Algunas Propiedades de la Regresión Inversa Partida y un Nuevo Enfoque para el Método

María Eugenia Szretter Noste

Director: Dr. Víctor Yohai

Tesis para acceder al título de

Magíster en Estadística Matemática

Universidad de Buenos Aires

Resumen

En esta tesis se estudia el procedimiento de reducción de la dimensión para observaciones multivariadas conocido como Regresión Inversa Partida *(Sliced Inverse Regression, SIR)* propuesto por K. C. Li [18]. En ella demostramos que el algoritmo desarrollado por Li [18] para resolver el problema de la regresión inversa partida es equivalente a obtener las coordenadas discriminantes correspondientes a la partición *(slicing)* de las covariables de acuerdo a los valores que toma la variable respuesta en cada observación.

Por otro lado, probamos que dicho algoritmo proporciona los mismos resultados que estimar por máxima verosimilitud el subespacio que contiene a las medias de los grupos derivados de la partición, suponiendo normalidad y una misma matriz de covarianza.

La utilidad de este enfoque subyace en poder enmarcar al estimador en una estrategia de máxima verosimilitud, lo cual permite aplicarle al mismo los resultados conocidos para este método general de estimación. A la vez, permite encontrar una vía alternativa a las ya propuestas (Gather et al. [10], [11]) para obtener un método robusto de estimación.

Abstract

In this thesis we study the procedure of dimension reduction for multivariate observations known as Sliced Inverse Regression *(SIR)* presented by K. C. Li [18]. We prove that the algorithm developed by Li [18] to solve the problem of sliced inverse regression is equivalent to that obtained with the discriminant coordinates corresponding to the groups formed by slicing the covariates in accordance to the values of the dependent variable of each observation.

On other hand, we prove that this algorithm provides the same results as those obtained by the maximum likelihood method of the subspace that contains the means of the groups induced by the slicing of the observations, under the assumption of normality and equal covariance matrix.

The usefulness of this approach lies on the possibility of viewing the estimators within a maximum likelihood strategy. This enables to apply to the obtained estimators the well known properties of this general estimation method. Also, it allows to search for an alternative way of finding a robust estimator, different from those proposed so far (Gather et al.[10], [11]).

Índice

1. Introducción

Los modelos de regresión establecen una relación entre una variable de respuesta y y un conjunto de varias variables explicativas, que supondremos almacenadas en un vector $\mathbf{x}$ p-dimensional. El más simple de los modelos de regresión es el modelo de regresión lineal. En algunas aplicaciones, los modelos de regresión lineal ajustan satisfactoriamente a los datos y entonces, los métodos estadísticos clásicos o robustos permiten estimar los parámetros del modelo, y hacer inferencia a partir de ellos.

Sin embargo, en la mayoría de las aplicaciones, cualquier modelo paramétrico constituye sólo una aproximación al modelo subyacente, y la búsqueda de un modelo adecuado no es sencilla. Cuando no hay ningún modelo paramétrico disponible que proporcione un modelo adecuado de los datos, las técnicas de regresión no paramétricas surgen como alternativas más flexibles. Como denominador común, estas técnicas no paramétricas explotan la idea de suavizado local, que solamente utiliza las propiedades de continuidad o diferenciabilidad local de la función de regresión verdadera. El éxito del suavizado local depende de la presencia de una cantidad suficiente de observaciones alrededor de cada punto de interés en el espacio, para que éstas puedan proveer la información adecuada para la estimación. Para problemas de una dimensión existen muchas técnicas disponibles para atacar el problema de regresión no paramétrica. Sin embargo, cuando la dimensión p de $\mathbf{x}$, aumenta, el número total de observaciones requeridas para que las técnicas de suavizado local funcionen crece exponencialmente con p. Entonces, a menos que dispongamos de una muestra de tamaño gigantesco, impracticable en la mayoría de las aplicaciones, los diferentes métodos de suavizado (como los estimadores por núcleos o estimadores de vecinos más cercanos) fallan debido a la escasez de observaciones en las zonas de interés. Esto es lo que se conoce como "la maldición de la dimensionalidad".

Una manera de solucionar esta dificultad es utilizar modelos donde la variable y dependa de $\mathbf{x}$ a través de proyecciones en unas pocas direcciones. En esta línea de trabajo, K. C. Li [18] propone en 1991 el modelo de regresión inversa partida, que es presentado en la Sección 2. Dada una muestra de observaciones $(\mathbf{x}_i', y_i)'$ $(1 \leq i \leq N)$, este modelo permite reducir la cantidad de covariables utilizando sólo algunas pocas combinaciones lineales. Estas combinaciones bastan para explicar la mayor parte de la relación entre y y el vector $\mathbf{x}$.

El modo de proceder que propone Li [18] es el siguiente. En una primera etapa se estiman las pocas direcciones de las cuales depende y, y en una segunda etapa se puede proceder a estimar la relación no paramétrica entre la respuesta y y las combinaciones lineales encontradas. Como se ha reducido la dimensión del espacio de covariables, las técnicas usuales de suavizado pueden aplicarse con mejores resultados eliminando "la maldición de la dimensionalidad" .

En la Sección 2, se describe el modelo propuesto por Li [18], se estudia el algoritmo que él propone para estimar las direcciones y sus principales características. En la Sección 2.3, probamos que el algoritmo propuesto por Li [18] es equivalente a calcular las coordenadas discriminantes para ciertos grupos de observaciones (slices) convenientemente elegidos según el valor de la variable dependiente y.

En la Sección 3, presentamos un nuevo enfoque para obtener las direcciones que utiliza el modelo propuesto por Li [18]. Esto se hace asumiendo que los diferentes grupos (slices) que utiliza el método de Li tienen distribución normal

multivariada con una misma matriz de covarianza y con medias pertenecientes a un subespacio de dimension dada.

En la Sección 4 probamos que bajo ciertos supuestos, los estimadores de máxima verosimilitud para este modelo coinciden con la solución que se obtiene usando el algoritmo propuesto por Li.

En la Sección 5, presentamos las conclusiones de este trabajo, y finalmente destinamos al Apéndice, en la Sección 6, las demostraciónes auxiliares que involucran resultados de álgebra lineal, así como de diferenciación vectorial y propiedades de esperanza condicional.

2. Regresión Inversa Partida

(Sliced Inverse Regression) K. C. Li [18].

2.1. El modelo

Supongamos que queremos estimar, a partir de una muestra $(\mathbf{x}_i', y_i)'$, $1 \leq i \leq N$, una relación no paramétrica entre $\mathbf{x}$ e y, donde la dimensión p del vector $\mathbf{x}$ es grande. Como se mencionó en la sección anterior, esto no es posible salvo que N sea muy grande, lo cual generalmente no es factible. Para superar este problema, Li [18] propone el siguiente modelo no paramétrico donde y depende de $\mathbf{x}$ sólo a través de un reducido número, K, de combinaciones lineales

$$y = f\left(\beta_1'\mathbf{x}, \ldots, \beta_K'\mathbf{x}, \varepsilon\right) \tag{1}$$

donde y es la variable respuesta, $\mathbf{x}$ es el vector p-dimensional de variables explicativas, ε el error, que es independiente de las $\mathbf{x}$, β_i son vectores desconocidos en $\mathbb{R}^p$ y f es una función arbitraria, $f : \mathbb{R}^{K+1} \to \mathbb{R}$.

En este caso, las K combinaciones lineales $\beta_1'\mathbf{x}, \ldots, \beta_K'\mathbf{x}$ capturan todo lo que es necesario conocer de las covariables $\mathbf{x}$ para relacionarlas con y. Si K es pequeño y logramos estimar los vectores β_i eficientemente, podemos alcanzar el objetivo de reducir el número de variables y hacer factible una estimación no paramétrica.

Li, en su paper, propone un método para estimar las direcciones β_i del modelo (1). Una vez que éstas sean estimadas, se puede reducir el número de covariables y el problema de estimar a la función f usando métodos no paramétricos resulta factible. Cuánto menor sea K, menor número de observaciones será necesario para estimar f.

Por otro lado, cuando se está haciendo un análisis exploratorio, es útil visualizar los datos graficándolos. Esto es posible cuando el número de covariables es 2. Otra ventaja de utilizar el modelo propuesto por Li en el caso de $K = 2$, es poder hacer un gráfico tridimensional de y en función de las dos combinaciones relevantes. Si $K > 2$, se pueden hacer gráficos bidimensionales de dispersión entre la y y cada una de las K combinaciones lineales $\beta_i'\mathbf{x}_i$. En este sentido, el modelo (1) permite una simplificación del análisis exploratorio de los datos antes de aplicar otros métodos que permitan la construcción de modelos, selección de variables, análisis de heteroscedasticidad, etc.

Si miramos el modelo (1) vemos que cambiando f adecuadamente, la expresión (1) puede ser reparametrizada a través de cualquier conjunto de K direcciones que pertenezcan al subespacio generado por $\{\beta_1, \ldots, \beta_K\}$. Luego, lo que puede ser identificado es el subespacio y no cada dirección $\beta_1, \ldots, \beta_K$ individualmente, a menos que se impongan restricciones sobre f.

Definición 2.1 *Bajo el modelo (1) llamaremos una* ***dirección de efectiva reducción de la dimensión*** *(****dirección edr****) a cualquier combinación lineal de los β_i, y llamaremos* ***espacio edr*** *al espacio generado por los $\{\beta_1, \ldots, \beta_K\}$, es decir a gen $\{\beta_1, \ldots, \beta_K\}$.*

Sea Ψ =Var($\mathbf{x}$) la matriz de covarianza de $\mathbf{x}$, y consideremos la versión estandarizada de $\mathbf{x}$:

$$\mathbf{z} = \Psi^{-1/2}\left[\mathbf{x} - E\left(\mathbf{x}\right)\right].$$

Podemos reescribir el modelo (1)

$$y = f\left(\eta_1'\mathbf{z}, \ldots, \eta_K'\mathbf{z}, \varepsilon\right)$$

con $\eta_i = \Psi^{1/2}\beta_i$ las **direcciones edr estandarizadas**. A cualquier vector que pertenezca al subespacio generado por los η_i lo llamaremos una dirección edr estandarizada.

El algoritmo que propone Li está basado en la idea de **regresión inversa**: hacer una regresión de las $\mathbf{x}$ versus las y. En vez de modelar la regresión directa $E\left(y \mid \mathbf{x}\right)$ se propone hacer una regresión de $\mathbf{x}$ en función de y, es decir estudiar la $E\left(\mathbf{x} \mid y\right)$. Sabemos que $E\left(\mathbf{x} \mid y\right)$ describe una curva en $\mathbb{R}^p$, y además que su esperanza coincide con la de las $\mathbf{x}$. Si imponemos la siguiente condición sobre la distribución del vector $\mathbf{x}$, Li prueba en su artículo que la curva de regresión inversa centrada:

$$E\left(\mathbf{x} \mid y\right) - E\left(\mathbf{x}\right)$$

yace en un subespacio de dimensión K de $\mathbb{R}^p$.

Condición A. *Dados $\beta_{\mathbf{1}}, \ldots, \beta_{\mathbf{K}}$ las direcciones edr, para todo $\mathbf{b} \in \mathbb{R}^p$, la esperanza condicional $E\left(\mathbf{b}'\mathbf{x} \mid \beta_{\mathbf{1}}'\mathbf{x}, \ldots, \beta_{\mathbf{K}}'\mathbf{x}\right)$ es lineal en $\beta_1'\mathbf{x}, \ldots, \beta_{\mathbf{K}}'\mathbf{x}$; es decir, para algunas constantes $c_0, c_1, \ldots, c_K$,*

$$\mathrm{E}\left(\mathbf{b}'\mathbf{x} \mid \beta_{\mathbf{1}}'\mathbf{x}, \ldots, \beta_{\mathbf{K}}'\mathbf{x}\right) = c_0 + c_1\beta_1'\mathbf{x} + \cdots + c_K\beta_K'\mathbf{x}.$$

Observación 2.1 *La condición es satisfecha por las distribuciones elípticamente simétricas (en particular, por la normal multivariada).*

Teorema 2.1 *(Li [18], 1991) : Bajo el modelo (1) y la condición anterior, con probabilidad 1 la curva de regresión inversa centrada $\mathrm{E}\left(\mathbf{x} \mid y\right) - \mathrm{E}\left(\mathbf{x}\right)$ está contenida en el subespacio lineal K dimensional generado por $\Psi\beta_1, \ldots, \Psi\beta_K$ donde Ψ denota la matriz de covarianza de las $\mathbf{x}$.*

Demostración. Sin pérdida de generalidad, supondremos que $\mathrm{E}\left(\mathbf{x}\right) = \mathbf{0}$. Sea $\mathbf{b}$ en el complemento ortogonal en $\mathbb{R}^p$ del espacio generado por $\left\{\Psi\beta_k\right\}_{1 \le k \le K}$, es decir,

$$\mathbf{b}'\Psi\beta_k = 0, \quad 1 \le k \le K.$$

Queremos ver que $\mathbf{b}'\mathrm{E}\left(\mathbf{x} \mid y\right) = 0$ con probabilidad 1. Usando la Proposición 6.1 del Apéndice con $A = \mathbf{b}'\mathbf{x}$, $B = y$, $C = (\beta_1{}'\mathbf{x}, \ldots, \beta_K{}'\mathbf{x})$ resulta

$$\mathbf{b}'\mathrm{E}\left(\mathbf{x} \mid y\right) = \mathrm{E}\left(\mathbf{b}'\mathbf{x} \mid y\right) = \mathrm{E}\left(\mathrm{E}\left[\mathbf{b}'\mathbf{x} \mid \beta_1{}'\mathbf{x}, \ldots, \beta_K{}'\mathbf{x}, y\right] \mid y\right).$$

Luego, en virtud de que bajo el modelo (1) $(\beta_1{}'\mathbf{x}, \ldots, \beta_K{}'\mathbf{x})$ y ε son independientes, por la Proposición 6.2 del Apéndice resulta que

$$\mathrm{E}\left(\mathrm{E}\left[\mathbf{b}'\mathbf{x} \mid \beta_1{}'\mathbf{x}, \ldots, \beta_K{}'\mathbf{x}, y\right] \mid y\right) = \mathrm{E}\left(\mathrm{E}\left[\mathbf{b}'\mathbf{x} \mid \beta_1{}'\mathbf{x}, \ldots, \beta_K{}'\mathbf{x}\right] \mid y\right).$$

Entonces, basta ver que

$$\mathrm{E}\left[\mathbf{b}'\mathbf{x} \mid \beta_1{}'\mathbf{x}, \ldots, \beta_K{}'\mathbf{x}\right] = 0,$$

o equivalentemente,

$$\mathrm{E}\left(\left(\mathrm{E}\left[\mathbf{b}'\mathbf{x} \mid \beta_1{}'\mathbf{x}, \ldots, \beta_K{}'\mathbf{x}\right]\right)^2\right) = 0.$$

Pero

$$\begin{aligned}
\mathrm{E}\left(\left(\mathrm{E}\left[\mathbf{b}'\mathbf{x} \mid {\beta_1}'\mathbf{x}, \ldots, {\beta_K}'\mathbf{x}\right]\right)^2\right) &= \mathrm{E}\left\{\left(\mathrm{E}\left[\mathbf{b}'\mathbf{x} \mid \beta_k\mathbf{x}'s\right]\right)\left(\mathrm{E}\left[\mathbf{b}'\mathbf{x} \mid \beta_k\mathbf{x}'s\right]\right)\right\} \\
&= \mathrm{E}\left\{\mathrm{E}\left[\mathrm{E}\left[\mathbf{b}'\mathbf{x} \mid \beta_k\mathbf{x}'s\right] \cdot \mathbf{b}'\mathbf{x} \mid \beta_k\mathbf{x}'s\right]\right\} \\
&= \mathrm{E}\left\{\mathrm{E}\left[\mathbf{b}'\mathbf{x} \mid \beta_k\mathbf{x}'s\right] \cdot \mathbf{b}'\mathbf{x}\right\} \\
&= \mathrm{E}\left\{\left[c_0 + \sum_{k=1}^{K} c_k\beta_k'\mathbf{x}\right] \cdot \mathbf{x}'\mathbf{b}\right\},
\end{aligned}$$

donde la última igualdad es consecuencia de la Condición A. Como $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, resulta

$$\begin{aligned}
\mathrm{E}\left(\left(\mathrm{E}\left[\mathbf{b}'\mathbf{x} \mid {\beta_1}'\mathbf{x}, \ldots, {\beta_K}'\mathbf{x}\right]\right)^2\right) &= \sum_{k=1}^{K} c_k\beta_k'\mathrm{E}\left\{\mathbf{x}\mathbf{x}'\right\}\mathbf{b} \\
&= \sum_{k=1}^{K} c_k\beta_k'\Psi\mathbf{b} \\
&= \mathbf{0},
\end{aligned}$$

lo que completa la prueba del teorema. ■

Este teorema permite vincular el subespacio al cual pertenece la $\mathrm{E}(\mathbf{x} \mid y) - \mathrm{E}(\mathbf{x})$ con las direcciones edr. Si el vector $\mathbf{x}$ tiene matriz de covarianza identidad, el subespacio que contiene a la curva de regresión inversa será directamente el subespacio edr buscado. Esto es lo que afirma el siguiente corolario.

Corolario 2.1 *Supongamos que* $\mathbf{x}$ *ha sido estandarizada a* $\mathbf{z}$ *mediante*

$$\mathbf{z} = \Psi^{-1/2}\left[\mathbf{x} - \mathrm{E}(\mathbf{x})\right].$$

Luego, bajo el modelo (1) se puede también escribir

$$y = f\left(\eta_1'\mathbf{z}, \ldots, \eta_K'\mathbf{z}, \varepsilon\right)$$

con $\eta_i = \Psi^{1/2}\beta_i$. *Si además se cumple la Condición A., la curva inversa estandarizada* $\mathrm{E}(\mathbf{z} \mid y)$ *está contenida en el subespacio lineal generado por las direcciones edr standarizadas* $\eta_1, \ldots, \eta_K$.

Demostración. En virtud del Teorema 2.1,

$$\mathrm{E}(\mathbf{x} \mid y) - \mathrm{E}(\mathbf{x}) \in gen\{\Psi\beta_1, \ldots, \Psi\beta_K\}.$$

Luego, $\Psi^{-1/2}\left[\mathrm{E}(\mathbf{x} \mid y) - \mathrm{E}(\mathbf{x})\right] \in gen\{\Psi^{-1/2}\Psi\beta_1, \ldots, \Psi^{-1/2}\Psi\beta_K\}$ es decir

$$\Psi^{-1/2}\left[\mathrm{E}(\mathbf{x} \mid y) - \mathrm{E}(\mathbf{x})\right] \in gen\{\Psi^{1/2}\beta_1, \ldots, \Psi^{1/2}\beta_K\},$$

o, equivalentemente

$$\mathrm{E}\left(\Psi^{-1/2}\left[\mathbf{x} - \mathrm{E}(\mathbf{x})\right] \mid y\right) = \mathrm{E}(\mathbf{z} \mid y) \in gen\{\eta_1, \ldots, \eta_K\}. \ ■$$

Este teorema sugiere un algoritmo para hallar las *direcciones edr* a partir de una muestra de datos $(\mathbf{x}_i, y_i)$ con $1 \leq i \leq N$. Este algoritmo se describirá en la próxima subsección.

2.2. Algoritmo de Li

Li propone un algoritmo computacionalmente simple para estimar las direcciones edr. El algoritmo consiste de los siguientes pasos.

Algoritmo de Li

1. Estandarizar $\mathbf{x}_i$ por una transformación afín para obtener $\mathbf{z}_i = \widehat{\Psi}^{-1/2}\left[\mathbf{x}_i - \overline{\mathbf{x}}\right]$, con $\widehat{\Psi}$ la matriz de covarianza muestral de las $\mathbf{x}_i$.

2. Dividir el rango de y en H fetas (slices) $I_1, ..., I_H$. Sea n_h la cantidad de observaciones cuyo valor de y está en la feta I_h.

3. Para cada feta, calcular $\widehat{\mathbf{m}}_h$ el promedio de las observaciones $\mathbf{z}_i$ cuyos valores y_i caen en la feta $I_h, \widehat{\mathbf{m}}_h = \frac{1}{n_h}\sum_{y_i \in I_h} \mathbf{z}_i$.

