UNIVERSIDAD NACIONAL DEL LITORAL
FACULTAD DE INGENIERÍA QUÍMICA

Tesis presentada como parte de los requisitos de la Universidad Nacional del Litoral para la obtención del Grado Académico de

Doctor en Matemática

En el campo de: **Estadística**

Título de la tesis:

Modelos lineales generalizados: regresión de rango reducido y reducción suficiente de dimensiones.

Institución donde se realizó:
Instituto de Matemática Aplicada del Litoral
(CONICET-UNL)

Autor:
Sabrina Lorena Duarte

Directores de Tesis:
Dra. Liliana Forzani y Dra. Efstathia Bura

Defendida ante el jurado compuesto por:
Dr. Jorge Adrover
Dr. Mariela Sued
Dr. Víctor Yohai

Año de presentación: 2016

A mamá, papá, Bruno y Mauro.

AGRADECIMIENTOS

En estas lineas quiero expresar mi más sincero agradecimiento a todas las personas que con su apoyo y su ayuda desinteresada hicieron posible la realización de esta tesis.

A mis directores Lili y Effie, por su enseñanza, por apoyarme constantemente y permitirme aprender de sus conocimientos brindándome su tiempo de manera generosa. Gracias por toda la paciencia que me han tenido durante estos años.

A mis padres, por todo el esfuerzo que hicieron para que pueda realizar una carrera universitaria, por su confianza y su apoyo incondicional...Gracias infinitas mamá y papá.

A Fernando, por muchas cosas, pero principalemte por el aguante, por comprenderme y acompañarme durante todos estos años.

A mis amigas, por hacer más linda la vida y por cada vez que renovaron mis energías y fueron mi cable a tierra cuando estaba trabajando arduamente en la tesis. Ana Laura, Evelyn, Andrea, Gise Romi, Analia, Roxi, Belén, Marina, Luz, Flor y Xime...Las adoro.

Al CONICET por haber apoyado económicamente la realización de esta tesis. Gracias a todo el IMAL por tanta calidez y contención personal y profesional. Gracias a mis compañeros del IMAL, por crear siempre ese agradable ambiente de trabajo y estudio...especialmente, a Jose y Estefi, por esos primeros años de doctorado donde hemos compartido buenos mates, lindas charlas y valiosos consejos...

A todos los profesores quienes durante estos diez años de carrera matemática me han enseñado todo lo que sé, gracias por compartir toda su sabiduría conmingo.

A los miembros del jurado, por haber aceptado el compromiso de evaluar esta tesis y a aquellos que han tenido la tarea de leerla, gracias por sus valiosos comentarios y correcciones.

Finalmente, quiero agradecer al grupo de estádistica con el que trabajamos diariamente. Son un grupo humano hermoso, con el cual he aprendido y sigo aprendiendo mucha estadística y mucho compañerismo. En especial, quiero agradecer a Pame y Diego, por su gran ayuda con el Latex y el Matlab.

ÍNDICE GENERAL

RESUMEN

La respuesta a muchos de los problemas de interés en ciencias experimentales requieren el estudio de una o varias variables $\mathbf{Y}$ (respuesta/s) en función de otras variables $\mathbf{X}$ (predictores). Desde el punto de vista de la estadística, esto significa estudiar la distribución condicional de un vector $\mathbf{Y} \in \mathbb{R}^r$, dado el vector $\mathbf{X} \in \mathbb{R}^p$. Cuando el número de predictores p es grande, casi todos los métodos usados para estudiar esta relación incluye algún tipo de redución en la dimensión de $\mathbf{X}$. Componentes principales es el método de reducción más popular entre las ciencias aplicadas. Más recientes son los métodos estadísticos de reducción establecidos bajo el paradigma de reducción suficiente de dimensiones. Su premisa es la obtención de una redución de los predictores $R(\mathbf{X}) \in \mathbb{R}^d$ con $d \leq p$ sin que ésta pierda información acerca de la respuesta en el sentido que $\mathbf{Y}|\mathbf{X} \sim \mathbf{Y}|R(\mathbf{X})$.

Reducción suficiente de dimensiones es un área muy actual de la estadística. En esta tesis trabajamos bajo el enfoque inicializado en **[Cook, 2007]**, el cual está basado en la suposición de un modelo parámetrico para la regresión inversa $\mathbf{X}|\mathbf{Y}$. El atractivo de este enfoque es que cuando la respuesta es univariada, $\mathbf{X}|Y$ consta de p regresiones univariadas las cuales son sencillas de modelar, contrariamente a lo que ocurre cuando se modela Y dado $\mathbf{X}$. Bajo este enfoque se han estudiado reducciones suficientes para la regresión de $\mathbf{Y}$ en $\mathbf{X}$ asumiendo diferentes modelos para la distribución de $\mathbf{X}|Y$. En especial, se ha estudiado en detalle los modelos $\mathbf{X}|(Y = y) \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Delta})$ **[Cook and Forzani, 2008]**, $\mathbf{X}|(Y = y) \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Delta}_y)$ **[Cook and Forzani, 2009]** y $\mathbf{X}|(Y = y)$ con distribución perteneciente a una familia exponencial a p parámetros naturales con predictores condicionalmente independientes **[Cook and Li, 2009]**. Este último modelo ha permitido estudiar aquellos problemas o conjuntos de datos que contienen variables predictoras de tipo discreto o mezcla de variables discretas y continuas, aunque el supuesto de independencia condicional es muy restrictivo.

El objetivo de esta tesis es desarrollar una metodología de reducción suficiente de dimensiones asumiendo que la distribución de $\mathbf{X}|Y$ pertenece a una familia exponencial a k parámentos

naturales con posiblemente $k \geq p$. Para este modelo identificamos la reducción suficiente minimal, obtenemos estimadores de máxima verosimilitud para dicha reducción, estudiamos las distribuciones asintóticas de las reducciones y presentamos test de dimensiones para el valor d. Además, mostramos ejemplos y simulaciones para ilustrar las conexiones, diferencias y ventajas de nuestro método con los ya existentes.

Para poder desarrollar este trabajo es necesario, en primer lugar, estudiar en detalle los modelos lineales generalizados multivariados, es decir modelos donde la respuesta multivariada, dado los predictores, pertenecen a una familia de exponenciales. En particular, es necesario completar el trabajo **Yee and Hastie, 2003** que adapta la idea del modelo de regresión lineal de rango reducido a este contexto. De esta forma, cuando la matriz de coeficientes de la regresión no es de rango completo, es posible obtener estimadores de la regresión más eficientes. Sin embargo, para poder aplicar estos resultados a nuestro contexto de reducción suficiente debemos probar que los estimadores propuestos por **Yee and Hastie, 2003** son asintóticamente normales y encontrar su varianza asintótica.

