Una propuesta plug-in robusta para el problema de Discriminación bajo un modelo de Componentes Principales Comunes para las matrices de Covarianza

María Elena García García
Universidad de Ciencias Exactas de Buenos Aires,
Argentina

Tesis de Licenciatura para la carrera de Matemática Aplicada

Introducción

El análisis discriminante es uno de los métodos más usados en Estadística Multivariada y tiene como fin obtener reglas de clasificación de individuos. A diferencia del análisis de cluster, en el análisis discriminante se trabaja con poblaciones cuyos individuos pertenecen a distintos grupos y dicho grupo es conocido. Al contar con esta información se pueden medir los errores que se están cometiendo y tratar de minimizarlos. La regla de clasificación que se obtenga podrá ser utilizada posteriormente para asignar nuevas observaciones a alguno de estos grupos.
El problema de clasificación es importante dado que aparece frecuentemente en diversas ámbitos cotidianos:

- El médico que está haciendo un diagnóstico para decirle al paciente si sufre una enfermedad o no.
- Una entidad financiera que tiene que decidir si le otorga un crédito a un nuevo cliente.
- Una empresa que va a enviar una carta ofreciendo un producto a un grupo seleccionado de clientes suponiendo que estos podrán adquirirlo.
- Una universidad que tiene que decidir a que estudiante otorgarle una beca, etc.

Fisher fue el que introdujo el término discriminante en 1936, para distinguir entre 2 especies de flores de Iris: "Setosaz "Versicolor". En el capítulo 1 haremos un repaso del método y de las reglas que llevan su nombre suponiendo que las poblaciones tienen varianzas iguales o diferentes. Si los datos con los que se trabajan, siguen una distribución normal, estos métodos son los mejores en el sentido de minimizar los errores de mala clasificación (ver Welch (1939), Smith (1947)) cuya estimación también se verá en este capítulo.

Cuando se está bajo el supuesto de un modelo paramétrico y no se puede suponer igualdad de las matrices de varianza, Flury y Schmid (1992) propusieron poner restricciones sobre las mismas, para así tener que estimar menos parámetros y mejorar la variabilidad de los mismos. En este trabajo, supondremos que estamos bajo el modelo de componentes principales (CPC) comunes cuya definición se dará en el capítulo 2.1.

En el capítulo 3, se describen los algoritmos que permiten el cálculo de los estimadores de los ejes principales y sus varianzas bajo el modelo CPC.

Como veremos en el capítulo 4, la regla discriminante obtenida puede ser altamente afectada ante la presencia de unos pocos outliers por lo cual haremos una revisión de algunos de los métodos robustos existentes. En el capítulo 5 veremos el caso particular de la regla discriminante propuesta por Flury para después presentar nuestra propuesta basada en estimadores de escala y posición robustos. Finalmente, mediante un estudio de Monte Carlo compararemos los resultados de las reglas discriminantes basadas en el estimador de escala clásico y en el de Stahel–Donoho que tiene alto punto de ruptura en altas dimensiones, en el caso en que el modelo CPC sea válido .

Capítulo 1

Análisis Discriminante

El análisis discriminante es un método de Estadística Multivariada que consiste en determinar una regla de clasificación a partir de conjuntos de observaciones. Este método está dentro de los que se denominan de aprendizaje supervisado, ya que para cada observación se conoce la respuesta que se desea obtener. Dada una regla, podemos comparar el resultado obtenido contra el verdadero, supervisarlo y optimizarlo. El análisis de clusters no es supervisado dado que no hay grupos definidos a priori.

Uno de los ejemplos más conocidos dentro del análisis discriminante es el de las flores Iris presentado por Fisher (1936). A partir de las mediciones en 50 flores de tipo Setosa y 50 Versicolor con respecto a:

- largo del sépalo
- ancho del sépalo
- largo del pétalo
- ancho del pétalo,

Fisher (1936) dio la regla para clasificar ambos conjuntos y ante una nueva observación saber a qué grupo asignarlo. Fisher (1936) presentó el problema como la búsqueda de una combinación lineal de las variables que maximizara el cociente entre la diferencia de las medias y el desvío standard.

$$\frac{(a'\mu_1 - a'\mu_2)^2}{a'\Sigma a}$$

Como se puede encontrar en Johnson (1998), el máximo se alcanza para cualquier vector proporcional a $a_0 = (\mu_1 - \mu_2)^{\mathrm{T}}\Sigma^{-1}$. La regla que Fisher (1936) propuso para clasificar nuevas

observaciones consiste en asignar la misma al grupo para el cual la distancia a la media de ese grupo (de los proyectados) es menor.

Calculemos estos valores para el ejemplo de Iris para separar entre Setosa y Versicolor. Los grupos tienen $n_1 = n_2 = 50$ observaciones y sus matrices de covarianza estimadas son:

$$S_1 = \begin{pmatrix} 12{,}42 & 9{,}92 & 1{,}64 & 1{,}03 \\ 9{,}92 & 14{,}37 & 1{,}17 & 0{,}93 \\ 1{,}64 & 1{,}17 & 3{,}02 & 0{,}61 \\ 1{,}03 & 0{,}93 & 0{,}61 & 1{,}11 \end{pmatrix}, \qquad S_2 = \begin{pmatrix} 26{,}64 & 8{,}52 & 18{,}29 & 5{,}58 \\ 8{,}52 & 9{,}85 & 8{,}27 & 4{,}12 \\ 18{,}29 & 8{,}27 & 22{,}08 & 7{,}31 \\ 5{,}58 & 4{,}12 & 7{,}31 & 3{,}91 \end{pmatrix}$$

y los estimadores de posición son:

$$\widehat{\mu}_1 = \begin{pmatrix} 50{,}06 & 34{,}28 & 14{,}62 & 2{,}46 \end{pmatrix} \qquad \widehat{\mu}_2 = \begin{pmatrix} 59{,}36 & 27{,}70 & 42{,}60 & 13{,}26 \end{pmatrix}$$

Fisher (1936) supuso que las varianzas eran iguales por lo cual el estimador utilizado es:

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

Eligiendo la primera coordenada de $a = 1$ la combinación lineal obtenida es:

$$y = x_1 + 5{,}90x_2 + 7{,}13x_3 - 10{,}10x_4$$

La discriminación del gráfico 1.1 se realizó teniendo en cuenta sólo la información referente al ancho y al largo del pétalo. En dicho caso $y = x_3 - 0{,}9x_4$.

Aunque Fisher (1936) introdujo el método buscando la combinación lineal que maximiza las distancias entre las medias, la forma más usual de definir este método es definiendo regiones de clasificación. La solución en este caso es la que se puede observar en la Figura 1.2. Para este caso fue posible encontrar 2 regiones que separaban perfectamente los datos. En general, esto no va a ser posible y lo que se debería buscar es minimizar los errores de mala clasificación.

Tomemos otro ejemplo con 2 poblaciones, ahora se tienen familias que:

- Poseen cortadoras de cesped montables
- No las poseen

Un fabricante de este tipo de cortadoras de cesped va a lanzar una campaña de venta y quiere identificar las familias factibles de comprarlas. Para la clasificación cuenta con la siguiente información:

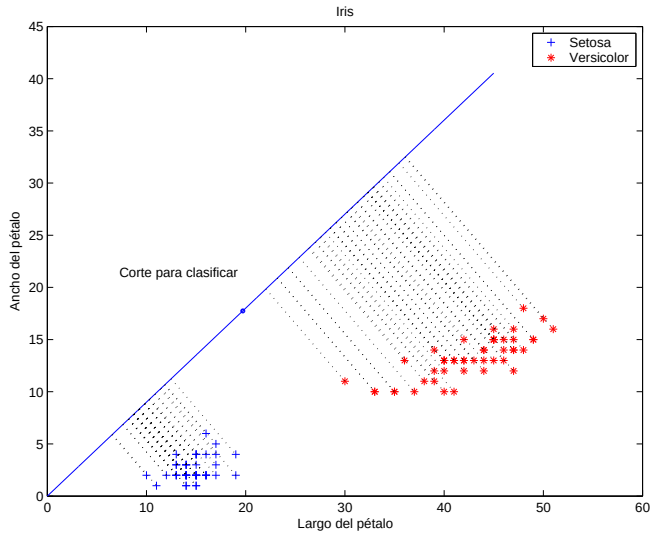

Figura 1.1: Regla Discriminante de Fisher para los datos de las Flores de Iris.

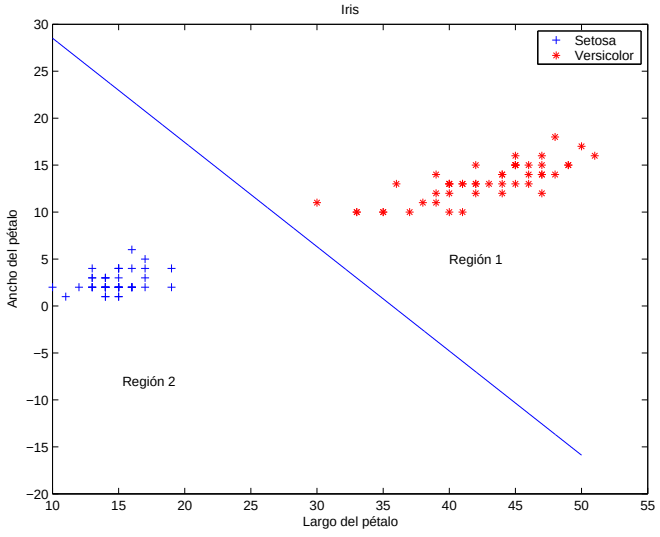

Figura 1.2: Regiones de clasificación para las Flores de Iris.

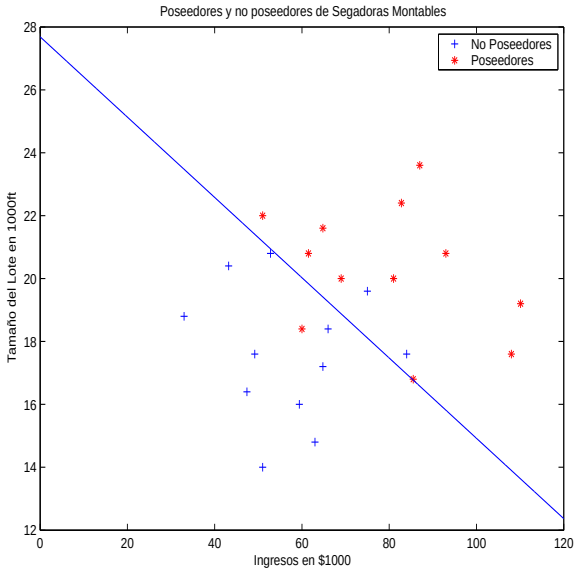

Figura 1.3: Propietarios de cortadoras de cesped montables.

- Ingresos de la familia
- Tamaño del lote

En la Figura 1.3 se puede observar que los poseedores de este tipo de cortadoras tienden a tener ingresos más altos. La recta está definiendo dos regiones. Las personas cuyas observaciones pertenezcan al semiplano superior, serán clasificadas como poseedoras, o posibles compradoras de este producto por lo que serían las familias a las que este fabricante se querría dirigir.

A continuación formalizaremos estos conceptos para ver como se deberían definir las regiones.

1.1. Dos grupos: Distribuciones conocidas

Supongamos que tenemos una población $\mathcal{P}$ con una proporción π_1 de observaciones del grupo $\mathcal{G}_1$ y $\pi_2 = (1 - \pi_1)$ observaciones del grupo $\mathcal{G}_2$. Sea $f_i(x)$ la probabilidad o la densidad de las observaciones x pertenecientes al grupo $\mathcal{G}_i$, $i = 1, 2$. Se da la siguiente regla de clasificación: Asignar x a $\mathcal{G}_i$ si x pertenece a R_i, donde R_1 y R_2 son excluyentes y $R_1 \bigcup R_2 = R$, el espacio muestral para $\mathcal{P}$. Los errores que podríamos cometer se indican en la Figura 1.4 y son:

- Asignar x a $\mathcal{G}_2$ cuando pertenece $\mathcal{G}_1$
- Asignar x a $\mathcal{G}_1$ cuando pertenece $\mathcal{G}_2$

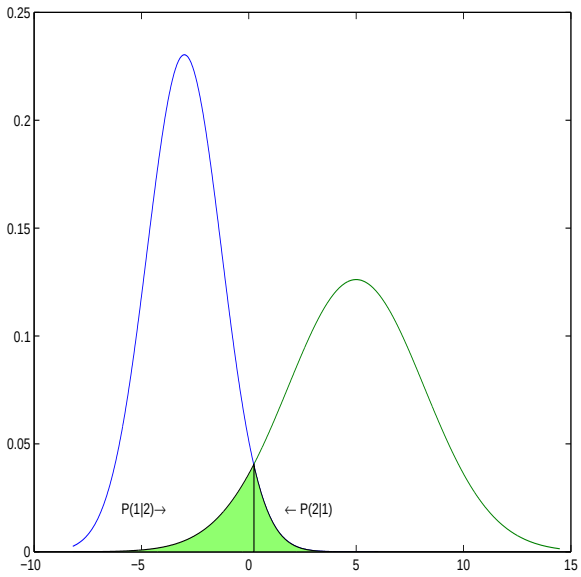

Figura 1.4: Errores de clasificación.

Si conocemos las densidades de cada grupo, podemos calcular las probabilidades asociadas a los mismos:

$$\mathrm{P}(2|1) = \int_{R_2} f_1(x)dx \quad y \quad \mathrm{P}(1|2) = \int_{R_1} f_2(x)dx \tag{1.1}$$

Debemos entonces definir criterios para elegir R_1 y R_2.

El siguiente lema, que puede hallarse en Seber (1984, pp. 281), nos permitirá construir las regiones R_1 y R_2.

Lema 1 . *La integral* $\int_{R_1} g(x)\ dx$ *se minimiza respecto de* R_1 *cuando*

$$R_1 = R_{01} = \{x : g(x) < 0\}$$

.

A continuación, detallaremos algunos criterios para la selección de las regiones.

a) *Minimización de la probabilidad total de mala clasificación*

Una forma de elegir R_1 y R_2 consiste en elegir las regiones que minimizan la probabilidad total de mala clasificación la cual está dada por:

$$\mathrm{P}(R, f) \;=\; \sum_{i=1}^{2} \Pr\left(\text{asignar incorrectamente } x \text{ a } \mathcal{G}_i\right) =$$

$$
\begin{aligned}
&= \sum_{i=1}^{2}\sum_{\substack{j=1\\ j\neq i}}^{2} \Pr\left(\text{asignar } x \text{ a } \mathcal{G}_i | x \in \mathcal{G}_j\right) \Pr\left(x \in \mathcal{G}_j\right) \\
&= \mathrm{P}(1|2)\pi_2 + \mathrm{P}(2|1)\pi_1 \qquad (1.2)
\end{aligned}
$$

A partir de (1.1) y (1.2) obtenemos

$$
\begin{aligned}
\mathrm{P}(R,f) &= \mathrm{P}(1|2)\pi_2 + \mathrm{P}(2|1)\pi_1 = \pi_2 \int_{R_1} f_2(x)dx + \pi_1 \int_{R_2} f_1(x)dx \\
&= \pi_2 \int_{R_1} f_2(x)dx + \pi_1 \int_{R-R_1} f_1(x)dx = \pi_2 \int_{R_1} f_2(x)dx + \pi_1 \left(1 - \int_{R_1} f_1(x)dx\right) \\
&= \int_{R_1} (\pi_2 f_2(x) - \pi_1 f_1(x))dx + \pi_1
\end{aligned}
$$

Con lo cual, usando el Lema 1 con $g = \pi_2 f_2(x) - \pi_1 f_1(x)) \leq 0$ resulta que

$$
R_{01} = \left\{ x : \quad \frac{f_1(x)}{f_2(x)} \geq \frac{\pi_2}{\pi_1} \right\} . \qquad (1.3)
$$

La regla de clasificación debida a Welsh (1939) asigna x a $\mathcal{G}_1$ si $x \in R_{01}$.

b) *Minimización del costo total de mala clasificación.*

Si suponemos que clasificar incorrectamente una observación del grupo 1 tiene un costo diferente a clasificar mal una del grupo 2, deberíamos minimizar el costo total de mala clasificación definido por

$$
C_T = C(1|2)\mathrm{P}(1|2)\pi_2 + C(2|1)\mathrm{P}(2|1)\pi_1
$$

Utilizando nuevamente el Lema 1, en forma análoga a la anterior, se obtiene

$$
R_{01} = \left\{ x : \quad \frac{f_1(x)}{f_2(x)} \geq \frac{\pi_2 C(2|1)}{\pi_1 C(1|2)} \right\} \qquad (1.4)
$$

que se reduce a (1.3) si los costos son iguales.

c) *Maximización de la probabilidad a posteriori.*

Si fueran conocidas las probabilidades a priori π_i, sería posible aplicar el teorema de Bayes y obtener la probabilidad a posteriori como

$$
q_i(x) = \mathrm{P}(\mathcal{G}_i|x) = \frac{\pi_i f_i(x)}{\sum_j \pi_j f_j(x)}
$$

Un criterio razonable asigna x a $\mathcal{G}_1$ si $q_1(x) > q_2(x)$, o sea, se asigna x al grupo que maximiza la probabilidad a posteriori. Por lo tanto, se da la siguiente regla de clasificación

$$\begin{aligned} x \in R_i \quad & \text{si} \quad q_i(x) > q_j(x) \\ \text{o sea} & \\ x \in R_i \quad & \text{si} \quad \pi_i f_i(x) > \pi_j f_j(x) \ . \end{aligned}$$

Como puede observarse esta regla de clasificación coincide con la definida en (1.3).

1.1.1. Distribución Normal con matrices de covarianza iguales

Supongamos que $X|\mathcal{G}_i \sim \mathbf{N}_p(\mu_i, \Sigma)$, es decir que

$$f_i(x) = (2\pi)^{-p/2} \left|\Sigma\right|^{-1/2} \exp\left[\frac{-1}{2}(x-\mu_i)^{\mathrm{T}}\Sigma^{-1}(x-\mu_i)\right], \quad x \in I\!R^p \ .$$

La condición (1.4) para densidades normales queda

$$\exp\left[(\mu_i-\mu_j)^{\mathrm{T}}\Sigma^{-1}x - \frac{1}{2}(\mu_i-\mu_j)^{\mathrm{T}}\Sigma^{-1}(\mu_i+\mu_j)\right] > \frac{\pi_j C(j|i)}{\pi_i C(i|j)} \ .$$

Llamando $k_{ij} = \frac{\pi_j C(j|i)}{\pi_i C(i|j)}$ y tomando logaritmo, obtenemos que clasificamos a x en $\mathcal{G}_i$ si

$$(\mu_i-\mu_j)^{\mathrm{T}}\Sigma^{-1}x > \frac{1}{2}(\mu_i-\mu_j)^{\mathrm{T}}\Sigma^{-1}(\mu_i+\mu_j) + \log(k_{ij}) \quad \forall j \neq i \tag{1.5}$$

La regla de discriminación es lineal en x y en el caso de dos poblaciones, puede escribirse de la siguiente forma: asigne x a $\mathcal{G}_1$ si $\alpha^{\mathrm{T}}x > \frac{1}{2}\left(\alpha^{\mathrm{T}}\mu_1 + \alpha^{\mathrm{T}}\mu_2\right) + \log(k_{12})$ donde $\alpha = \Sigma^{-1}(\mu_1-\mu_2)$.

1.1.2. Distribución Normal con matrices de covarianza distintas

Ahora supongamos que $X|\mathcal{G}_i \sim N_p(\mu_i, \Sigma_i)$, es decir que la condición (1.4) pasa a ser

$$\frac{\left|\Sigma_i\right|^{-1/2}}{\left|\Sigma_j\right|^{-1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(x-\mu_i) + \frac{1}{2}(x-\mu_j)^{\mathrm{T}}\Sigma_j^{-1}(x-\mu_j)\right] > k_{ij}$$

con $k_{ij} = \frac{\pi_j C(j|i)}{\pi_i C(i|j)}$. Nuevamente, tomando el logaritmo, clasificamos a x en $\mathcal{G}_i$ si

$$-\frac{1}{2}\log\frac{\left|\Sigma_i\right|}{\left|\Sigma_j\right|} - \frac{1}{2}(x-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(x-\mu_i) + \frac{1}{2}(x-\mu_j)^{\mathrm{T}}\Sigma_j^{-1}(x-\mu_j) > \log(k_{ij}) \quad \forall j \neq i \tag{1.6}$$

Como vemos la regla de discriminación es, a diferencia de la anterior, cuadrática en x.