4. Realizar un análisis de componentes principales (pesadas) para los $\widehat{\mathbf{m}}_h$ del modo siguiente: armar la matriz de covarianza pesada $\widehat{\Phi} = \sum_{h=1}^{H} \frac{n_h}{N} \widehat{\mathbf{m}}_h \widehat{\mathbf{m}}_h'$, luego hallar sus autovalores y autovectores. O, en términos de las observaciones originales,

$$\widehat{\Phi} = \widehat{\Psi}^{-1/2} \sum_{h=1}^{H} \frac{n_h}{N} \left(\frac{1}{n_h} \sum_{y_i \in I_h} \left[\mathbf{x}_i - \overline{\mathbf{x}}\right] \right) \left(\frac{1}{n_h} \sum_{y_i \in I_h} \left[\mathbf{x}_i - \overline{\mathbf{x}}\right] \right)' \widehat{\Psi}^{-1/2}. \quad (2)$$

5. Sean $\widehat{\eta}_1, ..., \widehat{\eta}_K$ los autovectores correspondientes a los K mayores autovalores de $\widehat{\Phi}$. La estimación de las direcciones edr según este algorimo resultan ser
$$\widehat{\beta}_k = \widehat{\Psi}^{-1/2} \widehat{\eta}_k, k = 1, \ldots, K. \quad (3)$$

Como las $\mathbf{z}_i$ tienen matriz de covarianza próxima a la identidad y esperanza próxima a $\mathbf{0}$, y cada $\widehat{\mathbf{m}}_h$ se puede interpretar como una estimación de $E(\mathbf{z}|y \in I_h)$, de acuerdo al Teorema 2.1 estos vectores pertenecen al subespacio de dimensión K de las direcciones edr. En el paso 4 se estima este subespacio como el generado por las primeras K componentes principales. Esto se justifica dado que este subespacio es el más próximo en la métrica l_2 a los vectores $\widehat{\mathbf{m}}_h, 1 \leq h \leq H$. Finalmente, en el paso 5 se retransforman las direcciones edr encontradas a la escala original.

A continuación enunciamos una lista de resultados y consideraciones que pueden encontrarse en Li [18].

- El número de fetas H en que se divida el rango de y puede afectar la varianza asintótica del estimador, sin embargo no parece haber demasiada variabilidad en la aplicación a tamaños muestrales fijos, o en estudios de simulación. Esta arbitrariedad parece ser menos importante que la determinación del parámetro de suavizado (ancho de ventana) para los estimadores no paramétricos de regresión.

- Los estimadores de las direcciones edr resultan consistentes con orden de convergencia $n^{1/2}$.

- En el caso en que las covariables se distribuyen normalmente, el valor de K puede ser aproximado a partir de los datos.

- En los pasos 2 y 3 se obtiene una estimación simple de la curva de regresión inversa $\mathrm{E}(\mathbf{z} \mid y)$. Utilizando métodos no paramétricos más sofisticados como estimadores de núcleos, vecinos más cercanos o splines es posible obtener mejores estimadores para la curva de regresión inversa estandarizada. Sin embargo, el algoritmo de Li resulta especialmente atractivo por su simplicidad.

El esquema de estimación y la aplicación de la regresión inversa partida (SIR) se discuten en detalle en Li [18] y en Chen y Li [5]. El SIR ha sido investigado en numerosos artículos. Enfoques que ponen el énfasis en sus propiedades asintóticas se pueden ver en Li [18], Hsing y Carroll [15], Kötter [17], Zhu y Fang [22], Gannoun y Saracco [8]. La estimación de la dimensión K del espacio edr como un paso previo a la estimación del SIR es el objetivo de Li [18], Cook y Weisberg [6] y Schott [20] . Las condiciones de diseño que involucran a la distribución de las covariables han sido también un tópico importante en la discusión del método, en los artículos de Härdle y Tsybakov [13] y Hall y Li [14].

Un análisis detallado del efecto de los outliers en la regresión inversa partida puede leerse en Gather, Hilker y Becker [11]. En él se muestra cuan severamente puede ser influenciado el SIR por los outliers en los datos. En Gather, Hilker y Becker [10] se presenta una versión robusta del SIR (el DAME) que propone reemplazar los estimadores clásicos involucrados en el cálculo por sus versiones robustas manteniendo el innovador esquema de estimación original del SIR, pero ofreciendo una mejor resistencia a la presencia de outliers en el espacio de covariables.

Como fue notado por Li [18] en su artículo y en los comentarios finales del mismo y profundizado por Cook y Weisberg [6] en algunos casos la curva de regresión inversa puede yacer en un subespacio propio del espacio edr, en tal caso el procedimiento de regresión inversa estará condenado a perderse algunas de las direcciones de interés (algunos de los β). Sin embargo, si además de estudiar la $\mathrm{E}(\mathbf{x} \mid y)$ se estudian momentos condicionales de mayor orden de $\mathbf{x}$ dado y, podemos reducir la posibilidad de perder direcciones importantes. Hay más de una manera de implementar esto, lo que da lugar a distintos procedimientos SIR-II: Li[18] y SAVE: Cook y Weisberg [6].

Algunos artículos investigan al SIR en relación con otros modelos: Carroll y Li [4] usan el SIR en el marco de un modelo de regresión generalizado con errores en las covariables, Bura y Cook [3] proponen un modelo lineal multivariado para la curva de regresión inversa, y Chen y Li [5] comparan el SIR con el modelo de regresión lineal múltiple.

2.3. Método de Coordenadas Discriminantes

El método de coordenadas discriminantes es un método de reducción de dimensión clásico del Análisis Multivariado, que se aplica a observaciones clasificadas en grupos. Supongamos que tenemos N observaciones $p-$dimensionales, de las cuales n_i pertenecen al $i-$ésimo grupo $\left(i = 1, \ldots, H,\ N = \sum_{i=1}^{H} n_i\right)$. Sea

$\mathbf{x}_{ij}$ la $j-$ésima observación del grupo $i-$ésimo, y definimos

$$\overline{\mathbf{x}}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij},$$

el promedio de las observaciones correspondientes al $i-$ésimo grupo y

$$\overline{\mathbf{x}}_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{n_i}\mathbf{x}_{ij},$$

el promedio de todas las observaciones. Definamos las dos siguientes matrices: la matriz *"entre grupos"*

$$\begin{aligned} B &= \frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{n_i}\left(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}\right)\left(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}\right)' \\ &= \frac{1}{N}\sum_{i=1}^{H} n_i\left(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}\right)\left(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}\right)' \end{aligned} \tag{4}$$

y la matriz *"dentro de los grupos"*

$$W = \frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij}-\overline{\mathbf{x}}_{i\bullet}\right)\left(\mathbf{x}_{ij}-\overline{\mathbf{x}}_{i\bullet}\right)'. \tag{5}$$

Llamemos S_i a la matriz de covarianza muestral insesgada correspondiente a las observaciones del grupo i dada por

$$S_i = \frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij}-\overline{\mathbf{x}}_{i\bullet}\right)\left(\mathbf{x}_{ij}-\overline{\mathbf{x}}_{i\bullet}\right)'$$

y S a la matriz de covarianza muestral de todas las observaciones dada por

$$S = \frac{1}{N-1}\sum_{i=1}^{H}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij}-\overline{\mathbf{x}}_{\bullet\bullet}\right)\left(\mathbf{x}_{ij}-\overline{\mathbf{x}}_{\bullet\bullet}\right)'.$$

Entonces tenemos que

$$\begin{aligned} W &= \frac{1}{N}\sum_{i=1}^{H}\left(n_i-1\right)S_i \\ S\left(N-1\right) &= \sum_{i=1}^{H}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij}-\overline{\mathbf{x}}_{\bullet\bullet}\right)\left(\mathbf{x}_{ij}-\overline{\mathbf{x}}_{\bullet\bullet}\right)' \\ &= \sum_{i=1}^{H}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij}-\overline{\mathbf{x}}_{i\bullet}\right)\left(\mathbf{x}_{ij}-\overline{\mathbf{x}}_{i\bullet}\right)' + \sum_{i=1}^{H} n_i\left(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}\right)\left(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}\right)' \\ &= N\left(W+B\right) \end{aligned}$$

Las matrices B y W son los análogos multivariados de las habituales sumas de cuadrados *"entre"* y *"dentro"* de los grupos utilizadas en ANOVA. El grado de

clustering dentro de los grupos puede ser determinado por la relación entre las "magnitudes" de las matrices B y W.

En un primer paso, queremos reducir las observaciones multivariadas $\mathbf{x}_{ij}$ a observaciones univariadas $z_{ij} = \mathbf{c}'\mathbf{x}_{ij}$ de tal manera que la combinación lineal de las $p-$variables originales dada por $\mathbf{c}$ permita que se distingan los H grupos de la manera más clara posible. Para esto calculamos las sumas de cuadrados entre los grupos, dada por

$$B_{\mathbf{c}} = \frac{1}{N}\sum_{i=1}^{H} n_i \left(\overline{z}_{i\bullet} - \overline{z}_{\bullet\bullet}\right)^2 = \mathbf{c}'B\mathbf{c}$$

y la suma de cuadrados dentro de los grupos

$$W_{\mathbf{c}} = \frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{n_i} \left(z_{ij} - \overline{z}_{i\bullet}\right)^2 = \mathbf{c}'W\mathbf{c}.$$

Para que la proyección univariada distinga bien los grupos es preciso que la suma de cuadrados $B_{\mathbf{c}}$, que mide la variabilidad entre los grupos, sea grande comparada con la dispersión existente dentro de los grupos medida por $W_{\mathbf{c}}$. Luego, habrá una buena discriminación en la proyección univariada dada por $\mathbf{c}$ cuando el cociente

$$\frac{B_{\mathbf{c}}/\left(H-1\right)}{W_{\mathbf{c}}/\left(N-H\right)}$$

sea grande. Buscamos entonces el vector $\mathbf{c}$ tal que maximice

$$\frac{B_{\mathbf{c}}\left(N-H\right)}{W_{\mathbf{c}}\left(N-1\right)} = \frac{\left(N-H\right)}{\left(N-1\right)}\frac{\mathbf{c}'B\mathbf{c}}{\mathbf{c}'W\mathbf{c}}$$

sujeto a que $\mathbf{c} \in \mathbb{R}^p,\ \mathbf{c} \neq \mathbf{0}$. Esto es equivalente a maximizar

$$g\left(\mathbf{c}\right) = \frac{\mathbf{c}'B\mathbf{c}}{\mathbf{c}'W\mathbf{c}}.$$

Como

$$g\left(\lambda\mathbf{c}\right) = g\left(\mathbf{c}\right)\ \ \forall\ \lambda \in \mathbb{R},$$

existen infinitos máximos, todos en el mismo subespacio de dimensión 1. Por lo tanto, para determinar $\mathbf{c}$ impondremos la restricción

$$\mathbf{c}'W\mathbf{c} = 1.$$

Llamaremos $\mathbf{c}_1$ al argumento donde se alcanza dicho máximo. Entonces $\mathbf{c}_1$ es tal que

$$\mathbf{c}_1 = \arg\max_{\mathbf{c}\neq\mathbf{0}} \frac{\mathbf{c}'B\mathbf{c}}{\mathbf{c}'W\mathbf{c}}$$

sujeto a que $\mathbf{c}'W\mathbf{c} = 1$. Para garantizar unicidad en la elección de $\mathbf{c}_1$ debemos además, pedir, por ejemplo, que la primer coordenada no nula de $\mathbf{c}_1$ sea positiva. Supondremos que éste es el caso en la determinación de $\mathbf{c}_1$ y la de los vectores $\mathbf{c}_k$ que elegimos a continuación.

Como una única combinación lineal puede no ser suficiente para discriminar entre los distintos grupos podemos buscar una nueva combinación lineal de las

$\mathbf{x}_{ij}$ originales. Nuevamente, esta combinación lineal será elegida de modo tal de agregar la mayor cantidad de información posible para discriminar entre los H grupos. Para esto, buscaremos $\mathbf{c}_2$ en $\mathbb{R}^p$ maximizando $g(\mathbf{c})$ sujeto a que (a) $\mathbf{c}'W\mathbf{c} = 1$, (b) $\mathbf{c}_1'\mathbf{x}$ y $\mathbf{c}'\mathbf{x}$ sean no correlacionadas. La condición (b) garantiza no repetir con la segunda combinación lineal información contenida en la primera. La condición (b) la podemos expresar como

$$\begin{aligned}\widehat{\text{cov}}\left(\mathbf{c}_1'\mathbf{x}, \mathbf{c}'\mathbf{x}\right) &= \mathbf{c}_1'\widehat{\text{cov}}\left(\mathbf{x}\right)\mathbf{c} \\ &= 0\end{aligned}$$

donde $\widehat{\text{cov}}(\mathbf{x})$ es un estimador de la covarianza de $\mathbf{x}$. Un estimador de esta matriz de covarianza está dado por la matriz W. Entonces, la condición (b) equivale a

$$\mathbf{c}_1'W\mathbf{c}_2 = 0.$$

Luego $\mathbf{c}_2$ se elige como aquel vector $\mathbf{c}$ que cumple

$$\text{máx}\,\frac{\mathbf{c}'B\mathbf{c}}{\mathbf{c}'W\mathbf{c}}$$

sujeto a que

$$\mathbf{c}'W\mathbf{c} = 1,\ \mathbf{c}'W\mathbf{c}_1 = 0.$$

Podríamos seguir agregando combinaciones lineales de las variables originales. Una vez que se tienen $\mathbf{c}_1, \ldots, \mathbf{c}_k$ el siguiente vector $\mathbf{c}_{k+1}$ se elige de modo que

$$\text{máx}\,\frac{\mathbf{c}'B\mathbf{c}}{\mathbf{c}'W\mathbf{c}} \tag{6}$$

sujeto a que

$$\mathbf{c}'W\mathbf{c} = 1, \begin{cases} \mathbf{c}'W\mathbf{c}_1 = 0 \\ \mathbf{c}'W\mathbf{c}_2 = 0 \\ \quad\vdots \\ \mathbf{c}'W\mathbf{c}_k = 0. \end{cases} \tag{7}$$

Estas condiciones aseguran que la nueva combinación lineal tiene varianza muestral 1 y no está correlacionada muestralmente con ninguna de las anteriores.

En general, interesarán las primeras K direcciones $\mathbf{c}_1, \ldots, \mathbf{c}_K$. Estas direcciones, que son la solución al problema algebraico planteado en (6) y (7) dan lugar a las K proyecciones $\mathbf{c}_1'\mathbf{x}_{ij}, \ldots, \mathbf{c}_K'\mathbf{x}_{ij}$ que se denominan *coordenadas discriminantes*. Como se verá en las siguientes proposiciones, las direcciones resultan ser los K autovectores asociados a los K mayores autovalores de $W^{-1}B$, es decir $\mathbf{c}_1, \ldots, \mathbf{c}_K$ que se pueden elegir de modo que $\mathbf{c}_s'W\mathbf{c}_t = \delta_{st}$.

Veamos que el método SIR coincide con el método de coordenadas discriminantes, si los grupos se eligen de modo adecuado.

Para ello necesitaremos los siguientes resultados.

Proposición 2.1 *Sea $D \in \mathbb{R}^{p\times p}$ matriz definida positiva, $A \in \mathbb{R}^{p\times p}$ matriz simétrica.*

i. λ es un autovalor de $D^{-1}A$ si y sólo si λ es un autovalor de $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ Como esta última matriz es simétrica, dichos autovalores son todos reales.

ii. Sean $\mathbf{v}_1, \ldots, \mathbf{v}_p$ *autovectores ortonormales de la matriz* $D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ *asociados a los autovalores* $\lambda_1 \geq \cdots \geq \lambda_p$. *Sean* $\mathbf{w}_i = D^{-\frac{1}{2}} \mathbf{v}_i$, *entonces los* $\mathbf{w}_i$ *resultan autovectores de la matriz* $D^{-1}A$ *asociados a los mismos autovalores* λ_i *y satisfacen*

a) $\mathbf{v}_i' \mathbf{v}_i = 1$, *o, equivalentemente* $\mathbf{w}_i' D \mathbf{w}_i = 1$

b) $\mathbf{v}_i' \mathbf{v}_j = 0$, *o,equivalentemente* $\mathbf{w}_i' D \mathbf{w}_j = 0 \qquad \forall\ i \neq j.$

iii. Sean $\mathbf{v}_1, \ldots, \mathbf{v}_p$ *autovectores ortonormales de la matriz* $D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ *asociados a los autovalores* $\lambda_1 \geq \cdots \geq \lambda_p$. *Sean* $\mathbf{z}_i = D^{\frac{1}{2}} \mathbf{v}_i$, *entonces los* $\mathbf{z}_i$ *resultan autovectores de la matriz* AD^{-1} *asociados a los mismos autovalores* λ_i *y satisfacen*

a) $\mathbf{v}_i' \mathbf{v}_i = 1$, *o, equivalentemente* $\mathbf{z}_i' D^{-1} \mathbf{z}_i = 1$

b) $\mathbf{v}_i' \mathbf{v}_j = 0$, *o,equivalentemente* $\mathbf{z}_i' D^{-1} \mathbf{z}_j = 0 \qquad \forall\ i \neq j.$

iv. Si A *es definida positiva y* λ *es un autovalor de* $D^{-1}A$, *resulta que* $\lambda > 0$.

v. Si $A - D$ *es definida positiva y* λ *es un autovalor de* $D^{-1}A$, *resulta que* $\lambda < 1$.

Demostración. Ver el Apéndice, sección 6.1. ■

El siguiente teorema sobre optimización de formas cuadráticas permitirá resolver el problema de coordenadas discriminantes, es decir, encontrar los $\mathbf{c}_i$ que cumplen (6) sujeto a (7).

Teorema 2.2 *Sean* $A,\ C \in \mathbb{R}^{p \times p}$ *matrices simétricas,* A *es definida positiva. Sean* $\lambda_1 \geq \cdots \geq \lambda_p$ *los autovalores de la matriz* $A^{-1}C$, $\lambda_i = \lambda_i\left(A^{-1}C\right)$ *y* $\mathbf{v}_1, \ldots, \mathbf{v}_p$ *los autovectores correspondientes. Entonces,*

i.
$$\sup_{\substack{\mathbf{v} \in \mathbb{R}^p \\ \mathbf{v} \neq \mathbf{0}}} \frac{\mathbf{v}' C \mathbf{v}}{\mathbf{v}' A \mathbf{v}} = \lambda_1$$
y el supremo se alcanza en $\mathbf{v} = \mathbf{v}_1$.

ii. El
$$\sup_{\substack{\mathbf{v} \in \mathbb{R}^p \\ \mathbf{v} \neq \mathbf{0}}} \frac{\mathbf{v}' C \mathbf{v}}{\mathbf{v}' A \mathbf{v}}$$
sujeto a que $\mathbf{v}_1' A \mathbf{v} = 0$ *se alcanza en* $\mathbf{v} = \mathbf{v}_2$ *y vale* λ_2.

iii. En general,
$$\sup_{\substack{\mathbf{v} \in \mathbb{R}^p \\ \mathbf{v} \neq \mathbf{0}}} \frac{\mathbf{v}' C \mathbf{v}}{\mathbf{v}' A \mathbf{v}}$$
sujeto a que $\mathbf{v}_i' A \mathbf{v} = 0$ *para* $1 \leq i \leq k$ *se alcanza en* $\mathbf{v} = \mathbf{v}_{k+1}$ *y vale* λ_{k+1}.

Demostración. Ver Seber [21] A.7.4 y A.7.5. ■

Corolario 2.2 (Aplicación a coordenadas discriminantes) *Se tienen observaciones* $(\mathbf{x}_{ij})_{1\leq i\leq H, 1\leq j\leq n_i}$ *clasificadas en* H *grupos. Sean las matrices* B *y* W *definidas como en (4) y (5) con* $N = \sum_{i=1}^{H} n_i$. *Las direcciones* $\mathbf{c}_1, \ldots, \mathbf{c}_K \in \mathbb{R}^p$ *que resuelven el siguiente problema*

$$\mathbf{c}_1 = \arg\max \frac{\mathbf{c}' B \mathbf{c}}{\mathbf{c}' W \mathbf{c}}$$

sujeto a que

$$\mathbf{c}_1' W \mathbf{c}_1 = 1$$

y

$$\mathbf{c}_l = \arg\max \frac{\mathbf{c}' B \mathbf{c}}{\mathbf{c}' W \mathbf{c}}$$

sujeto a que

$$\mathbf{c}' W \mathbf{c} = 1, \begin{cases} \mathbf{c}' W \mathbf{c}_1 = 0 \\ \mathbf{c}' W \mathbf{c}_2 = 0 \\ \vdots \\ \mathbf{c}' W \mathbf{c}_{l-1} = 0 \end{cases}$$

para $1 \leq l \leq K$, *corresponden a los autovectores asociados a los* K *mayores autovalores de la matriz* $W^{-1}B$.

Demostración. Se deduce del Teorema 2.2 y de la Proposición 2.1 iii. ■

En el esquema propuesto por Li [18], las observaciones se agrupan según el valor que toma la variable y. Esto quiere decir que los grupos (slices) se arman según el valor que toma en ellos la función f. Una vez armados los grupos, para que exista una función suave f capaz de tomar valores distintos en los distintos grupos de acuerdo al modelo (1),

$$y = f\left(\beta_1' \mathbf{x}, \ldots, \beta_K' \mathbf{x}, \varepsilon\right)$$

las direcciones edr deberían ser capaces de separarlos. Este argumento heurístico es confirmado por el siguiente teorema.