INTRODUCCIÓN

En el contexto de regresión lineal multivariada, los modelos de rango reducido han sido ampliamente estudiados en la literatura estadística. Estos modelos permiten obtener estimadores más eficientes en los casos donde es factible asumir que la matriz de coeficientes no es de rango completo. Esto generalmente ocurre cuando el número de respuestas y predictores es grande o hay presencia de correlación entre las variables respuesta. Sin embargo, resultados asintóticos, sus utilidades y aplicaciones han sido limitadas a las familias gaussianas.

En la primera parte de la tesis estudiamos los modelos lineales generalizados multivariados de rango reducido propuestos en [**Yee and Hastie, 2003**]. El objetivo principal de estos modelos es permitir que los potenciales beneficios de la regresión de rango reducido puedan ser transportados a una amplia clase de modelos, en particular a una amplia gama de tipos de respuesta $\mathbf{Y}$ en la regresión.

En el Capítulo 1 introducimos los modelos lineales generalizados multivariados, presentamos ejemplos de familias de distribuciones que modelan la estructura de los errores de estos modelos, estimadores de máxima verosimilitud de los parámetros del modelo y su distribución asintótica. Luego, en los Capítulos 2 y 3 abordamos la teoría de rango reducido.

Por un lado, en el Capítulo 2 exponemos cómo se aplican estas ideas a los modelos de regresión lineal multivariado clásicos. Presentamos una forma novedosa de obtener la distribución asintótica de los estimadores de máxima verosimilitud propuesta en [**Cook et al., 2015**] y además proponemos nuevos estimadores de rango reducido, sus varianzas asintóticas y la relación entre ellas.

Motivados por los resultados asintóticos obtenidos en el caso que los errores tengan una distribución normal, el objetivo del Capítulo 3 es extender los mismos a los modelos lineales generalizados de rango reducido. De esta forma, completamos los resultados expuestos en [**Yee and Hastie, 2003**] obteniendo la distribución asintótica de los estimadores de máxima verosimilitud presentados por estos autores. Además, este resultado nos permite obtener intervalos de confianza asintóticos para los parámetros y así complementar los ejemplos dados en

dicho trabajo. Finalmente, adaptamos al contexto de modelos lineales generalizados los demás estimadores de rango reducido propuestos en el Capítulo 2 para el caso del modelo lineal normal, obtenemos su varianza asintótitca y la relación entre ellas.

La segunda parte de la tesis está dedicada al estudio de reducción suficiente de dimensiones (SDR) para la regresión de la respuesta $\mathbf{Y} \in \mathbb{R}^r$ en los predictores $\mathbf{X} \in \mathbb{R}^p$. El objetivo principal de esta metodología es obtener una reducción $R(\mathbf{X}) \in \mathbb{R}^d$ con $d \leq p$ tal que $R(\mathbf{X})$ no pierda información acerca de la respuesta en el sentido que $\mathbf{Y}|\mathbf{X} \sim \mathbf{Y}|R(\mathbf{X})$ **[Cook, 2007]**. La mayoría de las metodologías SDR están basadas en el enfoque de regresión inversa de $\mathbf{X}|\mathbf{Y}$. Este abordaje tiene la ventaja de que en el caso que la respuesta sea univariada, la regresión de $\mathbf{X}|Y$ consisten en p regresiones univariadas que, por medio de gráficos, es mas simple de modelar.

Existen dos tipo de enfoques en las metodologías SDR, por un lado aquellos métodos basados en momentos de $\mathbf{X}$ (como SIR **[Li, 1991]**, SAVE **[Cook and Weisberg, 1991]** y Sir Parcial **[Chiaromonte et al., 2002]**, entre otros) y por otro lado aquellos basados en modelos para la regresión inversa $\mathbf{X}|\mathbf{Y}$, lo cuales surgieron para superar las limitaciones de los primeros (**[Cook, 2007]**, **[Cook and Forzani, 2008]**, **[Cook and Forzani, 2009]**, entre otros). Todos estos métodos obtienen solamente reducciones lineales, es decir, $R(\mathbf{X}) = \boldsymbol{\alpha}^T\mathbf{X}$ con $\boldsymbol{\alpha} \in \mathbb{R}^{p\times d}$ y exceptuando Sir Parcial, todos requieren que los predictores $\mathbf{X}$ sean de tipo continuo. En el Capítulo 5 exponemos conceptos básicos y los métodos conocidos.

En el Capítulo 6 presentamos un nuevo método de reducción suficiente de dimensiones que permite identificar reducciones suficientes en regresiones con predictores que pueden ser todos continuos, todos categóricos o una mezcla de variables continuas o categóricas. Nos basamos en el enfoque basado en modelos y asumimos que la distribución de $\mathbf{X}|Y$ pertenece a una familia exponencial con k parámetros, generalizando el trabajo de **[Cook and Li, 2009]**. Identificamos la reducción suficiente minimal (es decir, la reducción más pequeña) para la regresión de Y dado $\mathbf{X}$ y mostramos que ésta no es necesariamente lineal en los predictores pero sí en $\mathbf{T}(\mathbf{X})$, el estadístico suficiente y minimal de la familia. En algunos casos, por ejemplo cuando $\mathbf{X}|Y$ está distribuido como Bernoulli multivariada, la reducción minimal suficiente es no lineal en los predictores.

En este contexto, proponemos dos enfoques para obtener estimadores de la reducción suficiente minimal. Por un lado, como contamos con un modelo para la regresión inversa de $\mathbf{X}|\mathbf{Y}$, es posible obtener estimadores de máxima verosimilitud. Estos se logran ajustando modelos lineales generalizados de rango reducido, estudiados en detalle en el Capítulo 3. Por otro lado, basándonos en los estimadores presentados en Capítulo 3, es posible obtener un segundo

estimador de la reducción suficiente minimal. Este tiene la ventaja de ser obtenido a través de cálculos sencillos, pero es menos eficiente que el primero.

Finalmente, presentamos ejemplos y simulaciones para ilustrar las conexiones, diferencias y ventajas de nuestro método con los presentados en el Capítulo 5 y estudiamos en detalle el caso en que $\mathbf{X}|Y$ es un vector Bernoulli multivariado.