1.2. Discriminación con poblaciones paramétricas desconocidas

1.2.1. Matrices de Varianza iguales

En general cuando se enfrenta un problema de discriminación no se conocen las ditribuciones o alguno de los parámetros de las distribuciones. Podría saberse que $X \in \mathcal{G}_i$ tienen una distribución Normal pero no conocerce los parámetros μ_i, Σ_i. Wald (1944) y Anderson (1984) sugirieron reemplazar los parámetros poblacionales por los muestrales: Si se tienen n_i observaciones de $\mathcal{G}_i$ que se guardan en una matriz X_i, cuyas filas corresponden a las n_i observaciones y las columnas a las p variables

$$X_i = \begin{pmatrix} x_{i11} & x_{i12} & \dots & x_{i1p} \\ & & \dots & \\ x_{in_i1} & x_{in_i2} & \dots & x_{in_ip} \end{pmatrix} \qquad = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \dots \\ X_{in_i} \end{pmatrix}$$

Los estimadores de la media y de la matriz de covarianza están dados por

$$\widehat{\mu}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i} \qquad S_i = \frac{\sum_{j=1}^{n_i} (X_{ij} - \widehat{\mu}_i)(X_{ij} - \widehat{\mu}_i)^{\mathrm{T}}}{n_i - 1}$$

Al estar bajo el supuesto que los datos tienen la misma matriz de covarianza, podemos utilizar como estimador de la misma

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

Con lo cual reemplazando estos parámetros en (1.5), la regla obtenida clasifica x en $\mathcal{G}_i$ si

$$(\widehat{\mu}_i - \widehat{\mu}_j)^{\mathrm{T}} S^{-1} x - \frac{1}{2}(\widehat{\mu}_i - \widehat{\mu}_j)^{\mathrm{T}} S^{-1}(\widehat{\mu}_i + \mu_j) > \log(k_{ij}) \quad \forall j \neq i \ , \tag{1.7}$$

donde $k_{ij} = \frac{\pi_j C(j|i)}{\pi_i C(i|j)}$. Si tenemos $k = 2$ poblaciones, podríamos escribir esta regla de clasificación como

$$x \in \widehat{R_1} \qquad \text{si} \qquad L(x) = \widehat{\alpha}^{\mathrm{T}} x + \widehat{\alpha}_0 > \log(k_{12})$$

con

$$\widehat{\alpha} = S^{-1}(\widehat{\mu}_1 - \widehat{\mu}_2) \qquad \widehat{\alpha}_0 = -\frac{1}{2}\widehat{\alpha}^{\mathrm{T}}(\widehat{\mu}_1 + \widehat{\mu}_2) \ .$$

Observemos que $\widehat{\alpha}$ es el coeficiente obtenido en la propuesta de Fisher y provee un estimador consistente del vector $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$, definido anteriormente. $L(x)$ se denomina la función discriminante lineal.

1.2.2. Matrices de Varianza distintas

Al igual que en la sección anterior, podría saberse que $X \in \mathcal{G}_i$ tienen una distribución Normal con parámetros μ_i, Σ_i desconocidos y todos diferentes. Nuevamente se reemplaza en la regla (1.6) los parámetros por sus estimadores obteniéndose

$$Q(x) = -\frac{1}{2}\log\frac{|S_i|}{|S_j|} - \frac{1}{2}(x-\widehat{\mu}_i)^{\mathrm{T}} S_i^{-1}(x-\widehat{\mu}_i) + \frac{1}{2}(x-\widehat{\mu}_j)^{\mathrm{T}} S_j^{-1}(x-\widehat{\mu}_j) > \log(k_{ij}) \qquad (1.8)$$

con $k_{ij} = \frac{\pi_j C(j|i)}{\pi_i C(i|j)}$.

Si indicamos por $D_i(x) = (x-\widehat{\mu}_i)^{\mathrm{T}} S_i^{-1}(x-\widehat{\mu}_i)$ la distancia de Mahalanobis entre x y la media estimada $\widehat{\mu}_i$ para $\mathcal{G}_i$, la condición anterior puede escribirse como:

$$-D_i(x) + D_j(x) > 2\log(k_{ij}) + \log\frac{|S_i|}{|S_j|} \qquad (1.9)$$

Es decir, la observación x es clasificada en el grupo de cuya media está más próxima en el sentido de la distancia de Mahalanobis con una diferencia mayor a una constante. La regla cuadrática fue obtenida bajo normalidad por Smith (1947).

1.3. Análisis de errores

Una vez obtenida una relgla de clasificación, se debe poder validar su capacidad de asignación correcta o incorrecta a los distintos grupos. Para ello estudiaremos los distintos tipos de errores.

1. *Tasas de errores óptimos*

   En la sección 1.1, donde se suponía que la distribución de los grupos era conocida, vimos cuáles eran los posibles errores y cómo se calculaban. También vimos como encontrar la región que hacía mínimo dicho error. Se denomina tasa de error óptimo al error cometido al utilizar la región óptima R_{oj}, $j = 1, 2$.

   $$e_{i,opt} = \mathrm{P}(j|i) = \int_{R_{0j}} f_i(x|\theta_i)$$

   y (ver 1.2)

   $$e_{opt} = \pi_1 e_{1,opt} + \pi_2 e_{2,opt} = \mathrm{P}(R_o, f) \qquad (1.10)$$

2. *Tasas de errores verdaderos (actual errors)*

   Partimos de una muestra para estimar las regiones, y calculamos el error de estas regiones estimadas, $\widehat{R}_{0j}$, al cual denominaremos error verdadero

$$e_{i,act} = \mathrm{P}(j|i) = \int_{\widehat{R}_{0j}} f_i(x|\theta_i)$$

$$e_{act} = \pi_1 e_{1,act} + \pi_2 e_{2,act} = \mathrm{P}(\widehat{R}_o, f) \tag{1.11}$$

   Si un número elevado de observaciones son clasificadas usando $\widehat{R}_o$, entonces alrededor de $100e_{act}\,\%$ observaciones serán mal clasificadas. Observemos que $e_{opt} \leq e_{act}$

3. *Tasas esperadas de errores verdaderos*

   Al estar partiendo de una muestra, es de interés calcular la tasa esperada de los errores verdades:

$$\mathbf{E}[e_{act}] = \pi_1 \mathbf{E}[e_{1,act}] + \pi_2 \mathbf{E}[e_{2,act}] \tag{1.12}$$

En la mayoría de los casos, no se conocerán algunos parámetros de las densidades por lo cual no será posible calcular los errores anteriores. Por lo que será necesario estimar dichos errores.

1. *Estimaciones "plug–in"o de reemplazo*

   Partiendo del error verdadero, se pueden reemplazar los parámetros desconocidos por estimaciones de los mismos, y así obtener una estimación de los errores actuales.

$$\hat{e}_{i,act} = \mathrm{P}(j|i) = \int_{\widehat{R}_{0j}} f_i(x|\widehat{\theta}_i)$$

$$\hat{e}_{act} = \mathrm{P}(\widehat{R}_o, \widehat{f}) \tag{1.13}$$

2. *Tasas de errores aparentes*

   Esta tasa de error se calcula reclasificando la muestra a partir de la cual se obtuvo la regla discriminante y contando los individuos mal clasificados en cada grupo. Llamemos m_i a las observaciones del grupo $\mathcal{G}_i$ clasificadas incorrectamente y n_i al total de individuos de dicho grupo. Luego,

$$e_{i,app} = m_i/n_i \tag{1.14}$$

   y el error aparente total sería:

$$e_{app} = \pi_1 e_{1,app} + \pi_2 e_{2,app}\ . \tag{1.15}$$

Si π_1 y π_2 son desconocidos, y las $n = n_1 + n_2$ observaciones corresponden a una muestra aleatoria de una población $\mathcal{P}$, entonces se podrían estimar las proporciones π_i por $\widehat{\pi}_i = n_i/(n_1 + n_2)$ y la estimación de la tasa de error aparente quedará entonces

$$\hat{e}_{app} = \pi_1 \hat{e}_{1,app} + \pi_2 \hat{e}_{2,app} = \frac{n_1}{n_1 + n_2}\frac{m_1}{n_1} + \frac{n_2}{n_1 + n_2}\frac{m_2}{n_2} = \frac{m_1 + m_2}{n_1 + n_2}\,. \tag{1.16}$$

El problema de este error es que al utilizar la misma población tanto para obtener la regla como para validarla tiende a producir valores "optimistas". Lo ideal para calcular los errores aparentes sería tener una muestra suficientemente grande para poder dividirla en dos submuestras:

- **Muestra de entrenamiento**: para obtener la regla
- **Muestra de validación**: la cual se reclasificaría con la regla obtenida anteriormente y se calcularían los errores aparentes, obteniendo de esta forma estimaciones no sesgadas.

En general el número de observaciones disponibles no es tan elevado como para optar por este procedimiento, y si lo fuera, en general, se prefieren utilizar todos los datos en las estimaciones.

3. *Validación cruzada*

La necesidad de mejorar las estimaciones de los errores, dio origen a otros procedimientos. La técnica de validación cruzada, propuesta por Lachenbruch y Mickey (1968) consiste en:

Paso 1: Estimar la regla discriminante con todas las observaciones.

Para validar esta regla se procedería de la siguiente manera:

Paso 2: Se parte la muestra de entrenamiento en dos submuestras de tamaño $n-k_1$ y k_1. Con la primera se obtiene la regla discriminante con la cual se clasifican el resto de los k_1 individuos y se cuenta la cantidad de clasificaciones erróneas en cada grupo.

Paso 3: Se repite el paso 2 un número elevado de veces o tantas como sea posible es decir $\binom{n}{k_1}$. Finalmente, se calcula la media de las tasas de mala clasificación observada en cada grupo, $e_c(i)$ y la tasa de error actual total:

$$e_c = \pi_1 e_{1,c} + \pi_2 e_{2,c} \tag{1.17}$$

El valor más usual para k_1, y que fue el inicialmente propuesto, es uno, por lo cual se conoce la regla como "dejando uno afuera"("leaving one out"). En este caso, el error de mala clasificación en cada grupo está dado por:

$$e_{i,c} = \frac{a_i}{n_i} \tag{1.18}$$

y el error total:

$$e_c = \pi_1 e_{1,c} + \pi_2 e_{2,c} = \frac{a_1 + a_2}{n_1 + n_2} \tag{1.19}$$

Actualmente, gracias a los avances computacionales, es posible optar por otras opciones como $k_1 \simeq n/2$ conocida como "half cross-validation"o como la llama Andersen (1993) "double cross validation". Efron (1983) concluye que este método reduce sustancialmente el sesgo en las tasas aparentes pero que al mismo tiempo presenta una gran variabilidad, especialmente para $k_1 = 1$, por lo cual el autor propone un método bootstrap que reduce el sesgo y presenta menos dispersión.

4. *Tasas obtenidas por "bootstrap"*

Dado que el error $e_{i,app}$ es sesgado, Efron (1983) sugiere estimar el sesgo usando una técnica "bootstrap". En esta técnica una nueva muestra de tamaño $n_i \quad (i = 1, 2)$ se extrae con reposición de las n_i observaciones originales. La nueva muestra no es tan sólo una permutación aleatoria de los datos originales dado que la muestra es con reposición. De estas dos nuevas muestras se obtiene una nueva regla de clasificación. Si el muestraje hubiera sido mixto, el submuestreo también. Llamemos m_i^* al número de observaciones mal clasificadas de la nueva muestra por la nueva regla y m_i^{**} a las mal clasificadas de la muestra original. Se define:

$$d_i = \frac{m_i^{**} - m_i^*}{n_i} \tag{1.20}$$

Se repite el proceso un número elevado de veces (digamos alrededor de 100 veces) y sea $\overline{d_i}$ la media de los d_i de dichas réplicas. Las estimaciones "bootstrap"de las probabilidades de mal clasificación por grupo son:

$$e_{boot}(i) = \frac{m_i}{n_i} + \overline{d_i} \tag{1.21}$$

y la probabilidad total

$$e_{boot} = \pi_1 e_{boot}(1) + \pi_2 e_{boot}(2) \tag{1.22}$$

McLachlan (1980) mostró que la estimación del sesgo para $e_{i,app}$ dada por $\overline{d_i}$ es eficiente.

El inconveniente de las estimaciones por reemplazo o "plug–in" $\hat{e}_{i,act}$ es que dependen fuertemente de la correcta especifiación de las densidades f_i. Además, tienen un mal comportamiento para muestras pequeñas.

Por último, se tiene la siguiente propiedad : $\mathbf{E}[\hat{e}_{act}] \leq e_{opt} < \mathbf{E}_{\widehat{\theta}_i}[e_{act}] = \mathbf{E}[e_{act}]$ cuya demostración se puede encontrar en el Seber (1984, pp. 290).

1.3.1. Más de dos grupos

Supongamos que tenemos una población $\mathcal{P}$ compuesta por k grupos independientes y sea π_i la proporción de observaciones pertenecientes al Grupo $\mathcal{G}_i$ $(1 \leq i \leq k, \sum_{i=1}^{k} \pi_i = 1)$. Lo expuesto anteriormente para el caso de 2 poblaciones se generaliza para el caso k. Si f_i es la densidad de las observaciones del grupo $\mathcal{G}_i$, buscamos una partición $R_1, R_2, \ldots, R_k$ del espacio R y y la regla de clasificación asignará una observación x al grupo $\mathcal{G}_i$ si $x \in R_i$. La probabilidad de clasificar incorrectamente un individuo en $\mathcal{G}_j$ cuando en realidad pertenece a $\mathcal{G}_i$ está dada por:

$$\mathrm{P}(j|i) = \int_{R_j} f_i(x)dx \tag{1.23}$$

Supongamos que el costo asociado de clasificar incorrectamente un individuo es diferente entre los grupos, llamemos $C(j|i)$ al costo de asignar incorrectamente una observación del grupo $\mathcal{G}_i$ a $\mathcal{G}_j$. Por definición, $C(i|i) = 0$.

Luego, el costo de mal clasificar $x \in \mathcal{G}_i$ está dada por:

$$C(i) = \sum_{j=1, j\neq i}^{k} C(j|i)\mathrm{P}(j|i) \tag{1.24}$$

y el costo total de mala clasificación es:

$$C_T = \sum_{i=1}^{k} \pi_i C(i) \tag{1.25}$$

Las regiones de clasificación que minimizan C_T, están definidas asignando x al grupo $\mathcal{G}_i$, $1 \leq i \leq k$, para la cual

$$\sum_{j=1, j\neq i}^{k} \pi_j f_j(x) C(i|j) \tag{1.26}$$

es mínima (la demostración se puede encontrar en Anderson (1984)).

Si suponemos que el costo de mal clasificar una observación en otro grupo es para todos igual, (1.26) se va a minimizar cuando el término omitido $\pi_i f_i(x)$ sea máximo. Con lo cual, en el caso de k poblaciones la regla de discriminación que minimiza el costo total de mala clasificación consiste en asignar x a $\mathcal{G}_i$ cuando

$$\pi_i f_i(x) > \pi_j f_j(x) \quad \forall j \neq i \; .$$

Es interesante observar, que como en el caso de dos poblaciones, esta regla de clasificación coincide con la que maximiza la probabilidad a posteriori.

Las reglas de discriminación para el caso normal se obtienen como en el caso de dos poblaciones.

Capítulo 2

Restricciones en las matrices de covarianza de las poblaciones

Cuando se utiliza la técnica de análisis discriminante se llega a la regla discriminante lineal o cuadrática según se suponga que las matrices de covarianza de los poblaciones son todas iguales (LD) o todas distintas (QD).

En las situaciones en las que hay muchos parámetros a estimar, una forma de reducir la varianza de los estimadores es agregando restricciones válidas al espacio de parámetros. Los resultados obtenidos al utilizar el método de discriminación lineal, cuando este es válido, superan los obtenidos bajo una discriminación cuadrática sin restricciones en las matrices de covarianza (Seber, 1984). Aún en situaciones en las cuales el método sea teóricamente incorrecto, este puede superar al método correcto, cuadrático, en términos de tasas de errores esperados. Marks y Dunn (1974) compararon los resultados de las funciones discriminantes lineal y cuadrática en los casos asintóticos y para muestras pequeñas tanto en el caso de diferencias proporcionales y no proporcionales en Σ_i. El estudio mostró que en el caso de muestras chicas $n_1, n_2 < 25$, la función cuadrática se comporta peor que la lineal si $\Sigma_1 \approx \Sigma_2$ y p es moderadamente grande ($p > 6$). Flury (1988), propuso una jerarquía de las matrices de covarianza comenzando desde la igualdad de las mismas, finalizando con la desigualdad y en los casos intermedios proponiendo restricciones. Con esta jerarquización, mostró en otro trabajo (Flury y Schmid, 1992) , que poner condiciones válidas en los modelos reduce la variabilidad de los estimadores. Aunque su trabajo es asintótico, hace mención al caso de muestras pequeñas para el cual espera que usando un modelo parsimonioso se mejoren aún más los resultados. Debido a la dificultad matemática de este problema, la mayoría de los resultados mostrados son en base a simulaciones. Una de las excepciones es el trabajo de O'Neill (1984).

Por lo mencionado anteriormente es de interés el estudio de modelos intermedios. A continuación se da la jerarquía de similaridad entre k matrices de covarianza $\Sigma_1 \dots \Sigma_k$ de dimensión $p \times p$, con Σ_i definidas positivas, que Flury (1988) detalla en su libro:

N1. Todas las Σ_i son iguales.
El número de parámetros a ser estimados es de $p(p+1)/2$.

N2. Las Σ_i son proporcionales, es decir, $\exists\ \rho_2, \ldots, \rho_k > 0$ tales que

$$\Sigma_i = \rho_i \Sigma_1 \quad i = 2, \ldots, k$$

El número de parámetros a ser estimados es de $p(p+1)/2 + k - 1$.

N3. El modelo de Componentes Principales Comunes (CPC)

$$\Sigma_i = \beta \Lambda_i \beta^{\mathrm{T}} \qquad i = 2, \ldots, k$$

donde β es una matriz ortogonal de orden p y $\Lambda_i = \mathrm{diag}(\lambda_{i1}, \ldots, \lambda_{ip})$. El número de parámetros a ser estimados es de $p(p-1)/2$ para las matrices ortogonales β y kp para la matrices diagonales, o sea un número total de parámetros igual a $p(p-1)/2 + kp$.

N4. El modelo parcial CPC (ver más detalles en Flury (1984)).

N5. $\Sigma_1 \ldots \Sigma_k$ matrices de covarianza arbitrarias.
El número de parámetros a ser estimados bajo esta hipótesis es $kp(p+1)/2$.

En la siguiente sección ampliaremos sobre el modelo CPC que es en el que se basa esta tesis. Antes introduciremos el modelo de Componentes Principales para una muestra, para después generalizarlo a k muestras resultando en el modelo CPC de interés.

2.1. Componentes Principales Comunes

El modelo CPC es una generalización de componentes principales para varias muestras. El supuesto principal es que la transformación de componentes principales es idéntica en las k poblaciones consideradas, mientras que las varianzas asociadas a las componente pueden variar entre los grupos. Antes de definir los estimadores bajo un modelo CPC, veremos el caso de una población.

2.1.1. Componentes Principales

En el caso de una única población, el modelo de componentes principales se puede introducir de 3 formas diferentes:

1. Un método que transforma variables correlacionadas en otras no correlacionadas.

2. Un método que busca transformaciones lineales que se vayan quedando con la mayor variabilidad.

3. Es un método de reducción de la dimensión que puede ayudar a la interpretación de los datos.

El primer criterio por sí sólo, no define de manera única a las componentes principales: hay muchas formas de transformar p variables correlacionadas en p no correlacionadas. Por ejemplo, si hacemos la descomposición de Cholesky de la matriz de covarianza $\Sigma = TT'$ y transformamos las variables mediante la transformación lineal asociada a T. Por esta razón, la forma más frecuente de introducirlas es una combinación de los dos primeros criterios. Es decir, buscar transformaciones lineales de las variables originales de forma tal que resulten no correlacionadas y a su vez puedan explicar la mayor variabilidad posible de los datos. El tercer criterio podría ponerse como objetivo, una vez encontradas las componentes, elegir el subconjunto de transformaciones para las cuales la pérdida de información sea la menor posible. Este concepto fue el utilizado por Pearson (1901) en uno de los primeros acercamiento a este tema aunque fue Hotelling (1933) quien lo desarrolló.