Teorema 2.3 *Sea una muestra* $(\mathbf{x}_i', y_i)'$, $1 \leq i \leq N$, *donde* $y_i \in \mathbb{R}$ *y* $\mathbf{x}_i \in \mathbb{R}^p$. *Dividimos a la muestra en* H *grupos de acuerdo al valor que tome la variable* y. *Si la matriz de covarianza muestral de al menos uno de los grupos, es definida positiva, y también lo es*

$$B = \frac{1}{N} \sum_{i=1}^{H} n_i \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right) \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right)'$$

entonces el método de buscar K *coordenadas discriminantes y el algoritmo SIR con* K *direcciones edr proporcionan el mismo resultado.*

Demostración. El método SIR produce como resultado los vectores $\widehat{\beta}_k = \widehat{\Psi}^{-1/2} \widehat{\eta}_k$ según (3), siendo $\widehat{\eta}_k$ los autovectores de la matriz (2) correspondientes

a los K mayores autovalores. Observemos que, con la notación de coordenadas discriminantes, podemos escribir a esta matriz del siguiente modo

$$\begin{aligned}\sum_{h=1}^{H} \frac{n_h}{N} \widehat{\mathbf{m}}_h \widehat{\mathbf{m}}_h' &= \sum_{h=1}^{H} \frac{n_h}{N} \widehat{\Psi}^{-1/2} \left(\overline{\mathbf{x}}_{h\bullet} - \overline{\mathbf{x}}\right) \left(\overline{\mathbf{x}}_{h\bullet} - \overline{\mathbf{x}}\right)' \widehat{\Psi}^{-1/2} \\ &= \widehat{\Psi}^{-1/2} \left[\frac{1}{N} \sum_{h=1}^{H} n_h \left(\overline{\mathbf{x}}_{h\bullet} - \overline{\mathbf{x}}\right) \left(\overline{\mathbf{x}}_{h\bullet} - \overline{\mathbf{x}}\right)'\right] \widehat{\Psi}^{-1/2} \\ &= \widehat{\Psi}^{-1/2} B \widehat{\Psi}^{-1/2}\end{aligned}$$

donde la última igualdad utiliza la definición de la matriz B de coordenadas discriminantes dada en (4). Luego, como los $\widehat{\eta}_k$ son los autovectores correspondientes a los K mayores autovalores de dicha matriz, utilizando la notación de la proposición 2.1, tomando

$$\begin{aligned}\mathbf{v}_k &= \widehat{\eta}_k \\ D &= \widehat{\Psi} \\ A &= B\end{aligned}$$

resulta que por la Proposición 2.1i. y 2.1ii., los $\widehat{\beta}_k = \widehat{\Psi}^{-1/2}\widehat{\eta}_k$ son los autovectores de la matriz $\widehat{\Psi}^{-1}B$ asociados a los K mayores autovalores.

Como por hipótesis, la matriz de covarianza muestral de al menos uno de los grupos es definida positiva, resulta que para dicho grupo la matriz W_i lo es, y como las matrices de covarianza muestral de los restantes son al menos semidefinidas positivas y

$$\begin{aligned}W &= \frac{1}{N} \sum_{i=1}^{H} \sum_{j=1}^{n_i} \left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_{i\bullet}\right) \left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_{i\bullet}\right)' \qquad (8) \\ &= \frac{1}{N} \sum_{i=1}^{H} n_i W_i\end{aligned}$$

valdrá lo mismo para W. Además, como

$$\widehat{\Psi} = B + W \qquad (9)$$

y las tres, $\widehat{\Psi}$, B y W, son por definición matrices semidefinidas positivas, resulta que $\widehat{\Psi}$ también es definida positiva.

Para ver que ambos procedimientos coinciden habría que verificar que los autovectores correspondientes a los mayores autovalores de las matrices $\widehat{\Psi}^{-1}B$ y $W^{-1}B$ coinciden.

Sea $\mathbf{v}$ un autovector de $\widehat{\Psi}^{-1}B$ de autovalor λ, en virtud de la Proposición 2.1 i. como $\widehat{\Psi} > 0$, resulta que todos los autovalores de $\widehat{\Psi}^{-1}B$ son reales. Como por hipótesis $B > 0$ en virtud de la Proposición 2.1 iv. resulta $\lambda > 0$. Como por (9) resulta

$$\widehat{\Psi} - B = W$$

que también es definida positiva, entonces por la Proposición 2.1 v. resulta que

$\lambda < 1$. Veamos que $\mathbf{v}$ también es autovector de $W^{-1}B$.

$$\begin{aligned} W^{-1}B\mathbf{v} &= W^{-1}\widehat{\Psi}\widehat{\Psi}^{-1}B\mathbf{v} \\ &= W^{-1}\widehat{\Psi}\lambda\mathbf{v} \\ &= W^{-1}\left(B+W\right)\lambda\mathbf{v} \\ &= \lambda W^{-1}B\mathbf{v} + \lambda\mathbf{v} \end{aligned}$$

que es equivalente a

$$W^{-1}B\mathbf{v}\left(1-\lambda\right) = \lambda\mathbf{v}.$$

Como $\lambda \neq 1$,

$$W^{-1}B\mathbf{v} = \frac{\lambda}{1-\lambda}\mathbf{v}$$

y $\mathbf{v}$ resulta ser autovector de autovalor $\frac{\lambda}{1-\lambda}$ de $W^{-1}B$. La función

$$g\left(\lambda\right) = \frac{\lambda}{1-\lambda}$$

es derivable si $\lambda \in (0,1)$ y

$$g'\left(\lambda\right) = \frac{\lambda}{\left(1-\lambda\right)^2} > 0, \text{ para } \lambda \in (0,1),$$

lo que implica que en dicho intervalo la función g es creciente. Luego el autovector correspondiente al mayor autovalor de $\widehat{\Psi}^{-1}B$ también será el autovector de mayor autovalor para $W^{-1}B$. Lo mismo vale para el autovector correspondiente al $k-$ésimo autovalor de $\widehat{\Psi}^{-1}B$.

Recíprocamente, sea $\mathbf{w}$ autovector de $W^{-1}B$ de autovalor δ. Como $\mathbf{w}$ también será autovector de $W^{-\frac{1}{2}}BW^{-\frac{1}{2}}$ que es simétrica y definida positiva (estamos suponiendo que $B, W > 0$) resulta $\delta > 0$. Queremos ver que $\mathbf{w}$ es autovector de $\widehat{\Psi}^{-1}B$.

$$\widehat{\Psi}^{-1}B\mathbf{w} = \widehat{\Psi}^{-1}WW^{-1}B\mathbf{w} = \widehat{\Psi}^{-1}W\delta\mathbf{w} = \widehat{\Psi}^{-1}\left(\widehat{\Psi}-B\right)\delta\mathbf{w}$$

Esto implica

$$\begin{aligned} \widehat{\Psi}^{-1}B\mathbf{w} &= \left[\mathbb{I} - \left(\widehat{\Psi}\right)^{-1}B\right]\delta\mathbf{w} \\ \widehat{\Psi}^{-1}B\mathbf{w} &= \frac{\delta}{1+\delta}\mathbf{w} \quad \blacksquare \end{aligned}$$

La pregunta que queda por responder es cuándo se cumplirán las hipótesis del teorema anterior. La siguiente proposición responde a esta inquietud.

Proposición 2.2 *Sea S la matriz de covarianza muestral correspondiente a $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, vectores aleatorios independientes e idénticamente distribuidos con $n > p$, entonces S es definida positiva con probabilidad 1 si y sólo si $P\left(\mathbf{x}_1 \in \mathcal{F}_s\right) = 0$ para toda variedad lineal $\mathcal{F}_s$ de dimensión s en $\mathbb{R}^p$, $\forall\ 0 \leq s < p$.*

Demostración. Ver Eaton y Perlman [7]. ■

En virtud de la proposición anterior, si suponemos, por ejemplo, que las observaciones en cada grupo $(\mathbf{x}_{hj})_{j=1,\ldots,n_h}$ son vectores aleatorios i.i.d. con una distribución que no se concentra en ningún hiperplano $(p-1)$ dimensional, y además pedimos $n_h > p$ entonces las matrices de covarianza muestral W_h resultarán definidas positivas con probabilidad 1, es decir que cumpliremos uno de los requisitos del Teorema 2.3 con probabilidad 1.

3. Estimación de las direcciones edr utilizando un enfoque de máxima verosimilitud

3.1. Motivación

El siguiente ejemplo motivará el modelo propuesto en esta sección para estimar las direcciones edr. Se generaron $N = 1000$ observaciones $(\mathbf{x}_i', y_i)'$ independientes, donde $\mathbf{x}_i' = (x_{i1}, x_{i2})$ tiene distribución $\mathrm{N}_2(\alpha, \Sigma)$ con

$$\alpha = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \ \Sigma = \begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix},$$

la respuesta y_i está dada por

$$y_i = (-x_{1i} + 3x_{2i})^3 + 6 + 0{,}2\varepsilon_i$$

y los ε_i tiene distribución $N(0,1)$. Dividimos a la muestra en $H = 10$ grupos, según el método SIR. Los datos $\mathbf{x}_i$ generados y las medias de cada grupo están graficados en la Figura 1. También graficamos en esta figura las elipses

$$(\mathbf{x} - \mathbf{m}_i)' S_i^{-1} (\mathbf{x} - \mathbf{m}_i) \leq \chi^2_{2,0,975}$$

donde $\mathbf{m}_i$ y S_i son la media y covarianza muestral de cada grupo y $\chi^2_{p,\alpha}$ es el cuantil α de una distribución chi cuadrado con p grados de libertad.

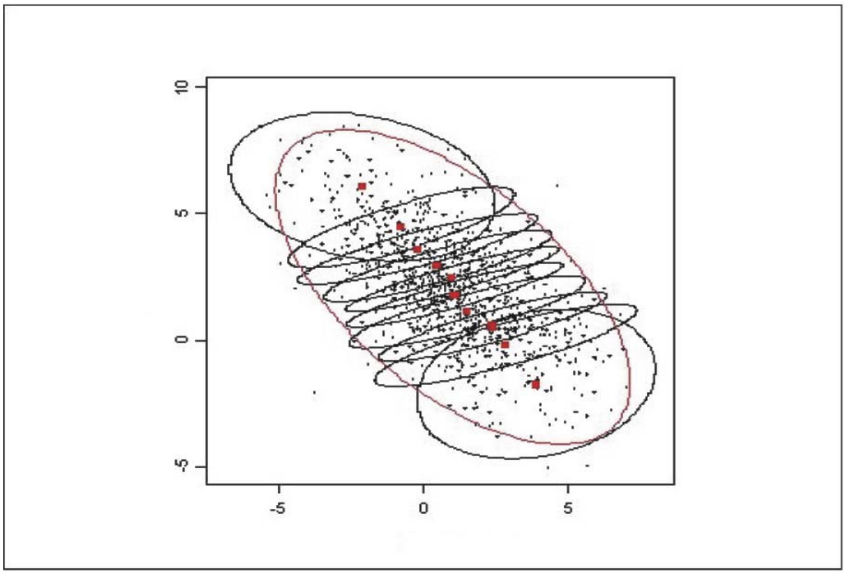

Figura 1

Vemos que para 8 de los 10 grupos las elipses son muy parecidas. Las otras dos elipsen restantes corresponden, a los grupos donde la variable respuesta toma los valores extremos.

Esto sugiere un método alternativo para estimar el subespacio de direcciones edr del modelo SIR: suponer que las observaciones del grupo h tienen una distribución normal multivariada p-dimensional con una media α_h y una matriz de covarianza común Σ. Es decir

$$\mathbf{x}_{hj} \sim N_p(\alpha_h, \Sigma),\ 1 \leq j \leq n_h, 1 \leq h \leq H \tag{10}$$

y de acuerdo al Teorema 2.1 y el Lema 3.1 que enunciamos abajo, las medias $\alpha_h, 1 \leq h \leq H$ estarán en una variedad lineal de $\mathbb{R}^p$ de dimensión K. Luego, se pueden estimar los α_h y Σ por máxima verosimilitud sujetos a esta condición.

Lema 3.1

i. Sean $\mathbf{z}$ *un vector aleatorio en* $\mathbb{R}^p$, V *un subespacio de dimensión* K *en* $\mathbb{R}^p$ *e* y *una variable aleatoria tales que* $\mathrm{E}(\mathbf{z} \mid y) \in V$. *Sea* $[a, b)$ *un intervalo en* $\mathbb{R}$ *contenido en el soporte de la variable* y, *entonces* $\mathrm{E}(\mathbf{z} \mid y \in [a, b)) \in V$.

ii. Sean $(\mathbf{x}_i, y_i)_{1 \leq i \leq N}$ *observaciones independientes correspondientes al modelo (1). Se las clasifica en* H *fetas de acuerdo al valor que toma la variable* y, *es decir rango*$(y) = \bigcup_{h=1}^{H} I_h$ *con* $I_h = [a_{h-1}, a_h)$, $a_0 = -\infty$, $a_1 < a_2 < \cdots < a_{H-1}$, $a_H = +\infty$. *Luego* $\mathbf{x}_{hj}$ *será la* j*-ésima observación del grupo* h *siempre que el valor de la variable* y *asociado pertenezca a la feta* I_h. *Si además* $\mathbf{x}_{hj} \sim N_p(\alpha_h, \Sigma)$ *resulta que las* α_h *pertenecerán a una variedad lineal de dimensión* K.

Demostración.

i. Sea $y \in [a, b)$, es decir $a \leq y < b$. Esto ocurre si y sólo si $\mathbb{I}_{[a,b)}(y) = 1$, o sea $\mathbb{I}_{[a,b)} \circ y = 1$ con $\mathbb{I}$ la función indicadora. Sea $\mathbf{c} \in V^{\perp}$, calculemos

$$\mathbf{c}'\mathrm{E}\left(\mathbf{z} \mid \mathbb{I}_{[a,b)} \circ y\right) = \mathrm{E}\left(\mathbf{c}'\mathbf{z} \mid \mathbb{I}_{[a,b)} \circ y\right) = \mathrm{E}\left(\mathrm{E}\left[\mathbf{c}'\mathbf{z} \mid y, \mathbb{I}_{[a,b)} \circ y\right] \mid \mathbb{I}_{[a,b)} \circ y\right)$$

donde la última igualdad es válida por propiedad de la esperanza condicional (ver Apéndice, Proposición 6.1). Como $\mathbb{I}_{[a,b)} \circ y$ es una función de y, resulta que $\mathrm{E}\left[\mathbf{c}'\mathbf{z} \mid y, \mathbb{I}_{[a,b)} \circ y\right] = \mathrm{E}\left[\mathbf{c}'\mathbf{z} \mid y\right]$ luego

$$\begin{aligned}\mathbf{c}'\mathrm{E}\left(\mathbf{z} \mid \mathbb{I}_{[a,b)} \circ y\right) &= \mathrm{E}\left(\mathrm{E}\left[\mathbf{c}'\mathbf{z} \mid y\right] \mid \mathbb{I}_{[a,b)} \circ y\right) \\ &= \mathrm{E}\left(\mathbf{c}'\mathrm{E}\left[\mathbf{z} \mid y\right] \mid \mathbb{I}_{[a,b)} \circ y\right) = 0\end{aligned}$$

por lo que $\mathrm{E}\left(\mathbf{z} \mid \mathbb{I}_{[a,b)} \circ y\right) \in V$. Notemos que no es necesario que el conjunto $[a, b)$ sea un intervalo, podría ser cualquier boreliano de la recta.

ii. Por el Teorema 2.1 bajo el modelo (1) y la Condición A. resulta que $\mathrm{E}(\mathbf{x} \mid y) - \mathrm{E}(\mathbf{x}) \in V$ donde V es un subespacio de dimensión K. En virtud de la parte i. de este lema, lo mismo sucederá con las $\mathrm{E}(\mathbf{x} \mid y \in I_h) - \mathrm{E}(\mathbf{x})$. Luego, las $\alpha_h = \mathrm{E}(\mathbf{x} \mid y \in I_h)$ pertenecerán a una variedad lineal, $V + \mathrm{E}(\mathbf{x})$ de dimensión K. ■

3.2. Algunos resultados sobre proyecciones

Recordemos la definición de *norma inducida por una matriz simétrica*:

Definición 3.1 *Sea $M \in \mathbb{R}^{p\times p}$ una matriz simétrica, M definida positiva (por lo tanto, M^{-1} existe y es definida positiva). Definimos el siguiente producto escalar en $\mathbb{R}^p$*

$$\langle \mathbf{x}, \mathbf{y} \rangle_M = \mathbf{x}' M^{-1} \mathbf{y}, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^p.$$

Este producto escalar induce una norma en $\mathbb{R}^p$

$$\|\mathbf{x}\|_M = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_M} = \sqrt{\mathbf{x}' M^{-1} \mathbf{x}}, \qquad \mathbf{x} \in \mathbb{R}^p$$

y permite definir una distancia entre dos vectores

$$d_M(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_M = \sqrt{(\mathbf{x} - \mathbf{y})' M^{-1} (\mathbf{x} - \mathbf{y})}, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^p.$$

Definición 3.2 *Sea E un espacio euclídeo y V un subespacio de dimensión finita. Dado $\mathbf{x} \in E$, llamamos proyección ortogonal de $\mathbf{x}$ en V según la norma inducida por la matriz M, $p_M(\mathbf{x}, V)$, a un elemento de V cuya distancia d_M a $\mathbf{x}$ sea mínima. Es decir,*

$$p_M(\mathbf{x}, V) = \arg \min_{\mathbf{y} \in V} \|\mathbf{x} - \mathbf{y}\|_M .$$

El elemento $p_M(\mathbf{x}, V)$ existe y es único, lo cual garantiza la buena definición. Más aún, si $\{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$ es una base ortonormal de V con la norma inducida por M resulta

$$\begin{aligned} p_M(\mathbf{x}, V) &= \sum_{i=1}^{K} \langle \mathbf{x}, \mathbf{e}_i \rangle_M \mathbf{e}_i = \sum_{i=1}^{K} \left(\mathbf{e}_i' M^{-1} \mathbf{x} \right) \mathbf{e}_i = \sum_{i=1}^{K} \mathbf{e}_i (\mathbf{e}_i' M^{-1} \mathbf{x}) \\ &= \left(\sum_{i=1}^{K} \mathbf{e}_i \mathbf{e}_i' \right) M^{-1} \mathbf{x} = [\mathbf{e}_1 \ldots \mathbf{e}_K] \begin{bmatrix} \mathbf{e}_1' \\ \vdots \\ \mathbf{e}_K' \end{bmatrix} M^{-1} \mathbf{x}. \end{aligned}$$

Vale que $\|\mathbf{x}\|_M^2 = \|p_M(\mathbf{x}, V)\|_M^2 + \|\mathbf{x} - p_M(\mathbf{x}, V)\|_M^2$.

En el caso de vectores aleatorios se tiene las siguiente definición.

Definición 3.3 *Sea $\mathbf{x}$ un vector aleatorio en $\mathbb{R}^p$ con $Var(\mathbf{x}) = \Sigma$. La norma inducida por la matriz Σ en $\mathbb{R}^p$ se denomina la distancia de Mahalanobis.*

3.3. Cálculo de los EMV

Sean $(\mathbf{x}_{ij})_{1 \le i \le H, 1 \le j \le n_i}$ vectores aleatorios en $\mathbb{R}^p$ independientes con distribución $N_p(\alpha_i, \Sigma)$. El primer subíndice, i, indica el grupo, el segundo, j, el índice de la observación dentro del mismo. Supondremos que los vectores de medias pertenecen a una variedad lineal $V + \mathbf{a}$ en $\mathbb{R}^p$ de dimensión K. Es decir, $\alpha_i \in V + \mathbf{a} \subset \mathbb{R}^p$, $1 \le i \le H$, con V un subespacio de dimensión K. Buscamos

los EMV de los α_i y de Σ, bajo esa restricción. La función de densidad de cada observación es

$$f_{\mathbf{x}_{ij}}\left(\mathbf{x}_{ij};\alpha_i,\Sigma\right)=\frac{1}{(2\pi)^{\frac{p}{2}}\left|\Sigma\right|^{\frac{1}{2}}}e^{-\frac{1}{2}(\mathbf{x}_{ij}-\alpha_i)'\Sigma^{-1}(\mathbf{x}_{ij}-\alpha_i)},$$

la función de verosimilitud basada en la muestra es

$$L\left(\alpha_1,\ldots,\alpha_H,\Sigma\right)=\prod_{i=1}^{H}\prod_{j=1}^{n_i}f_{\mathbf{x}_{ij}}\left(\mathbf{x}_{ij};\alpha_i,\Sigma\right)$$

y su logaritmo

$$\ln L\left(\alpha_1,\ldots,\alpha_H,\Sigma\right)=\sum_{i=1}^{H}\sum_{j=1}^{n_i}\left\{-\frac{p}{2}\ln(2\pi)-\frac{1}{2}\ln\left|\Sigma\right|-\frac{1}{2}\left(\mathbf{x}_{ij}-\alpha_i\right)'\Sigma^{-1}\left(\mathbf{x}_{ij}-\alpha_i\right)\right\}. \tag{11}$$

Buscamos los valores de los vectores $\alpha_1,\ldots,\alpha_H\in\mathbb{R}^p$ y de $\Sigma\in\mathbb{R}^{p\times p}$ que maximicen el $\ln L\left(\alpha_1,\ldots,\alpha_H,\Sigma\right)$ sujeto a las restricciones:

- existe un subespacio V de dimensión K en $\mathbb{R}^p$ y $\mathbf{a}\in\mathbb{R}^p$ tales que $\alpha_i\in V+\mathbf{a}$, $\forall i$.
- Σ es definida positiva ($\Sigma>0$).