NOTACIONES Y DEFINICIONES

Reproduciremos en este apartado las notaciones y definiciones que serán utilizadas a lo largo de esta tesis con el objetivo que la lectura resulte fluida.

Álgebra

Notación. Utilizaremos la notación $\mathbb{R}^{m\times n}$ para indicar el conjunto de todas las matrices reales de dimensión $m \times n$. La traspuesta de una matriz o de un vector $\mathbf{z}$ se denota como $\mathbf{z}^T$. Para $\mathbf{M} \in \mathbb{R}^{m\times n}$, span($\mathbf{M}$) $\subset \mathbb{R}^m$ es el subespacio generado por los vectores columnas de $\mathbf{M}$.

Definición. Sea $\mathbf{M}$ una matriz simétrica de dimensión $n \times n$. Esta matriz $\mathbf{M}$ se dice semi-definida positiva, y lo denotamos como $\mathbf{M} \geq 0$, si cumple que para todos los vectores no nulos $\mathbf{z} \in \mathbb{R}^n$

$$\mathbf{z}^T\mathbf{M}\mathbf{z} \geq 0.$$

Si la ecuación anterior satisface la desigualdad de forma estricta, es decir $\mathbf{z}^T\mathbf{M}\mathbf{z} > 0$ para todo vector no nulo $\mathbf{z} \in \mathbb{R}^n$, se dice que $\mathbf{M}$ es definida positiva. Para dos matrices $\mathbf{A}$ y $\mathbf{B}$ simétricas de igual dimensiones, se dice que $\mathbf{A} \geq \mathbf{B}$ (o $\mathbf{A} > \mathbf{B}$) si y sólo si $\mathbf{A} - \mathbf{B} \geq 0$ (o $\mathbf{A} - \mathbf{B} > 0$).

Definición. El rango de las columnas de una matriz $\mathbf{M} \in \mathbb{R}^{m\times n}$ es el número máximo de columnas linealmente independientes que esta contiene. De forma análoga, se define el rango de las filas de $\mathbf{M}$ como el número máximo de filas linealmente independientes. Para cualquier matriz $\mathbf{M}$ el rango de las filas es igual al rango de las columnas y este número es llamado el rango de $\mathbf{M}$ y denotado por rank($\mathbf{M}$).

Notación. Denotamos los siguientes conjuntos abiertos de matrices como

$$\begin{aligned} \Pi^{m\times n} &= \{\mathbf{M} \in \mathbb{R}^{m\times n}, \text{ tal que } \mathbf{M} \text{ es de rango completo}\} \text{ y} \\ S^m &= \{\mathbf{M} \in \mathbb{R}^{m\times m}, \text{ tal que } \mathbf{M} \text{ es simétrica}\}. \\ S^m_+ &= \{\mathbf{M} \in \mathbb{R}^{m\times m}, \text{ tal que } \mathbf{M} \text{ es simétrica y definida positiva}\}. \end{aligned}$$

Definición. La transformación lineal vec : $\mathbb{R}^{m\times n} \to \mathbb{R}^{mn}$ vectoriza cualquier matriz apilando sus columnas. A su vez, vech : $S^m \to \mathbb{R}^{m(m+1)/2}$ vectoriza las matrices simétricas considerando de sus columnas, los elementos de la diagonal y debajo de la diagonal. El producto kronecker de las matrices $\mathbf{A} : m \times n$ y $\mathbf{B} : p \times q$, denotado con $\mathbf{A} \otimes \mathbf{B}$, es la matriz $mp \times nq$

$$\begin{pmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{pmatrix}.$$

El uso de las inversas generalizadas de las matrices aparece constantemente en aplicaciones de la estadística. Además es conocido el hecho de que existen distintos tipo de matrices inversas generalizadas dependiendo de las condiciones exigidas. La siguiente es la definición clásica:

Definición. Sea $\mathbf{M}$ una matriz de dimensión $m \times n$. Se dice que una matriz $\mathbf{M}^{\ddagger} : n \times m$ es una inversa generalizada si verifica que

$$\mathbf{M}\mathbf{M}^{\ddagger}\mathbf{M} = \mathbf{M}.$$

De la defición es claro que en el caso que una matriz sea invertible, su inversa generalizada es la propia inversa y es única. Una de las clases de inversa generalizada que se destaca es la inversa generalizada de Moore-Penrose que, entre otras, goza de la propiedad de unicidad. Dicha clase se encuentra bajo la siguiente definición:

Definición. Sea $\mathbf{M}$ una matriz de dimensión $m \times n$. Se dice que la matriz $\mathbf{M}^{\dagger} : n \times m$ es la inversa generalizada de Moore-Penrose si verifica que

(a) $\mathbf{M}\mathbf{M}^{\dagger}\mathbf{M} = \mathbf{M}$

(b) $\mathbf{M}^{\dagger}\mathbf{M}\mathbf{M}^{\dagger} = \mathbf{M}^{\dagger}$

(c) $(\mathbf{M}^{\dagger}\mathbf{M})^T = \mathbf{M}^{\dagger}\mathbf{M}$

(d) $(\mathbf{M}\mathbf{M}^{\dagger})^T = \mathbf{M}\mathbf{M}^{\dagger}$.

Otras matrices que se utilizan en este contexto son las matrices de permutación $\mathbf{K}_{mn}$, de expansión $\mathbf{E}_m$ y de duplicación $\mathbf{D}_m$ (ver por ejemplo, [**Magnus, 1988**]):

Definición. Sea $\mathbf{M} : m \times n$ una matriz, se define $\mathbf{K}_{mn}$ a la única matriz de permutación de dimensiones $mn \times mn$ tal que $\text{vec}(\mathbf{M}^T) = \mathbf{K}_{mn}\text{vec}(\mathbf{M})$.

Definición. La matriz de duplicación $\mathbf{D}_m$ es la matriz de dimensiones $m^2 \times \frac{1}{2}m(m+1)$ con la propiedad que

$$\text{vec}(\mathbf{M}) = \mathbf{D}_m \text{vech}(\mathbf{M})$$

para cualquier matriz $\mathbf{M} : m \times m$ simétrica.

Una propiedad importante de esta matriz es que su inversa generalizada de Moore-Penrose $\mathbf{D}_m^\dagger \doteq \mathbf{E}_m$ cumple con la propiedad que

$$\mathbf{E}_m \text{vec}(\mathbf{M}) = \text{vech}(\mathbf{M})$$

para cualquier matriz $\mathbf{M} : m \times m$ simétrica.

