT$, with $\alpha>0.632$. In Section 4.2 we will see that *APP* for rules $R_n(\gamma)$, $\gamma\in[0, 1]$, is higher than *APP* for Fisher's rule. Consequently, the overfitting problem for these rules is less important and we do not consider *BT*632+ in our study.

More recently, Fu et al. propose a method based on cross-validation after bootstrap (*BCV*) that has a lower relative mean squared error than *LOOBT* and *BT*632 for small training samples. In this procedure *B* bootstrap samples $M_b^*, b=1,\ldots,B$ are obtained from M_n. Let CV_b be the true error rate estimator obtained using the cross-validation method on sample M_b^*. The final true error rate estimator is now $BCV=B^{-1}\sum_{b=1}^{B} CV_b$.

4.2 Theoretical results for the *APP* of restricted rules

Here we obtain some properties of *APP* for the restricted rules. In particular, in Theorem 1 we prove that *APP* is less optimistic for the rules $R_n(\gamma)$, $\gamma\in[0, 1]$, than for Fisher's rule. The proof of the Theorem is deferred to the Appendix in order to improve the readability of the paper.

The result will be proved for known Σ. Under this condition a simple transformation allows us to further assume that $\Sigma=I$. Recall also that throughout the paper equal a priori probabilities are being considered. In these conditions, the apparent error rate of rule $R_n(\gamma)$, $\gamma\in[0, 1]$, is $(App(\gamma)+APP_2(\gamma))/2$, where

$$APP_1(\gamma)=\frac{1}{n_1}\sum_{i=1}^{n} I_{[(X_i-(c_1\bar{X}_1+c_2\bar{X}_2)+c\delta^*_\gamma)'\delta^*_\gamma<0]}\; I_{[Y_i=1]}$$

and

$$APP_2(\gamma)=\frac{1}{n_2}\sum_{i=1}^{n} I_{[(X_i-(c_1\bar{X}_1+c_2\bar{X}_2)+c\delta^*_\gamma)'\delta^*_\gamma>0]}\; I_{[Y_i=2]}$$

are the apparent error rates of populations Π_1 and Π_2, respectively.

This is the result:

Theorem 1 *If $n_1=n_2$ then, for any $\gamma\in[0, 1]$,*

$$E(APP(\gamma))\geq E(APP(0))\geq E(APP(Fisher)).$$

Remark 2 *In Fernández et al. it is proved that, if $n_1=n_2$, the true error rate of rules $R_n(\gamma)$, $\gamma\in[0, 1]$, is lower than that of Fisher's rule. Moreover, in all simulations performed the true error rate of rules $R_n(\gamma)$, is higher than their expected apparent error rates. As from Theorem 1, $E(APP(\gamma))\geq E(APP(Fisher))$, this suggests that if $n_1=n_2$ the bias of APP for rules $R_n(\gamma)$, $\gamma\in[0, 1]$ is lower than that for Fisher's rule.*

A possible explanation for this is that the restricted rules are less dependent on the training sample values than Fisher's rule, as they are built not only using the training sample but also the additional information available for the problem.

Consequently, the usual procedures designed to correct the bias of APP cannot be expected to perform well. This points to the need for new estimators of E_n, specific for these restricted rules.

4.3 New proposals

In this subsection we propose new true error rate estimation procedures based on resampling techniques for the restricted rules. These methods modify *LOOBT* and *BCV* to make them able to cope properly with the information included in the rule. We will denote as *BT*2 and *BT*3 the methods generated from *LOOBT*, and *BT*2*CV* and *BT*3*CV* the ones coming from *BCV*.

The additional information we are considering can be written as $\delta=\mu_1-\mu_2\in C$, where $C=\{z\in\mathbb{R}^p: a_j'z\geq0, j=1,\ldots,m\}$ is the appropriate cone of restrictions. Let us denote as $\bar{C}$ the following random cone generated by $\bar{\delta}=\bar{X}_1-\bar{X}_2$:

$$\bar{C}=\left\{z\in\mathbb{R}^p: \begin{matrix} a_j'z\geq0, & \text{if} & a_j'\bar{\delta}\geq0 \\ a_j'z\leq0, & \text{if} & a_j'\bar{\delta}<0 \end{matrix}, j=1,\ldots,m\right\}.$$

The true error rate estimator *BT*2 is computed in a way similar to *LOOBT* but considering bootstrap classification rules generated using projections onto cone $\bar{C}$ instead of *C* for each bootstrap training sample. In other words, for each bootstrap sample $M_b^*=\{(X_i^{*b},Y_i^{*b}),i=1,2,\ldots,n\}$ we compute the bootstrap version of the estimator of δ that we denote as δ_γ^{*b} (with $\gamma\in[0, 1]$) defined as the limit of the following iterative procedure similar to the one considered in Section 2. Let $\hat{\delta}_\gamma^{(0)b}=\bar{X}_1^{*b}-\bar{X}_2^{*b}$ and $\hat{\delta}_\gamma^{(i)b}=p_{S^{-1}}(\hat{\delta}_\gamma^{(i-1)b}/\bar{C})-\gamma p_{S^{-1}}(\hat{\delta}_\gamma^{(i-1)b}/\bar{C}^P)$ for $i=1, 2, \ldots$ Now we denote as $R_n^{*b}(\gamma)$ the bootstrap versions of the classification rules $R_n(\gamma)$ defined as:

$$\text{Classify } u \text{ in } \Pi_1 \text{ iff } (u-(c_1\bar{X}_1^{*b}+c_2\bar{X}_2^{*b})+c\delta_\gamma^{*b})'S^{-1}\delta_\gamma^{*b}\geq 0.$$

For each rule $R_n^{*b}(\gamma)$, $b=1, 2, \ldots, B$, we classify the observations in the original training sample that do not belong to the bootstrap sample M_b^*. The true error rate estimator *BT*2 is the proportion of observations wrongly classified.