Sea $N=\sum_{i=1}^{H}n_i$, el total de observaciones disponibles. Entonces

$$\begin{aligned}\ln L\left(\alpha_1,\ldots,\alpha_H,\Sigma\right)&=-\frac{c}{2}-\sum_{i=1}^{H}\sum_{j=1}^{n_i}\left\{\frac{1}{2}\ln\left|\Sigma\right|+\frac{1}{2}\left(\mathbf{x}_{ij}-\alpha_i\right)'\Sigma^{-1}\left(\mathbf{x}_{ij}-\alpha_i\right)\right\}\\&=-\frac{c}{2}-\frac{N}{2}\ln\left|\Sigma\right|-\frac{1}{2}\sum_{i=1}^{H}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij}-\alpha_i\right)'\Sigma^{-1}\left(\mathbf{x}_{ij}-\alpha_i\right)\\&=-\frac{1}{2}\left[g_1\left(\Sigma\right)+g_2\left(\alpha_1,\ldots,\alpha_H,\mathbf{a},\Sigma\right)\right]\end{aligned} \tag{12}$$

con

$$g_1\left(\Sigma\right)=c+N\ln\left|\Sigma\right|$$

y

$$g_2\left(\alpha_1,\ldots,\alpha_H,\mathbf{a},\Sigma\right)=\sum_{i=1}^{H}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij}-\alpha_i\right)'\Sigma^{-1}\left(\mathbf{x}_{ij}-\alpha_i\right)$$

donde c es una constante.

3.4. EMV de los parámetros de posición suponiendo Σ conocida

Comenzamos suponiendo Σ conocida, por lo tanto, necesitaremos minimizar a g_2. Es decir, debemos encontrar $\alpha_i\in\mathbb{R}^p, 1\leq i\leq H$, un subespacio V de dimensión K en $\mathbb{R}^p$, y $\mathbf{a}\in\mathbb{R}^p$, tales que $\alpha_i-\mathbf{a}\in V$ y minimicen

$$\sum_{i=1}^{H}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij}-\alpha_i\right)'\Sigma^{-1}\left(\mathbf{x}_{ij}-\alpha_i\right).$$

Podemos escribir $\alpha_i = \mu_i + \mathbf{a}$ para cada i con $\mu_i \in V$. De este modo la expresión a minimizar está dada por

$$\min_{\substack{\mu_i \in V,\ V \text{ un subespacio} \\ \dim(V)=K,\ \mathbf{a}\in\mathbb{R}^p}} \sum_{i=1}^{H}\sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \mu_i - \mathbf{a})' \Sigma^{-1} (\mathbf{x}_{ij} - \mu_i - \mathbf{a}). \tag{13}$$

Es inmediato que si $\mathbf{a}$, μ_i, y V es una solución para el problema (13), para cualquier $\mathbf{v} \in V$, reemplazando $\mathbf{a}$ por $\mathbf{a} + \mathbf{v}$ y μ_i por $\mu_i - \mathbf{v}$, se obtiene otra solución . Por lo tanto, para que el vector $\mathbf{a}$ quede unívocamente determinado debemos imponer una restricción sobre $\mathbf{a}$ o sobre los μ_i' s.

Comenzaremos buscando $\mathbf{a}$ suponiendo conocidos los μ_i. Como $\mathbf{x}_{ij}$ tiene distribución $N_p(\mu_i + \mathbf{a}, \Sigma)$, resulta que $\mathbf{x}_{ij} - \mu_i \sim N_p(\mathbf{a}, \Sigma)$, en este caso el problema de estimar $\mathbf{a}$ por máxima verosimilitud se reduce a encontrar EMV de la media de los datos de la muestra $(\mathbf{x}_{ij} - \mu_i)$ para $1 \leq i \leq H, 1 \leq j \leq n_i$ que resultan variables aleatorias independientes e idénticamente distribuidas con distribución $N_p(\mathbf{a}, \Sigma)$. Entonces el estimador resulta ser el promedio de dichas observaciones

$$\widehat{\mathbf{a}} = \frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \mu_i) = \frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{n_i} \mathbf{x}_{ij} - \frac{1}{N}\sum_{i=1}^{H} n_i\mu_i = \overline{\mathbf{x}}_{\bullet\bullet} - \frac{1}{N}\sum_{i=1}^{H} n_i\mu_i.$$

Llamemos $\overline{\mu}_P$ al promedio ponderado de los μ_i, es decir

$$\overline{\mu}_P = \frac{1}{N}\sum_{i=1}^{H} n_i\mu_i$$

Luego,

$$\begin{aligned} \widehat{\mathbf{a}} &= \widehat{\mathbf{a}}(\mu_1, \ldots, \mu_H) \\ &= \overline{\mathbf{x}}_{\bullet\bullet} - \overline{\mu}_P. \end{aligned} \tag{14}$$

Impongamos la restricción que garantiza la unicidad de la solución al problema (13): sin pérdida de generalidad, pediremos que $\mathbf{a}$ sea ortogonal al subespacio V con la noción de perpendicularidad inducida por la matriz Σ, de modo que al escribir $\alpha_i = \mu_i + \mathbf{a}$ estemos descomponiendo a cada vector α_i como suma de dos vectores perpendiculares entre sí según la distancia inducida por Σ. Esto se puede traducir en un sistema de $K = dim(V)$ ecuaciones lineales sobre $\mathbf{a}$.

Hemos visto que si $\mu_1, \ldots, \mu_H$ son conocidos, entonces

$$g_2(\mu_1, \ldots, \mu_H, \overline{\mathbf{x}}_{\bullet\bullet} - \overline{\mu}_P, \Sigma) \leqslant g_2(\mu_1, \ldots, \mu_H, \mathbf{a}, \Sigma) \ \ \forall \mathbf{a} \in \mathbb{R}^p. \tag{15}$$

Luego, los EMV de los μ_i, $1 \leq i \leq H$, se obtendrán minimizando la siguiente función

$$\begin{aligned} &g_2(\mu_1, \ldots, \mu_H, \overline{\mathbf{x}}_{\bullet\bullet} - \overline{\mu}_P, \Sigma) = h(\mu_1, \ldots, \mu_H, \Sigma) \\ &= \sum_{i=1}^{H}\sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_{\bullet\bullet} - \mu_i + \overline{\mu}_P)' \Sigma^{-1} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_{\bullet\bullet} - \mu_i + \overline{\mu}_P) \end{aligned} \tag{16}$$

sujetos a la restricción de que $\mu_1, \ldots, \mu_H \in V$ un subespacio de $\mathbb{R}^p$ de dimensión K. Comencemos por reescribir esta sumatoria. Sean $\mathbf{w}_{ij} = \mathbf{x}_{ij} - \overline{\mathbf{x}}_{\bullet\bullet}$, $\xi_i = \mu_i - \overline{\mu}_P$, de donde resulta $\overline{\mathbf{w}}_{i\bullet} = \overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}$. Luego tenemos

$$
\begin{aligned}
&\sum_{j=1}^{n_i} \left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_{\bullet\bullet} - \mu_i + \overline{\mu}_P\right)' \Sigma^{-1} \left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_{\bullet\bullet} - \mu_i + \overline{\mu}_P\right) \\
&= \sum_{j=1}^{n_i} \left(\mathbf{w}_{ij} - \xi_i\right)' \Sigma^{-1} \left(\mathbf{w}_{ij} - \xi_i\right) \\
&= \text{traza} \left\{ \sum_{j=1}^{n_i} \left(\mathbf{w}_{ij} - \xi_i\right)' \Sigma^{-1} \left(\mathbf{w}_{ij} - \xi_i\right) \right\} \\
&= \sum_{j=1}^{n_i} \text{traza} \left\{ \left(\mathbf{w}_{ij} - \xi_i\right)' \Sigma^{-1} \left(\mathbf{w}_{ij} - \xi_i\right) \right\} \\
&= \sum_{j=1}^{n_i} \text{traza} \left\{ \Sigma^{-1} \left(\mathbf{w}_{ij} - \xi_i\right) \left(\mathbf{w}_{ij} - \xi_i\right)' \right\} \\
&= \text{traza} \left\{ \sum_{j=1}^{n_i} \Sigma^{-1} \left(\mathbf{w}_{ij} - \overline{\mathbf{w}}_{i\bullet} + \overline{\mathbf{w}}_{i\bullet} - \xi_i\right) \left(\mathbf{w}_{ij} - \overline{\mathbf{w}}_{i\bullet} + \overline{\mathbf{w}}_{i\bullet} - \xi_i\right)' \right\} \\
&= \text{traza} \left\{ \Sigma^{-1} \sum_{j=1}^{n_i} \left(\left(\mathbf{w}_{ij} - \overline{\mathbf{w}}_{i\bullet}\right) \left(\mathbf{w}_{ij} - \overline{\mathbf{w}}_{i\bullet}\right)' + \left(\overline{\mathbf{w}}_{i\bullet} - \xi_i\right) \left(\overline{\mathbf{w}}_{i\bullet} - \xi_i\right)' \right) \right\} \\
&= \text{traza} \left\{ \Sigma^{-1} \sum_{j=1}^{n_i} \left(\mathbf{w}_{ij} - \overline{\mathbf{w}}_{i\bullet}\right) \left(\mathbf{w}_{ij} - \overline{\mathbf{w}}_{i\bullet}\right)' + \Sigma^{-1} \sum_{j=1}^{n_i} \left(\overline{\mathbf{w}}_{i\bullet} - \xi_i\right) \left(\overline{\mathbf{w}}_{i\bullet} - \xi_i\right)' \right\} \\
&= \text{traza} \left\{ \Sigma^{-1} \left[\sum_{j=1}^{n_i} \left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_{i\bullet}\right) \left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_{i\bullet}\right)' + n_i \left(\overline{\mathbf{w}}_{i\bullet} - \xi_i\right) \left(\overline{\mathbf{w}}_{i\bullet} - \xi_i\right)' \right] \right\}
\end{aligned}
$$

El primer sumando $\sum_{j=1}^{n_i} \left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_{i\bullet}\right) \left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_{i\bullet}\right)'$ no depende de μ_i ni de V. Luego, de acuerdo a (8)

$$
\begin{aligned}
&\sum_{j=1}^{n_i} \left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_{\bullet\bullet} - \mu_i + \overline{\mu}_P\right)' \Sigma^{-1} \left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_{\bullet\bullet} - \mu_i + \overline{\mu}_P\right) \\
&= \text{traza} \left\{ \Sigma^{-1} n_i W_i \right\} + \text{traza} \left\{ \Sigma^{-1} n_i \left(\overline{\mathbf{w}}_{i\bullet} - \xi_i\right) \left(\overline{\mathbf{w}}_{i\bullet} - \xi_i\right)' \right\} \\
&= \text{traza} \left\{ \Sigma^{-1} n_i W_i \right\} + n_i \left(\overline{\mathbf{w}}_{i\bullet} - \xi_i\right)' \Sigma^{-1} \left(\overline{\mathbf{w}}_{i\bullet} - \xi_i\right) \\
&= \text{traza} \left\{ \Sigma^{-1} n_i W_i \right\} + n_i \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} - \left(\mu_i - \overline{\mu}_P\right)\right)' \Sigma^{-1} \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} - \left(\mu_i - \overline{\mu}_P\right)\right).
\end{aligned}
$$

Por lo tanto, según (16) la expresión a minimizar es

$$\begin{aligned} h(\mu_1,\ldots,\mu_H,\Sigma) &= \sum_{i=1}^{H}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\overline{\mathbf{x}}_{\bullet\bullet}-\mu_i+\overline{\mu}_P)'\,\Sigma^{-1}\,(\mathbf{x}_{ij}-\overline{\mathbf{x}}_{\bullet\bullet}-\mu_i+\overline{\mu}_P) \\ &= \sum_{i=1}^{H}\operatorname{traza}\left\{\Sigma^{-1}W_i n_i\right\} \\ &+ \sum_{i=1}^{H} n_i\,(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}-(\mu_i-\overline{\mu}_P))'\,\Sigma^{-1}\,(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}-(\mu_i-\overline{\mu}_P)) \\ &= \operatorname{traza}\left\{\Sigma^{-1}\sum_{i=1}^{H}W_i n_i\right\} \\ &+ \sum_{i=1}^{H} n_i\,(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}-(\mu_i-\overline{\mu}_P))'\,\Sigma^{-1}\,(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}-(\mu_i-\overline{\mu}_P)) \\ &= \operatorname{traza}\left\{\Sigma^{-1}WN\right\}+\sum_{i=1}^{H} n_i\,\|\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}-(\mu_i-\overline{\mu}_P)\|_\Sigma^2\,, \quad (17) \end{aligned}$$

donde W está definido en (8). Cuando Σ es conocida, bastará minimizar el segundo sumando de la expresión anterior, sujeta a la restricción de que los $\mu_1,\ldots,\mu_H \in V$, con V subespacio de $\mathbb{R}^p$ de dimensión K. Llamemos V^* al subespacio donde se realiza dicho mínimo. Claramente, una vez que V^* esté fijo, los $(\mu_i-\overline{\mu}_P)$ que den lugar al mínimo surgirán de proyectar a $\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}$ sobre V^* de manera ortogonal según la distancia de Mahalanobis. Es decir,

$$\widehat{\mu}_i-\widehat{\overline{\mu}}_P = p_\Sigma\,(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet},V^*)\,. \quad (18)$$

Por la definición 3.2, resulta

$$\begin{aligned} \sum_{i=1}^{H} n_i\left\|\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}-\left(\widehat{\mu}_i-\widehat{\overline{\mu}}_P\right)\right\|_\Sigma^2 &= \sum_{i=1}^{H} n_i\,\|\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}-p_\Sigma\,(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet},V^*)\|_\Sigma^2 \\ &= \sum_{i=1}^{H} n_i\left(\|\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}\|_\Sigma^2-\|p_\Sigma\,(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet},V^*)\|_\Sigma^2\right) \end{aligned}$$

Luego, para hallar el subespacio, bastará maximizar el segundo término, es decir habrá que encontrar el subespacio V^* de dimensión K que maximice

$$\sum_{i=1}^{H} n_i\,\|p_\Sigma\,(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet},V^*)\|_\Sigma^2\,.$$

Sea $\mathbf{d}_1,\ldots,\mathbf{d}_K$ una base ortonormal de $\Sigma^{-1/2}V^*$ con la norma usual, y sea $D\in\mathbb{R}^{p\times K}, D=[\mathbf{d}_1\cdots\mathbf{d}_K]$. Luego $D'D=I_K$ y $\Sigma^{1/2}\mathbf{d}_1,\ldots,\Sigma^{1/2}\mathbf{d}_K$ es una base ortonormal de $\Sigma^{1/2}\Sigma^{-1/2}V^*=V^*$ con la norma inducida por Σ. Por la definición 3.2 de proyección ortogonal resulta

$$\begin{aligned} p_\Sigma\,(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet},V^*) &= \Sigma^{1/2}D\left[\Sigma^{1/2}D\right]'\,\Sigma^{-1}\,(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}) \quad (19) \\ &= \Sigma^{1/2}DD'\Sigma^{1/2}\Sigma^{-1}\,(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}) \\ &= \Sigma^{1/2}DD'\Sigma^{-1/2}\,(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet})\,. \end{aligned}$$

Con lo cual, tenemos

$$
\begin{aligned}
&\sum_{i=1}^{H} n_i \left\| p_\Sigma \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}, V^* \right) \right\|_\Sigma^2 \\
&= \sum_{i=1}^{H} n_i \left[\Sigma^{1/2} DD' \Sigma^{-1/2} \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right) \right]' \Sigma^{-1} \Sigma^{1/2} DD' \Sigma^{-1/2} \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right) \\
&= \sum_{i=1}^{H} n_i \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right)' \Sigma^{-1/2} DD' \Sigma^{1/2} \Sigma^{-1} \Sigma^{1/2} DD' \Sigma^{-1/2} \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right) \\
&= \sum_{i=1}^{H} n_i \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right)' \Sigma^{-1/2} DD' DD' \Sigma^{-1/2} \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right) \\
&= \sum_{i=1}^{H} n_i \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right)' \Sigma^{-1/2} DD' \Sigma^{-1/2} \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right) \\
&= \sum_{i=1}^{H} \operatorname{traza} \left\{ n_i \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right)' \Sigma^{-1/2} DD' \Sigma^{-1/2} \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right) \right\} \\
&= \sum_{i=1}^{H} \operatorname{traza} \left\{ \Sigma^{-1/2} DD' \Sigma^{-1/2} \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right) \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right)' n_i \right\} \\
&= \operatorname{traza} \left\{ \Sigma^{-1/2} DD' \Sigma^{-1/2} \sum_{i=1}^{H} n_i \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right) \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right)' \right\} \\
&= \operatorname{traza} \left\{ D' \left[\Sigma^{-1/2} \sum_{i=1}^{H} n_i \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right) \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right)' \Sigma^{-1/2} \right] D \right\} \\
&= N \operatorname{traza} \left\{ D' \left[\Sigma^{-1/2} \; B \Sigma^{-1/2} \right] D \right\}, \qquad (20)
\end{aligned}
$$

donde B está definida en (4). Luego la matriz D se obtiene maximizando

$$\operatorname{traza} \left\{ D' \left[\Sigma^{-1/2} \; B \Sigma^{-1/2} \right] D \right\} \qquad (21)$$

sujeta a $D'D = I_K$.

Necesitaremos el siguiente lema:

Lema 3.2 *Si $Q \in \mathbb{R}^{p \times p}$, es semidefinida positiva, $C \in \mathbb{R}^{p \times K}$ con $K < p$, tal que $C'C = I_K$, entonces*

$$\sum_{i=1}^{K} \lambda_i \left(C'QC \right) \leq \sum_{i=1}^{K} \lambda_i \left(Q \right)$$

donde $\lambda_i (A)$ son los autovalores ordenados de la matriz simétrica $A : \lambda_1 (A) \geq \lambda_2 (A) \geq \cdots \geq \lambda_p (A)$. La igualdad se alcanza si la matriz C tiene por columnas a los autovectores de norma uno asociados a los autovalores $\lambda_i (Q)$ para $i = 1, \ldots, K$.

Demostración. Ver el Apéndice, sección 6.1. ■

Volvamos al problema (21). De acuerdo al lema anterior tendremos

$$\text{traza}\left\{D'\left[\Sigma^{-1/2}B\Sigma^{-1/2}\right]D\right\} = \sum_{j=1}^{K}\lambda_j\left(D'\left[\Sigma^{-1/2}B\Sigma^{-1/2}\right]D\right) \tag{22}$$

$$\leq \sum_{j=1}^{K}\lambda_j\left(\Sigma^{-1/2}B\Sigma^{-1/2}\right) \tag{23}$$

la última desigualdad es válida en virtud del lema anterior. La igualdad se alcanza si la matriz D tiene por columnas a los autovectores de norma uno asociados a los K mayores autovalores de la matriz simétrica $\Sigma^{-1/2}B\Sigma^{-1/2}$.

Sea $\{\mathbf{t}_1,\ldots,\mathbf{t}_p\}$ una base ortonormal de $\mathbb{R}^p$ con la norma usual, con $\mathbf{t}_i$ el autovector asociado al autovalor $\lambda_i\left(\Sigma^{-1/2}B\Sigma^{-1/2}\right)$ de la matriz $\Sigma^{-1/2}B\Sigma^{-1/2}$, donde $\lambda_1\left(\Sigma^{-1/2}B\Sigma^{-1/2}\right)\geq\cdots\geq\lambda_P\left(\Sigma^{-1/2}B\Sigma^{-1/2}\right)\geq 0$. Luego, $D=[\mathbf{t}_1\cdots\mathbf{t}_K]$ tiene por columnas a los K autovectores de $\Sigma^{-1/2}B\Sigma^{-1/2}$ asociados a los K mayores autovalores. Una base del espacio V^* lo constituyen las K columnas de la matriz $\Sigma^{1/2}D$ por (19). Es decir, $\left\{\Sigma^{1/2}\mathbf{t}_1,\ldots,\Sigma^{1/2}\mathbf{t}_K\right\}$ serán una base del subespacio V^* buscado.