Definición. Sea la matriz $\mathbf{A} \in \mathbb{R}^{m\times n}$ de rango completo y $\mathbf{V} \in \mathbb{R}^{m\times m}$ simétrica y definida positiva. La matriz que caracteriza la transformación lineal que proyecta $\mathbb{R}^m$ sobre span($\mathbf{A}$) con respecto al producto interno definido por $\mathbf{V}$, la llamamos matriz de proyección y su forma es la siguiente:

$$\mathbf{P}_{\mathbf{A}(\mathbf{V})} = \mathbf{A}(\mathbf{A}^T\mathbf{V}\mathbf{A})^\ddagger \mathbf{A}^T\mathbf{V}. \tag{0.1}$$

En particular, $\mathbf{P}_\mathbf{A}$ denota la proyección sobre span($\mathbf{A}$) con respecto al producto interno estandar. Además, $\mathbf{Q}_{\mathbf{A}(\mathbf{V})} = \mathbf{I} - \mathbf{P}_{\mathbf{A}(\mathbf{V})}$ es la matriz de proyección en el complemento ortogonal del subespacio span($\mathbf{A}$).

Notar que en la ecuación (0.1) consideramos una inversa generalizada de $\mathbf{A}^T\mathbf{V}\mathbf{A}$ pues $\mathbf{A}$ puede ser de rango no completo, por lo que $\mathbf{A}^T\mathbf{V}\mathbf{A}$ no sería inversible. Como (0.1) no depende de que inversa generalizada se use (ver Lema A.10), dicha proyección está bien definida.

Estadística

Notación. Si $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converge a una distribución normal con media 0 y matriz de covarianza $\mathbf{V}$, escribimos su matriz de covarianza asintótica como avar$(\sqrt{n}\widehat{\boldsymbol{\theta}}) = \mathbf{V}$.

Notación. Dada las muestras $\mathbf{x}_1, \ldots, \mathbf{x}_n$ e $\mathbf{y}_1, \ldots, \mathbf{y}_n$ de los vectores aleatorios $\mathbf{X} \in \mathbb{R}^p$ e $\mathbf{Y} \in \mathbb{R}^r$ respectivamente, denotamos

- $\overline{\mathbf{x}} \doteq \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \mathbf{x}_i$: vector promedio para la muestra $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

- $S_{\mathbf{x}} \doteq \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$: matriz de covarianza muestral de $p \times p$ de $\mathbf{x}_1, \ldots, \mathbf{x}_n$. De la misma forma $S_{\mathbf{y}} \doteq \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})^T$ es la matriz de covarianza muestral de $r \times r$ de $\mathbf{y}_1, \ldots, \mathbf{y}_n$.
- $S_{\mathbf{xy}} \doteq \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{y}_i - \overline{\mathbf{y}})^T$: matriz de covarianza muestral de $p \times r$ de $\mathbf{x}_1, \ldots, \mathbf{x}_n$ e $\mathbf{y}_1, \ldots, \mathbf{y}_n$.
- $S_{\mathbf{yx}} = S_{\mathbf{xy}}^T$.
- $S_{\mathbf{y}|\mathbf{x}} \doteq S_{\mathbf{y}} - S_{\mathbf{yx}}S_{\mathbf{x}}^{-1}S_{\mathbf{xy}}$: matriz de covarianza muestral de $r \times r$ correspondiente al ajuste de $\mathbf{Y}$ en $\mathbf{X}$, también llamada matriz de residuos de la regresión lineal de $\mathbf{Y}$ en $\mathbf{X}$.
- $C_{\mathbf{yx}} \doteq S_{\mathbf{y}}^{-1/2}S_{\mathbf{yx}}S_{\mathbf{x}}^{-1/2}$: matriz de correlaciones canónicas muestrales de $r \times p$.
- $C_{\mathbf{xy}} \doteq C_{\mathbf{yx}}^T$.

CAPÍTULO 1

REGRESIÓN LINEAL MULTIVARIADA Y MODELO LINEAL GENERALIZADO MULTIVARIADO

El modelo de regresión lineal multivariado clásico se basa en que la variable respuesta $\mathbf{Y}$ se relaciona de forma lineal con las variables predictoras $\mathbf{X}$ y que los errores condicionados en $\mathbf{X}$ se distribuyen normalmente con media cero y varianza constante. En muchas ocasiones, sin embargo, nos encontramos con que uno o varios de estos supuestos no se cumplen. Por ejemplo, es muy común que con ciertos tipos de conjuntos de datos, a medida que aumenta la media muestral aumente también su varianza observada. Estos problemas se pueden llegar a solucionar mediante la transformación de la variable respuesta. Sin embargo estas transformaciones no siempre consiguen corregir la falta de normalidad, la heterocedasticidad (varianza no constante) o la no linealidad de los datos. Además, muchas veces, resulta difícil interpretar los resultados obtenidos. Una alternativa a la transformación de la variable respuesta y a la falta de normalidad, es el uso de los modelos lineales generalizados. Los modelos lineales generalizados (GLM de las siglas en inglés de Generalized Linear Models) son una extensión de los modelos lineales que se pueden aplicar cuando los errores no son normales (binomial, Poisson, multinomial, entre otros). Estos modelos fueron propuestos en **Nelder and Wedderburn, 1972** y estudiados en trabajos como **Lindsey, 1997**, **McCulloch and Nelder, 2001** y **McCulloch and Searle, 2001** cuando la respuesta $\mathbf{Y}$ es univariada. Ciertos tipos de variables respuesta sufren indefectiblemente la violación de estos supuestos de los modelos gaussianos y los GLM ofrecen una alternativa para tratarlos. Específicamente, podemos considerar utilizar GLM cuando la variable respuesta es, por ejemplo, un conteo de casos (abundancia de una especie, número de embarazos, cantidad de llamadas telefónicas recibidas, etc.), una respuesta binaria (vivo o muerto, infectado o

no infectado, hombre o mujer, etc.) o una mezcla de predictores continuos y categóricos (por ejemplo, el peso, la altura y el sexo de las personas).

En este capítulo se presenta el modelo de regresión lineal multivariado clásico y los modelos lineales generalizados multivariados. Se expondrán conceptos y resultados conocidos que sevirán para comprender los próximos capítulos.