The heuristic under *BT*2 is that the “bootstrap world” should mirror the “real world”. In the “real world” the original training sample M_n is obtained from the populations Π_j, $j=1, 2$, that verify $\delta=\mu_1-\mu_2\in C$. In the “bootstrap world” the population is M_n, which verifies $\bar{\delta}=\bar{X}_1-\bar{X}_2\in\bar{C}$. Therefore, the bootstrap versions of the rules should be obtained replacing the cone C by $\bar{C}$.

Our second proposal to use the additional information in a way that the “bootstrap world” imitates the “real world” is to adapt the original training sample, instead of modifying the cone, as follows.

Assume that the original training sample M_n does not verify the restrictions, i.e., $\bar{\delta}=\bar{X}_1-\bar{X}_2\notin C$. For any $\gamma\in[0, 1]$ we can use δ_γ^*, the restricted estimator of δ, to obtain estimators for μ_i $i=1, 2$. As $\mu_1=(\mu_1+\mu_2+\delta)/2$ and $\mu_2=(\mu_1+\mu_2-\delta)/2$, we can consider $\mu_{\gamma 1}^*=(\bar{X}_1+\bar{X}_2+\delta_\gamma^*)/2$ and $\mu_{\gamma 2}^*=(\bar{X}_1+\bar{X}_2-\delta_\gamma^*)/2$ as estimators for μ_1 and μ_2, respectively. Now, we transform the original training sample in such a way that the difference of the new sample means belongs to C. The transformed training sample is $\{(W_i, Y_i), i=1, 2, \ldots, n\}$, where

$$W_i=X_i-\bar{X}_j+\mu_{\gamma j}^* \text{ if } Y_i=j, j=1,2.$$

In this way $\bar{W}_1-\bar{W}_2=\mu_{\gamma 1}^*-\mu_{\gamma 2}^*=\delta_\gamma^*\in C$. Now, the proposed estimator of the true error rate, that we denote as *BT*3, is *LOOBT* replacing the original training sample by the transformed one. In this way, the bootstrap samples are extracted from populations that verify the same property that is fulfilled by the populations from which the original training sample is extracted.

We also consider cross-validation after bootstrap versions of *BT*2 and *BT*3. They are denoted as *BT2CV* and *BT3CV*, respectively.

4.4 Estimators behavior: simulation study

The behavior of an estimator $\hat{E}$ of the true error rate E_n is analyzed through the distribution of the random variable $\hat{E}-E_n$. This distribution has been called deviation distribution of the error estimator by Braga-Neto and Dougherty (2004). As a global measure of the behavior of $\hat{E}$ we will use $E[(\hat{E}-E_n)^2]$. As usual, this measure can be decomposed in a variance and a bias component:

$$E[(\hat{E}-E_n)^2]=Var(\hat{E}-E_n)+[E(\hat{E}-E_n)]^2.$$

In this section we conduct a simulation study to compare the behavior of the estimators *APP*(γ), *CV*(γ), *LOOBT*(γ), *BT*632(γ), *BCV*(γ), *BT*2(γ), *BT*3(γ), *BT2CV*(γ) and *BT3CV*(γ) of the true error rate $E_n(\gamma)$ of the restricted classification rules $R_n(\gamma)$.

The purpose of this study is to propose a reasonable estimator of $E_n(\gamma)$ when the training sample does not fulfill the restrictions. For simplicity we consider $p=3$ and identity covariance matrix and study the positive orthant restrictions case, i.e., $\delta\in O_3^+=\{x\in\mathbb{R}^3: x_i\geq 0, i=1,2,3\}$. We generate training samples of size $n_1=n_2=10$, from populations Π_1, $N_3(\delta, \Sigma)$, and Π_2, $N_3(0, \Sigma)$, for different values of δ and Σ. Since in practice sample sizes and covariances are usually larger, we have also run the simulations with bigger sample sizes ($n_1=n_2=50$)

accordingly rescaling the covariance matrix, obtaining similar results. The simulations were performed for many values of δ both in the interior of the cone and on the frontier of O_3^+ and for several values of Σ. However, since there was no significative variation in the results, in order to save space we only present here the results obtained for $\Sigma=I$ and when δ is the vertex of the cone (0, 0, 0) or it is in the interior of O_3^+. To be more precise, we show the results for values of δ in the diagonal direction of the cone, i.e., $\delta=\lambda(1, 1, 1)$ with $\lambda\geq 0$. The values of λ have been chosen so that $||\delta||^2=0, 0.25, 0.5, \ldots, 2.5$. Notice that the values considered cover the situations where discrimination is easy since the distance between the means $||\delta||^2$ is large, and others where the samples from the populations are much more likely to overlap. Larger values of $||\delta||^2$ are not given since for those values the restrictions are almost always fulfilled and therefore the true error estimation procedures are equivalent.

For each scenario, we generate 1000 training samples for which the rules $R_n(\gamma)$, with $\gamma\in\{0, 0.5, 1\}$, are determined. For each of these three rules we compute *APP, CV, LOOBT, BT*632, *BCV, BT*2, *BT*3, *BT*2*CV* and *BT*3*CV*. The number of bootstrap replicas considered for the estimators involving bootstrap was $B=100$. The true error rate E_n for each training sample was computed using a test sample with 1000 observations from each of the two populations. Using this procedure we have 1000 values of the deviation distribution of each of the 9 error estimators. With these values we approximate the values of $(E[(\hat{E}-E_n)^2])^{\frac{1}{2}}$ and $E(\hat{E}-E_n)$ that we will denote as $A(\hat{E})$ and $B(\hat{E})$ respectively. For example, if we denote as $BT2^i(0.5)$ and $E_n^i(0.5)$ the values of *BT*2 and E_n obtained from the *i*-th training sample for rule $R_n(0.5)$ we can estimate $A(BT2(0.5))=(E[(BT2(0.5)-E_n(0.5))^2])^{\frac{1}{2}}$ and $B(BT2(0.5))=E(BT2(0.5)-E_n(0.5))$ by $\left(\frac{1}{1000}\sum_{i=1}^{1000}[BT2^i(0.5)-E_n^i(0.5)]^2\right)^{\frac{1}{2}}$ and $\frac{1}{1000}\sum_{i=1}^{1000}[BT2^i(0.5)-E_n^i(0.5)]$, respectively.