En general, este método no se lo utiliza como un fin en sí mismo sino que se usan los datos resultantes como información de partida en la aplicación de otros métodos. Por ejemplo, en regresión lineal una de las hipótesis es que los datos no estén correlacionados, en caso de no cumplirse la misma, el análisis discriminante podría utilizarse como paso previo.
Al ser un método de reducción de la dimensión, es ampliamente utilizado en el análisis de datos con muchas variables, problema que es frecuente encontrar en áreas como quimioterapia ('chemometrics'), ingeniería, genética, etc. El análisis de componentes principales es usualmente el primer paso en el análisis de datos, seguido por análisis discriminante, análisis de clusters y otras técnicas multivariadas.

A continuación veremos como encontrar las componentes principales.

Componentes Principales de una población

Supongamos que tenemos el vector aleatorio $X^{\mathrm{T}} = (x_1, x_2 \dots x_p)$ con matriz de covarianza Σ. Como ya mencionamos, las Componentes Principales son las combinaciones lineales de estas p variables aleatorias:

$$\begin{array}{rcl} Y_1 = & a_1^{\mathrm{T}} X = & a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p \\ Y_2 = & a_2^{\mathrm{T}} X = & a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p \\ \vdots & & \vdots \\ Y_p = & a_p^{\mathrm{T}} X = & a_{p1}x_1 + a_{p2}x_2 + \ldots + a_{pp}x_p \end{array} \tag{2.1}$$

donde la elección de los a_i se realiza de modo de maximizar la varianza que está dada por:

$$Var(Y_i) = a_i^{\mathrm{T}} \Sigma a_i \qquad i = 1, 2, \ldots, p \tag{2.2}$$

y además, resulten no correlacionadas. Es decir que también se pondrán condiciones sobre:

$$Cov(Y_i, Y_k) = a_i^{\mathrm{T}} \Sigma a_k \qquad i, k = 1, 2, \ldots, p \tag{2.3}$$

Geometricamente, estas combinaciones lineales representan la selección de un nuevo sistema de coordenadas, en el cual los ejes respresentan la dirección de máxima variabilidad. Las Componentes principales solo dependen de la matriz de covarianza Σ. Para obtenerlas no es necesario el supuesto de que la población tenga una distribución normal multivariada, aunque bajo esta condición las componentes tienen útiles interpretaciones en relación a los elipsoides de densidad constante. Además se pueden hacer inferencias de sus estimadores de máxima verosimilitud.

Primer Componente Principal: se elige como la combinación lineal:

$$Y_1 = a_1^{\mathrm{T}} X = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p$$

donde el a elegido es el que maximiza la varianza (2.2).
Dado que la misma crece al tomar múltiplos de un vector, el conjunto sobre el que se maximiza son los a tales que $\|a\| = 1$

$$Var(Y_1) = a_1^{\mathrm{T}} \Sigma a_1 = \max_{\{a \in I\!R^p:\ \ \|a\|=1\}} a^{\mathrm{T}} \Sigma a$$

Segunda Componente Principal:

$$Y_2 = a_2^{\mathrm{T}} X = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p$$

Nuevamente a se elige para que maximice la varianza (2.2). Pero ahora hay que agregar la condición para que resulte no correlacionado con Y_1.

$$Var(Y_2) = a_2^{\mathrm{T}} \Sigma a_2 = \max_{\{a \in I\!R^p : \|a\|=1\ ,\ a^{\mathrm{T}} \Sigma a_1 = 0\}} a^{\mathrm{T}} \Sigma a$$

$j-$ésima Componente Principal:

$$Y_j = a_j^{\mathrm{T}} X = a_{j1}x_1 + a_{j2}x_2 + \ldots + a_{jp}x_p$$

donde a_j se elije

$$Var(Y_j) = a_j^{\mathrm{T}} \Sigma a_j = \max_{\{a \in I\!R^p : \|a\|=1\ ,\ a^{\mathrm{T}} \Sigma a_i \text{ para } i<j\}} a^{\mathrm{T}} \Sigma a$$

Dado que Σ es una matriz simétrica y definida positiva, por el teorema de descomposción espectral $\Sigma = \beta\Lambda\beta^{\mathrm{T}}$ donde los valores de la matriz diagonal $\Lambda : \lambda_1 \geq \ldots \geq \lambda_p$ son los autovalores asociados a las columnas de β que son los autovectores de Σ. Veamos que esta elección es la que define las componentes principales. Dado que $\beta_1 \ldots \beta_p$ definen una base ortonormal de $I\!R^p$ $a = \sum_{i=1}^{p} \alpha_i\beta_i$, $\alpha_i \in I\!R$. Como β_j son ortogonales $\alpha^{\mathrm{T}}\alpha = 1$.
La varianza de $a^{\mathrm{T}}X$ es:

$$var(a^{\mathrm{T}}X) = a^{\mathrm{T}}\Sigma a = \sum(\alpha_i\beta_i^{\mathrm{T}}\Sigma\beta_i\alpha_i) = \sum_{i=1}^{p}\alpha_i^2\lambda_i \leq \lambda_1\sum_{i=1}^{p}\alpha_i^2 = \lambda_1 \ .$$

Por lo tanto, la primer componente principal es el auvector correspondiente al mayor autovalor. Veamos por inducción que si vale para $h-1$ entonces vale para h. Queremos maximizar la varianza de $a^{\mathrm{T}}X$ sujeto a que $\|a\| = 1$ y $a^{\mathrm{T}}\Sigma\beta_j = 0$ para $j < h$. Como β_j es autovector de Σ la segunda condición queda $a^{\mathrm{T}}\beta_j = 0$ por lo tanto, $a = \sum_{i=1}^{p}\alpha_i\beta_i = \sum_{i=h}^{p}\alpha_i\beta_i$ y

$$var(a^{\mathrm{T}}X) = a^{\mathrm{T}}\Sigma a = \sum_{i=h}^{p}(\alpha_i\beta_i^{\mathrm{T}}\Sigma\beta_i\alpha_i) = \sum_{i=h}^{p}(\alpha_i)^2\lambda_i \leq \lambda_h\sum_{i=h}^{p}(\alpha_i)^2 \leq \lambda_h$$

Por lo tanto, el $j-$ésimo autovector de Σ es la $j-$ésima componente principal.

Las Componentes principales de X, se definen como el vector aleatorio $p-$variado

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix} = \beta^{\mathrm{T}}X$$

donde β_j son los vectores característicos de Σ. La matriz de covarianza de Y está dada por

$$Cov(Y) = \beta^{\mathrm{T}}\Sigma\beta = \Lambda$$

La varianza total de X se define como :

$$\sigma^2_{total} = \sum_{j=1}^{p} var(X_j) = tr(\Sigma)$$

y la varianza generalizada como

$$\sigma^2_{gen} = \det(\Sigma) \ .$$

Estas dos medidas de dispersión multivariada son invariantes por las transformaciones ortogonales.

$$\sigma^2_{total} = tr(\Sigma) = tr(\beta\Lambda\beta^{\mathrm{T}}) = tr(\Lambda) = \sum_{1}^{p}\lambda_i = \sum_{1}^{p} var(Y_i)$$

y

$$\sigma^2_{gen} = \det(\Sigma) = \det(\beta)\det(\Lambda)\det(\beta^{\mathrm{T}}) = \det(\Lambda)$$

Por lo tanto, la proporción de la varianza explicada por la $k-$ésima componente es:

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \qquad k = 1, \ldots, p$$

Es de interés el caso en que una gran proporción (80 %, 90 %) de la varianza total de las p variables puede ser explicada por $k \leq 3$ componentes, ya que nos permitiría graficar con poca pérdida de variabilidad.

En el caso muestral, los estimadores de máxima verosimilitud de las componentes principales se pueden obtener reemplazando Σ por su estimador muestral.

2.1.2. Componentes principales muestrales

Desde el punto de vista analítico hay poca diferencia entre las componentes principales de una población y las de una muestra. Como mencionamos anteriormente, en el caso muestral se reeplaza Σ por un estadístico $\widehat{\Sigma}$ apropiado, al cual se le pide que sea definido positivo y simétrico, y se definen las componentes principales muestrales como la descomposición espectral de $\widehat{\Sigma}$. Usualmente se suele elegir como $\widehat{\Sigma}$ la matriz insesgada de covarianza muestral:

$$S = \frac{\sum_{j=1}^{n}(X_j - \overline{X})(X_j - \overline{X})^{\mathrm{T}}}{n-1}$$

donde

$$\overline{X} = \frac{\sum_{j=1}^{n} X_j}{n}$$

El uso de S se justifica por ser, excepto una constante $(n-1)/n$, el estimador de máxima verosimilitud en el caso normal. Supongamos que $X_1, \ldots, X_n$ es una muestra de una distribución normal $p-$variada $X \sim \mathbf{N}_p(\mu, \Sigma)$ y supongamos que la matriz de covarianza $\Sigma = \beta\Lambda\beta^{\mathrm{T}}$ es definida positiva. En dicho caso $S^* = (n-1)S/n$ es el estimador de máxima verosimilitud de Σ y si $n > p$, S es definida positiva con probabilidad 1 y los p autovalores de S son distintos. Sea

$$S = BLB' = \sum_{j=1}^{p} l_j b_j b_j' \tag{2.4}$$

la descomposición espectral de S, donde $B = (b_1, \ldots, b_p)$ es ortogonal y $L = \mathrm{diag}(l_1, \ldots, l_p)$. Cuando todos los autovalores de Σ son distintos, la descomposición espectral de Σ es única salvo por el signo de los autovectores.

El teorema de invarianza de los estimadores de máxima verosimilitud (Apéndice), implica que los estimadores de máxima verosimilitud de β y Λ, que indicaremos $\widehat{\beta}$ y $\widehat{\Lambda}$, sean los autovectores y autovalores de S^*, es decir,

$$\widehat{\beta} = B \tag{2.5}$$

$$\widehat{\Lambda} = \frac{n-1}{n} L \; . \tag{2.6}$$

La teoría asintótica de los estimadores de las componentes principales puede verse en Flury (1988).

Ahora veremos como encontrar las componentes principales comunes en el caso de tener k poblaciones.

2.1.3. Calculo de las CPC: Estimadores de Máxima Verosimilitud

Jolicoeur y Mosimann (1960) y Jolicoeur(1963b) analizaron las componentes principales de varios conjuntos de datos biométricos. Sus ejemplos están relacionados con las medidas de los caparazones de tortugas y con medidas de huesos humanos y animales. En cada uno de estos ejemplos separaron machos y hembras, observando que las componentes principales eran muy similares mientras que las varianzas entre el sexo femenino y el masculino eran muy distintas. Aunque notaron claramente estas diferencias, no llegaron a formalizar la definición de modelos apropiados como el CPC.
En el modelo CPC se está asumiendo que las componentes de las distintas poblaciones son las mismas aunque la variabilidad podría ser diferente en cada población. Si llamamos $\Sigma_1 \dots \Sigma_k$ a las matrices de covarianza correspondientes a k poblaciones donde se tuvieron en cuenta p atributos, es decir $\Sigma_i \in I\!R^{p\times p}$, la hipótesis del modelo CPC tal como se puede encontrar en Flury (1988) es

$$H_{CPC}: \quad \Sigma_i = \beta \Lambda_i \beta^{\mathrm{T}} \qquad i = 2, \dots, k$$

donde:

β es una matriz ortogonal de $p \times p$, y
$\Lambda_i = \mathrm{diag}(\lambda_{i1}, \dots, \lambda_{ip})$.

La diferencia de parámetros a ser estimados entre el modelo CPC y otro en el que todas las matrices de covarianza son diferentes, es de $p(p-1)(k-1)/2$. Por lo que este método va a resultar ventajoso en el caso en que p y k sean grandes.

Si la muestra $X_{i1}, \dots, X_{in_i}$ de la $i-$ésima población tiene una distribución normal p-variada: $X_i \sim N_p(\mu_i, \Sigma_i)$. Los estimadores insesgados de posición y escala son respectivamente:

$$S_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^{\mathrm{T}} \qquad \bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \qquad i = 1, \dots, k \; .$$

Si $N_i = n_i - 1 > p$, la distribución de S_i es Wishart con N_i grados de libertad y parámetro $\frac{\Sigma_i}{N_i}$ que indicaremos $S_i \sim W_p(N_i, \Sigma_i/N_i)$. Luego, la densidad de S_i es (Seber, 1984)

$$f(W) = \frac{1}{\Gamma_p\left(\frac{N_i}{2}\right)|\Sigma_i|^{\frac{N_i}{2}}}\left(\frac{N_i}{2}\right)^{\frac{N_i\,p}{2}} \operatorname{etr}\left(-\frac{N_i}{2}\Sigma_i^{-1}W\right)|W|^{\frac{N_i-p-1}{2}} \tag{2.7}$$

si W es definida positiva y 0 en otro caso, donde

$$\Gamma_p(y) = \pi^{\frac{p(p-1)}{4}}\prod_{j=1}^{p}\Gamma\left[y - \frac{1}{2}(j-1)\right]$$

es la función de gamma multivariada, y etr es función exponencial de la traza.
La función de verosimilitud de $\Sigma_1 \ldots \Sigma_k$ dados $S_1 \ldots S_k$ es, por lo tanto,

$$L(\Sigma_1, \ldots, \Sigma_k) = C \times \prod_{i=1}^{k}\operatorname{etr}\left(-\frac{N_i}{2}\Sigma_i^{-1}S_i\right)|\Sigma_i|^{-\frac{N_i}{2}} \tag{2.8}$$

donde la constante C no depende de los parámetros. Maximizar (2.8) es equivalente a minimizar

$$G(\Sigma_1, \ldots, \Sigma_k) = -2\log L + 2\log C = \sum_{i=1}^{k} N_i\left[\operatorname{tr}(\Sigma_i^{-1}S_i) + \log|\Sigma_i|\right]$$

Si suponemos que el modelo H_{CPC} es válido , y si $\beta = (\beta_1, \ldots, \beta_p)$, reemplazando obtenemos

$$G(\beta, \Lambda_1, \ldots, \Lambda_k) = \sum_{i=1}^{k} N_i \sum_{j=1}^{p}\log(\lambda_{ij}) + \frac{\beta_j^{\mathrm{T}} S_i \beta_j}{\lambda_{ij}}\ .$$

Como esta minimización está restringida a que $\beta_i^{\mathrm{T}}\beta_i = 1$ introducimos los multiplicadores de Lagrange δ_i y para las $p(p-1)/2$ condiciones $\beta_i^{\mathrm{T}}\beta_j = 0$ los multiplicadores δ_{ij}. Por lo tanto, la función a minimizar será:

$$g(\beta, \Lambda_1, \ldots, \Lambda_k) = \sum_{i=1}^{k} N_i \sum_{j=1}^{p}\log(\lambda_{ij}) + \frac{\beta_j^{\mathrm{T}} S_i \beta_j}{\lambda_{ij}} - \sum_{j=1}^{p}\delta_j(\beta_j^{\mathrm{T}}\beta_j - 1) - 2\sum_{h<j}^{p}\delta_{hj}\beta_h^{\mathrm{T}}\beta_j \tag{2.9}$$

Tomando las derivadas parciales con respecto de λ_{im} e igualando a cero resulta

$$N_i\left(\frac{1}{\lambda_{im}} - \frac{\beta_m^{\mathrm{T}} S_i \beta_m}{\lambda_{im}^2}\right) = 0 \Longrightarrow \lambda_{im} = \beta_m^{\mathrm{T}} S_i \beta_m \tag{2.10}$$

Lo que implica, para el máximo de la función de verosimilitud:

$$\operatorname{tr}(\Sigma_i^{-1}S_i) = \operatorname{tr}(\beta\Lambda_i^{-1}\beta^{\mathrm{T}}S_i) = \operatorname{tr}(\Lambda_i^{-1}\beta^{\mathrm{T}}S_i\beta) = \sum_{m=1}^{p}\frac{\beta_m^{\mathrm{T}} S_i \beta_m}{\lambda_{im}} = p$$

La ecuación (2.10) nos muestra que maximizar la verosimilitud es equivalente a buscar la matriz β que haría que $\beta^{\mathrm{T}} S_i \beta$ sea una matriz diagonal y por lo tanto, a minimizar la función

$$\Phi(\beta) = \prod_{i=1}^{k} \left[\frac{\det \ \mathrm{diag}(\beta^{\mathrm{T}} S_i \beta)}{\det(\beta^{\mathrm{T}} S_i \beta)} \right]^{N_i} \tag{2.11}$$

con $\beta \in \mathcal{O}(p)$, el grupo de matrices ortogonales de $p \times p$ alcance el valor mínimo "1"(desigualdad de Hadamard).

Derivando (2.9) respecto de β_j e igualando a cero se obtiene

$$\sum_{i=1}^{k} N_i \frac{S_i \beta_j}{\lambda_{ij}} - \delta_j \beta_j - \sum_{h \neq j} \delta_{jh} \beta_h = 0 \quad 1 \leq j \leq p \ . \tag{2.12}$$

A partir de (2.10) y (2.12), si se está bajo el supuesto que $X_{i1}, \ldots, X_{in_i}$ son i.i.d. $X_i \sim N_p(\mu_i, \Sigma_i)$ y que Σ_i cumplen el modelo CPC, es fácil deducir multiplicando (2.12) por β_m^{T}, $m \neq j$ que los estimadores de máxima verosimilitud de β y de Λ_i, $\widehat{\beta}$ y $\widehat{\Lambda}_i$, cumplen las siguientes ecuaciones:

$$\widehat{\beta}_m^{\mathrm{T}} \left[\sum_{i=1}^{k} N_i \frac{\widehat{\lambda}_{im} - \widehat{\lambda}_{ij}}{\widehat{\lambda}_{im} \widehat{\lambda}_{ij}} S_i \right] \widehat{\beta}_j \ = \ 0 \qquad \text{para } 1 \leq j, m \leq p \quad m \neq j \tag{2.13}$$

$$\widehat{\beta}_m^{\mathrm{T}} \widehat{\beta}_j \ = \ \begin{cases} 0 & \text{si } m \neq j \\ 1 & \text{si } m = j \end{cases} \tag{2.14}$$

$$\widehat{\Lambda}_i \ = \ \mathrm{diag}\left(\widehat{\beta}^{\mathrm{T}} S_i \widehat{\beta} \right) \tag{2.15}$$

Estas condiciones forman el sistema básico de ecuaciones en el análisis CPC y las mismas tienen que ser resueltas bajo las restricciones de ortogonalidad de los β.

Bajo la hipótesis de H_{CPC}, se puede esperar que el máximo de la función de verosimilitud sea único.

Flury y Gautschi (1986) propusieron el algoritmo FG para resolver este sistema de ecuaciones, el cual se describirá en el Capítulo 3.
Debido a la equivalencia con el problema de minimización (2.11) sobre el conjunto compacto $\mathcal{O}(p)$ podemos afirmar que el máximo de la función existe.

Asumiendo por el momento que los estimadores de máxima verosimilitud de $\beta = (\beta_1, \ldots, \beta_p)$ y λ_{ij} son únicos, el estimador de máxima verosimilitud de Σ queda

$$\widehat{\Sigma}_i = \widehat{\beta} \widehat{\Lambda_i} \widehat{\beta} \qquad i = 1, \ldots, k \ . \tag{2.16}$$

El máximo de la función de verosimilitud se obtiene de (2.8) como

$$L(\widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_k) = C \times \prod_{i=1}^{k} \exp\left(-\frac{pN_i}{2} \right) \left| \widehat{\Sigma}_i \right|^{-N_i/2} \tag{2.17}$$

y el máximo irrestricto es

$$L(S_1,\ldots,S_k) = C \times \prod_{i=1}^{k} \exp\left(-\frac{pN_i}{2}\right) |S_i|^{-N_i/2} \ . \tag{2.18}$$

Por lo tanto, el logaritmo de la razón de los estimadores de verosimilitud para testear H_{CPC} queda

$$X^2_{CPC} = \ -2\log\frac{L(\widehat{\Sigma}_1,\ldots,\widehat{\Sigma}_k)}{L(S_1,\ldots,S_k)} = \sum_{i=1}^{k} N_i \log\frac{\left|\widehat{\Sigma}_i\right|}{|S_i|} \tag{2.19}$$

Bajo la hipótesis nula, X^2_{CPC} tiene distribución asintótica chi–cuadrado con $(k-1)p$ grados de libertad, como se deduce de la teoría del cociente de verosimilitud, (Rao, 1973).