1.1. Modelo clásico de regresión lineal multivariado gaussiano

El modelo de regresión lineal multivariado para los predictores $\mathbf{X} \in \mathbb{R}^p$ y las respuestas $\mathbf{Y} \in \mathbb{R}^r$ asume que $\mathbf{Y}|\mathbf{X} = \mathbf{x} \sim \mathcal{N}(\boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{x}, \boldsymbol{\Sigma})$ lo cual permite expresar la variable respuesta $\mathbf{Y}$ como

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon} \tag{1.1}$$

con los errores $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ e independiente de $\mathbf{X}$ ([**Anderson, 2003**], [**Muirhead, 1982**], [**Eaton, 2007**], [**Srivastava and Khatri, 1979**]). El objetivo de este modelo es predecir $\mathbf{Y}$ en función de $\mathbf{X}$ a través de la matriz de coeficientes $\boldsymbol{\beta}$ de dimensiones $r \times p$.

Supongamos que contamos con n observaciones independientes $\mathbf{y}_1, \ldots, \mathbf{y}_n$ del vector $\mathbf{Y}$ y $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sus respectivos valores de $\mathbf{X}$. Sin pérdida de generalidad, asumimos que la muestra de predictores está centrada, de modo que $\overline{\mathbf{x}} = 0$. Los estimadores de máxima verosimilitud o mínimos cuadrados obtenidos bajo este modelo clásico de regresión lineal multivariada lo denotamos como $\widehat{\boldsymbol{\alpha}}_{OLS}$, $\widehat{\boldsymbol{\beta}}_{OLS}$ y $\widehat{\boldsymbol{\Sigma}}_{OLS}$ y están dados por:

$$\begin{aligned} \widehat{\boldsymbol{\alpha}}_{OLS} &= \overline{\mathbf{y}} \\ \widehat{\boldsymbol{\beta}}_{OLS} &= S_{\mathbf{yx}} S_{\mathbf{x}}^{-1} \\ \widehat{\boldsymbol{\Sigma}}_{OLS} &= S_{\mathbf{y}} - \widehat{\boldsymbol{\beta}}_{OLS} S_{\mathbf{xy}} = S_{\mathbf{y}} - S_{\mathbf{yx}} S_{\mathbf{x}}^{-1} S_{\mathbf{xy}} = S_{\mathbf{y}|\mathbf{x}}. \end{aligned} \tag{1.2}$$

El vector $\left(\text{vec}^T\left(\widehat{\boldsymbol{\beta}}_{OLS}\right), \text{vech}^T\left(\widehat{\boldsymbol{\Sigma}}_{OLS}\right)\right)^T$ es asintóticamente normal con la siguiente matriz de covarianza (ver [**Magnus and Neudecker, 1999**]):

$$\text{avar}\left(\sqrt{n}\begin{pmatrix} \text{vec}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) \\ \text{vech}\left(\widehat{\boldsymbol{\Sigma}}_{OLS}\right) \end{pmatrix}\right) = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma} & 0 \\ 0 & 2\mathbf{E}_r(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})\mathbf{E}_r^T \end{pmatrix}, \tag{1.3}$$

donde $\boldsymbol{\Sigma}_{\mathbf{X}} = \text{var}(\mathbf{X})$ en el caso que $\mathbf{X}$ sea aleatorio y en el caso que se condicione en $\mathbf{X}$ o sea fijo, $\boldsymbol{\Sigma}_{\mathbf{X}} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$ donde se asume que dicho límite existe.

1.2. Familias exponenciales multivariadas

En esta sección introducimos las familias exponenciales de distribuciones multivariadas, lo que nos permitirá definir los modelos lineales generalizados multivariados. Estas familias de distribuciones unifican bajo una estructura en particular muchas de las distribuciones conocidas, tanto discretas como continuas. Presentamos como ejemplos la distribución normal multivariada, Bernoulli multivariada y multinomial. Estas serán ultilizadas como distribuciones modelo para los conjuntos de datos que analizaremos y en particular, la distribución normal servirá de ejemplo para analizar los resultados dentro del contexto de modelos lineales generalizados y comparar estos con los ya conocidos.

Sea $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_r)$ un vector aleatorio de dimensión r con distribución $f(\cdot|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^s$, donde $\boldsymbol{\Theta}$ es un subconjunto abierto y conexo de $\mathbb{R}^s$. La familia de distribuciones $\mathcal{F} = \{f(\cdot|\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ se denomina familia exponencial a k-parámetros si las densidades o funciones de probabilidad puntual del vector $\mathbf{Y}$ del conjunto $\mathcal{F}$ tiene la forma

$$f(\mathbf{y}|\boldsymbol{\theta}) = e^{\mathbf{T}(\mathbf{y})^T \boldsymbol{\eta}(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta})} h(\mathbf{y}), \tag{1.4}$$

donde los parámetros naturales de la familia $\boldsymbol{\eta}(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), \ldots, \eta_k(\boldsymbol{\theta}))^T$ son funciones de $\boldsymbol{\theta}$ dos veces continuamente diferenciables con matriz Jacobiana de rango completo. $\mathbf{T}(\mathbf{y}) = (T_1(\mathbf{y}), T_2(\mathbf{y}), \ldots, T_k(\mathbf{y}))^T$ es un vector de funciones reales conocidas, $h(\mathbf{y}) \geq 0$ es una función real conocida y $\psi(\boldsymbol{\theta})$ es tal que $f(\cdot|\boldsymbol{\theta})$ sea una densidad. Vamos a asumir que el conjunto $\boldsymbol{\Lambda} = \{\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$ tal que $\lambda_i = \eta_i(\boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ contiene una esfera en $\mathbb{R}^k$ y además contiene $k+1$ puntos $\lambda^{(1)}, \ldots, \lambda^{(k+1)}$ tales que $\{\lambda^{(j)} - \lambda^{(1)} : j = 2, \ldots, k+1\}$ son linealmente independientes. Estas condiciones son suficientes para afirmar que $\mathbf{T}(\mathbf{y})$ es un estadístico minimal suficiente y completo de $\boldsymbol{\theta}$ para la familia. Denotamos a un vector aleatorio que sigue esta distribución como $\mathbf{Y} \sim \mathcal{F}_{\boldsymbol{\eta},\mathbf{T},\psi}$. El espacio de parámetro naturales

$$H = \{\boldsymbol{\eta} \in \mathbb{R}^k : e^{\psi(\boldsymbol{\eta})} = \int e^{\boldsymbol{\eta}^T \mathbf{T}(\mathbf{x})} h(\mathbf{x}) d\mathbf{x} < \infty\},$$

donde la integral es remplazada por una suma en el caso de vectores aleatorios discretos, es el conjunto más grande de valores que puede tomar $\boldsymbol{\eta}$ para los cuales la densidad (o función de probabilidad puntual) puede ser definida.