Again, in order to save space, as the results obtained for the three classification rules were similar, in Table 1 we only present the values for $\gamma=1$. For each value of $||\delta||^2$, the two lowest values of $A(\hat{E})$ appear in bold. Notice that the lowest values are the ones for *BT*2 and *BT*3 for almost all values of $||\delta||^2$.

In Figure 4 we represent the values of $A(\hat{E})$ and $B(\hat{E})$ depending on $||\delta||^2$ for the nine estimators of the true error rate that we are considering. As in other simulation studies, *APP* generally has the largest negative bias, *CV* has the lowest bias but it is the one with highest variance, and *LOOBT* shows a positive bias. Estimators *BCV, BT*2*CV, BT*3*CV* and *BT*632 exhibit a negative bias. The new estimators proposed in this paper *BT*2 and *BT*3, which modify the bootstrap in order to cope with the additional information incorporated to the rules, have similar behavior. This is somehow surprising since they are based on very different ideas. They are also the best estimators of the true error rate for the smallest values of $||\delta||^2$. These are obviously the most interesting situations in practice, as they correspond to scenarios where the discrimination is more difficult and where the additional information is more likely to play a key role in the rule.

In order to have a more thorough idea of their behavior, we also obtained kernel estimators of the density of the deviation distribution for each of the nine estimators of E_n. The kernel density estimators corresponding to scenario $||\delta||^2=0.3$ for the new estimators proposed in this paper, namely *BT*2, *BT*3, *BT*632, *BT*2*CV* and *BT*3*CV*, are represented in Figure 5. From this figure it is clear that the kernel estimators for *BT*2 and *BT*3 have the lowest values of bias and variance among the five represented in the graph. The estimators *BT*2*CV* and *BT*3*CV* have a similar variance component but are much more biased, while the *BT*632 has a higher variance component than the rest.

Therefore, we conclude with the recommendation of *BT*2 and *BT*3 as estimators of the true error rate of the restricted rules.

5 Application: bladder cancer data

The data considered in this application come from a bladder cancer project aimed to select classifiers in the context of an in vitro diagnostic tool for the disease. Our industrial and pharmaceutical partners in this

Table 1 Simulations results for the nine estimators under Σ=*I* and γ=1.

Estimator												**||δ²||**
		0	**0.25**	**0.5**	**0.75**	**1**	**1.25**	**1.5**	**1.75**	**2**	**2.25**	**2.5**
APP	*A*	0.144	0.141	0.141	0.137	0.133	0.133	0.129	0.131	0.121	0.122	0.115
	B	−0.015	−0.054	−0.071	−0.076	−0.076	−0.074	−0.080	−0.080	−0.075	−0.081	−0.072
CV	*A*	0.130	0.126	0.126	0.123	0.118	0.125	0.114	0.119	0.112	0.104	0.104
	B	0.001	0.008	0.001	0.006	0.003	0.009	0.000	0.006	0.008	−0.006	0.005
LOOBT	*A*	**0.085**	0.090	0.091	0.092	0.090	0.097	0.094	0.098	**0.094**	0.092	**0.091**
	B	−0.005	0.008	0.006	0.014	0.015	0.018	0.014	0.019	0.022	0.014	0.024
BT632	*A*	0.104	0.102	0.102	0.098	0.095	0.100	0.095	0.098	0.091	**0.091**	**0.087**
	B	−0.009	−0.015	−0.022	−0.019	−0.018	−0.016	−0.021	−0.017	−0.014	−0.021	−0.011
BCV	*A*	0.114	0.114	0.112	0.108	0.105	0.107	0.103	0.103	0.097	0.097	0.092
	B	−0.011	−0.033	−0.045	−0.046	−0.049	−0.048	−0.052	−0.049	−0.046	−0.053	−0.043
BT2	*A*	0.094	**0.079**	**0.086**	**0.088**	**0.088**	**0.093**	**0.093**	**0.096**	0.095	**0.091**	0.092
	B	−0.069	−0.010	−0.002	0.010	0.013	0.015	0.013	0.018	0.021	0.014	0.024
BT3	*A*	**0.093**	**0.078**	**0.084**	**0.087**	**0.087**	**0.094**	**0.092**	**0.096**	**0.094**	**0.091**	**0.091**
	B	−0.069	−0.012	−0.002	0.010	0.013	0.016	0.013	0.018	0.021	0.014	0.024
BT2CV	*A*	0.146	0.109	0.109	0.105	0.102	0.103	0.102	0.100	0.097	0.096	**0.091**
	B	−0.129	−0.074	−0.068	−0.059	−0.057	−0.055	−0.055	−0.052	−0.048	−0.054	−0.044
BT3CV	*A*	0.145	0.110	0.108	0.104	0.102	0.103	0.102	0.100	0.097	0.097	**0.091**
	B	−0.128	−0.076	−0.067	−0.058	−0.057	−0.055	−0.056	−0.053	−0.048	−0.054	−0.044