2.1.4. Teoría asintótica

La distribución asintótica para los estimadores de máxima verosimilitud $\widehat{\beta}$ y $\widehat{\Lambda}_i$ bajo el modelo CPC se puede encontrar en Flury (1984). A continuación daremos un breve resumen de las mismas. Los estimadores de β y de Λ_i tienden en distribución a una normal por la teoría de máxima verosimilitud. También usaremos el hecho de que la matriz de covarianza de los estimadores de máxima verosimilitud está dada en forma aproximada por la inversa de la matriz de información de Fisher. El problema es que la matriz ortogonal β contiene solo $p(p-1)/2$ parámetros independientes, lo que no posibilita el cálculo directo de la matriz de información. Supondremos que los p autovalores de Σ_i son todos distintos. De (2.8) tenemos que el logaritmo de la función de verosimilitud es

$$-\frac{1}{2}\sum_{i=1}^{k} N_i \sum_{j=1}^{p}\left(\log(\lambda_{ij}) + \frac{\beta_j^{\mathrm{T}} S_i \beta_j}{\lambda_{ij}}\right) \tag{2.20}$$

Sean $\lambda_{(i)} = (\lambda_{i1},\ldots,\lambda_{ip})^{\mathrm{T}}$, $s = p(p-1)/2$ y β^* un vector compuesto de s elementos funcionalmente independientes que determinan β . Llamemos $N = N_1 + \ldots + N_k$ y $r_i = N_i/N$. Entonces la matriz de información es:

	$\lambda_{(1)}^{\mathrm{T}}$	$\lambda_{(2)}^{\mathrm{T}}$	$\ldots$	$\lambda_{(k)}^{\mathrm{T}}$	$\beta^{*\mathrm{T}}$
$\lambda_{(1)}$	$\frac{1}{2}Nr_1\Lambda_1^{-2}$	0	$\ldots$	0	
$\lambda_{(2)}$	0	$\frac{1}{2}Nr_2\Lambda_2^{-2}$	$\ldots$	0	
$\vdots$	$\vdots$			$\vdots$	nG^{T}
$\lambda_{(k)}$	0	0	$\ldots$	$\frac{1}{2}Nr_k\Lambda_k^{-2}$	
β^*			NG		NA

(2.21)

donde A y G todavía no fueron determinados.

Como $\widehat{\beta}$ es un estimador consistente de β, la distribución asintótica de $\widehat{\beta}^{\mathrm{T}}\left[\sqrt{n_i}(S_i-\Sigma_i)\right]\widehat{\beta}$ es la misma que la distribución asintótica $\beta^{\mathrm{T}}\left[\sqrt{n_i}(S_i-\Sigma_i)\right]\beta$. Por lo tanto, el vector aleatorio $\sqrt{N_i}(\widehat{\lambda}_{(i)}-\lambda_{(i)})$ tiene la misma distribución asintótica que los elementos diagonales de $\sqrt{N_i}(\beta^{\mathrm{T}}S_i\beta-\Lambda_i)$. Usando que $\beta^{\mathrm{T}}S_i\beta\sim W_p(N_i,\Lambda_i/N_i)$, de la teoría de las matrices de Wishart (Muirhead, 1982) se deduce que los p elementos de $\sqrt{N_i}(\widehat{\lambda}_{(i)}-\lambda_{(i)})$ son asintóticamente normales e independientes con media cero y varianza $2\lambda_{ij}^2$. Además, los vectores $\sqrt{N_i}(\widehat{\lambda}_{(i)}-\lambda_{(i)})$, $\sqrt{N_h}(\widehat{\lambda}_{(h)}-\lambda_{(h)})$ $(h\neq i)$ son asintóticamente independientes, dado que las matrices S_i son independientes.

De (2.21), y suponiendo que r_i permanece constante al tender mín(N_i) a infinito, la distribución conjunta asintótica de los k vectores $\sqrt{N}(\widehat{\lambda}_{(i)}-\lambda_{(i)})$ es $N(0,V_\lambda)$ con V_λ dada por

$$V_\lambda=\left[\begin{bmatrix}\frac{1}{2}r_1\Lambda_1^{-2} & \dots & 0\\ \vdots & & \vdots\\ 0 & \dots & \frac{1}{2}r_k\Lambda_k^{-2}\end{bmatrix}-G^{\mathrm{T}}A^{-1}G\right]^{-1} \tag{2.22}$$

donde se utilizó la fórmula para la inversa de una matriz particionada. Por otra parte, por lo discutido anteriormente, V_λ es una matriz diagonal en bloques donde la diagonal está compuesta de k bloques de la forma $2r_i^{-1}\Lambda_i^2$. Teniendo en cuenta que en el máximo de la verosimilitud la matriz A es definida positiva, se deduce que $G=0$ y por lo tanto, los estimadores de λ_{ij} son asintóticamente independientes de $\widehat{\beta}$.

El siguiente teorema, resume los resultados anteriores.

Teorema 1 . *Las pk variables aleatorias* $\sqrt{N_i}(\widehat{\lambda}_{ij}-\lambda_{ij})$ *tienen distribución asintótica* $N(0,2\lambda_{ij}^2)$, *si (*mín $N_i\to\infty$*), independientes entre sí y de* $\widehat{\beta}$.

Ahora busquemos la distribución asintótica de $\widehat{\beta}$. De la función de verosimilitud (2.20) se deduce que la matriz NA se puede escribir como la suma de k matrices $N_1A_1,\dots,N_kA_k$ donde N_iA_i está asociada con la $i-$ésima muestra. Más aún, la matriz N_iA_i es exactamente igual a la matriz que se obtendría si hubiera una única muestra. De la sección 2.1.2 sabemos que, en el caso de una sola muestra $k=1$, los autovectores de S_i son los estimadores de máxima verosimilitud de β. La matriz de covarianza asintótica de $\sqrt{N}(\widehat{\beta}-\beta)$, como se obtendría en el caso de la i-ésima muestra, sale del siguiente Teorema que se encuentra en Flury (1988).

Teorema 2 . *Sea* S *una matriz aleatoria simétrica de* $p\times p$, *con distribución* $W_p(n,\Sigma/n)$ *donde* Σ *es simétrica y definida positiva. Sea* $S=BLB^{\mathrm{T}}$ *y* $\Sigma=\beta\Lambda\beta^{\mathrm{T}}$ *la descomposición espectral de* S *y* Σ, $L=diag(l_1,\dots,l_p)$ $\Lambda=diag(\lambda_1,\dots,\lambda_p)$. *Supongamos que todos los autovalores de* Σ *son distintos,* $\lambda_1>\ldots>\lambda_p$. *Entonces*

a) La distribución asintótica de

$$\sqrt{n}\begin{bmatrix} l_1-\lambda_1 \\ \vdots \\ l_p-\lambda_p \end{bmatrix} \tag{2.23}$$

es normal $p-$variada con media cero y matriz de covarianza $diag(2\lambda_1^2,\ldots,2\lambda_p^2)$. Más aún, $\{l_j : 1\leq j\leq p\}$ son asintóticamente independientes de B.

b) La distribución asintótica de $\sqrt{n}\ vec(B-\beta)$ es normal p^2-variada (singular) con esperanza cero y matriz de covarianza V_B dada por

$$V_B = \begin{bmatrix} \beta & 0 & \ldots & 0 \\ 0 & \beta & \ldots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \ldots & \beta \end{bmatrix} V_E \begin{bmatrix} \beta^{\mathrm{T}} & 0 & \ldots & 0 \\ 0 & \beta^{\mathrm{T}} & \ldots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \ldots & \beta^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \beta V_{11}\beta^{\mathrm{T}} & \ldots & \beta V_{1p}\beta^{\mathrm{T}} \\ \vdots & & \vdots \\ \beta V_{p1}\beta^{\mathrm{T}} & \ldots & \beta V_{pp}\beta^{\mathrm{T}} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1,j\neq 1}^{p}\theta_{1j}\beta_j\beta_j^{\mathrm{T}} & -\theta_{12}\beta_2\beta_1^{\mathrm{T}} & \ldots & -\theta_{1p}\beta_p\beta_1^{\mathrm{T}} \\ \vdots & & \vdots & \\ -\theta_{21}\beta_1\beta_2^{\mathrm{T}} & \sum_{j=1,j\neq 2}^{p}\theta_{2j}\beta_j\beta_j^{\mathrm{T}} & \ldots & -\theta_{2p}\beta_p\beta_2^{\mathrm{T}} \\ \vdots & \vdots & & \vdots \\ -\theta_{p1}\beta_1\beta_p^{\mathrm{T}} & -\theta_{p2}\beta_2\beta_p^{\mathrm{T}} & \sum_{j=1,j\neq p}^{p}\theta_{pj}\beta_j\beta_j^{\mathrm{T}} & \end{bmatrix}$$

con

$$\theta_{hj} = \begin{cases} 0 & h=j \\ \dfrac{\lambda_h\lambda_j}{(\lambda_h-\lambda_j)^2} & h\neq j \end{cases}$$

Por lo tanto, las matrices V_B para cada una de las muestras, que indicaremos V_i, serán iguales a

$$V_i = \begin{bmatrix} \sum_{j=1,j\neq 1}^{p}\theta_{1j}^{(i)}\beta_j\beta_j^{\mathrm{T}} & -\theta_{12}^{(i)}\beta_2\beta_1^{\mathrm{T}} & \ldots & -\theta_{1p}^{(i)}\beta_p\beta_1^{\mathrm{T}} \\ -\theta_{21}^{(i)}\beta_1\beta_2^{\mathrm{T}} & \sum_{j=1,j\neq 2}^{p}\theta_{2j}^{(i)}\beta_j\beta_j^{\mathrm{T}} & \ldots & -\theta_{2p}\beta_p\beta_2^{\mathrm{T}} \\ \vdots & \vdots & & \vdots \\ -\theta_{p1}^{(i)}\beta_1\beta_p^{\mathrm{T}} & -\theta_{p2}^{(i)}\beta_2\beta_p^{\mathrm{T}} & \ldots & \sum_{j=1,j\neq p}^{p}\theta_{pj}^{(i)}\beta_j\beta_j^{\mathrm{T}} \end{bmatrix} \tag{2.24}$$

con

$$\theta_{hj}^{(i)} = \begin{cases} 0 & h = j \\ r_i^{-1} \dfrac{\lambda_{ih}\lambda_{ij}}{(\lambda_{ih} - \lambda_{ij})^2} & h \neq j \end{cases}$$

El problema con las matrices V_i es la singularidad lo que nos impide invertirlas directamente. Sin embargo, tienen una propiedad destacable la cual se enuncia en el siguiente Lema.

Lema 2 . *Las $p^2 \times p^2$ matrices V_i definidas en (2.24) pueden ser diagonalizadas por la misma matriz ortogonal.*

Es decir que la propiedad de las Σ_i de tener autovectores idénticos es heredada por las V_i. Este lema se demuestra probando que V_iV_m=V_mV_i para todos los pares (i, m) usando la equivalencia de diagonalización simultánea y conmutatividad.

Los autovectores y autovalores de V_i se obtienen del siguiente Lema.

Lema 3 . *Los $s = p(p-1)/2$ autovalores normalizados de V_i asociados con los autovalores positivos están dados como sigue: para cada par (j, h) tal que $1 \leq j < h \leq p$, hay un autovector que tiene $\beta_h\sqrt{2}$ en la posición j y $-\beta_j\sqrt{2}$ en la posición h. El resto de los elementos son ceros y el autovalor asociado es $2\theta_{jh}^{(i)}$*

Para encontrar la distribución asintótica de $\widehat{\beta}$, la clave es diagonalizar las matrices V_i simultáneamente, reduciendo de esa manera el problema de matrices de dimensión $p^2 \times p^2$ a matrices de dimensión $s \times s$. Esto significa que los s parámetros β^* funcionalmente independientes son elegidos de forma tal que las matrices de información $N_i\ A_i$ sean diagonales simultáneamente.

Por el Lema 3 existe una matriz ortogonal $H = (H_1, H_2)$, de $p^2 \times p^2$, con H_1 de dimensión $p^2 \times s$ tal que

$$H^{\mathrm{T}} V_i H = \begin{bmatrix} \Gamma_i & 0 \\ 0 & 0 \end{bmatrix}, \qquad i = 1, \ldots, k$$

y tal que la matriz Γ_i es diagonal. Los elementos de Γ_i son los s autovalores no nulos de V_i y las s columnas de H_1 son los autovectores asociados, es decir,

$$\begin{aligned} \Gamma_i &= 2\mathrm{diag}(\theta_{12}^{(i)}, \theta_{13}^{(i)}, \ldots, \theta_{p-1,p}^{(i)}) \\ H_1 &= (h_{12}, h_{13}, \ldots, h_{p-1,p}) \end{aligned}$$

donde h_{jm} está definido en el Lema 3. Para los s parámetros funcionalmente independientes $\beta^* = H_1^{\mathrm{T}}\mathrm{vec}\beta$, la información para la i-ésima muestra es

$$N\, r_i A_i = 2\, N\, \left[\mathrm{diag}(\theta_{12}^{(i)}, \theta_{13}^{(i)}, \ldots, \theta_{p-1,p}^{(i)})\right]^{-1} .$$

La suma de estas k matrices de información es:

$$\begin{aligned} NA &= N\sum_{i=1}^{k} r_i A_i = 2N\text{diag}\left(\sum_{i=1}^{k} {\theta_{12}^{(i)}}^{-1}, \ldots, {\theta_{p-1,p}^{(i)}}^{-1}\right) \\ &= 2N\text{diag}\left(\theta_{12}^{-1}, \ldots, \theta_{p-1,p}^{-1}\right) \end{aligned} \tag{2.25}$$

donde

$$\theta_{jm} = \left[\sum_{i=1}^{k} {\theta_{jm}^{(i)}}^{-1}\right]^{-1}, \qquad j \neq m \tag{2.26}$$

es la media armónica de $\theta_{jm}^{(1)}$ a $\theta_{jm}^{(k)}$.

La inversa de A representa la matriz de covarianza asintótica del vector aleatorio $\sqrt{n}H_1^{\mathrm{T}}\ \text{vec}(\widehat{\beta} - \beta)$. Transformando para obtener $\widehat{\beta}$, obtenemos la matriz de covarianza asintótica de $\sqrt{n}\ \text{vec}(\widehat{\beta} - \beta)$,

$$V_\beta = H \begin{bmatrix} A^{-1} & 0 \\ 0 & 0 \end{bmatrix} H^{\mathrm{T}} = H_1 A^{-1} H_1^{\mathrm{T}} = 2 \sum_{1 \leq j < m \leq p} \theta_{jm} h_{jm} h_{jm}^{\mathrm{T}} \, .$$

Escribiendo a los vectores h_{jm} en términos de β_j y β_m como se indica en el Lema 3, obtenemos

$$\begin{aligned} V_\beta &= \theta_{12} \begin{bmatrix} \beta_2\beta_2^{\mathrm{T}} & -\beta_2\beta_1^{\mathrm{T}} & 0 & \ldots & 0 \\ -\beta_1\beta_2^{\mathrm{T}} & \beta_1\beta_1^{\mathrm{T}} & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix} + \theta_{13} \begin{bmatrix} \beta_3\beta_3^{\mathrm{T}} & 0 & -\beta_3\beta_1^{\mathrm{T}} & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ -\beta_1\beta_3^{\mathrm{T}} & 0 & -\beta_1\beta_1^{\mathrm{T}} & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix} + \ldots \\ &+ \theta_{p-1,p} \begin{bmatrix} 0 & \ldots & 0 & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \ldots & \beta_p\beta_p^{\mathrm{T}} & -\beta_p\beta_{p-1}^{\mathrm{T}} \\ 0 & \ldots & -\beta_{p-1}\beta_p^{\mathrm{T}} & \beta_{p-1}\beta_{p-1}^{\mathrm{T}} \end{bmatrix} . \end{aligned}$$

Estos resultados se resumen en el siguiente Teorema.

Teorema 3 *La distribución asintótica de* $\sqrt{N}\ vec(\widehat{\beta} - \beta)$ *es normal con media cero y matriz de covarianza* V_β *dada por*

	$\widehat{\beta}_1^{\mathrm{T}}$	$\widehat{\beta}_2^{\mathrm{T}}$	$\ldots$	$\widehat{\beta}_p^{\mathrm{T}}$
$\widehat{\beta}_1$	$\sum_{j=1,j\neq 1}^{p} \theta_{1j}\beta_j\beta_j^{\mathrm{T}}$	$-\theta_{12}\beta_2\beta_1^{\mathrm{T}}$	$\ldots$	$-\theta_{1p}\beta_p\beta_1^{\mathrm{T}}$
$\widehat{\beta}_2$	$-\theta_{21}\beta_1\beta_2^{\mathrm{T}}$	$\sum_{j=1,j\neq 2}^{p} \theta_{2j}\beta_j\beta_j^{\mathrm{T}}$	$\ldots$	$-\theta_{2p}\beta_p\beta_2^{\mathrm{T}}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\widehat{\beta}_p$	$-\theta_{p1}\beta_1\beta_p^{\mathrm{T}}$	$-\theta_{p2}\beta_2\beta_p^{\mathrm{T}}$	$\ldots$	$\sum_{j=1,j\neq p}^{p} \theta_{pj}\beta_j\beta_j^{\mathrm{T}}$

donde los θ_{jm} están definidos como en (2.26) y los β_j son los autovectores en común de las k matrices Σ_i.

Capítulo 3

Métodos Numéricos: Algoritmo FG

3.1. Algoritmo de Jacobi

El método más antiguo para la diagonalización de matrices simétricas fue creado por Jacobi (1846). Sea $A = (a_{mj})$ la matriz simétrica de $p \times p$ que queremos diagonalizar y sea $= BLB'$ su descomposición espectral. El método de Jacobi consiste en pre y post multiplicar la matriz A por una secuencia de matrices ortogonales que vayan eliminando los elementos fuera de la diagonal. Estas matrices representan rotaciones ortogonales que se denominan rotaciones de Jacobi y son de la forma:

$$J = J(m, j, \theta) = \begin{pmatrix} 1 & \dots & 0 & & & 0 \\ & \vdots & & \vdots & & \\ 0 & \dots & c & \dots & -s & \dots \\ & \vdots & & \vdots & & \\ 0 & \dots & s & \dots & c & \dots \\ & \vdots & & \vdots & & \\ 0 & \dots & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ m \\ \\ j \\ \\ \end{matrix}$$

donde $c = \cos\theta$ y $s = \sin\theta$. La matriz de Jacobi es esencialmente una matriz diagonal de $p \times p$ donde los elementos (m, m), (m, j), (j, m), (j, j) se reemplazaron por c, $-s$, s y c, respectivamente.

Veamos como se debe elegir θ de forma que se vayan eliminando los elementos no diagonales de A. Dados m, j $1 \leq m < j \leq p$, consideremos la transformación $H = J^{\mathrm{T}}AJ$ con $J = J(m, j, \theta)$. La matriz H tiene los mismos elementos que la matriz A salvo en las filas y columnas m, j

$$h_{mm} = c^2 a_{mm} + s^2 a_{jj} + 2c\, s a_{mj} \tag{3.1}$$

$$h_{jj} = c^2 a_{jj} + s^2 a_{mm} - 2c\, s a_{mj} \tag{3.2}$$

$$h_{mj} \quad = \quad h_{jm} = (c^2 - s^2)a_{mj} + c\,s\,(a_{jj} - a_{mm}) \tag{3.3}$$

Llamemos $t = \tan(\theta) = s/c$. Como el objetivo es que los elementos no diagonales sean nulos igualamos (3.3) a 0 y dividimos la igualdad por c^2. Si suponemos que $a_{mj} \neq 0$, obtenemos

$$h_{mj}/c^2 = (1-t^2)a_{mj} + t(a_{jj} - a_{mm}) = 0 \quad \Leftrightarrow \quad (1-t^2) + t\frac{a_{jj} - a_{mm}}{a_{mj}} = 0 \Leftrightarrow$$

$$t^2 + \frac{a_{mm} - a_{jj}}{a_{mj}}t - 1 \quad = \quad 0$$

Las 2 raíces reales t_1, t_2 solución de esta ecuación, cumplen que $t_1 t_2 = -1$, por lo tanto los ángulos de rotación correspondientes θ_1 y θ_2 difieren en $\pi/2$. La raíz que se suele elegir es la que cumple que $|\theta| = |\tan^{-1}(t)| \leq \pi/4$.