La siguiente propiedad resulta útil para el cálculo de la esperanza y la varianza del estadístico suficiente de una familia exponencial de distribuciones (ver [**van der Vaart, 1998**], Sección 4.2).

Proposición 1.1. *Para todo* $\boldsymbol{\eta} \in H$,

$$\mathrm{E}_{\boldsymbol{\eta}}(\mathbf{T}(\mathbf{Y})) = \left(\frac{\partial \psi(\boldsymbol{\eta})}{\partial \eta_j}\right) \qquad \text{y} \qquad \mathrm{var}_{\boldsymbol{\eta}}(\mathbf{T}(\mathbf{Y})) = \left(\frac{\partial^2 \psi(\boldsymbol{\eta})}{\partial \eta_j \partial \eta_k}\right).$$

En particular, $\mathrm{E}_{\boldsymbol{\eta}}(T_i(\mathbf{Y})) = \dfrac{\partial \psi(\boldsymbol{\eta})}{\partial \eta_i}$ *y* $\mathrm{cov}_{\boldsymbol{\eta}}(T_i(\mathbf{Y}), T_j(\mathbf{Y})) = \dfrac{\partial^2 \psi(\boldsymbol{\eta})}{\partial \eta_i \partial \eta_j}$.

En las siguientes subsecciones presentamos algunos ejemplos.

1.2.1. Distribución normal multivariada

Sea $\mathbf{Y}$ un vector aleatorio de dimensión r con distribución normal multivariada, $\mathbf{Y} \sim \mathcal{N}_r(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Veamos que esta distribución pertenece a una familia exponencial a k parámetros, expresando su densidad como en (1.4):

$$\begin{aligned}
f(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{r/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})} \\
&= \frac{1}{(2\pi)^{r/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}\mathrm{tr}(\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})} \\
&= e^{\mathrm{vec}^T(\mathbf{y}\mathbf{y}^T)\mathrm{vec}(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}) + \mathbf{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}(\mathrm{tr}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) + \log|\boldsymbol{\Sigma}| + r\log(2\pi))} \\
&= e^{\mathrm{vech}^T(\mathbf{y}\mathbf{y}^T)\mathbf{D}_r^T\mathbf{D}_r\mathrm{vech}(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}) + \mathbf{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}(\mathrm{tr}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) + \log|\boldsymbol{\Sigma}| + r\log(2\pi))} \\
&= e^{\mathbf{T}(\mathbf{y})^T\boldsymbol{\eta} - \psi(\boldsymbol{\eta})}
\end{aligned}$$

donde

$$\begin{aligned}
\mathbf{T}(\mathbf{y}) &= (\mathbf{y}, -\frac{1}{2}\mathbf{D}_r^T\mathbf{D}_r\mathrm{vech}(\mathbf{y}\mathbf{y}^T)) \\
\boldsymbol{\eta} &= (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \mathrm{vech}\boldsymbol{\Sigma}^{-1}) \qquad (1.5) \\
\psi(\boldsymbol{\eta}) &= \frac{1}{2}(\mathrm{tr}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) + \log|\boldsymbol{\Sigma}| + r\log(2\pi)) \qquad (1.6) \\
k &= r + \frac{r(r+1)}{2}. \qquad (1.7)
\end{aligned}$$

1.2.2. Distribución bernoulli multivariada

Veamos ahora un caso discreto. Consideremos una serie de r eventos binarios diferentes no necesariamente independientes, cada uno con dos posibles resultados como por ejemplo: éxito o error, presencia o ausencia, defectuoso o no defectuoso. Sea el vector aleatorio $\mathbf{Y}$ que registra cada uno de los resultados de esos eventos, es decir, $\mathbf{Y} = (Y_1, \ldots, Y_r)$ donde $Y_i \in \{0, 1\}$, $i = 1, \ldots, r$. Luego, $\mathbf{Y}$ es un vector aleatorio con distribución Bernoulli multivariada. Por ejemplo, una situación real que puede enmarcarse bajo esta distribución es registrar si r especies de animales diferentes están presentes o no en un determinado habitad.

Denotamos las diferentes probabilidades con $p_{j_1,j_2,\ldots,j_r} = P(Y_1 = j_1, Y_2 = j_2, \ldots, Y_r = j_r)$ con $j_1, \ldots, j_r = 0, 1$. La forma más general de la función de probabilidad puntual es

$$\begin{aligned} p(\mathbf{y} = (y_1, \ldots, y_r)) &= P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_r = y_r) \\ &= p_{0,0,\ldots,0}^{[\prod_{i=1}^r (1-y_i)]} p_{1,0,\ldots,0}^{[y_1 \prod_{i=2}^r (1-y_i)]} p_{0,1,\ldots,0}^{(1-y_1) y_2 \prod_{i=3}^r (1-y_i)} \\ &\quad \ldots p_{0,1,\ldots,1}^{[(1-y_1)\prod_{i=2}^r y_i]} p_{1,\ldots,1}^{[\prod_{i=1}^r y_i]}. \end{aligned} \tag{1.8}$$

Ahora, (1.8) puede ser escrita como familia exponencial de la forma **[Dai et al., 2013]**

$$p(\mathbf{y}) = \exp\left[\sum_{q=1}^{r}\left(\sum_{1\le i_1<i_2<\ldots<i_q\le r} \eta^{i_1 i_2 \ldots i_q} B^{i_1 i_2 \ldots i_q}(\mathbf{x})\right) - \log\frac{1}{p_{0,0,\ldots,0}}\right],$$

donde

- el estadístico suficiente está dado por

  $$\begin{aligned} \mathbf{T}(\mathbf{y}) &= (B^1(\mathbf{y}), B^2(\mathbf{y}), \ldots, B^r(\mathbf{y}), B^{12}(\mathbf{y}), B^{13}(\mathbf{y}), \ldots, B^{1r}(\mathbf{y}), \ldots, B^{12\ldots r}(\mathbf{y})) \\ &= (y_1, y_2, \ldots, y_r, \ldots, y_1 y_2, \ldots, y_1 y_3, \ldots, y_1 y_q, \ldots, y_1 \ldots, y_r) \end{aligned}$$

  con función de interacción B definida como

  $$B^{i_1 i_2 \ldots i_q}(\mathbf{y}) = y_{i_1} y_{i_2} \ldots y_{i_q}.$$

- $\boldsymbol{\eta} = (\eta^1, \eta^2, \ldots, \eta^r, \eta^{12}, \eta^{13}, \ldots, \eta^{1r}, \ldots, \eta^{12\ldots r})^T$ es el vector de parámetros naturales que está definido de la siguiente forma

  $$\eta^{i_1 i_2 \ldots i_q} = \log\frac{\prod p(\text{cantidad par de ceros entre las } i_1, \ldots, i_q \text{ y las restantes son ceros})}{\prod p(\text{cantidad impar de ceros entre las } i_1, \ldots, i_q \text{ y las restantes son ceros})}.$$