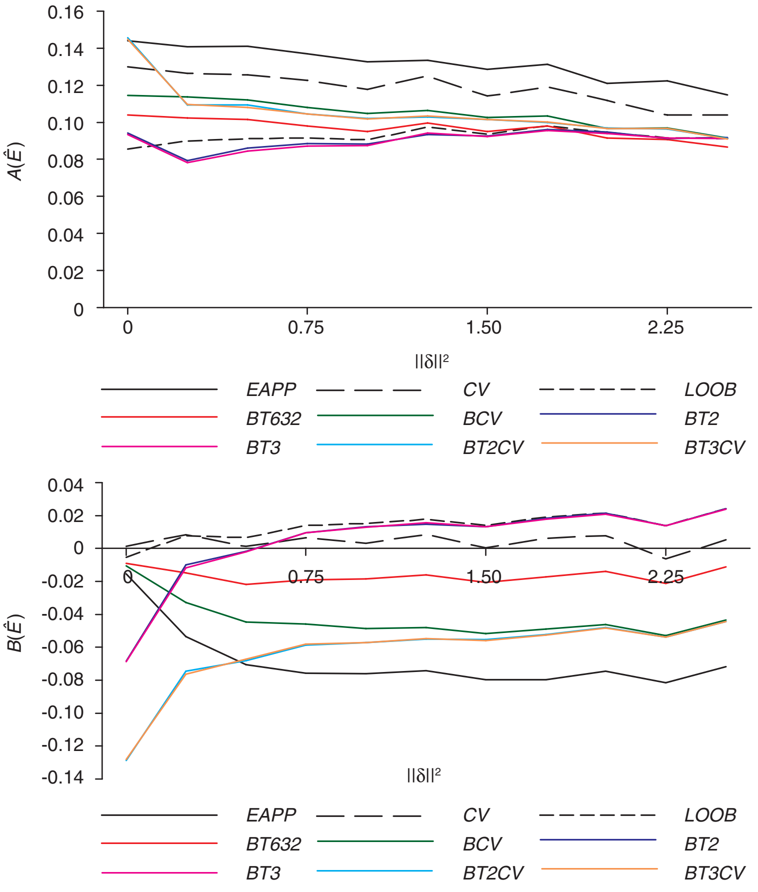

Figure 4 $A(\hat{E})$ and $B(\hat{E})$ for the true error rate estimators for Σ=*I* and γ=1.

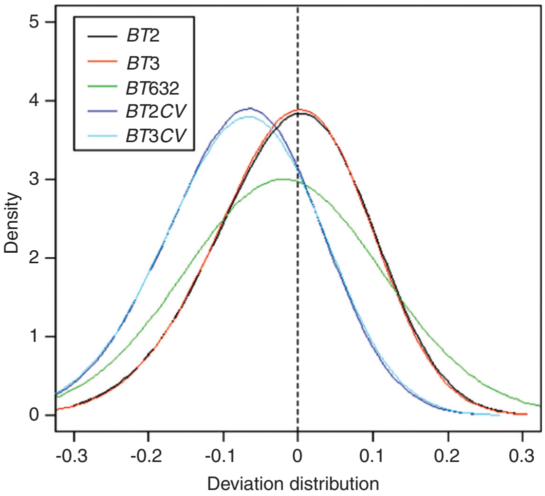

Figure 5 Kernel estimators for the density function of $\hat{E}-E_n$ for several estimators for $||\delta||^2=0.3$.

research are Proteomika S.L. and Laboratorios SALVAT, S.A. For intellectual property reasons, the names of the proteins used in this study are not disclosed in the paper.

Patients were classified in five levels based on cytoscopy. First level is control level (i.e., negative result of cytoscopy, therefore considered as absence of bladder cancer) and the other four are denoted as Ta, T1G1, T1G3 and T2, each of them corresponding to increasingly advanced levels of cancer. This combines the TNM staging (see UICC, 2009) and the grading. To be more precise, stage T describes the size of the tumor and whether it has spread and grade G refers to the appearance of the cells under the microscope. For this application, and in order to keep the populations balanced we will consider the control level as population Π_1 and levels T1G3 and T2 as population Π_2.

As usual in this kind of research, an initial database was provided. The purpose of this pilot study was to confirm or discard the associations among the proteins and the illness in order to establish a larger multicenter study. The data set D_1 contained information on 41 patients from Π_1 and 32 from Π_2 and 11 proteins together with the real stage of the illness the patients belonged to. This is the initial data set and the one we will use to build the rules. In the usual statistical terminology this is the training set.

We started considering all 11 available proteins but we obtained that, when these proteins were used together for classifying, the results were not as good as when a smaller number of proteins was considered. This was confirmed when simple studies made using t-tests told us that not all proteins were relevant. Moreover, our industrial and pharmaceutical partners informed us that, based on previous knowledge, three of the 11 proteins were expected to be more informative than the rest. Therefore, based on that previous knowledge and on the results of our studies, we decided to use four proteins in this study to discriminate among Π_1 and Π_2. It is clear that this sort of selection procedure can only be done when there is some prior information and the total number of variables is small. Obviously, we are aware that nowadays in many disciplines it is common to collect high dimensional data in a limited number of samples, and it becomes necessary to use variable selection algorithms. This is a very important issue that will be considered further in the discussion Section.

We will denote the four proteins selected as P_1, P_2, P_3 and P_4. For each of these four proteins it was expected that higher values on the proteins were related to more advanced stages of the illness, i.e., $\delta=\mu_2-\mu_1\in O^+$. As usual, the values of the proteins levels have been transformed logarithmically so that the variables are approximately normally distributed. The mean values in each of the populations and the pooled covariance matrix obtained from this data set appear in Table 2. From this table it is obvious that the additional information was not fulfilled by the training set so the classifications rules $R_n(\gamma)$ are relevant in this problem. Table 3 contains the values for the restricted estimator δ_γ^* appearing in these rules for $\gamma\in\{0, 1\}$.