Como medida de desvío de la matriz A respecto de la diagonalidad, se toma la suma de cuadrados de los elementos no diagonales,

$$\text{off}(A) = 2 \sum_{1 \leq m < j \leq p} a_{mj}^2$$

Si la matriz A es diagonal off(A)=0. Se compara entonces off(A) con off($J^{\mathrm{T}}AJ$) = off(H), donde $J = J(m, j, \theta)$

$$\text{off}(A) - \text{off}(H) = h_{mm}^2 + h_{jj}^2 - a_{mm}^2 - a_{jj}^2 \; .$$

Como $h_{mm}^2 + h_{jj}^2 + 2h_{mj} = a_{mm}^2 + a_{jj}^2 + 2m_{mj}$ se obtiene off(A) − off(H) = $2(a_{mj}^2 - h_{mj}^2)$. Por lo tanto, la reducción máxima se obtiene si se elige el ángulo de rotación θ tal que $h_{mj} = 0$. Más aún, con una única rotación de Jacobi $J(m, j, \theta)$ la suma de elementos no diagonales se puede reducir en $2a_{mj}^2$.

El procedimiento de ir anulando los elementos no diagonales iterativamente, es el que se conoce como el Algoritmo de iteración clásico de Jacobi. En cada paso, se eligen los índices m y j correspondientes al máximo $|a_{mj}|$, logrando así la máxima reducción posible en cada paso. Es evidente que cada paso de anulación de elementos no diagonales, puede destruir algunos de los ceros no diagonales introducidos en los pasos anteriores, y el proceso debe iterarse hasta que off(A) $< \epsilon$ con ϵ suficientemente chico. En el libro de Golub y Van Loan (1983, pp. 297-298) se puede encontrar una discusión de la convergencia del método .

Para evitar la búsqueda del máximo valor absoluto no diagonal entre los $(p-1)p/2$ elementos no diagonales, se pueden elegir los pares de rotación ciclicamente, por ejemplo, en el orden $(1,2)$, $(1,3)$, ..., $(1,p)$, $(2,3)$,..., $(p-1,p)$. Cada conjunto de $(p-1)p/2$ rotaciones de este esquema cíclico se denomina un *sweep* (barrido). Aunque con el método de Jacobi cíclico se desperdicia tiempo anulando elementos que son ceros o cercanos a cero, las propiedades de convergencia son tan buenas como las del método clásico y se ahorra tiempo al no necesitar ordenar los elementos no diagonales (ver Golub y Van Loan , 1983, pp. 299-300).

El algoritmo FG es una generalización del método de Jacobi cíclico. Una demostración de la convergencia del método puede verse en Flury (1988).

3.2. Algoritmo FG

El algoritmo FG se utiliza para resolver el sistem básico de ecuaciones del análisis CPC.

Sea F una matriz simétrica definida positiva, en lugar de off(F) como medida de la desviación de F de la diagonalidad, usaremos:

$$\varphi(F) = \frac{\det(\text{diag}(F))}{\det(F)}$$

la cual fue mencionada al definir los estimadores de máxima verosimilitud bajo el modelo CPC.

Por la desigualdad de Hadamard, $\det(A) \leq \det(\text{diag}(A))$ con A simétrica y definida positiva, por lo tanto $\varphi(F) \geq 1$ con igualdad si F es diagonal. Además $\varphi(A)$ es monótona creciente a medida que F es "inflada" de diag(F) a F. Esto es lo que se enuncia en el siguiente Lema cuya demostración se puede encontrar en Flury (1988).

Lema 4 . *Sea $F = f_{mj}$ una matriz simétrica, definida positiva de dimensión $p \times p$, entonces*

$$d(\alpha) = \det \begin{bmatrix} f_{11} & \alpha f_{12} & \dots & \alpha f_{1p} \\ \alpha f_{21} & f_{22} & \dots & \alpha f_{2p} \\ \vdots & \vdots & & \vdots \\ \alpha f_{p1} & \alpha f_{p2} & \dots & f_{pp} \end{bmatrix}$$

es una función decreciente de α para $\alpha \in [0, 1]$. Si F no es diagonal entonces, $d(\alpha)$ es estrictamente decreciente.

Consideremos el caso de k matrices simétricas definidas positivas $F_1, \dots, F_k$ y pesos positivos $n_1, \dots, n_k$. Se define como medida simultánea de desviación de la diagonal de las matrices $F_1, \dots, F_k$ con pesos $n_1, \dots, n_k$ a la función

$$\Phi(F_1, \dots, F_k; n_1, \dots, n_k) = \prod_{i=1}^{k} [\varphi(F_i)]^{n_i}$$

Sea $F_i = B^{\mathrm{T}} A_i B$ es decir $A_i = B F_i B^{\mathrm{T}}$ para una matriz ortogonal B. Definamos

$$\Phi_0(A_1, \dots, A_k; n_1, \dots, n_k) = \min_{B \in \mathcal{O}(p)} \Phi(B^{\mathrm{T}} A_1 B, \dots, B^{\mathrm{T}} A_k B; n_1, \dots, n_k)$$

como medida de diagonalización simultánea de $A_1, \dots, A_k$, donde $\mathcal{O}(p)$ es el grupo de matrices ortogonales de $p \times p$. $\Phi_0 \geq 1$, y la igualdad se va a alcanzar si las matrices A_i se pueden diagonalizar simultaneamente por la misma transformación ortogonal. En la sección 2.1.3, se

mostró que la matriz $B_0 = (b_1, \ldots, b_p) \in \mathcal{O}(p)$, para la cual se obtiene el mínimo, verifica el siguiente sistema de ecuaciones

$$b_m^{\mathrm{T}} \left(\sum_{i=1}^{k} n_i \frac{\lambda_{im} - \lambda_{ij}}{\lambda_{im}\lambda_{ij}} A_i \right) b_j = 0 \qquad 1 \leq m < j \leq p \tag{3.4}$$

donde

$$\lambda_{ih} = b_h^{\mathrm{T}} A_i b_h \qquad 1 \leq i \leq k \quad 1 \leq h \leq p\,. \tag{3.5}$$

El algoritmo FG resuelve estas ecuaciones iterativamente minimizando la función Φ. El mínimo de Φ siempre existe dado que el conjunto de las matrices ortogonales de orden p es compacto. Por esta misma razón, el máximo también existe y la matriz B para la cual se alcanza el máximo cumple también las ecuaciones (3.4) y (3.5). Esto es importante en la elección del punto inicial del algoritmo. Otros puntos críticos del sistema (3.4) y (3.5) podrían además existir.

El algoritmo FG se compone de 2 algoritmos separados que se denominan F y G, respectivamente, los cuales minimizan Φ por iteraciones en 2 niveles.

En el nivel más externo (nivel F), cada par (b_m, b_j) de vectores de la aproximación de B a la solución B_0 es rotada para que las correspondientes ecuaciones (3.4) se satisfagan. Cada barrido (*sweep*) del algoritmo F se compone de las rotaciones cíclicas de todos los $p(p-1)/2$ pares de columnas de B.

En el nivel interno (nivel G), se busca por iteraciones una matriz ortogonal de 2×2 que resuelva un sistema de dimensión 2 análogo al (3.4). Esta matriz determina la rotación de un par de vectores que están siendo ajustados en el nivel F.

3.2.1. El algoritmo F

Indiquemos por $\Phi(B) = \Phi(B^{\mathrm{T}}A_1B, \ldots, B^{\mathrm{T}}A_kB; n_1, \ldots, n_k)$ la desviación simultanea de $B^{\mathrm{T}}A_1B, \ldots, B^{\mathrm{T}}A_kB$ de la diagonal como función de B, considerando A_i y n_i, $1 \leq i \leq k$ fijos. El algoritmo F produce una sucesión convergente de matrices ortogonales $B^{(0)}, B^{(1)}, \ldots$ tales que $\Phi(B^{(f+1)}) \leq \Phi(B^{(f)})$.

A continuación se detallan los pasos del algoritmo:

Paso F_0**:** Sea $B = (b_1, \ldots, b_p) \in \mathcal{O}(p)$ una aproximación inicial a la matriz ortogonal que minimiza Φ, se puede tomar $B \leftarrow I_p$, Inicializamos $f = 0$.

Paso F_1**:** Asignar $B^{(f)} \leftarrow B$ $f \leftarrow f + 1$.

Paso F_2**:** Repetir los pasos F_{21} a F_{23}, en orden cíclico, para todos los pares (m, j) con $1 \leq m < j \leq p$

Paso F_{21}**:** Asignar $T_i(2 \times 2) \leftarrow (b_m, b_j)^{\mathrm{T}} A_i (b_m, b_j)$, $(i = 1, \ldots, k)$.

Paso F_{22}**:** Ejecutar el algoritmo G sobre $(T_1, \ldots, T_k)$ con pesos $(n_1, \ldots, n_k)$ para obtener una matriz ortogonal de 2×2

$$J = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$$

Paso F_{23}**:** Rotar las columnas b_m y b_j de B por el ángulo α, es decir, asignar $B \leftarrow BJ$, donde $J = J(m, j, \alpha)$ es una matriz de rotación de Jacobi.

Paso F_3**:** Si para algún $\epsilon_F > 0$, $\left|\Phi\left(B^{(f-1)}\right) - \Phi(B)\right| < \epsilon_F$, terminar. Sino, comenzar el siguiente barrido en F_1.

3.2.2. El algoritmo G

Este algoritmo resuelve la ecuación

$$q_1^{\mathrm{T}} \left(\sum_{i=1}^{k} n_i \frac{\delta_{i1} - \delta_{i2}}{\delta_{i1}\delta_{i2}} T_i \right) q_2 = 0 \tag{3.6}$$

donde $T_1, \ldots, T_2$ son matrices definidas positivas de 2×2, $n_i > 0$ son constantes positivas

$$\delta_{ij} = q_i^{\mathrm{T}} T_i q_j \qquad i = 1, \ldots, k \quad j = 1, 2 \tag{3.7}$$

y $Q = (q_1, q_2)$ es la matriz ortogonal de 2×2 buscada. Las iteraciones del algoritmo producen una serie de matrices ortogonales $Q^{(0)}, Q^{(1)}, \ldots$ que convergen a una solución de (3.6).

Los pasos del algoritmo son los que se dan a continuación:

Paso G_0**:** Se define Q de 2×2 como una aproximación inicial a la solución de (3.6), por ejemplo, $Q \leftarrow I_2$. Se inicializa $g \leftarrow 0$.

Paso G_1**:** Asignar $Q^{(g)} \leftarrow Q$ y $g \leftarrow g + 1$.

Paso G_2**:** Calcular δ_{ij} dados en (3.7) usando la matriz Q actual. Asignar

$$T(2 \times 2) \leftarrow \sum_{i=1}^{k} n_i \frac{\delta_{i1} - \delta_{i2}}{\delta_{i1}\delta_{i2}} T_i$$

Paso G_3**:** Calcular la matriz ortogonal que diagonaliza T

$$\begin{pmatrix} c & -s \\ s & c \end{pmatrix}$$

donde $c = \cos\alpha$, $s = \sin\alpha$, eligiendo la solución para la cual $|\alpha| \leq \pi/4$. Asignar

$$Q \leftarrow \begin{pmatrix} c & -s \\ s & c \end{pmatrix}$$

Paso G_4: Si $\|Q^{(g-1)} - Q\| < \epsilon_G$, donde $\|.\|$ indica una norma matricial y ϵ_G es una constante positiva pequeña, parar. En otro caso, comenzar otra iteración en G_1.

La idea de utilizar dos algoritmos y la conexión con (3.4) puede verse en Flury (1988, pp. 182-183).

Capítulo 4

Medidas de Robustez

4.1. Función de Influencia

La función de influencia (IF) fue definida por Hampel (1968, 1974) para investigar el comportamiento infinitesimal de funcionales reales $T(G)$. Aunque la IF es principalmente una herramienta heurística, tiene una interpretación intuitiva ya que mide la alteración de un estimador provocada al agregar a una muestra grande, una nueva observación en el punto x. Para profundizar sobre este tema se puede consultar Hampel (1986).

La definición de función de influencia de un estimador usa la relación entre un estimador y el funcional asociado. Se dice que un estimador T_n está asociado a un funcional $T(F)$ si depende de la muestra a través de la distribución empírica: $T_n = T(F_n)$.

Definición. *La función de influencia, IF : $\mathbb{R}^m \to \mathbb{R}^p$, del funcional T en F se define puntualmente por*

$$IF(x; T, F) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon\Delta_x) - T(F)}{\epsilon} \tag{4.1}$$

para los puntos $x \in \chi \subset \mathbb{R}^m$ donde el limite exista, siendo Δ_x la distribución discreta en el punto.

La función de influencia verifica las siguientes propiedades:

1. El valor esperado en F de IF es cero, $\int \mathrm{IF}(x; T, F)dF(x) = 0$.
2. Para G "próxima"a F se puede usar la aproximación de primer orden de von Mises de T en F, $T(G) \simeq T(F) + \int \mathrm{IF}(x; T, F)dG(x)$, (Mallows (1975) estudió términos de orden superior).

En particular, se tiene el desarrollo

$$T_n = T(F_n) \simeq T(F) + \int \mathrm{IF}(x; T, F) dF_n(x) = T(F) + \frac{1}{n} \sum_{i=1}^{n} \mathrm{IF}(x_i; T, F)$$

3. Si $X_1, \ldots, X_n$ son i.i.d. $X_i \sim F$, entonces $\sqrt{n}(T_n - T(F))$ es asintóticamente normal con media cero y matriz de covarianza dada por

$$\mathrm{V}(T, F) = \lim_{n \to \infty} n \, \textsc{Var}(T_n) = \int \mathrm{IF}(x; T, F) \mathrm{IF}(x; T, F)^{\mathrm{T}} dF(x) \; .$$

4. Si $\beta(\theta)$ es una transformación de los parámetros, $\beta : \Theta \subset I\!R^p \to \beta(\Theta) \subset I\!R^k$, con matriz Jacobiana $B(\theta) = \partial\beta(\theta)/\partial\theta$ y $\beta(T_n)$ es un estimador de $\beta(\theta)$ entonces, $\mathrm{IF}(x; \beta(T), F) = B(T(F))\mathrm{IF}(x; T, F)$ y $\mathrm{V}(\beta(T), F) = B(T(F))\mathrm{V}(T, F)B(T(F))^{\mathrm{T}}$

Estas propiedades se pueden verificar utilizando las derivadas de Gateaux, cuya definición está en Hampel (1986).

4.1.1. Medidas derivadas de la función de influencia

Desde el punto de vista de la robustez hay por lo menos 3 valores calculados a partir de la IF los cuales fueron introducidos por Hampel (1968, 1974).

- El primero y el más importante es la *sensibilidad a errores groseros*, la cual se define como el supremo de la IF sobre los valores que puede tomar x y sobre los cuales la función está definida

$$\gamma^* = \sup_x \|\mathrm{IF}(x; T, F)\| \; .$$

  Este valor da una aproximación del máximo efecto que podrá causar una pequeña contaminación sobre el estimador. Es una acotación superior del sesgo asintótico del mismo. Una propiedad deseable es que los estimadores tengan $\gamma^*(T, F)$ acotada, en cuyo caso se dice que el funcional T es B-robusto (bias) en F (Rousseeuw, 1981a). Fijar una cota sobre $\gamma^*(T, F)$ es el primer objetivo para obtener un estimador robusto, lo cual en general se contradice con el objetivo de eficiencia asintótica. Por lo tanto, se buscarán estimadores B-óptimos que no puedan ser mejorados simultáneamente con respecto a γ^* y a $\mathrm{V}(T, F)$. En la mayoría de los casos existirá un valor positivo de γ^* para los estimadores Fisher consistentes ($T(F_\theta) = \theta$).

- La segunda medida está relacionada con pequeños cambios en las observaciones. El estimador se modifica con los cambios en las observaciones, estos cambios se pueden deber a errores en las mediciones o redondeos que no deberán afectar demasiado a los estimadores.

El efecto de considerar una observación y en vez de la observación x se puede medir por $\mathrm{IF}(y;T,F) - \mathrm{IF}(x;T,F)$, con y en alguna vecindad de x. El efecto estandarizado de moverse alrededor de x puede ser descripto aproximadamente por una diferencia normalizada o simplemente por la pendiente de IF en el punto. Una medida del peor cambio alrededor de un punto sería la *sensibilidad a cambios locales* dada por

$$\lambda^* = \sup_{x \neq y} \frac{\|\mathrm{IF}(y;T,F) - \mathrm{IF}(x;T,F)\|}{\|x - y\|}$$

Un viejo concepto de robustez rechazaba los outliers por completo, los estimadores con esta propiedad datan de la época de Bernoulli en 1769 (ver Stigler, 1980). En el sentido de las función de influencia, esto significa que IF se anula fuera de cierta región. De hecho, si IF es idénticamente cero en alguna región, la contaminación en dichos puntos no tendrá ningún efecto. Si la distribución subyacente F es simétrica (y poniendo el centro de simetría en cero), podemos definir **el punto de rechazo** como

$$\rho^* = \inf\{r > 0 : \ \mathrm{IF}(x;T,F) = 0 \quad \|x\| > r\}$$

Si no existe dicho r, entonces ρ^* es ∞. Todos los puntos con norma mayor a ρ^* son rechazados. Una característica deseable es que ρ^* sea acotado.

4.1.2. Funciones de Influencia Parcial

Las funciones de influencia parcial fueron introducidas por Pires y Branco (2002) para generalizar el concepto de función de influencia de una muestra al caso de varias muestras. Aunque la función de influencia ya había sido utilizada para estimadores que dependían de más de una muestra por Campbell (1978), Radhakrishan and Kshirsagar (1981), Radhakrishan (1983) y Hampel (1986), Pires y Branco (2002) dieron la definición precisa de la misma.

Definición. *Llamemos F al producto $F = F_1 \times \ldots \times F_k$. Las funciones de influencia parcial del funcional $T(F)$ se definen como*

$$PIF_{i_0}(x;T,F) = \lim_{\epsilon \to 0} \frac{T(F_{x,\epsilon,i_0}) - T(F)}{\epsilon}$$

donde

$$F_{x,\epsilon,i_0} = F_1 \times F_2 \times \ldots \times F_{i_0-1} \times F_{i_0,\epsilon,x} \times F_{i_0+1} \ldots F_k$$

con $F_{i_0,\epsilon,x} = (1-\epsilon)F + \epsilon\ \Delta_x$.

En analogía a la función de influencia, las cuatro propiedades nombradas para ésta siguen siendo válidas.

1. El valor esperado en F_i de PIF_i es cero, $\int \mathrm{PIF}_i(x;T,F)dF_i(x) = 0$.

2. Para el caso k=2, el desarrollo de von Mises del funcional es

$$T(G_1,G_2) \simeq T(F_1,F_2) + \int \mathrm{PIF}_1(x;T,F)dG_1(x) + \int \mathrm{PIF}_2(x;T,F)dG_2(x)\ .$$

   En particular, para las distribuciones empíricas F_{1,n_1}, F_{2,n_2} (con n_1 y n_2 observaciones respectivamente)

$$\begin{aligned} T(F_{1,n_1},F_{2,n_2}) &\simeq T(F_1,F_2) + \int \mathrm{PIF}_1(x;T,F)dF_{1,n_1}(x) + \int \mathrm{PIF}_2(x;T,F)dF_{2,n_2} \\ &= T(F) + \frac{1}{n_1}\sum_{i=1}^{n_1} \mathrm{PIF}_1(x_{1i};T,F) + \frac{1}{n_2}\sum_{i=1}^{n_2} \mathrm{PIF}_2(x_{2i};T,F) \end{aligned}$$

3. Si x_{ij}, $1 \le j \le n_i$ son i.i.d. $x_{ij} \sim F_i$, independientes entre sí, entonces $\sqrt{(n_1+n_2)}(T(F_{1,n_1},F_{2,n_2}) - T(F))$ es asintóticamente normal con esperanza cero y matriz de covarianza dada por

$$\begin{aligned} \mathrm{V}(T,F) &= \lim_{\substack{n_1+n_2\to\infty \\ \frac{n_1}{n_2}=w}} (n_1+n_2)\textsc{Var}(T(F_{1,n_1},F_{2,n_2})) \\ &= \lim_{\substack{n_1+n_2\to\infty \\ \frac{n_1}{n_2}=w}} (n_1+n_2)\frac{n_1\textsc{Var}(\mathrm{PIF}_1)}{n_1^2} + (n_1+n_2)\frac{n_2\textsc{Var}(\mathrm{PIF}_2)}{n_2^2} \\ &= \frac{1}{w_1}\mathrm{V}_1 + \frac{1}{w_2}\mathrm{V}_2 \end{aligned}$$

   con $w_1 = 1 - w_2 = n_1/(n_1+n_2)$ y $\mathrm{V}_i = \int \mathrm{PIF}_i(x;T,F)\mathrm{PIF}_i(x;T,F)^{\mathrm{T}}dF_i(x)$.