Los EMV de μ_i y de $\mathbf{a}$ deben cumplir, por lo tanto, con las ecuaciones encontradas en (18)y (14)

$$\widehat{\mu}_i-\widehat{\overline{\mu}}_P = p_\Sigma\left(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet},V^*\right) \tag{24}$$

$$\widehat{\mathbf{a}} = \overline{\mathbf{x}}_{\bullet\bullet}-\widehat{\overline{\mu}}_P. \tag{25}$$

Recordando que para garantizar unicidad de los estimadores pedíamos que $\widehat{\mathbf{a}}$ fuera ortogonal al subespacio V^* con la noción de perpendicularidad inducida por la matriz Σ, podemos ver que tomando

$$\widehat{\mu}_i = p_\Sigma\left(\overline{\mathbf{x}}_{i\bullet},V^*\right)$$

resulta que

$$\widehat{\overline{\mu}}_P = \frac{1}{N}\sum_{i=1}^{H}n_i\widehat{\mu}_i = \frac{1}{N}\sum_{i=1}^{H}n_i p_\Sigma\left(\overline{\mathbf{x}}_{i\bullet},V^*\right)$$

$$= p_\Sigma\left(\sum_{i=1}^{H}\frac{n_i}{N}\overline{\mathbf{x}}_{i\bullet},V^*\right) = p_\Sigma\left(\overline{\mathbf{x}}_{\bullet\bullet},V^*\right),$$

por lo tanto se satisface la ecuación (24). De (25) resulta

$$\widehat{\mathbf{a}} = \overline{\mathbf{x}}_{\bullet\bullet}-\widehat{\overline{\mu}}_P = \overline{\mathbf{x}}_{\bullet\bullet}-p_\Sigma\left(\overline{\mathbf{x}}_{\bullet\bullet},V^*\right) = p_\Sigma\left(\overline{\mathbf{x}}_{\bullet\bullet},V^{*\perp}\right)$$

de modo que también se cumple la condición que garantiza la unicidad de los EMV: $\widehat{\mathbf{a}}$ es ortogonal a V^*, con la norma inducida por Σ, y los $\widehat{\mu}_i$ pertenecen al subespacio V^*.

Además

$$\widehat{\alpha}_i = \widehat{\mu}_i+\widehat{\mathbf{a}} = \widehat{\mu}_i+\overline{\mathbf{x}}_{\bullet\bullet}-\widehat{\overline{\mu}}_P$$

es decir,

$$\widehat{\alpha}_i = p_\Sigma\left(\overline{\mathbf{x}}_{i\bullet}-\overline{\mathbf{x}}_{\bullet\bullet},V^*\right)+\overline{\mathbf{x}}_{\bullet\bullet} \tag{26}$$

con V^* el subespacio generado por $\left\{\Sigma^{1/2}\mathbf{t}_1, \ldots, \Sigma^{1/2}\mathbf{t}_K\right\}$ siendo $\{\mathbf{t}_1, \ldots, \mathbf{t}_K\}$ los autovectores ortonormales de $\Sigma^{-1/2}B\Sigma^{-1/2}$ asociados a los K mayores autovalores. Notemos que todos estos estimadores dependen de la matriz Σ, que hasta ahora suponíamos fija. En la sección que sigue usaremos la siguiente notación que explicita esta dependencia

$$\begin{aligned}
\widehat{\alpha}_i &= \widehat{\alpha}_i(\Sigma), \qquad 1 \le i \le H \\
\widehat{\mu}_i &= \widehat{\mu}_i(\Sigma), \qquad 1 \le i \le H \\
\widehat{\overline{\mu}}_P &= \widehat{\overline{\mu}}_P(\Sigma), \\
\widehat{\mathbf{a}} &= \widehat{\mathbf{a}}(\Sigma) \\
V^* &= V^*(\Sigma).
\end{aligned}$$

3.5. EMV de la matriz de covarianza

Ahora busquemos Σ que maximice la verosimilitud. De acuerdo a (12) tenemos

$$\begin{aligned}
\ln L\left(\widehat{\alpha}_1(\Sigma), \ldots, \widehat{\alpha}_H(\Sigma), \Sigma\right) &= -\frac{1}{2}\left[g_1(\Sigma) + g_2\left(\widehat{\alpha}_1(\Sigma), \ldots, \widehat{\alpha}_H(\Sigma), \overline{\mathbf{x}}_{\bullet\bullet} - \widehat{\overline{\mu}}_P(\Sigma), \Sigma\right)\right] \\
&= -\frac{1}{2}\left[g_1(\Sigma) + h\left(\widehat{\mu}_1(\Sigma), \ldots, \widehat{\mu}_H(\Sigma), \Sigma\right)\right]
\end{aligned}$$

con

$$g_1(\Sigma) = c + N \ln|\Sigma|$$

y

$$g_2(\alpha_1, \ldots, \alpha_H, \mathbf{a}, \Sigma) = \sum_{i=1}^{H}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \alpha_i)' \Sigma^{-1}(\mathbf{x}_{ij} - \alpha_i).$$

Usando (17) resulta

$$\begin{aligned}
&\ln L\left(\widehat{\alpha}_1(\Sigma), \ldots, \widehat{\alpha}_H(\Sigma), \Sigma\right) \\
&= -\frac{1}{2}\left[c + N\ln|\Sigma| + \operatorname{traza}\left\{\Sigma^{-1}NW\right\} + \sum_{i=1}^{H} n_i \left\|\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} - \left(\widehat{\mu}_i(\Sigma) - \widehat{\overline{\mu}_P}(\Sigma)\right)\right\|_\Sigma^2\right] \\
&= -\frac{1}{2}\left[c + N\ln|\Sigma| + N\operatorname{traza}\left\{\Sigma^{-1}W\right\}\right. \\
&\left. + \sum_{i=1}^{H} n_i\left(\|\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\|_\Sigma^2 - \|p_\Sigma(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}, V^*)\|_\Sigma^2\right)\right].
\end{aligned}$$

Además, se verifica

$$\begin{aligned}\sum_{i=1}^{H} n_i \left\|\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right\|_{\Sigma}^2 &= \sum_{i=1}^{H} \operatorname{traza}\left\{n_i \left\|\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right\|_{\Sigma}^2\right\} \\ &= \sum_{i=1}^{H} \operatorname{traza}\left\{n_i \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right)' \Sigma^{-1} \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right)\right\} \\ &= \sum_{i=1}^{H} \operatorname{traza}\left\{\Sigma^{-1} n_i \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right)\left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right)'\right\} \\ &= \operatorname{traza}\left\{\Sigma^{-1} \sum_{i=1}^{H} n_i \left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right)\left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right)'\right\} \\ &= \operatorname{traza}\left\{\Sigma^{-1} N B\right\} = N \operatorname{traza}\left\{\Sigma^{-1} B\right\}.\end{aligned}$$

Por (20) y (22) y por la Proposición 2.1 resulta

$$\begin{aligned}\sum_{i=1}^{H} n_i \left\|p_{\Sigma}\left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}, V^*\right)\right\|_{\Sigma}^2 &= N \sum_{j=1}^{K} \lambda_j \left(\Sigma^{-1/2}\ B \Sigma^{-1/2}\right) \\ &= N \sum_{j=1}^{K} \lambda_j \left(\Sigma^{-1} B\right).\end{aligned}$$

Luego, la función de log-verosimilitud evaluada en los EMV de posición calculados para en Σ es

$$\begin{aligned}&\ln L\left(\widehat{\alpha}_1(\Sigma), \ldots, \widehat{\alpha}_H(\Sigma), \Sigma\right) \\ &= -\frac{1}{2}\left[c + N \ln|\Sigma| + N \operatorname{traza}\left\{\Sigma^{-1} W\right\} + N \operatorname{traza}\left\{\Sigma^{-1} B\right\} - N \sum_{j=1}^{K} \lambda_j \left(\Sigma^{-1} B\right)\right] \\ &= -\frac{N}{2}\left[\frac{c}{N} + \ln|\Sigma| + \operatorname{traza}\left\{\Sigma^{-1} W\right\} + \operatorname{traza}\left\{\Sigma^{-1} B\right\} - \sum_{j=1}^{K} \lambda_j \left(\Sigma^{-1} B\right)\right] \\ &= -\frac{N}{2}\left[\frac{c}{N} + \ln|\Sigma| + \operatorname{traza}\left\{\Sigma^{-1} W\right\} + \sum_{j=K+1}^{p} \lambda_j \left(\Sigma^{-1} B\right)\right] \\ &= -\frac{N}{2}\left[c_0 + \ln|\Sigma| + \operatorname{traza}\left\{\Sigma^{-1} W\right\} + \sum_{j=K+1}^{p} \lambda_j \left(\Sigma^{-1} B\right)\right]\end{aligned}$$

donde B y W están dadas por (4) y (5) respectivamente y $c_0 = c/N$.

Luego, tenemos que minimizar con respecto a Σ la función

$$f(\Sigma) = \ln|\Sigma| + \operatorname{traza}\left\{\Sigma^{-1} W\right\} + \sum_{h=K+1}^{p} \lambda_h \left(\Sigma^{-1} B\right).$$

Si evaluamos la función en $B^{1/2}\Sigma B^{1/2}$, tenemos

$$\begin{aligned}
f\left(B^{1/2}\Sigma B^{1/2}\right) &= \ln\left|B^{1/2}\Sigma B^{1/2}\right| + \text{traza}\left\{\left(B^{1/2}\Sigma B^{1/2}\right)^{-1} W\right\} \\
&+ \sum_{h=K+1}^{p} \lambda_h\left(\left(B^{1/2}\Sigma B^{1/2}\right)^{-1} B\right) \\
&= \ln|B\Sigma| + \text{traza}\left\{B^{-1/2}\Sigma^{-1}B^{-1/2}W\right\} \\
&+ \sum_{h=K+1}^{p} \lambda_h\left(B^{-1/2}\Sigma^{-1}B^{-1/2}B\right) \\
&= \ln|B| + \ln|\Sigma| + \text{traza}\left\{\Sigma^{-1}\left(B^{-1/2}WB^{-1/2}\right)\right\} \\
&+ \sum_{h=K+1}^{p} \lambda_h\left(\Sigma^{-1}\left(B^{-1/2}BB^{-1/2}\right)\right) \\
&= \ln|B| + \ln|\Sigma| + \text{traza}\left\{\Sigma^{-1}\left(B^{-1/2}WB^{-1/2}\right)\right\} \\
&+ \sum_{h=K+1}^{p} \lambda_h\left(\Sigma^{-1}\right) \\
&= \ln|B| + g\left(\Sigma\right),
\end{aligned}$$

con

$$g\left(\Sigma\right) = \ln|\Sigma| + \text{traza}\left\{\Sigma^{-1}\left(B^{-1/2}WB^{-1/2}\right)\right\} + \sum_{h=K+1}^{p} \lambda_h\left(\Sigma^{-1}\right).$$

Para encontrar el mínimo de la función f bastará hallar el mínimo de g sujeto a que $\Sigma > 0$, digamos Σ_0 y luego tomar $B^{1/2}\Sigma_0 B^{1/2}$. Descomponemos espectralmente a la matriz $\Sigma = C\Theta C'$ con $\Theta = \text{diag}(\theta_1, \ldots, \theta_p)$, $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_p \geq 0$ los autovalores de Σ y $C = [\mathbf{c}_1 \cdots \mathbf{c}_p]$ la matriz ortogonal que contiene la base ortonormal de los autovectores asociados por columnas. Es decir, $CC' = C'C = I$. Entonces, llamando $A = B^{-1/2}WB^{-1/2}$ resulta

$$\begin{aligned}
\ln|\Sigma| &= \ln\left(\prod_{h=1}^{p}\theta_h\right) = \sum_{h=1}^{p}\ln\left(\theta_h\right) \\
\text{traza}\left\{\Sigma^{-1}A\right\} &= \text{traza}\left(C\Theta^{-1}C'A\right) = \text{traza}\left(\Theta^{-1}C'AC\right).
\end{aligned}$$

Como

$$\left(C'AC\right)_{ij} = \left(\begin{bmatrix}\mathbf{c}_1' \\ \vdots \\ \mathbf{c}_p'\end{bmatrix} A\left[\mathbf{c}_1 \cdots \mathbf{c}_p\right]\right)_{ij} = \left(\begin{bmatrix}\mathbf{c}_1' \\ \vdots \\ \mathbf{c}_p'\end{bmatrix}\left[A\mathbf{c}_1 \cdots A\mathbf{c}_p\right]\right)_{ij} = \mathbf{c}_i'A\mathbf{c}_j,$$

obtenemos

$$\left(\Theta^{-1}C'AC\right)_{ij} = \frac{1}{\theta_i}\left[0\cdots 0\underbrace{1}_{i}0\cdots 0\right]\begin{bmatrix}\mathbf{c}_1'A\mathbf{c}_j \\ \vdots \\ \mathbf{c}_p'A\mathbf{c}_j\end{bmatrix} = \frac{1}{\theta_i}\left(\mathbf{c}_i'A\mathbf{c}_j\right),$$

de donde,

$$\text{traza}\left\{\Sigma^{-1}A\right\} = \sum_{h=1}^{p} \frac{\mathbf{c}_h' A \mathbf{c}_h}{\theta_h}.$$

Como los autovalores de Σ^{-1} son $(1/\theta_h)_{1\leq h\leq p}$, que ordenados de mayor a menor dan $\theta_p^{-1} \geq \cdots \geq \theta_1^{-1}$, deducimos que la suma de los $p-K$ autovalores más pequeños de Σ^{-1} resulta ser

$$\sum_{h=K+1}^{p} \lambda_h\left(\Sigma^{-1}\right) = \frac{1}{\theta_1} + \cdots + \frac{1}{\theta_{p-K}}.$$

Luego,

$$g\left(C,\Theta\right) = \sum_{h=1}^{p} \ln\left(\theta_h\right) + \sum_{h=1}^{p} \frac{\mathbf{c}_h' A \mathbf{c}_h}{\theta_h} + \sum_{h=1}^{p-K} \frac{1}{\theta_h}. \tag{27}$$

Buscamos el mínimo de g sujeto a las restricciones

$$\begin{aligned} &\mathbf{c}_h'\mathbf{c}_h = 1,\ \forall 1 \leq h \leq p \\ &\mathbf{c}_h'\mathbf{c}_j = 0,\ \forall 1 \leq h < s \leq p \\ &\theta_1 \geq \theta_2 \geq \cdots \geq \theta_p \geq 0. \end{aligned} \tag{28}$$

Usando multiplicadores de Lagrange ρ_h para la primera de las restricciones y $2\rho_{hs}$ para la siguiente resulta que la expresión a minimizar es

$$\begin{aligned} g^*\left(C,\Theta\right) = &\sum_{h=1}^{p} \ln\left(\theta_h\right) + \sum_{h=1}^{p} \frac{\mathbf{c}_h' A \mathbf{c}_h}{\theta_h} + \sum_{h=1}^{p-K} \frac{1}{\theta_h} \\ &- \sum_{h=1}^{p} \rho_h\left(\mathbf{c}_h'\mathbf{c}_h - 1\right) - 2 \sum_{1\leq h<s\leq p} \rho_{hs}\mathbf{c}_h'\mathbf{c}_s. \end{aligned}$$

Derivando con respecto a θ_h

$$\frac{\partial}{\partial\theta_h} g^*\left(C,\Theta\right) = \begin{cases} \dfrac{1}{\theta_h} - \dfrac{1}{\theta_h^2} - \dfrac{\mathbf{c}_h' A \mathbf{c}_h}{\theta_h^2} & \text{si} \quad 1 \leq h \leq p-K \\ \dfrac{1}{\theta_h} - \dfrac{\mathbf{c}_h' A \mathbf{c}_h}{\theta_h^2} & \text{si} \quad p-K < h \leq p \end{cases}$$

e igualando a cero, obtenemos

$$\frac{1}{\theta_h} - \frac{1}{\theta_h^2} - \frac{\mathbf{c}_h' A \mathbf{c}_h}{\theta_h^2} = 0, \quad 1 \leq h \leq p-K$$

o equivalentemente

$$\theta_h = 1 + \mathbf{c}_h' A \mathbf{c}_h, \quad 1 \leq h \leq p-K \tag{29}$$

y

$$\frac{1}{\theta_h} - \frac{\mathbf{c}_h' A \mathbf{c}_h}{\theta_h^2} = 0, \quad p-K < h \leq p$$

o equivalentemente

$$\theta_h = \mathbf{c}_h' A \mathbf{c}_h, \quad p - K < h \leq p. \tag{30}$$

Recordando las reglas de derivación matricial que figuran en el apéndice, en la Sección 6.3, derivando con respecto a $\mathbf{c}_h$, $(1 \leq h \leq p)$ (tomamos $\rho_{hs} = \rho_{sh}$ para $h > s$) obtenemos

$$\frac{\partial}{\partial \mathbf{c}_h} g^* (C, \Theta) = 2\frac{A\mathbf{c}_h}{\theta_h} - 2\rho_h \mathbf{c}_h - 2\sum_{\substack{s=1 \\ s \neq h}}^{p} \rho_{hs} \mathbf{c}_s. \tag{31}$$

Igualando (31) a cero, se obtiene

$$\frac{A\mathbf{c}_h}{\theta_h} - \rho_h \mathbf{c}_h - \sum_{\substack{s=1 \\ s \neq h}}^{p} \rho_{hs} \mathbf{c}_s = 0. \tag{32}$$

Si premultiplicamos a (32) por $\mathbf{c}_j$ con $j \neq h$, tenemos

$$\frac{\mathbf{c}_j' A \mathbf{c}_h}{\theta_h} - \rho_h \mathbf{c}_j' \mathbf{c}_h - \sum_{\substack{s=1 \\ s \neq h}}^{p} \rho_{hs} \mathbf{c}_j' \mathbf{c}_s = 0,$$

o equivalentemente

$$\frac{\mathbf{c}_j' A \mathbf{c}_h}{\theta_h} = \rho_{hj}.$$

De modo similar, tomando derivadas parciales de g^* con respecto a $\mathbf{c}_j$, igualando a cero y luego premultiplicando por $\mathbf{c}_h$ tenemos

$$\frac{\mathbf{c}_h' A \mathbf{c}_j}{\theta_j} = \rho_{jh}.$$

Como $\mathbf{c}_h' A \mathbf{c}_j$ es un escalar (y, por lo tanto, igual a su traspuesto) y $\rho_{jh} = \rho_{hj}$, resulta

$$\theta_j = \theta_h \text{ ó } \mathbf{c}_h' A \mathbf{c}_j = 0.$$

Suponiendo que todos los θ_j son distintos entre sí, tenemos

$$\mathbf{c}_h' A \mathbf{c}_j = 0, \forall\ h \neq j. \tag{33}$$

Este sistema de ecuaciones, junto con (28), (29) y (30) puede escribirse del siguiente modo,

$$\begin{aligned} C'C &= I \\ C'AC &= \text{diag}(\theta_1 - 1, \ldots, \theta_{p-K} - 1, \theta_{p-K+1}, \ldots, \theta_p). \end{aligned} \tag{34}$$

El sistema de ecuaciones (34) determina a la matriz Σ_0 que cumple con ser punto crítico de g de tal forma que los autovectores de Σ_0 coinciden con los de la matriz $A = B^{-1/2} W B^{-1/2}$, pero con los autovalores

$$\theta_h = \begin{cases} \mathbf{c}_h' A \mathbf{c}_h + 1 & \text{si} \quad 1 \leq h \leq p - K \\ \\ \mathbf{c}_h' A \mathbf{c}_h & \text{si} \quad p - K < h \leq p. \end{cases} \tag{35}$$

Sabemos que para cada h entre 1 y p existe un i_h entre 1 y p tal que

$$\mathbf{c}_h' A \mathbf{c}_h = \lambda_{i_h}(A).$$

Nos resta probar que $\mathbf{c}_h' A \mathbf{c}_h = \lambda_h(A)$, donde notamos con $\lambda_h(A)$ como hasta ahora a los autovalores de la matriz A en orden decreciente $\lambda_1(A) \geq \cdots \geq \lambda_p(A)$. Reemplazando en la expresión (27) que tenemos para la función g resulta

$$\begin{aligned}
g(C,\Theta) &= \sum_{h=1}^{p} \ln(\theta_h) + \sum_{h=1}^{p} \frac{\mathbf{c}_h' A \mathbf{c}_h}{\theta_h} + \sum_{h=1}^{p-K} \frac{1}{\theta_h} \\
&= \sum_{h=1}^{p-K} \ln(\mathbf{c}_h' A \mathbf{c}_h + 1) + \sum_{h=p-K+1}^{p} \ln(\mathbf{c}_h' A \mathbf{c}_h) + \sum_{h=1}^{p-K} \frac{\mathbf{c}_h' A \mathbf{c}_h}{\mathbf{c}_h' A \mathbf{c}_h + 1} \\
&+ \sum_{h=p-K+1}^{p} \frac{\mathbf{c}_h' A \mathbf{c}_h}{\mathbf{c}_h' A \mathbf{c}_h} + \sum_{h=1}^{p-K} \frac{1}{\mathbf{c}_h' A \mathbf{c}_h + 1} \\
&= \sum_{h=1}^{p-K} \ln(\mathbf{c}_h' A \mathbf{c}_h + 1) + \sum_{h=p-K+1}^{p} \ln(\mathbf{c}_h' A \mathbf{c}_h) + \sum_{h=1}^{p-K} \frac{\mathbf{c}_h' A \mathbf{c}_h + 1}{\mathbf{c}_h' A \mathbf{c}_h + 1} \\
&+ \sum_{h=p-K+1}^{p} 1 \\
&= \sum_{h=1}^{p-K} \ln(\mathbf{c}_h' A \mathbf{c}_h + 1) + \sum_{h=p-K+1}^{p} \ln(\mathbf{c}_h' A \mathbf{c}_h) + p.
\end{aligned}$$

Como

$$\ln(x+1) = \ln\left(x\left(1+\frac{1}{x}\right)\right) = \ln(x) + \ln\left(1+\frac{1}{x}\right)$$

resulta

$$g(C,\Theta) = \sum_{h=1}^{p} \ln(\mathbf{c}_h' A \mathbf{c}_h) + \sum_{h=1}^{p-K} \ln\left(1 + \frac{1}{\mathbf{c}_h' A \mathbf{c}_h}\right) + p. \tag{36}$$

Sabemos que $\theta_1, \ldots, \theta_p$ son los valores que minimizan a la función g. Buscamos ver a qué índice i_h corresponde la igualdad $\mathbf{c}_h' A \mathbf{c}_h = \lambda_{i_h}(A)$. Claramente, la expresión

$$\sum_{h=1}^{p-K} \ln\left(1 + \frac{1}{\lambda_{i_h}(A)}\right)$$

alcanza su mínimo valor cuando $\{i_1, \ldots, i_{p-K}\} = \{1, \ldots, p-K\}$. Luego, de acuerdo a (28) y a (35) los valores de θ_h que minimizan a la función g resultan ser

$$\theta_h = \begin{cases} \lambda_h(A) + 1 & \text{si} \quad 1 \leq h \leq p-K \\ \\ \lambda_h(A) & \text{si} \quad p-K < h \leq p \end{cases}$$

y para todo h se tiene que $\mathbf{c}_h' A \mathbf{c}_h = \lambda_h(A)$.