Table 2 Mean for each group and pooled covariance matrix from D_1.

					Means
	N	**log(P_1)**	**log(P_2)**	**log(P_3)**	**log(P_4)**
Π_1	41	1.416	1.356	3.879	1.417
Π_2	32	1.409	0.976	4.348	1.578

$$S=\begin{pmatrix} 1.065 & 0.455 & -0.154 & 0.106 \\ 0.455 & 0.515 & -0.052 & -0.053 \\ -0.154 & -0.052 & 0.544 & 0.148 \\ 0.106 & -0.053 & 0.148 & 0.450 \end{pmatrix}$$

Table 3 Values of δ_γ^* for the $R_n(\gamma)$ rules built from D_1.

	$\boldsymbol{\delta_\gamma^*}$
$R_n(0)$	(0.328, 0, 0.430, 0.123)
$R_n(1)$	(0.664, 0.380, 0.392, 0.084)

We use this data set D_1 as a training set to build the rules. In this bladder cancer research, a second data set D_2 containing measures on the same 11 proteins and the real illness stage for a different set of 118 patients was received in a later stage. We use this second set D_2 as a test set.

Before obtaining an estimator of the true error rate of the rules, it will be useful to compare the 4 rules considered in the paper with regard to the performance measures introduced in Section 3. Rules are built from D_1 and evaluated with D_2.

We can see in Table 4 that $R_n(1)$ takes the lowest calibration score. Refinement scores are very similar for all the rules, being $R_n(0)$ the one that takes the lowest value.

In Figure 6 we represent the ROC curves for the rules built from D_1. We can see that the best ROC curve is that of $R_n(1)$. Consequently, $R_n(1)$ has the largest AUC (see Table 4).

We use D_2 set as a test set in order to obtain an estimator of the true error rate of the rules. In this way we will be able to compare the estimators of the true error rate previously defined with another value obtained from an independent sample and evaluate the behavior of the true error rate estimators in this application.

Table 5 contains the results obtained with the nine estimators of the true error rate considered in the paper and the independent estimation obtained from D_2. The bootstrap values have been obtained generating B=100 bootstrap samples as in Subsection 4.4. There are several questions that are worth noticing. One of them is the fact that, as mentioned in Subsection 4.2, *APP* increases with γ, which is logical since the rules with higher values of γ are less dependent from the original training sample. Moreover, for the data in the application, *APP* is higher than the independent estimation of the true error obtained from D_2 for rule $R_n(1)$. This is not usual for Fisher's rule although it may happen more frequently for the restricted rules since *APP*

Table 4 Performance measures of the rules.

Measure	***Fisher***	***SmDLDA***	$\boldsymbol{R_n(0)}$	$\boldsymbol{R_n(1)}$
CS	0.082	0.083	0.048	0.028
RS	0.170	0.171	0.163	0.170
AUC	0.652	0.652	0.681	0.708

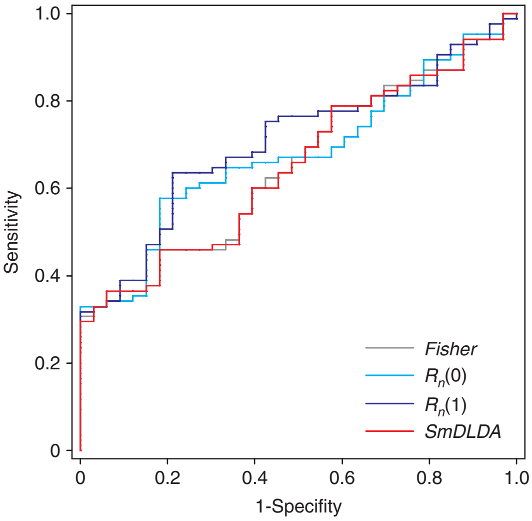

Figure 6 ROC curves of the rules.

Table 5 Estimations of the true error rate of the rules.

Estimator	*Fisher*	*SmDLDA*	$R_n(0)$	$R_n(1)$
APP	30.14%	30.14%	32.88%	50.68%
CV	36.99%	36.99%	36.99%	49.32%
LOOBT	36.49%	36.22%	39.05%	48.93%
BT632	34.15%	33.98%	36.78%	49.57%
BCV	29.37%	30.75%	32.95%	45.15%
BT2	–	–	35.53%	34.76%
BT3	–	–	41.77%	42.80%
BT2CV	–	–	29.05%	29.33%
BT3CV	–	–	34.67%	37.03%
Estimation from D_2	39.83%	39.83%	36.44%	33.90%

usually increases and the true error decreases with γ. Notice, however, that from the results obtained in the simulations Section *APP* still has a negative bias as estimator of the true error rate. We can observe that, even in this not standard case, the *BT2* estimator, which had the second best behavior in the simulations, has a very good performance for the values of γ considered.

In conclusion, we can see that, as in previous Sections, the combination of rule $R_n(1)$ with the estimator *BT2* yields good results in this case.

6 Discussion

In the classical problem of discrimination between two normal populations with equal covariance matrices, Fernández et al. defined new classification rules that take into account the additional information that is frequently available in classification problems (restricted rules) and showed that these rules have lower misclassification error than the usual Fisher's rule. These rules, that we denote as $R_n(\gamma)$, $\gamma \in [0, 1]$, are obtained from Bayes rule considering estimators of the mean vector of the populations that take into account the additional information of the problem by projecting the sample means on the cone defined by the additional information.