4. La propiedad referente a transformaciones también es válida reemplazando IF por PIF_i

En Pires y Branco (2002), se muestra que el siguiente desarrollo es válido bajo condiciones de regularidad del funcional T

$$N^{\frac{1}{2}}\{T(F_N) - T(F)\} = \sum_{i=1}^{k} \frac{1}{(\tau_i n_i)^{\frac{1}{2}}} \sum_{j=1}^{n_i} \mathrm{PIF}(x_{ij},T,F) + o_p(1)$$

donde F_N representa la distribución empirica de k muestras independientes x_{ij}, $1 \le j \le n_i$, $1 \le i \le k$. Por lo tanto, la varianza asintótica de los estimadores puede ser evaluada por

$$\mathrm{asvar}(T,F) = \sum_{i=1}^{k} \tau_i^{-1} E_{F_i}\left(\mathrm{PIF}(x_{ij},T,F)\mathrm{PIF}(x_{ij},T,F)^{\mathrm{T}}\right)$$

Boente, Pires y Rodrigues (2002) obtuvieron la función de influencia para los estimadores *plug–in* bajo el modelo CPC, es decir, al utilizar matrices de escala robusta en lugar de las matrices de covarianza clásica en las ecuaciones de máxima verosimilitud para el caso normal.

4.2. Estimadores Robustos de Posición y Escala Multivariados

Para poder encontrar estimadores robustos de las componentes principales, Maronna (1976) y Campbell (1980) propusieron usar M–estimadores de la matriz de escala en las ecuaciones que las definen en lugar de las matrices de covararianza muestrales.

4.2.1. M–estimadores

En el caso univariado, se define como M–estimador de posición a la solución $\widehat{\mu}$ de una ecuación de la forma:

$$\sum_{i=1}^{n} \Psi\left(\frac{x_i - \mu}{s}\right) = 0 \tag{4.2}$$

donde s es un estimador de escala robusto para x. Las distintas propuestas surgen de la selección de Ψ. Huber (1964) propuso como estimador robusto para $\widehat{\mu}$ al que corresponde a la función de escores

$$\Psi_{\mathrm{H},a}(x) = \begin{cases} -a & \text{si } x < -a \\ x & \text{si } |x| < a \\ a & \text{si } x > a \end{cases}$$

donde a se elige de modo que el estimador resultante tenga una eficiencia del 95 % respecto del de máxima verosimilitud. En el caso normal, el valor de a es cercano a 1.5. Con esta función Ψ, (4.2) debe ser resuelto iterarivamente.

Los M–estimadores propuestos por Maronna (1976), Huber (1977a, b) son solución iterada del sistema

$$t^{(k+1)} = t^{(k+1)}(X) = \frac{\sum_{i=1}^{n} w_1(D_i^{(k)}) x_i}{\sum_{i=1}^{n} w_1(D_i^{(k)})} \tag{4.3}$$

$$V^{(k+1)} = V^{(k+1)}(X) = \frac{1}{n}\sum_{i=1}^{n} w_2(D_i^{(k)})(x_i - t^{(k)})(x_i - t^{(k)})^{\mathrm{T}} \tag{4.4}$$

donde $D_i^{(k)} = \left((x_i - t^{(k)})^{\mathrm{T}} V^{(k)^{-1}} (x_i - t^{(k)})\right)^{1/2}$ es la distancia de Mahalanobis de la $i-$ésima observación.
Una posible elección de las funciones de peso son las funciones de peso asociadas a la función de Huber, $w_1(t) = \Psi_{\mathrm{H},a}/t$, o sea,

$$w_1(D) = w_H(D) = I(D \leq a) + (a/D)I(D > c)$$

y

$$w_2(D^2) = \{w_1(D)\}^2 / c$$

con c una constante para que el estimador resulte Fisher–consistente, $a = \sqrt{\chi^2_d(\beta)}$ y $\beta = 0{,}90$.

Para las simulaciones el estimador de escala se divide por el factor 0,9363 para que resulte consistente bajo normalidad.

El problema de los M–estimadores con función de peso asociada a una función de scores monótona, es que su punto de ruptura disminuye al aumentar la dimensión del espacio siendo siempre inferior a $1/(p+1)$ (con p la dimensión del espacio) si la función de escores asociada a la función de peso es monótona. Para resolver este problema se introdujeron otras familias de estimadores robustos entre las que podemos mencionar el estimador de elipsoide de mínimo volumen (Rousseeuw y van Zomeren, 1990), el de mínimo determinante (MCD, Rousseeuw, 1985), el de Donoho (1982)–Stahel (1981) y los S-, MM- y $\tau-$estimadores (Lopuhaä, 1990). Todas estas propuestas pueden alcanzar un punto de ruptura igual a $\frac{1}{2}$. Entre ellos, sólo el de elipsoide de mínimo volumen converge a una tasa menor, $n^{\frac{1}{3}}$, mientras que los otros convergen a una tasa del orden de $\sqrt{n}$.

4.2.2. Estimador MCD

Recientemente, Croux y Haesbroeck (2000) usaron estimadores con punto de ruptura positivo como el método del determinante de minima covarianza (MCD) (Rousseeuw 1984) y estimadores S (Davies 1987, Rousseeuw and Leroy 1987) para definir componentes principales robustas. El estimador MCD tiene punto de ruptura cercano a 1/2 en cualquier dimensión pero tiene poca eficiencia bajo el modelo normal. Se define como la media y la matriz de covarianza de las h observaciones de una submuestra de tamaño h cuya matriz de covarianza muestral minimiza el determinante. El valor de h se elige igual a un valor entre el 50 % y el 75 % del tamanño de muestra total.

4.2.3. Estimador de Stahel-Donoho

El estimador de Stahel-Donoho fue el primer estimador robusto equivariante de posición y escala con un punto de ruptura alto en cualquier dimensión. El estimador se define como el promedio ponderado y la matriz de covarianza ponderada, donde el peso asignado a cada punto es función de una medida de atipicidad del mismo. Así, los puntos con medida de atipicidad grande reciben pesos chicos. La medida se basa en la idea de que si un punto es un dato multivariado atípico, entonces debe existir una proyección unidimensional en la cual sea un outlier. Sea $X=\{\ x_1, \ldots, x_n\ \}$ un conjunto de n puntos en $I\!R^p$. Sean m y s estimadores de posición y dispersión univariada. Se define para todo $y \in I\!R^p$ la atipicidad r como

$$r(y, X) = \sup_{a \in I\!R^p,\ a \neq 0} r_1(y, a, X) \tag{4.5}$$

donde

$$r_1(y, a, X) = \frac{|a^{\mathrm{T}} y - m(a^{\mathrm{T}} X)|}{s(a^{\mathrm{T}} X)} \tag{4.6}$$

Sea w una función de peso $R_+ \to R_+$, el estimador Stahel-Donoho (SDE) de posición y dispersión $(t(X), V(X))$ se define como

$$t = t(X) = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{4.7}$$

$$V = V(X) = \beta \frac{\sum_{i=1}^{n} w_i (x_i - t)(x_i - t)^{\mathrm{T}}}{\sum_{i=1}^{n} w_i} \tag{4.8}$$

con $w_i = w(r(x_i, X))$ y β una constante de calibración para obtener Fisher–consistencia.

Los estimadores de posición y escala multivariados resultan ser afín equivariantes si los estimadores univariados m y s lo son. Por otra parte, si deseamos estimadores Fisher–consistentes cuando $x_i \sim N_p(0, \mathrm{I}_p)$, la constante β debe elegirse igual a

$$\beta = \frac{p\, E\left(w\left(W_p\right)\right)}{E\left(w\left(W_p\right) W_p\right)},$$

donde $W_p \sim \chi^2_p$.

Stahel (1981) mostró que (t, V) tiene un punto de ruptura asintótica $\frac{1}{2}$ para modelos multivariados continuos si los estimadores de posición y escala univariados m y s tienen punto de ruptura asintótica $\frac{1}{2}$. Donoho (1982) encontró el punto de ruptura de (t, V) para muestras finitas, para el caso en que se eligen como estimadores de posición y escala univariados la mediana y la MAD (mediana de la desviación absoluta).

En el estudio de simulación, los pesos son calculados utilizando la función de Huber, o sea, la función $w(t) = w_H\left(\sqrt{t}\right)$ con w_H la función de peso de Huber dada por:

$$w_H(r) = I_{[0,c]}(r) + I_{(c,\infty)}(r) \left(\frac{c}{r}\right)^2$$

donde $c = \sqrt{\chi^2_{0{,}95,p}}$. Por otra parte, para obtener estimadores afín equivariantes, se eligieron como estimadores univariados $m(y_1, \ldots, y_n) = \underset{1 \le i \le n}{\text{median}}\,(y_i)$, la mediana de las observaciónes, y $s(y_1, \ldots, y_n) = \dfrac{1}{\Phi^{-1}(0{,}75)}\,\text{MAD}(y_1, \ldots, y_n)$, la mediana de los desvíos absolutos respecto de la mediana, escalada de modo a resultar Fisher–consistente cuando $y_i \sim N(0, 1)$.

Al utilizar la función de Huber $w(t) = w_H\left(\sqrt{t}\right)$ y para obtener Fisher–consistencia para datos normales, debe elegirse la constante de consistencia $\beta = p\,\dfrac{c_0}{c_2}$, donde

$$\begin{aligned} c_0 &= E\left(w(W_p)\right) = P\left(W_p < c^2\right) + c^2\, E\left(\frac{1}{W_p} I_{(c^2,\infty)}(W_p)\right) \\ &= P\left(W_p < c^2\right) + c^2\,\frac{1}{2\,(p-2)}\,\left(1 - P\left(W_{p-1} < c^2\right)\right) \quad \text{si} \quad p \neq 2 \\ c_2 &= E\left(w(W_p)W_p\right) = E\left(W_p I_{(0,c^2)}(W_p)\right) + c^2\, E\left(I_{(c^2,\infty)}(W_p)\right) \\ &= p\, P\left(W_{p+2} < c^2\right) + c^2\,\left(1 - P\left(W_p < c^2\right)\right)\,. \end{aligned}$$

4.2.4. Algoritmo para el cálculo aproximado del estimador de Stahel–Donoho para muestras finitas

Stahel (1981) propuso un algoritmo para el cálculo aproximado de (t, V) basado en submuestras. Se define $\tilde{r}$ como r en (4.5), pero el supremo se calcula para los vectores $\mathbf{a}$ en un conjunto finito $\mathcal{A}$ que se define a continuación. Para cada submuestra $\widetilde{\mathbf{X}}$ de tamaño p de $X = \{x_1, \ldots, x_n\}$, sea $\mathbf{a}$ la dirección ortogonal al hiperplano conteniendo $\widetilde{\mathbf{X}}$. Llamemos $\mathcal{A}$ el conjunto formado por todas estos vectores $\mathbf{a}$. Dado que $\mathcal{A}$ es grande salvo que p y n sean chicos se reemplaza $\mathcal{A}$ por una muestra aleatoria de tamaño N, $\mathcal{A}_N$ y se trabaja con r_N la medida de atipicidad resultante, en lugar del supremo en (4.5).

Maronna y Yohai (1981) mencionan en su trabajo que experimentos numéricos mostraron que para $p = 4$, eligiendo como estimadores de posición y escala univariados la mediana y el promedio de los k_1 y k_2 valores más chicos de la desviación absoluta alrededor de la mediana (una especie de MAD levemente modificada) donde

$$k_1 = p - 1 + \left[\frac{n+1}{2}\right] \quad y \quad k_2 = p - 1 + \left[\frac{n+2}{2}\right] \tag{4.9}$$

con $[t]$ la parte entera de t, valores de N entre 500 y 1000 hacen que r_N este muy próximo del valor óptimo r. Para $p = 6$, estos autores obtuvieron buenos resultados con $N = 1000$.

Estas elecciones de los parámetros hacen que se alcance la cota superior del punto de ruptura encontrada por Davies (1987) para muestras finitas.

4.2.5. Función de influencia para el estimador de Stahel–Donoho:

En esta sección vamos a ver algunas propiedades de las funciones elípticas, para después buscar la función de influencia de los estimadores de Stahel–Donoho en este caso. Los resultados que se resumen a continuación se pueden encontrar en Gervini (2002).

Para analizar las propiedades asinóticas de los estimadores de Stahel–Donoho, Gervini (2002) se restringe a las distribuciones elípticas. Estas son suficientemente flexibles para acomodarse a distribuciones con colas pesadas sin momentos finitos, pero que poseen parámetros de posición y dispersión finitos.

Se dice que un vector $X \in I\!R^p$ tiene un distribución elíptica, $F_{\mu,\Sigma}$, si su función de densidad es de la forma

$$f(x) = |\Sigma|^{-1/2}\, h\left((x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu_i)\right), \quad x \in I\!R^p$$

con h una función no negativa e integrable, $\mu \in I\!R^p$ es el parámetro de posición y Σ pertenece a la familia de matrices simétricas y definidas positivas.

El vector $Y = \Sigma^{1/2}(X-\mu)$, donde $\Sigma^{1/2}$ es la única raíz cuadrada simétrica de Σ tiene densidad $f(y) = h(\|y\|^2)$ y por lo tanto, su distribución es esférica. Una definición más general de distribución esférica, que no necesita de la existencia de funciones de densidad, requiere que la distribución de Y sea invariante por rotaciones ortogonales.

Tanto en Bilodeau y Brenner (1999), Hampel (1986), Muirhead (1982) se pueden encontrar muchas propiedades de las distribuciones esféricas, entre ellas:

- $F_{a'X} = F_{\|a\|X_1}$ para todo $a \in I\!R^p$ con F_{X_1} la distribución marginal de la primer coordenada de X que resulta simétrica alrededor de cero.
- X se puede factorizar como $X = RU$, con $R = \|X\|$ estocásticamente independiente de $U = \frac{X}{\|X\|}$. Más aún, U tiene distribución uniforme en $\mathcal{S}^{p-1}$ la esfera unitaria de $I\!R^p$. Por lo tanto, $E(U) = 0$ y $E(UU') = p^{-1}\mathrm{I}_p$

Dos modelos elípticos de importancia son:

- La Normal multivariada $N_p(\mu, \Sigma)$ donde

$$h(x) = \frac{1}{2\pi^{\frac{p}{2}}} \exp\left(-\frac{t}{2}\right)$$

- La distribución student multivariada con ν grados de libertad:

$$h(x) = \frac{\Gamma(\frac{\nu+p}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{p}{2}}} \left(\frac{\nu}{\nu+t}\right)^{\frac{\nu+p}{2}}$$

En el caso de estimadores de Stahel–Donoho, t y V, nos restringimos a aquellos que son equivariantes bajo transformaciones afines. Por lo tanto, en el caso de distribuciones elípticas $F_{\mu,\Sigma}$ tenemos:

$$\begin{aligned} T(F_{\mu,\Sigma}) &= \Sigma^{\frac{1}{2}} T(F_{0,\mathrm{I}_p}) + \mu \\ V(F_{\mu,\Sigma}) &= \Sigma^{\frac{1}{2}} V(F_{0,\mathrm{I}_p}) \Sigma^{\frac{1}{2}} \end{aligned}$$

Lo que implica que las funciones de influencia sean:

$$\begin{aligned} \mathrm{IF}(z, T, F_{\mu,\Sigma}) &= \Sigma^{\frac{1}{2}} \mathrm{IF}(\Sigma^{-\frac{1}{2}}(z-\mu), T, F_{0,\mathrm{I}_p}) \\ \mathrm{IF}(z, V, F_{\mu,\Sigma}) &= \Sigma^{\frac{1}{2}} \mathrm{IF}(\Sigma^{-\frac{1}{2}}(z-\mu), V, F_{0,\mathrm{I}_p}) \Sigma^{\frac{1}{2}} \end{aligned}$$

donde F_{0,I_p} es esférica. Esto justifica que en el caso de funciones elípticas se estudie sólo el caso F_{0,I_p}.

Gervini (2002) muestra que, bajo ciertas condiciones de regularidad, la función de influencia para los estimadores de Stahel–Donoho es de la forma:

$$\begin{aligned} \mathrm{IF}(z, T, F) &= \left(\frac{c_0(F)}{c_1(F)} g_1(\|z\|) + \frac{w(\|z\|^2/s_0^2)\|z\|}{c_0(F)} \right) \frac{z}{\|z\|} \\ \mathrm{IF}(z, V, F) &= \left(\frac{c_2(F)}{c_0(F)} g_2(\|z\|) + \frac{w(\|z\|^2/s_0^2)\|z\|^2}{c_0(F)} \right) \left(\frac{zz^T}{\|z\|} - \frac{I}{p} \right) \\ &+ \left(\frac{c_2(F)}{c_0(F)} g_3(\|z\|) + \frac{w(\|z\|^2/s_0^2)\|z\|^2 - c_3(F)}{c_0(F)} \right) \frac{1}{p} \mathrm{I}_p \end{aligned}$$

con

$$\begin{aligned} c_0(F) &= E(w(R^2/s_0^2)) \\ c_1(F) &= -2E((w'(R^2/s_0^2)R^2/s_0) \\ c_2(F) &= -2E((w'(R^2/s_0^2)R^4/s_0) \\ c_3(F) &= -2E((w'(R^2/s_0^2)R^2) \end{aligned}$$

donde w' es la derivada de los pesos w. Las funciones g_1, g_2 y g_3, en el caso en que los estimadores univariados sean la mediana y la MAD normalizada (para resultar Fisher–consistente) están dadas por:

$$\begin{aligned} g_1(x) &= \frac{\Gamma(\frac{p}{2})}{\sqrt{\pi}\Gamma(\frac{p+1}{2})\ 2\ s_0 f_1(0)} \\ g_2(x) &= \frac{F_{\beta(\frac{1}{2},\frac{p-1}{2})}\left(\frac{\Phi^{-1}(0{,}75)^2}{x^2}\right) - F_{\beta(\frac{3}{2},\frac{p-1}{2})}\left(\frac{\Phi^{-1}(0{,}75)^2}{x^2}\right)}{(p-1)2\Phi^{-1}(0{,}75) f_1(\Phi^{-1}(0{,}75))} \\ g_3(x) &= \frac{\frac{1}{2} - F_{\beta(\frac{1}{2},\frac{p-1}{2})}\left(\frac{\Phi^{-1}(0{,}75)^2}{x^2}\right)}{2\Phi^{-1}(0{,}75) f_1(\Phi^{-1}(0{,}75))} \end{aligned}$$

donde $R = \|X\|$, $X \sim F$ esférica, f_1 es la distribución marginal de X_1 y $\beta\,(p, q)$ es la distribución Beta de parámetros p y q. Si $F = N_p(0, \mathrm{I}_p)$ entonces $s_0 = 1.$.