Llamando Ω a la matriz diagonal que contiene a los autovalores de $A = B^{-1/2} W B^{-1/2}$ en orden decreciente,

$$\Omega = \operatorname{diag}(\lambda_1(A), \ldots, \lambda_p(A))$$

resulta que la descomposición espectral de la matriz A es

$$A = B^{-1/2} W B^{-1/2} = C \Omega C'. \tag{37}$$

Si llamamos Σ_0 a la matriz que resulta ser el punto crítico de g, entonces

$$\Sigma_0 = C \Omega^* C'$$

con $\Omega^* = \text{diag}(1 + \lambda_1(A), \ldots, 1 + \lambda_{p-K}(A) \ , \lambda_{p-K+1}(A), \ldots, \lambda_p(A))$. Llamemos $I^* \in \mathbb{R}^{p \times p}$ a la matriz

$$I^* = \begin{bmatrix} I_{(p-K)} & 0_{(p-K)\times K} \\ 0_{K\times(p-K)} & 0_{K\times K} \end{bmatrix}, \tag{38}$$

donde $0_{m\times n}$ es una matriz de $m \times n$ ceros. Luego, I^* es una matriz diagonal que tiene $p - K$ unos y K ceros, en ese orden y $\Omega^* - \Omega = I^*$. Por lo tanto, el EMV de Σ que estamos buscando resulta ser

$$\begin{aligned} \widehat{\Sigma} &= B^{1/2} \Sigma_0 B^{1/2} = B^{1/2} C \Omega^* C' B^{1/2} = B^{1/2} C (\Omega + I^*) C' B^{1/2} \\ &= B^{1/2} C \Omega C' B^{1/2} + B^{1/2} C I^* C' B^{1/2} \\ &= B^{1/2} B^{-1/2} W B^{-1/2} B^{1/2} + B^{1/2} C I^* C' B^{1/2} \\ &= W + B^{1/2} C I^* C' B^{1/2} \qquad (39) \\ &= W + B^{1/2} C_1 C_1' B^{1/2}, \qquad (40) \end{aligned}$$

donde C_1 está formada por las primeras $p - K$ columnas de C, C es la matriz que surge de la descomposición espectral de $B^{-1/2} W B^{-1/2} = C \Omega C'$ y la matriz I^* está definida en (38). Por otro lado, de acuerdo a (26) se tendrá

$$\begin{aligned} \widehat{\alpha}_h &= \widehat{\alpha}_h \left(\widehat{\Sigma} \right) \\ &= p_{\widehat{\Sigma}} (\overline{\mathbf{x}}_{h\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}, V^*) + \overline{\mathbf{x}}_{\bullet\bullet} \\ &= \widehat{\Sigma}^{1/2} D D' \widehat{\Sigma}^{-1/2} (\overline{\mathbf{x}}_{h\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}) + \overline{\mathbf{x}}_{\bullet\bullet} \qquad (41) \end{aligned}$$

con $V^* = V^* \left(\widehat{\Sigma} \right)$ el subespacio generado por $\left\{ \widehat{\Sigma}^{1/2} \mathbf{t}_1, \ldots, \widehat{\Sigma}^{1/2} \mathbf{t}_K \right\}$ siendo $\{\mathbf{t}_1, \ldots, \mathbf{t}_K\}$ los autovectores ortonormales de $\widehat{\Sigma}^{-1/2} B \widehat{\Sigma}^{-1/2}$ asociados a los K mayores autovalores, y $D = [\mathbf{t}_1 \cdots \mathbf{t}_K]$.

Hemos hallado los estimadores de máxima verosimilitud para $\alpha_1, \ldots, \alpha_H, \Sigma$ fijando primero a Σ. Podríamos también haber procedido al revés. A continuación, sintetizamos ambos procedimientos, para obtener una expresión alternativa para el estimador de la matriz Σ. Para ello, introduzcamos la siguiente notación: sea

$$\alpha = (\alpha_1', \ldots, \alpha_H')',$$

luego la función de verosimilitud dada en (11) se puede escribir

$$L(\alpha_1, \ldots, \alpha_H, \Sigma) = L(\alpha, \Sigma). \tag{42}$$

Modo I

1. Fijamos Σ. Luego buscamos el máximo de (42). Llamamos

$$\arg \max_{\substack{\alpha_1, \ldots, \alpha_H \in V \\ \dim(V) = K \\ V \text{ variedad lineal}}} L(\alpha, \Sigma) = \widehat{\alpha}(\Sigma).$$

2. Ahora formamos la función

$$L\left(\widehat{\alpha}\left(\Sigma\right),\Sigma\right)$$

que sólo depende de la matriz Σ y maximizamos obteniendo

$$\arg\max_{\Sigma>0} L\left(\widehat{\alpha}\left(\Sigma\right),\Sigma\right)=\widehat{\Sigma}.$$

3. Finalmente, los EMV resultan ser

$$\widehat{\alpha}=\widehat{\alpha}\left(\widehat{\Sigma}\right),\ \widehat{\Sigma}. \tag{43}$$

Por otro lado, podríamos haber procedido al revés:

Modo II

1. Fijamos α, luego maximizamos a la función de verosimilitud en Σ, es decir, definimos

$$\arg\max_{\Sigma>0} L\left(\alpha,\Sigma\right)=\widetilde{\Sigma}\left(\alpha\right).$$

Para hacer esta cuenta, podemos ver que cuando las medias $\alpha=\left(\alpha_1',\ldots,\alpha_H'\right)'$ son conocidas, los $\mathbf{x}_{hj}-\alpha_h$ pueden considerarse una muestra aleatoria de distribución $N_p\left(\mathbf{0},\Sigma\right)$ y son independientes. Luego, la función (42) es la función de verosimilitud de una muestra $N_p\left(\mathbf{0},\Sigma\right)$ de tamaño $N=\sum_{h=1}^{H}n_h$, el total de observaciones disponibles. En este caso el estimador de máxima verosimilitud es bien conocido y resulta ser

$$\widetilde{\Sigma}\left(\alpha\right)=\frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(\mathbf{x}_{hj}-\alpha_h\right)\left(\mathbf{x}_{hj}-\alpha_h\right)'.$$

2. Formamos la función,

$$L\left(\alpha,\widetilde{\Sigma}\left(\alpha\right)\right)$$

que ahora sólo depende de α. Y buscamos el máximo de dicha función

$$\arg\max_{\substack{\alpha_1,\ldots,\alpha_H\in V\\ \dim(V)=K\\ V\text{ variedad lineal}}} L\left(\alpha,\widetilde{\Sigma}\left(\alpha\right)\right)=\widetilde{\alpha}.$$

3. Por último, tenemos los estimadores de máxima verosimilitud

$$\widetilde{\alpha},\ \widetilde{\Sigma}=\widetilde{\Sigma}\left(\widetilde{\alpha}\right). \tag{44}$$

Como la función de verosimilitud (42) tiene únicos máximos, resulta que los estimadores calculados en (43) y (44) coinciden, lo que implica, por ejemplo que si queremos expresar al EMV de la matriz Σ en función de los valores de α estimados obtengamos

$$\begin{aligned}\widehat{\Sigma}=\widetilde{\Sigma}&=\widetilde{\Sigma}\left(\widetilde{\alpha}\right)\\ &=\frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(\mathbf{x}_{hj}-\widetilde{\alpha}_h\right)\left(\mathbf{x}_{hj}-\widetilde{\alpha}_h\right)'\\ &=\frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(\mathbf{x}_{hj}-\widehat{\alpha}_h\right)\left(\mathbf{x}_{hj}-\widehat{\alpha}_h\right)'.\end{aligned} \tag{45}$$

3.6. EMV de las direcciones edr

Resta entonces estimar las direcciones edr $\beta_1, \ldots, \beta_K$, que cumplen el modelo (1). El Teorema 2.1 indica que la curva de regresión inversa centrada $\mathrm{E}(\mathbf{x} \mid y) - \mathrm{E}(\mathbf{x})$ está contenida en el subespacio lineal K dimensional generado por $\Psi\beta_1, \ldots, \Psi\beta_K$, donde Ψ denota la matriz de covarianza de las $\mathbf{x}$. Por otro lado, en las dos subsecciones anteriores encontramos un estimador de máxima verosimilitud de dicho subespacio: $V^* = gen\left\{\widehat{\Sigma}^{1/2}\mathbf{t}_1, \ldots, \widehat{\Sigma}^{1/2}\mathbf{t}_K\right\}$ siendo $\{\mathbf{t}_1, \ldots, \mathbf{t}_K\}$ los autovectores ortonormales de $\widehat{\Sigma}^{-1/2}B\widehat{\Sigma}^{-1/2}$ asociados a los K mayores autovalores. Luego, para obtener el EMV de las direcciones edr, debemos estimar por máxima verosimilitud a la matriz Ψ y tomar $\widehat{\beta}_1 = \widehat{\Psi}^{-1}\widehat{\Sigma}^{1/2}\mathbf{t}_1, \ldots, \widehat{\beta}_K = \widehat{\Psi}^{-1}\widehat{\Sigma}^{1/2}\mathbf{t}_K$.

Recordemos que para armar los grupos consideramos las H fetas disjuntas $I_1, \ldots, I_H$ en las que se divide el rango de y. Usaremos la siguiente relación

$$\mathrm{Var}(\mathbf{x}) = \mathrm{E}(\mathrm{Var}(\mathbf{x} \mid \ y \in I_h)) + \mathrm{Var}(\mathrm{E}(\mathbf{x} \mid y \in I_h)). \tag{46}$$

El modelo (10) puede ser escrito del siguiente modo

$$\mathbf{x} \mid y \in I_h \sim N_p(\alpha_h, \Sigma), \ h = 1, \ldots, H.$$

Luego, se tiene

$$\mathrm{Var}(\mathbf{x} \mid y \in I_h) = \Sigma$$

y entonces

$$\mathrm{E}(\mathrm{Var}(\mathbf{x} \mid y \in I_h)) = \Sigma.$$

Por otro lado,

$$\mathrm{E}(\mathbf{x} \mid y \in I_h) = \alpha_h.$$

Sin embargo, no hemos propuesto una distribución para las medias $(\alpha_h)_h$, de hecho, éstas figuran como parámetros del modelo que estimamos por máxima verosimilitud. Es por ello que no podemos estimar por máxima verosimilitud la $\mathrm{Var}(\mathrm{E}(\mathbf{x} \mid y \in I_h))$.

Si el número de observaciones elegidas en cada grupo (n_h) fuera proporcional a la probabilidad de que una observación cayera en ese grupo, entonces podríamos pensar a la variable $\mathrm{E}(\mathbf{x} \mid y \in I_h)$ como una variable aleatoria discreta con probabilidades puntuales n_h/N para cada valor. En tal caso, un estimador del segundo sumando de (46) sería

$$\sum_{h=1}^{H} \frac{n_h}{N} (\widehat{\alpha}_h - \widehat{\alpha})(\widehat{\alpha}_h - \widehat{\alpha})',$$

donde

$$\widehat{\alpha} = \frac{1}{N}\sum_{h=1}^{H} n_h \widehat{\alpha}_h, \tag{47}$$

y la estimación de $\Psi = \mathrm{Var}(\mathbf{x})$ se podría obtener como

$$\begin{aligned}\hat{\Psi} &= \hat{\Sigma} + \sum_{h=1}^{H} \frac{n_h}{N} (\widehat{\alpha}_h - \widehat{\alpha})(\widehat{\alpha}_h - \widehat{\alpha})' \\ &= \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h} (\mathbf{x}_{hj} - \widehat{\alpha}_h)(\mathbf{x}_{hj} - \widehat{\alpha}_h)' + \sum_{h=1}^{H} \frac{n_h}{N} (\widehat{\alpha}_h - \widehat{\alpha})(\widehat{\alpha}_h - \widehat{\alpha})', \end{aligned} \tag{48}$$

siendo válida la última igualdad en virtud de (45). La matriz de covarianza muestral de las N observaciones tiene como numerador la siguiente expresión

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj}-\overline{\mathbf{x}}_{\bullet\bullet})(\mathbf{x}_{hj}-\overline{\mathbf{x}}_{\bullet\bullet})'\\
&=\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj}-\widehat{\alpha}_h+\widehat{\alpha}_h-\overline{\mathbf{x}}_{\bullet\bullet})(\mathbf{x}_{hj}-\widehat{\alpha}_h+\widehat{\alpha}_h-\overline{\mathbf{x}}_{\bullet\bullet})'\\
&=\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj}-\widehat{\alpha}_h)(\mathbf{x}_{hj}-\widehat{\alpha}_h)'+\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\widehat{\alpha}_h-\overline{\mathbf{x}}_{\bullet\bullet})(\widehat{\alpha}_h-\overline{\mathbf{x}}_{\bullet\bullet})'\\
&+\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj}-\widehat{\alpha}_h)(\widehat{\alpha}_h-\overline{\mathbf{x}}_{\bullet\bullet})'+\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\widehat{\alpha}_h-\overline{\mathbf{x}}_{\bullet\bullet})(\mathbf{x}_{hj}-\widehat{\alpha}_h)'.
\end{aligned}$$

Calculemos cada uno de los sumandos por separado y veamos que los dos últimos sumandos valen 0. Como uno es el traspuesto del otro, bastará trabajar con uno de ellos, por ejemplo,

$$\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj}-\widehat{\alpha}_h)(\widehat{\alpha}_h-\overline{\mathbf{x}}_{\bullet\bullet})'=\sum_{h=1}^{H}n_h(\overline{\mathbf{x}}_{h\bullet}-\widehat{\alpha}_h)(\widehat{\alpha}_h-\overline{\mathbf{x}}_{\bullet\bullet})'.$$

En virtud de (41) que caracteriza a los $\widehat{\alpha}_h$, si tomamos $\{\mathbf{t}_1,\ldots,\mathbf{t}_p\}$ una base ortonormal de autovectores de $\widehat{\Sigma}^{-1/2}B\widehat{\Sigma}^{-1/2}$ ordenados de manera decreciente según el valor de sus autovalores $\lambda_1\geq\cdots\geq\lambda_p$ y $D=[\mathbf{t}_1\cdots\mathbf{t}_K]\in\mathbb{R}^{p\times K}$, tenemos

$$\begin{aligned}
\overline{\mathbf{x}}_{h\bullet}-\widehat{\alpha}_h&=\overline{\mathbf{x}}_{h\bullet}-\left[\widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_{h\bullet}-\overline{\mathbf{x}}_{\bullet\bullet})+\overline{\mathbf{x}}_{\bullet\bullet}\right]\\
&=\overline{\mathbf{x}}_{h\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}-\widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_{h\bullet}-\overline{\mathbf{x}}_{\bullet\bullet})\\
&=\left(I-\widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}\right)(\overline{\mathbf{x}}_{h\bullet}-\overline{\mathbf{x}}_{\bullet\bullet})
\end{aligned}$$

y

$$\begin{aligned}
\widehat{\alpha}_h-\overline{\mathbf{x}}_{\bullet\bullet}&=\left[\widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_{h\bullet}-\overline{\mathbf{x}}_{\bullet\bullet})+\overline{\mathbf{x}}_{\bullet\bullet}\right]-\overline{\mathbf{x}}_{\bullet\bullet}\\
&=\widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_{h\bullet}-\overline{\mathbf{x}}_{\bullet\bullet}).
\end{aligned}$$

Luego,

$$\begin{aligned}
&\sum_{h=1}^{H}n_h(\overline{\mathbf{x}}_{h\bullet}-\widehat{\alpha}_h)(\widehat{\alpha}_h-\overline{\mathbf{x}}_{\bullet\bullet})'\\
&=\left(I-\widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}\right)\sum_{h=1}^{H}n_h(\overline{\mathbf{x}}_{h\bullet}-\overline{\mathbf{x}}_{\bullet\bullet})(\overline{\mathbf{x}}_{h\bullet}-\overline{\mathbf{x}}_{\bullet\bullet})'\widehat{\Sigma}^{-1/2}DD'\widehat{\Sigma}^{1/2}\\
&=\left(I-\widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}\right)BN\widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}.
\end{aligned}$$

Sea $\Lambda^*\in\mathbb{R}^{K\times K},\Lambda^*=diag(\lambda_1,\ldots,\lambda_K)$, entonces

$$\widehat{\Sigma}^{-1/2}B\widehat{\Sigma}^{-1/2}D=D\Lambda^*,$$

o, equivalentemente

$$B\widehat{\Sigma}^{-1/2}D = \widehat{\Sigma}^{1/2}D\Lambda^*.$$

Entonces

$$\begin{aligned}
&N\left(I - \widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}\right)B\widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}\\
&= N\left(I - \widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}\right)\widehat{\Sigma}^{1/2}D\Lambda^*D'\widehat{\Sigma}^{-1/2}\\
&= N\widehat{\Sigma}^{1/2}D\Lambda^*D'\widehat{\Sigma}^{-1/2} - N\widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}\widehat{\Sigma}^{1/2}D\Lambda^*D'\widehat{\Sigma}^{-1/2}\\
&= N\widehat{\Sigma}^{1/2}D\Lambda^*D'\widehat{\Sigma}^{-1/2} - N\widehat{\Sigma}^{1/2}DD'D\Lambda^*D'\widehat{\Sigma}^{-1/2}\\
&= N\widehat{\Sigma}^{1/2}D\Lambda^*D'\widehat{\Sigma}^{-1/2} - N\widehat{\Sigma}^{1/2}D\Lambda^*D'\widehat{\Sigma}^{-1/2}\\
&= 0. \qquad (49)
\end{aligned}$$

Para el segundo sumando tenemos

$$\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(\widehat{\alpha}_h - \overline{\mathbf{x}}_{\bullet\bullet}\right)\left(\widehat{\alpha}_h - \overline{\mathbf{x}}_{\bullet\bullet}\right)' = \sum_{h=1}^{H} n_h\left(\widehat{\alpha}_h - \overline{\mathbf{x}}_{\bullet\bullet}\right)\left(\widehat{\alpha}_h - \overline{\mathbf{x}}_{\bullet\bullet}\right)',$$

y como

$$\widehat{\alpha}_h = \widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}\left(\overline{\mathbf{x}}_{h\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right) + \overline{\mathbf{x}}_{\bullet\bullet}$$

resulta que

$$\begin{aligned}
\widehat{\alpha} &= \frac{1}{N}\sum_{h=1}^{H} n_h\widehat{\alpha}_h\\
&= \widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}\frac{1}{N}\sum_{h=1}^{H} n_h\left(\overline{\mathbf{x}}_{h\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right) + \frac{1}{N}\sum_{h=1}^{H} n_h\overline{\mathbf{x}}_{\bullet\bullet}\\
&= \widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}\left(\overline{\mathbf{x}}_{\bullet\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right) + \overline{\mathbf{x}}_{\bullet\bullet}\\
&= \overline{\mathbf{x}}_{\bullet\bullet}. \qquad (50)
\end{aligned}$$

Entonces, la matriz de covarianza muestral de las observaciones se puede escribir del siguiente modo, en virtud de (49) y (50)

$$\begin{aligned}
&\frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(\mathbf{x}_{hj} - \overline{\mathbf{x}}_{\bullet\bullet}\right)\left(\mathbf{x}_{hj} - \overline{\mathbf{x}}_{\bullet\bullet}\right)'\\
&= \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(\mathbf{x}_{hj} - \widehat{\alpha}_h\right)\left(\mathbf{x}_{hj} - \widehat{\alpha}_h\right)' + \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(\widehat{\alpha}_h - \overline{\mathbf{x}}_{\bullet\bullet}\right)\left(\widehat{\alpha}_h - \overline{\mathbf{x}}_{\bullet\bullet}\right)'\\
&= \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(\mathbf{x}_{hj} - \widehat{\alpha}_h\right)\left(\mathbf{x}_{hj} - \widehat{\alpha}_h\right)' + \frac{1}{N}\sum_{h=1}^{H} n_h\left(\widehat{\alpha}_h - \widehat{\alpha}\right)\left(\widehat{\alpha}_h - \widehat{\alpha}\right)'\\
&= \hat{\Psi}
\end{aligned}$$

donde la última igualdad vale por (48). Por lo tanto, si el número de observaciones elegidas en cada grupo fuera proporcional a la probabilidad de que una observación cayera en dicho grupo, el estimador de máxima verosimilitud de Ψ sería la matriz de covarianza muestral de las observaciones, y, en este caso, los EMV de las direcciones edr se tomarían como $\widehat{\Psi}^{-1}\widehat{\Sigma}^{1/2}\mathbf{t}_1, \ldots, \widehat{\Psi}^{-1}\widehat{\Sigma}^{1/2}\mathbf{t}_K$ siendo $\{\mathbf{t}_1, \ldots, \mathbf{t}_K\}$ los autovectores ortonormales de $\widehat{\Sigma}^{-1/2}B\widehat{\Sigma}^{-1/2}$ asociados a los K mayores autovalores.