As projection is a contractive operator, these restricted rules can be viewed as shrinkage rules. Tong el al. have recently proposed to consider James-Stein type shrinkage estimators of the means to define discrimination rules known as SmDLDA rules. Considering scenarios similar to those in Tong et al., we have compared, via a simulation study, the performance of the restricted rules $R_n(\gamma)$ with that of SmDLDA and Fisher's rule using several of the most common criteria used in the literature. The criteria considered are the total misclassification probability (TMP), the area under ROC curve (AUC), the refinement (RS) and the well-calibratedness (CS). The results obtained show that the restricted rules compete well with SmDLDA, even for high dimensional situations, under any of the four criteria considered, showing better performance under many of the conditions considered in the simulations. The restricted rules also have the advantage that the shrinkage in these rules is not "blind" but motivated by the information at hand.

Another important issue for any classification rule is the estimation of the true error rate, i.e., the error rate of a given training sample. This problem has not been considered so far for the restricted rules. In this paper we check that the apparent error rate (*APP*) as estimator of the true error rate of the restricted rules has a different behavior than that of Fisher's rule. Namely, in Theorem 1 we prove that the expected apparent error of these rules is higher than that of Fisher's rule. As the true error rate of Fisher's rule is higher than that of the new rules, this means that these new rules do not suffer so much overfitting as Fisher's rule. Consequently, the usual procedures that try to reduce the bias of APP for estimating the true error rate such as *CV, LOOBT, BT*632 or *BCV* do not work as well as they should, and new estimators for the true error rate, specific for the restricted rules, are needed. We consider two methods based on different bootstrap procedures that take into account the additional information available on the problem. The first one, that we denote as *BT*2, adjusts the cone of restrictions to the training sample while the second one, denoted as *BT*3, adjusts the training sample to the cone of restrictions. The corresponding cross-validation after bootstrap versions of these procedures, *BT*2*CV* and *BT*3*CV*, are also considered.

Based on a simulation study we check that the new procedures *BT*2 and *BT*3 generally perform better as estimators of the true error rate, E_n, than the usual estimators designed for rules that do not account for additional information. Their performance is especially good for situations where the populations are not too separated. This is the scenario where the new rules are more interesting since it is the case where training samples not fulfilling the restrictions are more likely to appear.

We can also notice that for these rules it is not necessary to perform cross-validation after bootstrap, since *BT*2*CV* and *BT*3*CV* do not behave better than *BT*2 or *BT*3. Therefore, we conclude with the recommendation of estimators *BT*2 and *BT*3 to evaluate the true error rate of the discrimination rules defined in Fernández et al. (2006).

All the work in this paper has been applied to a real bladder cancer project. Four rules have been considered [Fisher's, SmDLDA, $R_n(0)$ and $R_n(1)$]. We observed that $R_n(1)$ is the one with best behavior with respect to three of the four evaluation measures considered (namely, true error rate, area under ROC curve and calibration score) and that all four rules have similar results for the refinement score. We also compared the estimations of the true error rate for the new estimators proposed and we exposed the good behavior of the new estimator *BT*2 for the restricted rules $R_n(0)$ and $R_n(1)$.

There is still work to be done for the restricted rules. A very important issue for a rule is variable selection. In many disciplines high dimensional data with small sample size are collected. These cases highlight the need for variable selection algorithms, with which to perform some kind of general variable selection strategy. A good recent proposal is that of Graf and Bauer (2009), who perform model selection based on FDR-thresholding optimizing the area under the ROC curve. We think that this proposal may be very useful for restricted rules. In any case, this is a very interesting area that we will explore in future works.

Acknowledgment: This research was partially supported by Spanish DGES grant MTM2012-37129. We thank the anonymous reviewers for several useful comments which improved this manuscript.

Appendix

The expected apparent error rate for Π_1 is

$$\begin{aligned}E(APP_1(\gamma))&=P((X_1-(c_1\bar{X}_1+c_2\bar{X}_2)+c\delta^*_\gamma)'\delta^*_\gamma<0,Y_1=1)\\&=E[P((X_1-(c_1\bar{X}_1+c_2\bar{X}_2)+c\delta^*_\gamma)'\delta^*_\gamma<0,Y_1=1/\bar{X}_1,\bar{X}_2)].\end{aligned}$$

For the proof of Theorem 1 we will need a previous result:

Lemma 3

$$\begin{aligned}&P((X_1-(c_1\bar{X}_1+c_2\bar{X}_2)+c\delta^*_\gamma)'\delta^*_\gamma<0,Y_1=1/\bar{X}_1,\bar{X}_2)\\&=\Phi\left(-\sqrt{\frac{n_1}{n_1-1}}\frac{(c_2\bar{\delta}+c\delta^*_\gamma)'\delta^*_\gamma}{\sqrt{\delta^{*\prime}_\gamma\delta^*_\gamma}}\right).\end{aligned}$$

Proof. In order to make the proof clearer and to remark the dependence of δ^*_γ on $\bar{X}_1$ and $\bar{X}_2$ during the proof we will write δ^*_γ as $\delta^*_\gamma(\bar{X}_1,\bar{X}_2)$. It is easy to check that

$$\begin{aligned}&(X_1-(c_1\bar{X}_1+c_2\bar{X}_2)+c\delta^*_\gamma(\bar{X}_1,\bar{X}_2))'\delta^*_\gamma(\bar{X}_1,\bar{X}_2)\\&=(X_1-\bar{X}_1)'\delta^*_\gamma(\bar{X}_1,\bar{X}_2)+(c_2\bar{\delta}+c\delta^*_\gamma(\bar{X}_1,\bar{X}_2))'\delta^*_\gamma(\bar{X}_1,\bar{X}_2)\end{aligned}$$

so that

$$\begin{aligned}&P((X_1-(c_1\bar{X}_1+c_2\bar{X}_2)+c\delta^*_\gamma(\bar{X}_1,\bar{X}_2))'\delta^*_\gamma(\bar{X}_1,\bar{X}_2)<0,Y_1=1/\bar{X}_1=t_1,\bar{X}_2=t_2)\\&=P((X_1-\bar{X}_1)'\delta^*_\gamma(\bar{X}_1,\bar{X}_2)<-(c_2\bar{\delta}+c\delta^*_\gamma(\bar{X}_1,\bar{X}_2))'\delta^*_\gamma(\bar{X}_1,\bar{X}_2),Y_1=1/\bar{X}_1=t_1,\bar{X}_2=t_2)\\&=P((X_1-\bar{X}_1)'\delta^*_\gamma(t_1,t_2)<-(c_2(t_1-t_2)+c\delta^*_\gamma(t_1,t_2))'\delta^*_\gamma(t_1,t_2),Y_1=1/\bar{X}_1=t_1,\bar{X}_2=t_2).\end{aligned}$$