4.2.6. Función de Influencia Parcial para los estimadores *plug–in* bajo el modelo CPC

Como hemos visto, las funciones de influencia son medidas de robustez con respecto a un dato atípico. Cuando se consideran varias poblaciones, las funciones de influencia parciales miden la resistencia de contaminaciones puntuales en cada población. Boente, Pires y Rodrigues (2002) obtuvieron las PIF de los estimadores robustos para el modelo CPC definidos en Boente y Orellana (2001). Estos estimadores se definían reemplazando las matrices de covarianza muestrales de cada población por estimadores robustos y consistentes V_i en las ecuaciones de máxima verosimilitud (2.13), (2.14) y (2.15). O sea, los estimadores robustos propuestos cumplen

$$\widehat{\beta}_m^{\mathrm{T}} \left(\sum_{i=1}^{k} N_i \frac{\widehat{\lambda}_{im} - \widehat{\lambda}_{ij}}{\widehat{\lambda}_{im}\widehat{\lambda}_{ij}} V_i \right) \widehat{\beta}_j = 0 \qquad 1 \le j \le p \quad 1 \le m \le p \quad m \ne j \tag{4.10}$$

$$\widehat{\beta}_m^{\mathrm{T}} \widehat{\beta}_j = \delta_{m,j} \tag{4.11}$$

$$\widehat{\lambda}_{im} = \widehat{\beta}_m^{\mathrm{T}} V_i \widehat{\beta}_m \qquad 1 \le i \le k \quad 1 \le m \le p \tag{4.12}$$

donde $\delta_{m,j}$ es la delta Kronecker.

Para una función de distribución dada $F = F_1 \times \ldots \times F_k$, sea $V_i(F_i)$ un funcional robusto de dispersión evaluado en la $i-$ésima muestra. Se definen los funcionales $\beta_V(F), \Lambda_{V,i}(F)(1 \le i \le k)$ asociados a los estimadores anteriores como la solución de

$$\mathrm{diag}\{\beta_V(F)^{\mathrm{T}} V_i(F) \beta_V(F)\} = \Lambda_{V,i}(F) \tag{4.13}$$

$$\beta_{V,\,j}^{\mathrm{T}} \left(\sum_{i=1}^{k} \tau_i \frac{\lambda_{V,im} - \lambda_{V,ij}}{\lambda_{V,im}\lambda_{V,ij}} V_i \right) \beta_{V,\,j} = 0 \quad m \ne j \tag{4.14}$$

$$\beta_{V,\,m}^{\mathrm{T}} \beta_{V,\,j} = \delta_{m,j} \; . \tag{4.15}$$

Cuando $V_i(F)$ es un estimador Fisher–consistente, es decir $V_i(F) = \Sigma_i$, las soluciones $(\Lambda_{V,i}(F), \beta_V(F))$ son Fisher–consistentes para (Λ_i, β).

El siguiente teorema da los valores de las funciones de influencia parcial para los estimadores *plug–in*.

Teorema 4 . *Sea $V_i(F)$ un funcional de dispersión tal que $V_i(F_i) = \Sigma_i$. Llamemos $\beta_1, \ldots, \beta_p$ los autovectores comunes asociados a los autovalores $\lambda_{i1}, \ldots, \lambda_{ip}$ de Σ_i. Si la función de influencia $IF(x, V_i, F_i)$ existe y si $\lambda_{i1} > \ldots, > \lambda_{ip}$, entonces, las funciones de influencia parcial de la solución $(\beta_V(F), \Lambda_{V,i}(F))$ de (4.13) a (4.15) están dadas por:*

$$PIF_i(x, \lambda_{V,\ell j}, F) = \delta_{\ell,i} \beta_j^{\mathrm{T}} IF(x, V_i, F_i) \beta_j$$

$$PIF_i(x, \beta_{V,\,j}, F) = \tau_i \sum_{m \ne j} \frac{\lambda_{ij} - \lambda_{im}}{\lambda_{im}\lambda_{ij}} \left\{ \sum_{\ell=1}^{k} \tau_\ell \frac{(\lambda_{\ell m} - \lambda_{\ell j})^2}{\lambda_{\ell m}\lambda_{\ell j}} \right\}^{-1} \left\{ \beta_j^{\mathrm{T}} IF(x, V_i, F_i)\, \beta_m \right\} \beta_m \; .$$

De este teorema se deduce que la función de influencia parcial para los estimadores clásicos está dada por:

$$\begin{aligned}
\mathrm{PIF}_i(x, \lambda_{S,l_j}, F) &= \delta_{l,i}\beta_j^{\mathrm{T}}(xx^{\mathrm{T}} - \sigma_i)\beta_j = \delta_{l,i}\{(\beta_j^{\mathrm{T}}x)^2 - \lambda_{ij}\} \\
\mathrm{PIF}_i(x, \beta_{S,\,j}, F) &= \tau_i \sum_{m\neq j} \frac{\lambda_{ij} - \lambda_{im}}{\lambda_{im}\lambda_{ij}} \left\{ \sum_{\ell=1}^{k} \tau_\ell \frac{(\lambda_{\ell m} - \lambda_{\ell j})^2}{\lambda_{\ell m}\lambda_{\ell j}} \right\}^{-1} \left\{ \beta_j^{\mathrm{T}} \left(xx^{\mathrm{T}} - \Sigma_i \right) \beta_m \right\} \beta_m \\
&= \tau_i \sum_{m\neq j} \frac{\lambda_{ij} - \lambda_{im}}{\lambda_{im}\lambda_{ij}} \left\{ \sum_{\ell=1}^{k} \tau_\ell \frac{(\lambda_{\ell m} - \lambda_{\ell j})^2}{\lambda_{\ell m}\lambda_{\ell j}} \right\}^{-1} \beta_j^{\mathrm{T}}x \;\beta_m^{\mathrm{T}}x \;\beta_m \;,
\end{aligned}$$

lo que muestra que no están acotadas.

Capítulo 5

Análisis Discriminante bajo el modelo CPC

5.1. Propuesta de Flury

Flury (1988) en su trabajo resalta la importancia que se le viene dando a las reglas de discriminación lineal y cuadrática pura pero no así a los casos intermedios, es decir, modelos que reduzcan de alguna manera los $p(p+1)/2$ parámetros del modelo cuadrático. Flury (1988) sugiere poner condiciones en la relación de las matrices de covarianza de las poblaciones. Tanto en el modelo proporcional como en el cuadrático, la regla discriminante sigue siendo cuadrática, pero los coeficientes están relacionados por restricciones que disminuyen la variabilidad de sus estimadores. Si las matrices Σ_i cumplen el modelo, es lógico utilizarlo. Resalta también que el mismo puede ser beneficioso por ser un modelo más parsimonioso. Como referencia, Flury (1988) resume los resultados analizados por Schmid (1987).

Schmid (1987) investigó el desempeño de los métodos de discriminación bajo el supuesto de distribución normal de las observaciones en los casos de: igualdad, proporcionalidad, CPC y desigualdad de las matrices de covarianza. La propuesta consistía en reemplazar en la regla cuadrática los parámetros desconocidos por los estimadores de máxima verosimilitud de $\widehat{\mu}_i$ y $\widehat{\Sigma}_i$ asociados a cada uno de esos modelos.

En ese trabajo, Schmid (1987) mostró que el modelo de discriminación CPC puede valer la pena en el caso de varios grupos y dimensión relativamente alta. La diferencia de parámetros entre el modelo CPC y el cuadrático es de $(k-1)p(p-1)/2$ lo que muestra que la ganancia en la disminución de parámetros se obtiene cuando k y p son grandes.

En el trabajo de Flury y Schmid (1992), los autores encontraron los valores asintóticos para las varianzas de los estimadores de la regla discriminante cuyos resultados se detallan a contiuación.

La regla obtenida en (1.6) se puede reescribir, en el caso de dos grupos, como:

$$q(x) = x^{\mathrm{T}} A x + b^{\mathrm{T}} x$$

con

$$A = -\frac{1}{2}(\Sigma_1^{-1} - \Sigma_2^{-1}) \tag{5.1}$$

$$b = \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2 \, . \tag{5.2}$$

Esta regla asigna x al grupo $\mathcal{G}_1$ si $q(x) \geq c(\mu_1, \mu_2, \Sigma_1, \Sigma_2)$ y a $\mathcal{G}_2$ si $q(x) \leq c(\mu_1, \mu_2, \Sigma_1, \Sigma_2)$

Como habíamos dicho en el capítulo 1 los parámetros Σ_i, μ_i usualmente son desconocidos. Flury y Schmid (1992) usan, en el caso sin restricciones, los estimadores usuales S_i de Σ_i mientras que cuando el modelo CPC es válido (discriminación CPC) utilizan los estimadores de máxima verosimilitud de $\Sigma_i = \beta\Lambda_i\beta^{\mathrm{T}}$ con β ortogonal de $p \times p$ y Λ_i=diag$(\lambda_1, \ldots, \lambda_p)$ y en ambos casos $\mu_i = \overline{x_i}$.

En ese trabajo, los autores trabajan bajo el supuesto de que las dos poblaciones eran independientes y con distribución normal, con matrices de covarianza diagonal, con lo cual, S_i, $i = 1, 2$, son matrices aleatorias independientes, $S_i \sim W_p(n_i, \Lambda/n_i)$ con $n_i \geq p$. Suponen además que si $r_i = \lim_{n\to\infty} \frac{n_i}{n}$, con $n = n_1 + n_2$, $0 < r_i < 1$.

Sea T_i el estimador de máxima verosimilitud de Λ_i obtenido bajo el modelo adecuado (cuadrático o CPC), $\overline{x}_i$, $i = 1, 2$, el promedio de las observaciones de la $i-$ésima población. Luego, $\overline{x}_i \sim N(\mu_i, \Lambda_i/(n_i + 1)$. Los Teoremas (5) y (6) que enunciaremos dan las aproximaciones asintóticas de las varianzas de :

$$\widehat{A} = -\frac{1}{2}(T_1^{-1} - T_2^{-1}) = (\widehat{a}_{jh}) \tag{5.3}$$

$$\widehat{b} = T_1^{-1}\overline{x}_1 - T_2^{-1}\overline{x}_2 = (\widehat{b}_1, \ldots, \widehat{b}_p)^{\mathrm{T}} \tag{5.4}$$

Teorema 5 . Discriminación Cuadrática: *Para n grande se tienen las siguientes aproximaciones*

a) $var(\widehat{a}_{jj}) \approx (2n)^{-1}(r_1^{-1}\lambda_{1j}^{-2} + r_2^{-1}\lambda_{2j}^{-2})$, $1 \leq j \leq p$

y

$var(\widehat{a}_{jh}) \approx (4n)^{-1}[(r_1\lambda_{1j}\lambda_{1h})^{-1} + (r_2\lambda_{2j}\lambda_{2h})^{-1}]$, $1 \leq j < h \leq p$

b) $var(\widehat{b}_j) \approx n^{-1}\sum_{j=1}^{2}(r_i\lambda_{ij})^{-1}\left(1 + \lambda_{ij}^{-1}\mu_{ij}^2 + \sum_{h=1}^{p}\lambda_{ih}^{-1}\mu_{ih}^2\right)$, $1 \leq j \leq p$.

Teorema 6 . Discriminación CPC: *Para n grande y si los elementos diagonales de Λ_i son todos distintos, se tienen las siguientes aproximaciones*

a) $var(\widehat{a}_{jj}) \approx (2n)^{-1}(r_1^{-1}\lambda_{1j}^{-2} + r_2^{-1}\lambda_{2j}^{-2})$, $1 \leq j \leq p$

y

$var(\widehat{a}_{jh}) \approx (4n)^{-1}\theta_{jh}[(\lambda_{1h}^{-1} - \lambda_{1j}^{-1}) - (\lambda_{2h}^{-1} - \lambda_{2j}^{-1})]^{-2}$ $j \neq h$, *donde*

$\theta_{jh} = [\theta_{jh}^{(1)^{-1}} + \theta_{jh}^{(2)^{-1}}]^{-1}$

y

$\theta_{jh}^{(i)} = r_i^{-1}\dfrac{\lambda_{ij}\lambda_{ih}}{(\lambda_{ij} - \lambda_{ih})^2} \qquad j \neq h.$

b) $var(\widehat{b}_j) \approx n^{-1}\left\{\sum_{j=1}^{2}(r_i\lambda_{ij})^{-1}\left(1 + 2\lambda_{ij}^{-1}\mu_{ij}^2\right) + \sum_{h=1,h\neq j}^{p}\theta_{hj}[\mu_{1h}(\lambda_{1h}^{-1} - \lambda_{1j}^{-1}) - \mu_{2h}(\lambda_{2h}^{-1} - \lambda_{2j}^{-1})]^{-2}\right\}$

Las demostraciones de estos teoremas se pueden encontrar en Flury y Schmid (1992).

Restringiendo los resultados asintóticos para $n_1 = n_2 = n/2$ es decir $r_1 = r_2 = 1/2$ y suponiendo $\mu_1 = 0$ y $\delta = \mu_2$, se obtiene

Varianzas asintóticas cuando el Modelo CPC es válido		
Coeficiente	**Modelo usado para discriminación**	**Varianza**
a_{jj} $j = 1, \ldots, p$	DIFF = CPC	$\lambda_{1j}^{-1} + \lambda_{2j}^{-1}$ $\lambda_{1j}^{-1} + \lambda_{2j}^{-1}$
a_{jh} $1 \leq j < h \leq p$	DIFF $\geq$ CPC	$\frac{1}{4}[\theta_{jh}^{(1)}(\lambda_{1h}^{-1} - \lambda_{1j}^{-1})^2 + \theta_{jh}^{(2)}(\lambda_{2h}^{-1} - \lambda_{2j}^{-1})^2]$ $\frac{1}{4}\frac{\theta_{jh}^{(1)}\theta_{jh}^{(2)}}{\theta_{jh}^{(1)} + \theta_{jh}^{(2)}}\left[(\lambda_{1h}^{-1} - \lambda_{1j}^{-1}) - (\lambda_{2h}^{-1} - \lambda_{2j}^{-1})\right]^2$
b_j $j = 1, \ldots, p$	DIFF $\geq$ CPC	$2\left(\lambda_{1j}^{-1} + \lambda_{2j}^{-1} + \lambda_{2j}^{-2}\delta_j^2 + \lambda_{2j}^{-1}\sum_{h=1}^{p}\lambda_{2h}^{-1}\delta_h^2\right)$ $2\left[\lambda_{1j}^{-1} + \lambda_{2j}^{-1} + \lambda_{2j}^{-2}\delta_j^2 + \lambda_{2j}^{-1}\left(\lambda_{2j}^{-1}\delta_j^2 + \sum_{h=1,h\neq j}^{p}\lambda_{2h}^{-1}\delta_h^2\phi_{jh}\right)\right]$

Flury y Schmid (1992) concluyeron que aunque en algunas circunstancias se estén trabajando con modelos incorrectos (por ejemplo, usar el CPC cuando en realidad sea cuadrático) el hecho de usar un modelo parsimonioso puede mejorar los resultados. La siguiente pregunta que se plantearon fue qué modelo utilizar para minimizar los errores de mala clasificación, lo que motivó el trabajo de Flury, Schmid y Narayanan (1994). En dicho trabajo, los autores utilizan el método de Monte Carlo para calcular las aproximaciones de las tasas esperadas de errores. Trabajaron con cuatro diseños diferentes para comparar los resultados obtenidos por las diferentes reglas de clasificación. El diseño cuatro de Flury, Schmid y Narayanan (1994) armado especialmente para poder aplicar CPC, muestra las ventajas sobre sus competidores. En el diseño 5 que era un modelo cuadrático, el modelo CPC mostró un buen desempeño para muestras de tamaño 60 y los autores resaltan que hasta muestras de tamaño 100, no existe ventaja significativa de la regla cuadrática usual sobre la generada por el modelo CPC.

5.2. Propuesta plug-in robusta

Dado que los estimadores de máxima verosimilitud se ven fuertemente afectados por las observaciones atípicas así como por desvíos de la distribución de los datos respecto de la distribución normal, nuestra propuesta robusta consiste en estimar los parámetros de posición y dispersión por estimadores robustos consistentes de posición y escala y construir a partir de ellos estimadores plug–in de las componentes principales comunes como se describió en la sección 4.2.6, utilizando el algoritmo FG.

Es decir, que partiendo de la regla

$$q(x) = x^{\mathrm{T}} A x + b^{\mathrm{T}} x$$

la cual asigna x al grupo $\mathcal{G}_1$ si $q(x) \geq c(\mu_1, \mu_2, \Sigma_1, \Sigma_2)$ y a $\mathcal{G}_2$ si $q(x) \leq c(\mu_1, \mu_2, \Sigma_1, \Sigma_2)$, donde A y b están definidos en (5.1) y (5.2) respectivamente.

Si las matrices de varianza de las dos poblaciones cumplen el modelo CPC, definimos como estimador robusto de la matriz Σ_i,

$$\widehat{\Sigma}_{i,\mathrm{CPC}} = \widehat{\beta}\widehat{\Lambda_i}\widehat{\beta}$$

donde $\widehat{\beta}$ y $\widehat{\Lambda_i}$ están definidos en (4.10) a (4.12) utilizando el estimador robusto de escala de Stahel-Donoho $V_i = V(\mathbf{X}_i)$ definido en (4.8), siendo $\mathbf{X}_i$ las observaciones de la $i-$ésima población. Es decir, utilizamos los estimadores *plug–in* asociados al estimador de Stahel–Donoho solución de

$$\begin{aligned} \widehat{\beta}_m^{\mathrm{T}} \left[\sum_{i=1}^{k} N_i \frac{\widehat{\lambda}_{im} - \widehat{\lambda}_{ij}}{\widehat{\lambda}_{im}\widehat{\lambda}_{ij}} V_i \right] \widehat{\beta}_j &= 0 \qquad \text{para } m \neq j \\ \widehat{\beta}_m^{\mathrm{T}} \widehat{\beta}_j &= \delta_{mj} \\ \widehat{\Lambda}_i &= \operatorname{diag}\left(\widehat{\beta}^{\mathrm{T}} V_i \widehat{\beta} \right) \end{aligned}$$

donde $\delta_{m,j}$ es la delta Kronecker. Como estimador del parámetro de posición μ_i utilizamos $t_i = t(\mathbf{X}_i)$ el correspondiente estimador robusto de posición Stahel-Donoho definidos en (4.7) para la población $i-$ésima.

Capítulo 6

Estudio de Simulación

Dada la dificultad desde el punto de vista teórico para validar las propiedades de robustez o para comparar los métodos multivariantes es frecuente recurrir a los estudios de simulación. Para que las conclusiones obtenidas de esta manera sean válidas hay que prestar mucha atención en el perfilamiento del problema de interés.

A continuación, daremos el delineamiento de Ana Pires (1995) del método Monte Carlos. Este método es el que utilizaremos para validar el comportamieno de la Regla plug–in robusta que propusimos en el capítulo anterior para el problema Discriminante bajo el modelo CPC para las matrices de covarianzas definido por Flury (1988).

6.1. Pasos del Método de Montecarlo

1. Identificación del problema de interés: Estimación de una regla discriminante para dos grupos, bajo la hipótesis de que el modelo CPC para las matrices de covarianzas es válido.
2. Identificación de los factores que pueden influenciar el desempeño de los diversos métodos:
   - *a*) Distribución de las poblaciones: tipo de distribución, medias y dispersiones.
   - *b*) Número de variables p
   - *c*) Dimensión de las muestras de entrenamiento
   - *d*) Probabilidades a priori π_1 y π_2

6.2. Reglas discriminantes a comparar

Para poder analizar el buen comportamiento o no de la regla propuesta, se compararán los errores aparentes obtenidos por las siguientes estimaciones:

- Métodos no robustos
  1. Clasificación por la regla cuadrática clásica
  2. Clasificación por la regla cuadrática clásica combinada con algoritmo FG
- Métodos robustos
  1. Método de estimadores de Maronna
  2. Método de estimadores de Maronna combinado con algoritmo FG
  3. Estimador de Stahel-Donoho
  4. Estimador de Stahel-Donoho con algoritmo FG

6.3. Parámetros elegidos para las simulaciones

La regla de discriminación se obtendrá bajo las siguientes condiciones:

- Dos poblaciones de tamaño $n_1 = n_2 = 100$
- La dimensión del espacio es $p = 4$.
- Los parámetros se estiman a partir de una muestra de desarrollo generada aleatoriamente bajo las condiciones definidas.
- El error aparente de mala clasificación se calcula mediante una muestra de validación generada con la misma contaminación que la muestra de desarrollo y sin contaminar.
- En el primer ejemplo también se incluye el error de validación cruzada dejando un dato afuera.

Este proceso se repite 1000 veces. De esta manera se calcula el promedio de los errores aparentes que se utilizarán para comparar las distintas reglas.