4. Comparación de los dos métodos de estimación del subespacio V: el algoritmo de Li y los estimadores obtenidos por máxima verosimilitud

Tomamos observaciones $(\mathbf{x}_i', y_i)'$ $(i = 1, ..., N)$. Comparemos los estimadores obtenidos en las Secciones 2 y 3. Para ello, dividimos el rango de y en H fetas $I_1, ..., I_H$, $I_h = [a_{h-1}, a_h)$. Sea n_h la cantidad de observaciones que caen en la feta I_h. Renombramos a cada observación según la feta a la que pertenece. Luego las observaciones de la feta I_h serán $\left(\mathbf{x}_{hj}', y_{hj}\right)'$ con $1 \leq j \leq n_h$. Entonces supondremos que $(\mathbf{x}_{hj})_{\ 1\leq h\leq H, 1\leq j\leq n_h}$ son vectores aleatorios en $\mathbb{R}^p$ independientes con distribución $N_p(\alpha_h, \Sigma)$ y supondremos que los vectores α_h, $1 \leq h \leq H$ pertenecen a una variedad lineal $V + \mathbf{a}$ en $\mathbb{R}^p$ de dimensión K. Ahora podemos estimar el subespacio de direcciones edr por máxima verosimilitud.

Primero comparemos los resultados de los estimadores del subespacio V obtenidos por ambos métodos. Llamaremos V_L^* y V_{MV}^* a los estimadores de los subespacios obtenidos por cada uno de los métodos, en el orden presentado.

4.1. Algoritmo de Li

Según el Teorema 2.1 los estimadores del subespacio V resultan ser $\widehat{\Psi}\widehat{\beta}_1, \ldots, \widehat{\Psi}\widehat{\beta}_K$, donde $\widehat{\Psi}$ denota la matriz de covarianza muestral de las $\mathbf{x}$, y los $\widehat{\beta}_i$ son los definidos en (3). Es decir, $\widehat{\beta}_k = \widehat{\Psi}^{-1/2}\widehat{\eta}_k$, $(k = 1, \ldots, K)$ donde $\widehat{\eta}_1, ..., \widehat{\eta}_K$ son los autovectores correspondientes a los K mayores autovalores de la matriz $\widehat{\Phi}$:

$$\widehat{\Phi} = \widehat{\Psi}^{-1/2} \sum_{h=1}^{H} \frac{n_h}{N} \left(\frac{1}{n_h} \sum_{y_i \in I_h} [\mathbf{x}_i - \overline{\mathbf{x}}] \right) \left(\frac{1}{n_h} \sum_{y_i \in I_h} [\mathbf{x}_i - \overline{\mathbf{x}}] \right)' \widehat{\Psi}^{-1/2}.$$

Con la notación de doble índice, tenemos

$$\begin{aligned} \widehat{\Phi} &= \widehat{\Psi}^{-1/2} \sum_{h=1}^{H} \frac{n_h}{N} \left(\frac{1}{n_h} \sum_{j=1}^{n_h} [\mathbf{x}_{hj} - \overline{\mathbf{x}}_{\bullet\bullet}] \right) \left(\frac{1}{n_h} \sum_{j=1}^{n_h} [\mathbf{x}_{hj} - \overline{\mathbf{x}}_{\bullet\bullet}] \right)' \widehat{\Psi}^{-1/2} \\ &= \widehat{\Psi}^{-1/2} \sum_{h=1}^{H} \frac{n_h}{N} \left(\overline{\mathbf{x}}_{h\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right) \left(\overline{\mathbf{x}}_{h\bullet} - \overline{\mathbf{x}}_{\bullet\bullet} \right)' \widehat{\Psi}^{-1/2} \\ &= \widehat{\Psi}^{-1/2} B \widehat{\Psi}^{-1/2} \end{aligned}$$

y

$$\widehat{\Psi} = \frac{1}{N} \sum_{h=1}^{H} \sum_{j=1}^{n_h} [\mathbf{x}_{hj} - \overline{\mathbf{x}}_{\bullet\bullet}] [\mathbf{x}_{hj} - \overline{\mathbf{x}}_{\bullet\bullet}]' = B + W$$

con las matrices B y W definidas en (4) y (5). Luego, los $\widehat{\eta}_i$ son los autovectores correspondientes a los K mayores autovalores de la matriz

$$\widehat{\Phi} = (B + W)^{-1/2} B (B + W)^{-1/2}.$$

Los generadores del subespacio V_L^* son

$$\widehat{\Psi}\widehat{\beta}_i = \widehat{\Psi}\widehat{\Psi}^{-1/2}\widehat{\eta}_i = \widehat{\Psi}^{1/2}\widehat{\eta}_i = (B + W)^{1/2} \widehat{\eta}_i, \ 1 \leq i \leq K.$$

Por la Proposición 2.1, resulta que los $(B+W)^{1/2}\,\widehat{\eta}_i$ son autovectores de la matriz $B\,(B+W)^{-1}$ asociados a los K mayores autovalores.

4.2. Estimadores de máxima verosimilitud

El método de máxima verosimilitud aplicado a este problema da lugar al subespacio estimado V^*_{MV} generado por $\left\{\widehat{\Sigma}^{1/2}\mathbf{t}_1,\ldots,\widehat{\Sigma}^{1/2}\mathbf{t}_K\right\}$ con $\{\mathbf{t}_1,\ldots,\mathbf{t}_K\}$ autovectores ortonormales asociados a los K mayores autovalores de la matriz simétrica $\widehat{\Sigma}^{-1/2}B\widehat{\Sigma}^{-1/2}$, donde $\widehat{\Sigma}$ se calcula según (40) es decir: $\widehat{\Sigma} = W + B^{1/2}CI^*C'B^{1/2}$, con C la matriz que tiene como columnas una base ortonormal de autovectores de $B^{-1/2}WB^{-1/2}$. Nuevamente, en virtud de la Proposición 2.1 resulta que los generadores del subespacio calculado por máxima verosimilitud, los $\widehat{\Sigma}^{1/2}\mathbf{t}_i$ son los autovectores de la matriz $B\widehat{\Sigma}^{-1}$ asociados a sus K mayores autovalores.

Teorema 4.1 *Las matrices $B\,(B+W)^{-1}$ y $B\widehat{\Sigma}^{-1}$ definidas anteriormente tienen los mismos p autovectores, y sus autovalores están relacionados del siguiente modo:*

$$\lambda_i\left(B\,(B+W)^{-1}\right) = \begin{cases} \dfrac{\lambda_i\left(B\widehat{\Sigma}^{-1}\right)}{1+\lambda_i\left(B\widehat{\Sigma}^{-1}\right)} & 1 \le i \le p-K \\ \\ \lambda_i\left(B\widehat{\Sigma}^{-1}\right) & p-K+1 \le i \le p \end{cases}$$

Demostración. De la definición de $\widehat{\Sigma}$

$$\widehat{\Sigma} = W + B^{1/2}CI^*C'B^{1/2}$$

postmultiplicando por la inversa de B

$$\begin{aligned}\widehat{\Sigma}B^{-1} &= WB^{-1} + B^{1/2}CI^*C'B^{1/2}B^{-1} \\ &= WB^{-1} + B^{1/2}CI^*C'B^{-1/2}.\end{aligned}$$

Consideremos la descomposición espectral de la matriz simétrica $B^{-1/2}WB^{-1/2}$ dada en (37)

$$B^{-1/2}WB^{-1/2} = C\Omega C'$$

con C una matriz ortogonal y Ω una matriz diagonal, de modo que $\omega_1 \ge \cdots \ge \omega_p$. Luego, premultiplicando por $B^{1/2}$

$$WB^{-1/2} = B^{1/2}C\Omega C'$$

y postmultiplicando por $B^{-1/2}$ tenemos

$$WB^{-1} = B^{1/2}C\Omega C'B^{-1/2}.$$

De aquí, resulta

$$\begin{aligned}\widehat{\Sigma}B^{-1} &= B^{1/2}C\Omega C'B^{-1/2} + B^{1/2}CI^*C'B^{-1/2} \\ &= B^{1/2}C\,(\Omega + I^*)\,C'B^{-1/2}.\end{aligned}$$

Invirtiendo la igualdad anterior tenemos (puesto que la matriz C es ortogonal)

$$B\widehat{\Sigma}^{-1} = B^{1/2}C\left(\Omega + I^*\right)^{-1}C'B^{-1/2}.$$

Equivalentemente, postmultiplicando por $B^{1/2}C$ obtenemos

$$B\widehat{\Sigma}^{-1}B^{1/2}C = B^{1/2}C\left(\Omega + I^*\right)^{-1}.$$

Como la matriz $\left(\Omega + I^*\right)^{-1}$ es diagonal, resulta que las columnas de la matriz $B^{1/2}C$ son los autovectores de $B\widehat{\Sigma}^{-1}$, y la diagonal de la matriz $\left(\Omega + I^*\right)^{-1}$ sus autovalores. Recordemos que Ω tiene los autovalores de la matriz $B^{-1/2}WB^{-1/2}$. Luego, los autovalores de $B\widehat{\Sigma}^{-1}$ son

$$\lambda_{p-i+1}\left(B\widehat{\Sigma}^{-1}\right) = \begin{cases} \left(1+\omega_i\right)^{-1} & 1 \leq i \leq p-K \\ \\ \omega_i^{-1} & p-K+1 \leq i \leq p \end{cases}$$

con $\omega_i = \lambda_i\left(B^{-1/2}WB^{-1/2}\right)$.

Por otro lado,

$$\begin{aligned}\left(B+W\right)B^{-1} &= I + WB^{-1} \\ &= I + B^{1/2}C\Omega C'B^{-1/2} \\ &= B^{1/2}C\left(I+\Omega\right)C'B^{-1/2}.\end{aligned}$$

Invirtiendo resulta

$$B\left(B+W\right)^{-1} = B^{1/2}C\left(I+\Omega\right)^{-1}C'B^{-1/2}.$$

Luego,

$$B\left(B+W\right)^{-1}B^{1/2}C = B^{1/2}C\left(I+\Omega\right)^{-1}$$

y las columnas de $B^{1/2}C$ también resultan ser los autovectores de $B\left(B+W\right)^{-1}$. Los autovalores son

$$\lambda_{p-i+1}\left(B\left(B+W\right)^{-1}\right) = \left(1+\omega_i\right)^{-1}, \quad 1 \leq i \leq p.$$

Claramente vale la relación entre los autovalores establecida en el enunciado. ■

Observación 4.1 *El subespacio que contiene a la curva de regresión inversa estimado por el algoritmo de Li a partir de una muestra, V_L^* y el que contiene a las medias de cada grupo estimado según máxima verosimilitud, V_{MV}^* coinciden. En particular, las estimaciones dadas por el algoritmo de Li y por el método de Máxima Verosimilitud planteado en la Sección 3 para las direcciones edr son iguales. Es decir: El subespacio V_L^* generado por $\left(B+W\right)^{1/2}\widehat{\eta}_i$, $1 \leq i \leq K$ con los $\widehat{\eta}_i$ autovectores ortonormales asociados a los mayores K autovalores de $\left(B+W\right)^{-1/2}B\left(B+W\right)^{-1/2}$ y el subespacio V_{MV}^* generado por $\left\{\widehat{\Sigma}^{1/2}\mathbf{t}_1, \ldots, \widehat{\Sigma}^{1/2}\mathbf{t}_K\right\}$ con $\{\mathbf{t}_1, \ldots, \mathbf{t}_K\}$ autovectores ortonormales asociados a los K mayores autovalores de la matriz simétrica $\widehat{\Sigma}^{-1/2}B\widehat{\Sigma}^{-1/2}$, donde $\widehat{\Sigma} = W + B^{1/2}CI^*C'B^{1/2}$ coinciden.*

Observación 4.2 *Los autovalores que sirven para evaluar la dimensión del subespacio K en el algoritmo de Li son los autovalores de la matriz*

$$\widehat{\Phi} = (B+W)^{-1/2} B (B+W)^{-1/2}.$$

que son los mismos que los de la matriz

$$B (B+W)^{-1}$$

en virtud de la Proposición 2.1.

5. Conclusiones

En esta tesis se buscó estudiar el procedimiento de reducción de dimensión para observaciones multivariadas conocido como Regresión Inversa Partida. Para este método se demostró que el algoritmo propuesto por Li (1991) para regresión inversa partida es equivalente a obtener las coordenadas discriminantes correspondiente a la partición de las observaciones de acuerdo a los valores de la variable respuesta. El resultado se presenta en el Teorema 2.3 de la Sección 2.3.

Por otro lado, probamos que el algoritmo SIR propuesto por Li es equivalente a estimar por máxima verosimilitud el subespacio donde están las medias de los diferentes elementos de la partición de las observaciones, suponiendo normalidad y una misma matriz de covarianza. La utilidad de este enfoque subyace en el hecho de que poder enmarcar un estimador dentro de una estrategia de estimación de máxima verosimilitud permite aplicarle a los estimadores resultantes todos los resultados válidos para un método tan clásico, a la vez que permite encontrar una vía alternativa a las ya propuestas (Gather et al. [10], [11]) para obtener un método robusto de estimación.

En esa línea de trabajo, propusimos un método alternativo para estimar el modelo SIR, suponiendo que las observaciones de cada grupo tienen una distribución normal multivariada p-dimensional, es decir,

$$\mathbf{x}_{hj} \sim N_p\left(\alpha_h, \Sigma\right),\ 1 \leq j \leq n_h, 1 \leq h \leq H. \tag{51}$$

Como por el Teorema 2.1 y el Lema 3.1, las medias condicionales $\alpha_h = \mathrm{E}\left(\mathbf{x}\,|y \in I_h\right)$, $1 \leq h \leq H$ estarán en una variedad lineal de $\mathbb{R}^p$ de dimensión K, propusimos encontrar los estimadores de los α_h y Σ por el método de máxima verosimilitud sujeto a la restricción de que las medias pertenezcan a una variedad lineal de dimensión K. Como el Teorema 2.1 establece un vínculo entre el subespacio al cual pertenece la $\mathrm{E}(\mathbf{x}|y) - \mathrm{E}(\mathbf{x})$ y las direcciones edr, que se da a través de la matriz de covarianza Ψ de las observaciones, este enfoque alternativo permite estimar las direcciones edr una vez que se tengan estimadores del subespacio al que pertenece la curva $\mathrm{E}(\mathbf{x}|y) - \mathrm{E}(\mathbf{x})$, si el número de observaciones elegidas en cada grupo (n_i) fuera proporcional a la probabilidad de que una observación cayera en dicho grupo.

Definiendo

$$B = \frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{n_i}\left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right)\left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right)'$$

y

$$W = \frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_{i\bullet}\right)\left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_{i\bullet}\right)',$$

los estimadores de máxima verosimilitud de Σ, α_h y β_k resultan ser

$$\hat{\Sigma} = W + B^{1/2}C\begin{bmatrix} I_{(p-K)} & 0_{(p-K)\times K} \\ 0_{K\times(p-K)} & 0_{K\times K} \end{bmatrix}C'B^{1/2}$$

con C la matriz ortogonal que surge de la descomposición espectral de

$$B^{-1/2}WB^{-1/2} = C\Omega C'$$

siendo Ω la matriz diagonal, que contiene a los autovalores ordenados en forma decreciente. Los EMV de α_h son

$$\widehat{\alpha}_h = \widehat{\Sigma}^{1/2} DD' \widehat{\Sigma}^{-1/2} \left(\overline{\mathbf{x}}_{h\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right) + \overline{\mathbf{x}}_{\bullet\bullet}$$

donde $D = [\mathbf{t}_1 \cdots \mathbf{t}_K] \in \mathbb{R}^{p \times K}$ con $\{\mathbf{t}_1, \ldots, \mathbf{t}_K\}$ los autovectores ortonormales de $\widehat{\Sigma}^{-1/2} B \widehat{\Sigma}^{-1/2}$ asociados a los K mayores autovalores y los estimadores de las direcciones edr resultan ser

$$\widehat{\beta}_k = \widehat{\Psi}^{-1} \widehat{\Sigma}^{1/2} \mathbf{t}_k,$$

$k = 1, \ldots, K$, donde $\widehat{\Psi}$ denota la matriz de covarianza muestral de las $\mathbf{x}$. Finalmente, demostramos que el subespacio que contiene a la curva de regresión inversa estimado por el algoritmo de Li a partir de una muestra, y el que contiene a las medias de cada grupo estimado por máxima verosimilitud siguiendo el modelo (51), coinciden. En particular, las estimaciones dadas por el algoritmo de Li y por el método de Máxima Verosimilitud para las direcciones edr coinciden, tal como se presentó en la Observación 4.1.

Finalmente, vale la pena notar que, ya que el procedimiento propuesto por Li es muy sensible a la presencia de outliers, (Gather et al. [11]), este enfoque que permite pensar el modelo SIR como un problema de estimación por máxima verosimilitud permitiría encontrar estimadores robustos, de manera análoga al procedimiento propuesto por García Ben, Martínez y Yohai [9] para hallar estimadores robustos y eficientes para el modelo lineal multivariado. Es en esta línea de trabajo que queda mucho por hacer.

6. Apéndice

6.1. Resultados de Álgebra Lineal utilizados

En esta sección incluimos las demostraciones pospuestas en la Sección 2.3, de Coordenadas Discriminantes y en la Sección 3.4, de cálculo de los Estimadores de Máxima Verosimilitud de posición.

Demostración de la Proposición 2.1.