Now, $(X_1-\bar{X}_1)'\delta^*_\gamma(t_1,t_2)\sim N\left(0,\frac{n_1-1}{n_1}\delta^*_\gamma(t_1,t_2)'\delta^*_\gamma(t_1,t_2)\right)$ is an ancillary statistic as its distribution does not depend on μ_1 or μ_2. As $(\bar{X}_1,\bar{X}_2)$ is sufficient and complete, from Basu's theorem we have that $(X_1-\bar{X}_1)'\delta^*_\gamma(t_1,t_2)$ and $(\bar{X}_1,\bar{X}_2)$ are independent. From this fact we have that

$$\begin{aligned}&P((X_1-\bar{X}_1)'\delta^*_\gamma(t_1,t_2)<-(c_2(t_1-t_2)+c\delta^*_\gamma(t_1,t_2))'\delta^*_\gamma(t_1,t_2),Y_1=1/\bar{X}_1=t_1,\bar{X}_2=t_2)\\&=P((X_1-\bar{X}_1)'\delta^*_\gamma(t_1,t_2)<-(c_2(t_1-t_2)+c\delta^*_\gamma(t_1,t_2))'\delta^*_\gamma(t_1,t_2),Y_1=1)\\&=\Phi\left(-\sqrt{\frac{n_1}{n_1-1}}\frac{(c_2(t_1-t_2)+c\delta^*_\gamma(t_1,t_2))'\delta^*_\gamma(t_1,t_2)}{\sqrt{\delta^*_\gamma(t_1,t_2)'\delta^*_\gamma(t_1,t_2)}}\right).\end{aligned}$$

See Lehmann and Casella (1998: p. 93) for the same argument in a similar situation. ■

In a similar way, for Π_2 we have

$$\begin{aligned}E(APP_2(\gamma))&=P((X_1-(c_1\bar{X}_1+c_2\bar{X}_2)+c\delta^*_\gamma)'\delta^*_\gamma>0,Y_1=2)\\&=E[P((X_1-(c_1\bar{X}_1+c_2\bar{X}_2)+c\delta^*_\gamma)'\delta^*_\gamma>0,Y_1=2/\bar{X}_1,\bar{X}_2)]\end{aligned}$$

and

$$P((X_1-(c_1\bar{X}_1+c_2\bar{X}_2)+c\delta^*_\gamma)'\delta^*_\gamma>0,\ \ Y_1=2/\bar{X}_1,\bar{X}_2)$$
$$=\Phi\left(-\sqrt{\frac{n_2}{n_2-1}}\frac{(c_1\bar{\delta}-c\delta^*_\gamma)'\delta^*_\gamma}{\sqrt{\delta^{*\prime}_\gamma\delta^*_\gamma}}\right).$$

Following the same lines we can also prove that for Fisher's rule

$$E(APP_1(Fisher))=E\left[P\left(\left(X_1-\frac{1}{2}(\bar{X}_1+\bar{X}_2)\right)'\bar{\delta}<0,\ \ Y_1=1/\bar{X}_1,\bar{X}_2\right)\right]$$
$$E(APP_2(Fisher))=E\left[P\left(\left(X_1-\frac{1}{2}(\bar{X}_1+\bar{X}_2)\right)'\bar{\delta}>0,\ \ Y_1=2/\bar{X}_1,\bar{X}_2\right)\right]$$

and

$$P\left(\left(X_1-\frac{1}{2}(\bar{X}_1+\bar{X}_2)\right)'\bar{\delta}<0,\ \ Y_1=1/\bar{X}_1,\bar{X}_2\right)=\Phi\left(-\frac{1}{2}\sqrt{\frac{n_1}{n_1-1}}\|\bar{\delta}\|\right)$$
$$P\left(\left(X_1-\frac{1}{2}(\bar{X}_1+\bar{X}_2)\right)'\bar{\delta}>0,\ \ Y_1=2/\bar{X}_1,\bar{X}_2\right)=\Phi\left(-\frac{1}{2}\sqrt{\frac{n_2}{n_2-1}}\|\bar{\delta}\|\right).$$

Now we are ready to prove Theorem 1:

Proof of Theorem. As $n_1=n_2$ we have that $c_1=c_2=\frac{1}{2}$ and $c=0$. Now, $\delta^*_0=p(\bar{\delta}/C)$ and $\delta^*_\gamma\in C$ so taking into account Theorem 1.3.2 in Robertson et al. (1988), $(\bar{\delta}-\delta^*_0)'\delta^*_0=0$ and $(\bar{\delta}-\delta^*_0)'\delta^*_\gamma\le 0$ From this,

$$\frac{\bar{\delta}'\delta^*_\gamma}{\sqrt{\delta^{*\prime}_\gamma\delta^*_\gamma}}\le\frac{\delta^{*\prime}_0\delta^*_\gamma}{\sqrt{\delta^{*\prime}_\gamma\delta^*_\gamma}}=\|\delta^*_0\|\cos(\delta^*_0,\delta^*_\gamma)\le\|\delta^*_0\|=\frac{\bar{\delta}'\delta^*_0}{\sqrt{\delta^{*\prime}_0\delta^*_0}}\le\|\bar{\delta}\|,$$

and the result follows from Lemma 3. ■

References

Beran, R. and L. Dümbgen (2010): "Least squares and shrinkage estimation under bimonotonicity constraints," Stat. Comput., 20(2), 177–189.