6.3.1. Elección de parámetros para Stahel–Donoho

- Parámetros de iteración:
  - Cantidad de direcciones para el cálculo de las proyecciones: 1000
- Pesos elegidos: pesos de Huber con $c = 3$ y $q = 2$:

$$w_H(r) = I(r <= c) + (c/r)^q I(r > c)$$

En el trabajo de Maronna y Yohai (1981) se muestra que el estimador de Donoho Stahel con los pesos de Huber y con con la constante de escala $c = \sqrt{\chi^2_4(0{,}95)} \approx 3$ es capaz de alcanzar una alta eficiencia tanto en el caso normal como para distribuciones de colas pesadas como la Cauchy a la vez que mantiene errores medios mucho más bajos que los estimadores S en los casos de datos contaminados de forma asimétrica.

La elección de la constante c, en el caso normal, representa la proporción de observaciones con máximo peso.

- Parámetros de posición y dispersión univariados:
  - Mediana
  - $\sigma = \frac{1}{\Phi^{-1}(0{,}75)} MAD \approx \frac{1}{0{,}674} MAD$. La constante $\frac{1}{\Phi^{-1}(0{,}75)}$ asegura la consistencia del estimador de escala, en el caso normal, al valor del desvío estándar.

6.3.2. Ejemplos

El trabajo de Flury y Smith (1992) muestra que la ventaja de utilizar el método CPC se debería poner en evidencia cuando los autovalores de las poblaciones guardan las siguientes relación:

$$\lambda_{1h}^{-1} - \lambda_{1j}^{-1} = \lambda_{2h}^{-1} - \lambda_{2j}^{-1} \qquad \text{para todo } (j, h).$$

Estas condiciones son las que tienen en cuenta Flury, Schmid y Narayanan (1994) en el diseño 4 de su trabajo para comparar las reglas discriminantes. Este mismo diseño será el que utilizaremos para validar la regla que estamos proponiendo.
Para comprobar el buen comportamiento de la misma bajo la presencia de datos atípicos se consideran dos tipos contaminaciones: puntuales o con mayor variabilidad en porcentajes del 10 % y 20 %.

Poblaciones Modeladas:

$$\mathcal{G}_1 \quad \sim N_4(0, \Sigma_1) \qquad \mathcal{G}_2 \sim N_4(\mu_2, \Sigma_2)$$

$$\mu_2 = (1,0,0,0) \qquad \Sigma_1 = \begin{pmatrix} 1/5 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 2/3 & 0 \\ 0 & 0 & 0 & 5/6 \end{pmatrix} \qquad \Sigma_2 = \begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 5 \end{pmatrix}$$

6.3.3. Ejemplo 1 :

Contaminación de la muestra de desarrollo $(1-\alpha) * N_4(\mu_i, \Sigma_i) + \alpha N_4(\mu_i, 9 * \Sigma_i)$

Errores Aparentes de la muestra de Validación contaminada en el mismo % que la de desarrollo				**Validación Cruzada**		
Porcentaje de contaminación:						
Estimador	$\alpha = 0$	$\alpha = 0{,}10$	$\alpha = 0{,}20$	$\alpha = 0$	$\alpha = 0{,}10$	$\alpha = 0{,}20$
Sin estimar	0.0997	0.1283	0.1562	0.0997	0.1283	0.1562
D-S	0.1049	0.1323	0.1645	0.1084	0.1378	0.1693
D-S CPC	0.1032	0.1310	0.1624	0.1112	0.1263	0.1549
M est	0.1044	0.1324	0.1641	0.1078	0.1380	0.1696
Mest CPC	0.1030	0.1313	0.1623	0.1181	0.1303	0.1535
Est insesg	0.1043	0.1553	0.2185	0.1070	0.1620	0.2252
Est insesg CPC	0.1027	0.1543	0.2196	0.1169	0.2165	0.3216

Errores Aparentes de la muestra de Validación no contaminada		
Porcentaje de contaminación:		
Estimador	$\alpha = 0{,}10$	$\alpha = 0{,}20$
Sin estimar	0.0997	0.0997
D-S	0.1061	0.1150
D-S CPC	0.1048	0.1135
M est	0.1061	0.1145
Mest CPC	0.1050	0.1133
Est insesg	0.1336	0.1903
Est insesg CPC	0.1328	0.1934

Como se puede observar en los cuadros, los errores aparentes calculados al utilizar estimadores robustos siempre mostraron una mejora cuando se utilizó la restricción CPC. El estimador de la regla cuadrática también muestra mejoras en los casos sin contaminación o con contaminación del 10 %. Sin embargo, con contaminación del 20 % la estimación ya no puede mejorarse y utilizar la hipótesis CPC empeora aún más el estimador. La ventaja de utilizar procedimientos robustos en lugar del procedimiento clásico se observa en el caso de ambas contaminaciones, donde los errores aparentes aumentan considerablemente. En el caso de los errores calculados por validación cruzada, la hipótesis CPC no los mejora cuando no hay datos atípicos, sí hay una leve mejoría cuando los datos están contaminados.

En el caso de contaminación del 10 % y muestra de validación contaminada, el estimador robusto de Stahel–Donoho es el que muestra mejores resultados mientras que en el caso de contaminación del 20 % ambos estimadores robustos muestran errores muy cercanos. En los caso en que la muestra de validación no estuvo contaminada, los estimadores robustos tuvieron comportamientos similares.

A continuación calculamos la tasa de crecimiento del error aparente de la muestra de validación al clasificarla con la regla obtenida con contaminación α en relación a la no contaminada. Como podemos observar, el error crece más rápidamente con α para los estimadores insesgados que para los robustos lo que muestra la ventaja de utilizar este tipo de estimadores.

Tasa de crecimiento del Error Aparente de la muestra no contaminada				
Porcentaje de contaminación de la muestra de desarrollo vs $\alpha = 0$				
Estimador	Error Aparente	$\alpha = 0{,}00$	$\alpha = 0{,}10$	$\alpha = 0{,}20$
D-S	0.1049	1.0000	1.0114	1.0963
D-S CPC	0.1032	1.0000	1.0155	1.0998
M est	0.1044	1.0000	1.0163	1.0967
Mest CPC	0.1030	1.0000	1.0194	1.1000
Est insesg	0.1043	1.0000	1.2809	1.8245
Est insesg CPC	0.1027	1.0000	1.2931	1.8832

6.3.4. Ejemplo 2 :

Contaminacón $(1-\alpha) * N_4(\mu_i, \Sigma_i) + \alpha b$

$$b = \begin{pmatrix} 10 & 0 & 0 & 0 \end{pmatrix}'$$

Errores Aparentes de la muestra de Validación contaminada en el mismo % que la de desarrollo			
Porcentaje de contaminación:			
Estimador	$\alpha = 0$	$\alpha = 0{,}10$	$\alpha = 0{,}20$
Sin estimar	0.0997	0.1405	0.1805
D-S	0.1049	0.1622	0.2390
D-S CPC	0.1032	0.1597	0.2360
M est	0.1044	0.2251	0.2576
Mest CPC	0.1030	0.2229	0.2531
Est insesg	0.1043	0.2531	0.2551
Est insesg CPC	0.1027	0.2501	0.2502

Errores Aparentes de la muestra de Validación no contaminada		
Porcentaje de contaminación:		
Estimador	$\alpha = 0{,}10$	$\alpha = 0{,}20$
Sin estimar	0.0997	0.0997
D-S	0.1249	0.1771
D-S CPC	0.1221	0.1735
M est	0.1972	0.2446
Mest CPC	0.1946	0.2411
Est insesg	0.2399	0.2443
Est insesg CPC	0.2369	0.2406

En el caso de contaminaciones puntuales, la restricción CPC siempre muestra una mejora del estimador en función del error aparente y el estimador de Stahel–Donoho es el que muestra mejor comportamiento con ambos niveles de contaminación.

Si se observa la tasa de crecimiento del error aparente de la muestra de validación al clasificarla con la regla obtenida con contaminación α en relación a la no contaminada, se nota el rápido crecimiento de la misma tanto para el M–estimador como para el estimador plug–in insesgado, llegando a duplicarse.

Tasa de crecimiento del Error Aparente de la muestra no contaminada				
Porcentaje de contaminación de la muestra de desarrollo vs $\alpha = 0$				
Estimador	Error Aparente	$\alpha = 0{,}00$	$\alpha = 0{,}10$	$\alpha = 0{,}20$
D-S	0.1049	1.0000	1.1907	1.6883
D-S CPC	0.1032	1.0000	1.1831	1.6812
M est	0.1044	1.0000	1.8889	2.3429
Mest CPC	0.1030	1.0000	1.8893	2.3408
Est insesg	0.1043	1.0000	2.3001	2.3423
Est insesg CPC	0.1027	1.0000	2.3067	2.3427

Apéndice

1.1. Estimadores de máxima verosimilitud

1.1.1. Invarianza de los EMV

Sea $f : \theta \to f(\theta)$ una función real en un dominio Θ, y sea $g : \theta \to g(\theta) = \phi$ biyectiva de Θ en Σ, la cual tiene inversa g^{-1}.
Llamemos $h(\theta) = f(g^{-1}(\widehat{\theta}))$ Si f alcanza un máximo en $\widehat{\theta}$, entonces h alcanza un máximo en $\widehat{\phi} = g(\widehat{\theta})$

Por lo tanto, si $\widehat{\theta}$ es un estimador de máxima verosimilitud de θ, entonces $\widehat{\lambda} = f(\widehat{\theta})$ es un estimador de máxima verosimilitud de λ.

Propiedad 1 . *Bajo condiciones muy generales los estimadores de máxima verosimilitud son fuertemente consistentes.*

Teorema 7 . *Sea $X_1, \ldots, X_2$ una muestra aleatoria de una distribución perteneciente a la familia $F(x, \theta)$ con $\theta \in \Theta, \Theta$ un abierto en $\mathbb{R}$ Supongamos que $F(x, \theta)$ tiene una función de densidad o frecuendia $p(x, \theta)$ que satisface:*

A: *El conjunto $S = \{x : p(x, \theta) > 0\}$ es independiente de θ*

B: *$\delta(X)$ es un estadístico tal que $E_\theta(|\delta(X)|) < \infty \quad \forall \theta \in \Theta$, entonces*

$$\frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \delta(x) p(x, \theta) dx = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \delta(x) \frac{\partial p(x, \theta)}{\partial \theta} dx$$

C: *El número de información de Fisher dado por:*

$$I(\theta) = E_\theta \left(\left(\frac{\partial \ln p(x, \theta)}{\partial \theta} \right)^2 \right) < \infty$$

D: *Llamemos* $\Psi(x,\theta) = \ln p(x,\theta)$ *y supongamos que*

$$\frac{\partial^3 \Psi(x,\theta)}{\partial \theta^3} \leq K \quad \forall x \in S$$

Si θ_n *es un estimador de máxima verosimilitud consistente de* θ *y* $q(\theta)$ *es una función derivable tal que* $q'(\theta) \neq 0 \quad \forall \theta$ *entonces* $\sqrt{n}\,(q(\theta_n) - q(\theta))$ *converge en distribución a una distribución normal con media cero y varianza* $[q'(\theta)]^2/I(\theta)$*, o sea,* $q(\theta_n)$ *es asintóticamente normal y eficiente (A.N.E)*

1.2. Propiedades de la distribución de Wishart

Teorema 8 *Si las matrices aleatorias de* $m \times m$ $A_1, \ldots, A_r$ *son independientes y* A_i *es* $W(n_i, \Sigma)$ *entonces* $\sum_{i=1}^{r} A_i$ *es* $W(n, \Sigma)$ *con* $n = \sum_{i=1}^{r} n_i$.

Teorema 9 *Si* A *es* $W_m(n, \Sigma)$ *y* M $k \times m$ *de rango* k *entonces* MAM^{T} *tiene distribución* $W_k(n, M\Sigma M^{\mathrm{T}})$.

1.3. El operador vectorial y el producto matricial Kronecker

1.3.1. Operador vectorial: vec

Se define el operador vec(A) como el operador que transforma una matriz en un vector: $vec : A \in I\!R^{m \times n} \longrightarrow vec(A) \in I\!R^{mn}$ Si $A = (a_1, \ldots, a_n)$ entonces

$$vec(A) = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

1.3.2. Operador Matricial de Kronecker

Sea $A = (a_{ij})$ una matriz de $p \times q$ y $B = (b_{ij})$ una matriz de $r \times s$. Se define el producto de Kronecker de A y B, $A \otimes B$, como la matriz de dimension $rp \times qs$

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1q}B \\ a_{21}B & a_{22}B & \dots & a_{2q}B \\ \vdots & & & \vdots \\ a_{p1}B & a_{p2}B & \dots & a_{pq}B \end{bmatrix}$$

Propiedad del operador de Kronecker

- Si A,B,C son matrices de $k \times l, l \times m, m \times n$ respectivamente, entonces $vec(ABC) = (C \otimes A)vec(B)$
- Si A,B,C,D son matrices de $m \times n, p \times q, n \times r, q \times s$ respectivamente, entonces $(A \otimes B)(C \otimes D) = (AB) \otimes (CD)$

Referencias

Ana M. V. N. Pires de Melo Parente (1995). Análise disciminante Novos Métodos Robustos de Estimacião. *Tese de Doutoramento* Universidade Técnica de Lisboa.

Anderson (1963). Asymptotic theory por principal component analysis. *Ann. Math. Statist.* , **34** , 122–148.

Anderson, T.W. (1984). An Introduction to Multivariate Statistical Methods *John Wiley* (2d ed.) New York.

Bilodeau y Brenner (1999). Theory of Multivariate Statistics. *Springer* , NY.

Boente, G. y Orellana, L. (2001). A robust approach to common principal components. In *Statistics in Genetics and in the Environmental Sciences* , Ed. L. T. Fernholz, S. Morgenthaler and W. Stahel, 117–47. Basel: Birkhauser.

Boente, Pires y I. Rodrigues (2002). Influence functions and outlier detection under the common principal components model: A robust approach. *Biometrika* , **89** , 4 , 861-875.

Bianco, A. and Boente, G. (2003). Robust estimators in semiparametric partly linear regression models. *J. Statist. Planning and Inference* .

Campbell (1978). The influence function as an aid in outlier detection in multivariate analysis. *Applied Statist.* , **27** , 251–258.

Campbell (1980). Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation, *Applied Statist.* , **29** , 231-237.

Croux, C., and Haesbroeck, G. (2000). Principal Components Analysis based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies. *Biometrika* , **87** , 603–618.

Davies (1987). Asympotic Behavior of S-Estimates of Multivariate Location Parameters and Dispersion Matrices. *Ann. Statist.* , **15** , 1269–1292.

Donoho (1982). Breakdown Properties of Multivariate Location Estimators. *Ph.D. qualifying paper* , Harvard University.

Efron (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation.

J. Amer. Statist. Assoc. , **78** , 316–331.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. of Eugenics* , **7** , 179–188.

Flury B. (1988). Common Principal Components & Related Multivariate Models. *John Wiley & Sons* .

Flury y Gautschi (1986). An Algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *Siam. J. on Scientific and Statist. Computing* , **7** , 169–184.

Flury B. y Schmid M. (1992). Quadratic Discriminant Functions with constrains on the Covariance Matrices: Some asymptotic Results. *Academic Press* , **40** , N 2.

Flury B., Schmid M. y Narayanan (1994). Error Rates in Quadratic Discrimination with constrains on the Covariance Matrices. *Journal of Classificarion* , **Vol 2** .

Gervini, D. (2002). The influence function of the Donoho–Stahel estimator of multivariate location and scale. *Stat. Probab. Letters* , **60** , 425–435.

Golub y Van Loan (1983). Matrix Computation. *The Johns Hopkins University Press* , Baltimore , MD .

Hampel (1968). Contributions to the theory of robust estimation *PH.D. Thesis* University of California , Berkeley.

Hampel (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* , **69** .

Hampel, Ronchetti, Rousseuw, Stahel (1986). Robust Statistics: The Approach Based on Influence Functions. *John Wiley & Sons* .

Hotelling (1933). Analysis of complex statistical variables into principal components. *J. Educ. Psychol.* , **24** , 417–441 , 498–520.

Huber (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* , **35** , 73–101.

Huber (1977). Robust Statistical Procedures . *CBMS-NSF Regional Conference Series in Applied Mathematics* , **27** , Published by Siam., Philadelphia, Penn.

Jacobi (1846). Concerning an easy process for solving equations occurring in the theory of secular disturbances. *J. fur reine un angewandre Mathematik* , **30** , 51–94.

Jolicoeur y Mosimann (1960). Size and shape variation in the painted turtle: a principal component analysis. *Growth* , **24** , 339–354.

Jolicoeur (1963b) The degree of generality of robustness in Martes Americana. *Growth* , **27** , 1–27.

Johson (1998). Applied Multivariate Statistical Analysis. *Prentice Hall* .

Lachenbruch y Mickey (1968). Estimation of error rates in discriminant analysis. *Technometrics* , **10** , 1–11.

Lopuhaä, H. P. (1990), Estimation of Location and Covariance with High Breakdown Point. *Ph. D. Thesis* . Delft University of Technology, Netherlands.

Mallows (1975). On some topics in robustness. *Technical Memorandum.* Bell Telephone Laboratories, Murray Hill, NJ [2.1b].

Marks y Dunn (1974). Discriminant functions when covariance matrices are unequal. *J. Amer. Statist. Assoc.* , **69** , 555–559.

Maronna (1976). Robust M–Estimators of multivariate location and scatter. *Ann. Statist.* , **4** , 51–67.

Maronna y Yohai (1981). The behavior of the Stahel-Donoho Robust Multivariate Estimator. *J. Amer. Statist. Assoc. March 1995* , Vol 90 , **429** .

McLachlan (1980). The efficiency of Efron's "bootstrap.ªpproach applied to error rate estimation in discriminant analysis. *J .Stat. Comput. Simulat.* , **11** , 273–279.

Muirhead (1982). Aspects of Multivarite Statistical theory. *John Wiley & Sons, Inc* .

O'Neill (1984). A Theoretical Method of Comparing Classification Rules under Non-Optimal Conditions with Application to the Estimates of Fisher's Linear and Quadratic Discriminant Rules under Unequal Covariance Matrices. *Technical Report* , **217** , Stanford University, Department of Statistics.

Pearson (1901). On lines an pleanes of closest fit to systems of points in space. *Philosophical Magazine* , ser 6 , **2** , 559–572.

Pires y Branco (2002). Partial influence functions. *J. Mult. Anal.* .

Radhakrishan (1983). Influence function for certain parameters in discriminant analysis, *Metron* , **30** , 183–194.

Radhakrishan and Kshirsagar (1981). Influence function for certain parameters in multivariate analysis. *Communications in Statistics-Theory and methods* , **10** , 515–529.

Rao (1973, 2nd ed.). Linear Statistical Inference and its Applications.*John Wiley* New York.

Rousseeuw, P. (1981 a). A new infinitesimal approach to robust estimation. *Z. Wahrsch verw. Geb* , **56** .

Rousseeuw, P. (1984). Least Median of Squares Regression. *J. Amer. Statist. Assoc.* , **79** , 871-880 .

Rousseeuw, P. (1985). Multivariate estimation with High Breakdown Point. *Mathematical Statistics and Applications* , (W. Grossmann, G. P.ug, I. Vincze and W. Werz, eds.), Dordrecht: Reidel, 283–297.

Rousseeuw, P., and Leroy, A. (1987). Robust Regression and Outlier Detection. *John Wiley* New York.

Rousseeuw, P. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *J. Amer. Statist. Assoc.,* **85** , 633–639.

Seber (1984). Multivariate Observations. *John Wiley & Sons* .

Schmid (1987). Anwndungen der Theorie proportionaler Kovarianzmatrizen und gemeinsamer Hauptkomponenten auf die quadratische Diskriminanzanalyse *PhD dissertation* , University of Berne, Department of Statistics, Berne, Switzerland .

Smith, C.A.B. (1947). Some examples of discrimination. *Ann. of Eugenics* , **13** , 272–282 .

Stahel (1981). Breakdown of Covariance Estimators, *Research Report* , **31** , Fachgruppe fur Statistik, E.T.H. Zurich.

Stigler, S.M. (1980). Sudies in the history of probability and statistics XXXVIII: R.H.Smith, a Victorian interested in robustness. *Biometrika* , **67** , 212–221.

Wald (1944). On a Statistical Problem Arising in Classification of an Individual into One of Two Groups. *Ann. Math. Statist.* , **15** , 145–162.

Welch (1939) Note on Discriminant Functions. *Biometrika* , **31** , 218–220.