Proposición 2.1 (página 16) *Sea $D \in R^{p\times p}$ matriz definida positiva, $A \in R^{p\times p}$ matriz simétrica.*

i. *λ es un autovalor de $D^{-1}A$ si y sólo si λ es un autovalor de $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ Como esta última matriz es simétrica, dichos autovalores son todos reales.*

ii. *Sean $v_1,\ldots,v_p$ autovectores ortonormales de la matriz $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ asociados a los autovalores $\lambda_1 \geq \cdots \geq \lambda_p$. Sean $w_i = D^{-\frac{1}{2}}v_i$, entonces los w_i resultan autovectores de la matriz $D^{-1}A$ asociados a los mismos autovalores λ_i y satisfacen*

   a) $v_i'v_i = 1$, *o, equivalentemente* $w_i'Dw_i = 1$.

   b) $v_i'v_j = 0$, *o,equivalentemente* $w_i'Dw_j = 0 \qquad \forall\ i \neq j$.

iii. *Sean $v_1,\ldots,v_p$ autovectores ortonormales de la matriz $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ asociados a los autovalores $\lambda_1 \geq \cdots \geq \lambda_p$. Sean $z_i = D^{\frac{1}{2}}v_i$, entonces los z_i resultan autovectores de la matriz AD^{-1} asociados a los mismos autovalores λ_i y satisfacen*

   a) $v_i'v_i = 1$, *o, equivalentemente* $z_i'D^{-1}z_i = 1$.

   b) $v_i'v_j = 0$, *o,equivalentemente* $z_i'D^{-1}z_j = 0 \qquad \forall\ i \neq j$.

iv. *Si A es definida positiva y λ es un autovalor de $D^{-1}A$, resulta que $\lambda > 0$.*

v. *Si $A - D$ es definida positiva y λ es un autovalor de $D^{-1}A$, resulta que $\lambda < 1$.*

Demostración.

i. λ es un autovalor de $D^{-1}A$ si y sólo si $\left|D^{-1}A - \lambda I\right| = 0$. Pero

$$\begin{aligned}\left|D^{-1}A - \lambda I\right| &= \left|D^{-\frac{1}{2}}D^{-\frac{1}{2}}AD^{-\frac{1}{2}}D^{\frac{1}{2}} - \lambda D^{-\frac{1}{2}}D^{\frac{1}{2}}\right| \\ &= \left|D^{-\frac{1}{2}}\left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}} - \lambda I\right)D^{\frac{1}{2}}\right| \\ &= \left|D^{-\frac{1}{2}}\right|\left|D^{-\frac{1}{2}}AD^{-\frac{1}{2}} - \lambda I\right|\left|D^{\frac{1}{2}}\right|.\end{aligned}$$

$\left|D^{-1}A - \lambda I\right| = 0$ si y sólo si $\left|D^{-\frac{1}{2}}AD^{-\frac{1}{2}} - \lambda I\right| = 0$, o, equivalentemente λ es un autovalor de $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$.

ii. En efecto,

$$D^{-1}A\mathbf{w}_i = D^{-1}AD^{-\frac{1}{2}}\mathbf{v}_i = D^{-\frac{1}{2}}D^{-\frac{1}{2}}AD^{-\frac{1}{2}}\mathbf{v}_i$$
$$= D^{-\frac{1}{2}}\lambda_i\mathbf{v}_i = \lambda_i\mathbf{w}_i.$$

Luego, los $\mathbf{w}_i$ resultan ser autovalores de $D^{-1}A$ asociados al mismo autovalor. Más aún,

(a) $\mathbf{w}_i'D\mathbf{w}_i = \mathbf{v}_i'D^{-\frac{1}{2}}DD^{-\frac{1}{2}}\mathbf{v}_i = \mathbf{v}_i'\mathbf{v}_i = \|\mathbf{v}_i\|^2 = 1.$

(b) $\mathbf{w}_i'D\mathbf{w}_j = \mathbf{v}_i'D^{-\frac{1}{2}}DD^{-\frac{1}{2}}\mathbf{v}_j = \mathbf{v}_i'\mathbf{v}_j = 0.$

iii. Como

$$D^{-\frac{1}{2}}AD^{-\frac{1}{2}}\mathbf{v}_i = \lambda_i\mathbf{v}_i$$
$$D^{-\frac{1}{2}}AD^{-\frac{1}{2}}D^{-\frac{1}{2}}D^{\frac{1}{2}}\mathbf{v}_i = \lambda_i\mathbf{v}_i.$$

Premultiplicando por $D^{\frac{1}{2}}$ resulta

$$AD^{-\frac{1}{2}}D^{-\frac{1}{2}}D^{\frac{1}{2}}\mathbf{v}_i = \lambda_iD^{\frac{1}{2}}\mathbf{v}_i$$
$$AD^{-1}\mathbf{z}_i = \lambda_i\mathbf{z}_i.$$

Luego, los $\mathbf{z}_i$ resultan ser autovalores de AD^{-1} asociados al mismo autovalor. Más aún

(a) $\mathbf{z}_i'D^{-1}\mathbf{z}_i = \mathbf{v}_i'D^{\frac{1}{2}}D^{-1}D^{\frac{1}{2}}\mathbf{v}_i = \mathbf{v}_i'\mathbf{v}_i = \|\mathbf{v}_i\|^2 = 1.$

(b) $\mathbf{z}_i'D^{-1}\mathbf{z}_j = \mathbf{v}_i'D^{\frac{1}{2}}D^{-1}D^{\frac{1}{2}}\mathbf{v}_j = \mathbf{v}_i'\mathbf{v}_j = 0.$

iv. Sea λ un autovalor de $D^{-1}A$. Por el ítem ii., λ es un autovalor de $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$, que es una matriz simétrica y definida positiva. Luego, $\lambda > 0$.

v. Sea λ un autovalor de $D^{-1}A$. Luego $1-\lambda$ resulta un autovalor de $I - D^{-1}A = D^{-1}D - D^{-1}A = D^{-1}(D-A)$. Como $D-A$ es definida positiva, resulta del ítem i. que $1-\lambda$ es autovalor de $D^{-\frac{1}{2}}(D-A)D^{-\frac{1}{2}}$, que es una matriz simétrica y definida positiva. Luego, $1-\lambda > 0$ si y sólo si $\lambda < 1$. ■

Demostración del Lema 3.2.

Lema 3.2 (página 30) *Si $Q \in R^{p\times p}$, es semidefinida positiva, $C \in R^{p\times K}$ con $K < p$, tal que $C'C = I_K$, entonces*

$$\sum_{i=1}^{K}\lambda_i(C'QC) \leq \sum_{i=1}^{K}\lambda_i(Q)$$

donde $\lambda_i(A)$ son los autovalores ordenados de la matriz simétrica $A: \lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_p(A)$. La igualdad se alcanza si la matriz C tiene por columnas a los autovectores de norma uno asociados a los autovalores $\lambda_i(Q)$ para $i = 1,\ldots,K$.

Demostración. Consideremos la descomposición espectral de la matriz Q,

$$Q = B\Lambda B'$$

con $B \in R^{p\times p}$ una matriz ortogonal, $BB' = B'B = I_p$ cuyas columnas contienen los autovectores ortogonales de Q y $\Lambda = \text{diag}(\lambda_1(Q), \ldots, \lambda_p(Q))$. Luego

$$C'QC = C'B\Lambda B'C.$$

Sea $D = B'C \in R^{p\times K}$, $D = (D_{ij})_{ij}$ entonces $D'D = C'BB'C = C'C = I_K$, por lo que la matriz D tiene la misma propiedad que C, pero con respecto a la matriz Λ que es diagonal. Como

$$\sum_{i=1}^{K} \lambda_i(C'QC) = \text{traza}(C'QC) = \text{traza}(D'\Lambda D),$$

llamando $E = D'\Lambda D$, $E = (E_{ij})_{ij}$ resulta

$$E_{ii} = \sum_{h=1}^{p}\sum_{s=1}^{p} D_{hi}\Lambda_{hs}D_{si} = \sum_{h=1}^{p} D_{hi}\Lambda_{hh}D_{hi} = \sum_{h=1}^{p} D_{hi}^2\lambda_h(Q),$$

por ser Λ una matriz diagonal. Por lo tanto,

$$\text{traza}(D'\Lambda D) = \sum_{i=1}^{K}\sum_{h=1}^{p} D_{hi}^2\lambda_h(Q) = \sum_{h=1}^{p}\lambda_h(Q)\left[\sum_{i=1}^{K} D_{hi}^2\right].$$

Llamando

$$f_h = \sum_{i=1}^{K} D_{hi}^2,$$

resulta

$$\sum_{h=1}^{p} f_h = \sum_{h=1}^{p}\sum_{i=1}^{K} D_{hi}^2 = \text{traza}(D'D) = \text{traza}(I_K) = K. \tag{52}$$

Luego, por (52) tenemos

$$\begin{aligned}
\text{traza}(D'\Lambda D) = \sum_{h=1}^{p}\lambda_h(Q) f_h &= \sum_{h=1}^{K}\lambda_h(Q) f_h + \sum_{h=K+1}^{p}\lambda_h(Q) f_h \\
&\leq \sum_{h=1}^{K}\lambda_h(Q) f_h + \lambda_K(Q)\sum_{h=K+1}^{p} f_h \\
&= \sum_{h=1}^{K}\lambda_h(Q) f_h + \lambda_K(Q)\left[K - \sum_{h=1}^{K} f_h\right] \\
&= \sum_{h=1}^{K}\lambda_h(Q) f_h + \lambda_K(Q)\left[\sum_{h=1}^{K}(1 - f_h)\right] \\
&\leq \sum_{h=1}^{K}\lambda_h(Q)[f_h + 1 - f_h] = \sum_{h=1}^{K}\lambda_h(Q). \quad \blacksquare
\end{aligned}$$

6.2. Propiedades de Esperanza Condicional

Proposición 6.1 *Sean A, B, C variables o vectores aleatorios. Entonces*

$$\mathrm{E}\left[A \mid B\right] = \mathrm{E}\left[\mathrm{E}\left[A \mid C, B\right] \mid B\right].$$

Demostración. Sean $g\left(B\right) = \mathrm{E}\left[\mathrm{E}\left[A \mid C, B\right] \mid B\right]$ y $w\left(C, B\right) = \mathrm{E}\left[A \mid C, B\right]$. Queremos probar que para toda h medible vale que $\mathrm{E}\left\{\left(A - g\left(B\right)\right) h\left(B\right)\right\} = 0$.

$$\begin{aligned}
\mathrm{E}\left\{g\left(B\right) h\left(B\right)\right\} &= \mathrm{E}\left\{\mathrm{E}\left(w\left(C, B\right) \mid B\right) h\left(B\right)\right\} \\
&= \mathrm{E}\left\{\mathrm{E}\left(w\left(C, B\right) h\left(B\right) \mid B\right)\right\} \\
&= \mathrm{E}\left\{w\left(C, B\right) h\left(B\right)\right\} \\
&= \mathrm{E}\left\{\mathrm{E}\left[A \mid C, B\right] h\left(B\right)\right\} \\
&= \mathrm{E}\left\{\mathrm{E}\left[Ah\left(B\right) \mid C, B\right]\right\} \\
&= \mathrm{E}\left\{Ah\left(B\right)\right\}. \quad \blacksquare
\end{aligned}$$

Proposición 6.2 *Sean X, T, variables (o vectores) aleatorias independientes y g, h, k funciones medibles. Entonces*

$$\mathrm{E}\left[g\left(X\right) \mid h\left(X\right), k\left(h\left(X\right), T\right)\right] = \mathrm{E}\left[g\left(X\right) \mid h\left(X\right)\right].$$

Demostración. Llamando U al miembro derecho de la igualdad y V al miembro izquierdo, por definición de esperanza condicional, para toda función medible p se tiene que

$$\mathrm{E}\left[Up\left(h\left(X\right)\right)\right] = \mathrm{E}\left[g\left(X\right) p\left(h\left(X\right)\right)\right]. \tag{53}$$

Para probar la igualdad, como U es función de $h\left(X\right)$, debemos ver que para toda función medible q vale

$$\mathrm{E}\left[g\left(X\right) q\left(h\left(X\right), k\left(h\left(X\right), T\right)\right)\right] = \mathrm{E}\left[Uq\left(h\left(X\right), k\left(h\left(X\right), T\right)\right)\right].$$

Llamemos

$$s(u) = \mathrm{E}\left(q\left(u, k(u, T)\right)\right).$$

Luego, por la independencia de X y T, se tiene

$$s\left(h(X)\right) = \mathrm{E}\left[q\left(h(X), k(h(X), T)\right) | X\right]. \tag{54}$$

Condicionando respecto de X y usando (54) se tiene

$$\begin{aligned}
\mathrm{E}(g(X)q\left(h(X), k(h(X), T)\right)) &= \mathrm{E}(\mathrm{E}\left[g(X)q\left(h(X), k(h(X), T)\right) | X\right]) \\
&= \mathrm{E}(g(X)\mathrm{E}\left[q\left(h(X), k(h(X), T)\right) | X\right]) \\
&= \mathrm{E}\left(g(X)s(h(X))\right) \\
&= \mathrm{E}(Us(h(X))). \qquad (55)
\end{aligned}$$

Como U es función de X, usando (53) también se tiene

$$\begin{aligned}
\mathrm{E}\left(Us(h(X))\right) &= \mathrm{E}(UE(q(h(X), k(h(X), T)|X)) \\
&= \mathrm{E}(E(Uq(h(X), k(h(X), T)|X)) \qquad (56) \\
&= \mathrm{E}(Uq(h(X), k(h(X), T))). \qquad (57)
\end{aligned}$$

A partir de (55) y de (57) se deduce la igualdad buscada. $\blacksquare$

6.3. Diferenciación vectorial

Proposición 6.3 *Sea* $\mathbf{x} \in \mathbb{R}^p$*,si* $\partial/\partial\mathbf{x}$ *denota el vector que en la i-ésima componente tiene a* $\partial/\partial\mathbf{x}_i$*,* $\mathbf{a} \in \mathbb{R}^p$ *y* $A \in \mathbb{R}^{p\times p}$ *es una matriz simétrica, tenemos que*

1. *Si* $f(\mathbf{x}) = \mathbf{a}'\mathbf{x}$*, entonces* $\frac{\partial}{\partial\mathbf{x}} f(\mathbf{x}) = \mathbf{a}$.
2. *Si* $f(\mathbf{x}) = \mathbf{x}'A\mathbf{x}$*, entonces* $\frac{\partial}{\partial\mathbf{x}} f(\mathbf{x}) = 2A\mathbf{x}$.

Demostración. Ver Seber [21], Apéndice A8. ■

Proposición 6.4 *Sean las matrices* $X \in \mathbb{R}^{m\times n}$*,* $A \in \mathbb{R}^{l\times m}$*,* $B \in \mathbb{R}^{n\times q}$*, asumimos que* A, B *son independientes de* $X = (x_{rs})$*, y sea* $\mathbf{e}_{i,n}$ *el vector canónico en* $\mathbb{R}^n$ *que tiene un uno en la coordenada i-ésima, y vale cero en las demás. Entonces*

i. $\frac{\partial}{\partial x_{rs}}[AXB] = A\mathbf{e}_{r,m}\mathbf{e}'_{s,n}B$.

ii. Si la matriz $X \in \mathbb{R}^{m\times m}$ *es no singular, entonces*

$$\frac{\partial}{\partial x_{rs}}\left[AX^{-1}B\right] = -AX^{-1}\mathbf{e}_{r,m}\mathbf{e}'_{s,m}X^{-1}B.$$

iii. Si la matriz $X \in \mathbb{R}^{m\times m}$ *es no singular,* $\mathbf{y} \in \mathbb{R}^m$*, entonces*

$$\frac{\partial}{\partial X}\left[\mathbf{y}'X^{-1}\mathbf{y}\right] = -X^{-1}\mathbf{y}\mathbf{y}'X^{-1}.$$

Demostración.

i. y ii. Para los dos primeros ítems, ver [12] Sección 4.5.

iii. Por el ítem anterior

$$\begin{aligned}\frac{\partial}{\partial x_{rs}}\left[\mathbf{y}'X^{-1}\mathbf{y}\right] &= -\mathbf{y}'X^{-1}\mathbf{e}_{r,m}\mathbf{e}'_{s,m}X^{-1}\mathbf{y}\\ &= traza\left(-\mathbf{y}'X^{-1}\mathbf{e}_{r,m}\mathbf{e}'_{s,m}X^{-1}\mathbf{y}\right)\\ &= -traza\left(\mathbf{e}_{r,m}\mathbf{e}'_{s,m}X^{-1}\mathbf{y}\mathbf{y}'X^{-1}\right).\end{aligned}$$

La matriz $\mathbf{e}_{r,m}\mathbf{e}'_{s,m}X^{-1}\mathbf{y}\mathbf{y}'X^{-1}$ es nula salvo en la fila r que contiene a la fila s de la matriz $X^{-1}\mathbf{y}\mathbf{y}'X^{-1}$. Luego, su traza será el elemento rr de la matriz resultante, es decir

$$\begin{aligned}\frac{\partial}{\partial x_{rs}}\left[\mathbf{y}'X^{-1}\mathbf{y}\right] &= -traza\left(\mathbf{e}_{r,m}\mathbf{e}'_{s,m}X^{-1}\mathbf{y}\mathbf{y}'X^{-1}\right)\\ &= -\left(X^{-1}\mathbf{y}\mathbf{y}'X^{-1}\right)_{sr},\end{aligned}$$

entonces

$$\frac{\partial}{\partial X}\left[\mathbf{y}'X^{-1}\mathbf{y}\right] = -X^{-1}\mathbf{y}\mathbf{y}'X^{-1}. \quad ■$$

Referencias

[1] Aragon, Y., y Saracco, J. (1997). Sliced Inverse Regression (SIR): An Appraisal of Small Sample Alternatives to Slicing. *Comput. Statist.,* **12**, 109-130.

[2] Brillinger, D. (1991). Comment on "Sliced Inverse Regression for Dimension Reduction" by K. C. Li. *Journal of the American Statistical Association,* **86**, 333.

[3] Bura, E. y Cook R. D. (2001). Estimating the Structural Dimension of Regressions via Parametric Inverse Regression. *J. Roy. Statist. Soc.* Ser. B, **63**, 393-410.

[4] Carroll, R. J., Li, K. C. (1992). Measurement Error Regression With Unknown Link: Dimension Reduction and Data Visualization. *J. Amer. Statist. Assoc.,* **87**, 1040-1050.

[5] Chen, C. H., Li, K. C. (1998) Can SIR be as Popular as Multiple Linear Regression? *Statist. Sinica,* **8**, 289-316.

[6] Cook, R. D. y Weisberg, S. (1991) Sliced Inverse Regression for Dimension Reduction: Comment. *Journal of the American Statistical Association,* **86**, 328 - 332.

[7] Eaton, M. L., y Pearlman, M. D. (1973). The non-singularity of generalized sample covariance matrices. *Ann. Statist.,* **1,** 710-717.

[8] Gannoun, A., y Saracco, J. (2003) An Asymptotic theory for Sir-α Method. *Statistica Sinica,* **13**, 297-310.

[9] García Ben, M., Martínez, E. J. y Yohai, V. J. (2004) Robust and Efficient Estimates for Multivariate Lineal Models, trabajo inédito.

[10] Gather, U., Hilker, T. y Becker, C. (2001) A robustified version of sliced inverse regression. *Statistics in Genetics and in the Environmental Sciences*, Proceedings of the Workshop on Statistical Methodology for the Sciences: Environmetrics and Genetics held in Ascona from May 23 to 28, Eds. L.T. Fernholz, S. Morgenthaler, W. Stahel, 147-157.

[11] Gather, U., Hilker, T. y Becker, C. (2002) A note on outlier sensitivity of Sliced Inverse Regression, *Statistics*, **13**(4), 271-281.

[12] Graham, A. *Kronecker Products and Matrix Calculus with Applications.* (1981). Elis Horwood Series.

[13] Härdle, W. y Tsybakov, A. B. (1991). Comment on "Sliced Inverse Regression for Dimension Reduction" by K. C. Li. *Journal of the American Statistical Association,* **86**, 333-335.

[14] Hall, P. y Li, K.C. (1993). On Almost Linearity of Low Dimensional Projections from High Dimensional Data. *Ann. Statist.,* **21**, 867-889.

[15] Hsing, T. y Carrol, R.J. (1992). An Asymptotic Theory for Sliced Inverse Regression. *Ann Statist.,* **20**, 1040-1061.

[16] Kent, J. (1991). Comment on "Sliced Inverse Regression for Dimension Reduction" by K. C. Li. *Journal of the American Statistical Association,* **86**, 336-337.

[17] Kötter, T. (1996). An Asymptotic Result for Sliced Inverse Regression. Comput. Statist. 11, 113-136.

[18] Li, K. C. (1991). Sliced Inverse Regression for Dimension Reduction, *Journal of the American Statistical Association,* **86**, 316-327.

[19] Saracco, J. (1997). An Asymptotic Theory for Sliced Inverse Regression. *Comm. Statist.-Theory Methods,* **26**, 2141-2171.

[20] Schott, J. R. (1994). Determining the Dimensionality in Sliced Inverse Regression. *Journal of the American Statistical Association*, **89**, 141-148.

[21] Seber, G. A. F. (1984). *Multivariate Analysis.* Wiley: New York.

[22] Zhu, L. X. y Fang, K. T. (1996). Asymptotics for Kernel Estimate of Sliced Inverse Regression. *Ann Statist.*, **24**, 1053-1068.