Braga-Neto, U. M. and E. R. Dougherty (2004): "Is cross-validation valid for small-sample microarray classification?," Bioinformatics, 20, 374–380.

Conde, D., M. A. Fernández, C. Rueda and B. Salvador (2012): "Classification of samples into two or more ordered populations with application to a cancer trial," Stat. Med., 31(28), 3773–3786.

Dettling, M. (2005): "Bagboosting for tumor classification with gene expression data," Bioinformatics, 20, 3583–3593.

Dudoit, S., J. Fridlyand and T. P. Speed (2002): "Comparison of discrimination methods for the classification of tumor using gene expression data," J. Am. Stat. Assoc., 97, 77–87.

Efron, B. (1983): "Estimating the error rate of a prediction rule: Improvement on cross-validation," J. Am. Stat. Assoc., 78, 316–331.

Efron, B. and R. Tibshirani (1997): "Improvement on cross-validation: the 632+bootstrap method," J. Am. Stat. Assoc., 92, 548–560.

Faraggi, D. and B. Reiser (2002): "Estimation of the area under the ROC curve," Stat. Med., 21(20), 3093–3106.

Fernández, M. A., C. Rueda and B. Salvador (2006): "Incorporating additional information to normal linear discriminant rules," J. Am. Stat. Assoc., 101, 569–577.

Fu, W. J., R. J. Carroll and S. Wang (2005): "Estimating misclassification error with small samples via bootstrap cross-validation," Bioinformatics, 21, 1979–1986.

Graf, A. C. and P. Bauer (2009): "Model selection based on FDR-thresholding optimizing the area under the ROC-curve," Stat. Appl. Genet. Mol. Biol., 8(1), 1–20.

Kim, J. H. (2009): "Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap," Comput. Stat. Data An., 53(11), 3735–3745.

Kim, J. and E. Cha (2006): "Estimating prediction errors in binary classification problem: Cross-validation versus bootstrap," Korean Commun. Stat., 13, 151–165.

Kim, K. I. and R. Simon (2011): "Probabilistic classifiers with high-dimensional data," Biostatistics, 12(3), 399–412.

Lachenbruch, P. and M. Mickey (1968): "Estimation of error rates in discriminant analysis," Technometrics, 10, 167–178.

Lee, J. W., J. B. Lee, M. Park and S. H. Song (2005): "An extensive comparison of recent classification tools applied microarray data," Comput. Stat. Data An., 48(4), 869–885.

Lehmann, E. L. and G. Casella (1998): Theory of Point Estimation, 2nd edition. New York: Springer-Verlag.

Lin, D., Z. Shkedy, D. Yekutieli, T. Burzykowski, H. W. H. Göhlmann, A. De Bondt, T. Perera, T. Geerts and L. Bijnens (2007): "Testing for trends in dose-response microarray experiments: a comparison of several testing procedures, multiplicity and resampling-based inference," Stat. Appl. Genet. Mol. Biol., 6(1), article 26.

Long, T. and R. D. Gupta (1998): "Alternative linear classification rules under order restrictions," Commun. Stat. A-Theor, 27, 559–575.

McLachlan, G. J. (1976): "The bias of the apparent error rate in discriminant analysis," Biometrika, 63, 239–244.

Molinaro, A. M., R. Simon and R. M. Pfeiffer (2005): "Prediction error estimation: a comparison of resampling methods," Bioinformatics, 15, 3301–3307.

Oh, M. S. and D. W. Shin (2011): "A unified Bayesian inference on treatment means with order constraints," Comput. Stat. Data An., 55(1), 924–934.

Pepe, M. S., H. Janes, G. Longton, W. Leisenring and P. Newcomb (2004): "Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker," Am. J. Epidemiol., 159, 882–890.

Pepe, M. S., T. Cai and G. Longton (2006): "Combining predictors for classification using the area under the receiver operating characteristic curve," Biometrics, 62(1), 221–229.

Robertson, T., F. T. Wright and R. L. Dykstra (1988): Order Restricted Statistical Inference, New York: Wiley.

Salvador, B., M. A. Fernández, I. Martn and C. Rueda (2008): "Robustness of classification rules that incorporate additional information," Comput. Stat. Data An., 52(5), 2489–2495.

Schiavo, R. A. and D. J. Hand (2000): "Ten more years of error rate research," Int. Stat. Rev., 68, 295–310.

Silvapulle, M. J. and P. K. Sen (2005): Constrained Statistical Inference, New Jersey: John Wiley & Sons.

Simmons, S. and S. D. Peddada (2007): "Order-restricted inference for ordered gene expression (ORIOGEN) data under heteroscedastic variances," Bioinformation, 1, 414–419.

Steele, B. M. and D. A. Patterson (2000): "Ideal bootstrap estimation of expected prediction error for k-nearest neighbor classifiers: applications for classification and error assessment," Stat. Comput., 10(4), 349–355.

Tong, T., L. Chen and H. Zhao (2012): "Improved mean estimation and its application to diagonal discriminant analysis," Bioinformatics, 28(4): 531–537.

UICC (2009): TNM Classification of Malignant Tumours, 7th edition. New Jersey: Wiley-Blackwell.

Wehberg, S. and M. Schumacher (2004): "A comparison of nonparametric error rate estimation methods in classification problems," Biometrical J., 46, 35–47.