  Explicado en más detalle: para cada subconjunto $\mathcal{I} \subseteq \{i_1, \ldots, i_q\}$ se considera el $p_{j_1,j_2,\ldots,j_r}$ con $j_1, \ldots, j_r = 0, 1$ que cumpla con los dos siguientes puntos:

  1. $j_i = 0$ si $i \in \mathcal{I}$ o $i \notin \{i_1, \ldots, i_q\}$
  2. $j_i = 1$ si $i \in \{i_1, \ldots, i_q\} - \mathcal{I}$.

  El factor $p_{j_1,j_2,\ldots,j_r}$ se agrega al numerador si el cardinal de $\mathcal{I}$ es par o nulo, y se agrega al denominador cuando el cardinal de $\mathcal{I}$ es impar.

- el término $\log \dfrac{1}{p_{0,0,\ldots,0}}$ como función de los parámetros naturales $\boldsymbol{\eta}$ tiene la siguiente forma:

$$\psi(\boldsymbol{\eta}) = \log\left[1 + \sum_{q=1}^{r}\left(\sum_{1\leq i_1<i_2<\ldots<i_q\leq r} \exp[S^{i_1 i_2 \ldots i_q}]\right)\right] \doteq \log\left[1 + b(\boldsymbol{\eta})\right] \tag{1.9}$$

con S siendo la suma

$$S^{i_1 i_2 \ldots i_q} = \sum_{1\leq s\leq q} \eta^{i_s} + \sum_{1\leq s<t\leq q} \eta^{i_s i_t} + \ldots + \eta^{i_1 i_2 \ldots i_q}$$

para $q = 1, \ldots, r$.

Por ejemplo, si consideramos el caso $r = 3$, los parámetros naturales tienen la forma

$$\eta^1 = \log \frac{p_{1,0,0}}{p_{0,0,0}} \qquad \eta^2 = \log \frac{p_{0,1,0}}{p_{0,0,0}} \qquad \eta^3 = \log \frac{p_{0,0,1}}{p_{0,0,0}}$$

$$\eta^{12} = \log \frac{p_{1,1,0}p_{0,0,0}}{p_{1,0,0}p_{0,1,0}} \qquad \eta^{13} = \log \frac{p_{1,0,1}p_{0,0,0}}{p_{1,0,0}p_{0,0,1}} \quad \eta^{23} = \log \frac{p_{0,0,0}p_{0,1,1}}{p_{0,1,0}p_{0,0,1}}$$

$$\eta^{123} = \log \frac{p_{1,1,1}p_{1,0,0}p_{0,1,0}p_{0,0,1}}{p_{1,1,0}p_{1,0,1}p_{0,1,1}p_{0,0,0}}$$

y las sumas S son

$$S^1 = \eta^1 \quad S^{12} = \eta^1 + \eta^2 + \eta^{12} \quad S^{123} = \eta^1 + \eta^2\eta^3 + \eta^{12} + \eta^{13} + \eta^{23} + \eta^{123}$$
$$S^2 = \eta^2 \quad S^{13} = \eta^1 + \eta^3 + \eta^{13}$$
$$S^3 = \eta^3 \quad S^{23} = \eta^2 + \eta^3 + \eta^{23}.$$

Una propiedad importante de esta distribución es que todas las componentes del vector aleatorio son independientes si y solo si $\eta^{i_1 i_2 \ldots i_q} = 0$, para todo $1 \leq i_1 < i_2 < \ldots < i_q \leq r$, y $q \geq 2$ [**Dai et al., 2013**]. En particular, cuando hablamos de correlación de grado 2, nos estamos refiriendo a que $\eta^{i_1 i_2 \ldots i_q} = 0$, para todo $1 \leq i_1 < i_2 < \ldots < i_q \leq r$, y $q \geq 3$. En dicho caso, el estadístico suficiente es

$$\mathbf{T}(\mathbf{y}) = (B^1(\mathbf{y}), B^2(\mathbf{y}), \ldots, B^r(\mathbf{y}), B^{12}(\mathbf{y}), \ldots, B^{(r-1)r}(\mathbf{y})) = (y_1, y_2, \ldots, y_r, y_1y_2, \ldots, y_{(r-1)y_r})$$

y los parámetros naturales son

$$\boldsymbol{\eta} = (\eta^1, \eta^2, \ldots, \eta^r, \eta^{12}, \ldots, \eta^{(r-1)r}).$$

Este caso será estudiado en detalle más adelante donde consideraremos el modelo Ising [**Ising, 1925**].

1.2.3. Distribución multinomial

La distribución multinomial es una extensión de la distribución Bernoulli univariada, donde la respuesta puede tomar más de dos valores. Supongamos ahora que la respuesta tiene r posibles resultados $E_1, E_2, \ldots, E_r$ con probabilidades $p_1, p_2, \ldots, p_r$ respectivamente de que ocurran y que el mismo experimento se repite un número m de veces. Sea $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_r)$ el vector aleatorio que cuenta cuantas veces ha ocurrido cada una de las posibles respuestas, es decir que $Y_i \in \{0, 1, \ldots, m\}$, con $i = 1, \ldots, r$ y $\sum_{i=1}^{r} Y_i = m$. Este tipo de familia de distribuciones es una familia exponecial a $k = r - 1$ parámetros. La función de probabilidad puntual de $\mathbf{Y}$ podemos expresarla como en (1.4) de la siguiente manera:

$$\begin{aligned}
P(y_1, \ldots, y_r) &= P(Y_1 = y_1, \ldots, Y_r = y_r) = \frac{m!}{y_1! y_2! \ldots y_r!} \; p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r} \; \mathbb{I}_{\{\sum_{i=1}^{r} y_i = m\}} \\
&= \exp\left(\sum_{i=1}^{r} y_i \log p_i \right) \; \mathbb{I}_{\{\sum_{i=1}^{r} y_i = m\}} \frac{m!}{y_1! y_2! \ldots y_r!} \\
&= \exp\left(\sum_{i=1}^{r-1} y_i \log p_i + y_r \log p_r \right) \; \mathbb{I}_{\{\sum_{i=1}^{r} y_i = m\}} \frac{m!}{y_1! y_2! \ldots y_r!} \\
&= \exp\left(\sum_{i=1}^{r-1} y_i \log p_i + (m - \sum_{j=1}^{r-1} y_j) \log(1 - \sum_{j=1}^{r-1} p_j) \right) \; \mathbb{I}_{\{\sum_{i=1}^{r} y_i = m\}} \frac{m!}{y_1! \ldots y_r!} \\
&= \exp\left(\sum_{i=1}^{r-1} \left(y_i \log \frac{p_i}{1 - \sum_{j=1}^{r-1} p_j} \right) + m \log(1 - \sum_{j=1}^{r-1} p_j) \right) \; \mathbb{I}_{\{\sum_{i=1}^{r} y_i = m\}} \frac{m!}{y_1! \ldots y_r!} \\
&= \exp(\mathbf{T}(\mathbf{y})^T \boldsymbol{\eta} - \psi(\boldsymbol{\eta})) h(\mathbf{y})
\end{aligned}$$

donde

$$\begin{aligned}
\mathbf{T}(\mathbf{y}) &= (y_1, \ldots, y_{r-1}) \\
\boldsymbol{\eta} &= (\eta_1, \ldots, \eta_{r-1}) = \left(\log \frac{p_1}{1 - \sum_{j=1}^{r-1} p_j}, \ldots, \log \frac{p_{r-1}}{1 - \sum_{j=1}^{r-1} p_j} \right) \\
\psi(\boldsymbol{\eta}) &= m \log \left(1 + \sum_{i=1}^{r-1} e^{\eta_i} \right) \\
k &= r - 1.
\end{aligned}$$

En el caso que $m = 1$, una forma compacta de expresar los datos multinomiales es considerar la variable Y que vale $1, 2, \ldots$ ó r indicando la categoría que resultó.

1.3. Modelos lineales generalizados multivariados

La unificación de muchos métodos estadísticos fue demostrado en **Nelder and Wedderburn, 1972** usando la idea de los modelos lineales generalizados. La extención al caso multivariado de los GLM (GLMM, de ahora en adelante) han sido tratados en **Fahrmeir and Tutz, 2001** y las referencias en dicho libro. Aquí reproducimos algunos de los conceptos más importantes.

1.3.1. Estructura del modelo

La construcción de este tipo de modelos involucra tres decisiones:

(a) ¿Cuál es la distribución de los datos (para valores fijos o aleatorios de los predictores y posiblemente después de una transformación de ellos)?

(b) ¿Cuáles serán los predictores?

(c) ¿Cuál es la función de la media que será modelada de forma lineal en los predictores?

(a) Distribución de $\mathbf{Y}|\mathbf{X}$

Sea $\mathbf{Y} \in \mathbb{R}^r$, $\mathbf{X} \in \mathbb{R}^p$ y suponemos que la distribución de $\mathbf{Y}|\mathbf{X}$ pertenece a una familia exponencial a k parámetros naturales:

$$f(\mathbf{y}|\mathbf{x}, \boldsymbol{\eta}_\mathbf{x}) = e^{\mathbf{T}(\mathbf{y})^T \boldsymbol{\eta}_\mathbf{x} - \psi(\boldsymbol{\eta}_\mathbf{x})} h(\mathbf{y}), \tag{1.10}$$

donde expresamos la distribución en la denominada forma canónica. De acuerdo a la estructura de datos que se quiere modelar, se puede considerar distribuciones como Bernoulli, multinomial, Poisson, normal, o una mezcla de ellas, que modelen de manera adecuada al tipo de problema que se quiere estudiar.

(b) Predictores

En la práctica, uno debe tomar decisiones acerca de cuáles serán los predictores y de que forma se incluirán en el modelo. Esto es similar a los modelos lineales multivariados.

(c) Función de enlace

Tradicionalmente, se quiere relacionar los paramétros de la distribución a varios predictores. En GLMM, se modela una transformación de la media $\boldsymbol{\mu}_\mathbf{x}$ como función lineal de los predictores:

$$\begin{aligned} \mathrm{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x}) &= \boldsymbol{\mu}_\mathbf{x} \\ g(\boldsymbol{\mu}_\mathbf{x}) &= \overline{\boldsymbol{\eta}} + \mathbf{D}\mathbf{x} \end{aligned}$$

donde $g(\cdot)$ es una función conocida, llamada función de enlace (ya que esta enlaza la media de $\mathbf{Y}$ y los predictores), $\overline{\boldsymbol{\eta}}$ es el vector ordenada al origen y $\mathbf{D}$ la matriz de coeficientes de la regresión. Generalmente, es común utilizar como función enlace aquella tal que

$$\boldsymbol{\eta}_{\mathbf{x}} = g(\boldsymbol{\mu}_{\mathbf{x}}) = \overline{\boldsymbol{\eta}} + \mathbf{D}\mathbf{x},$$

la cual se denomina enlace canónico. De los k parámetros naturales sólo vamos a modelar linealmente algunos de ellos. Es decir $k = k_1 + k_2$ donde k_1: número de parámetros a modelar linealmente y k_2: no se modelarán. Luego, conforme a esto denotamos $\boldsymbol{\eta}_{\mathbf{x}}^T = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)^T$ con $\boldsymbol{\eta}_1 : k_1 \times 1$ y $\boldsymbol{\eta}_2 : k_2 \times 1$. Esta separación de $\boldsymbol{\eta}_{\mathbf{x}}$ en $\boldsymbol{\eta}_1$ y $\boldsymbol{\eta}_2$ tiene que ver con que en general la media $\mathrm{E}(\mathbf{Y}|\mathbf{X})$ es sólo función de $\boldsymbol{\eta}_1$ y es por lo tanto el parámetro de mayor interés. Por ejemplo para el caso de la distribución normal multivariada con matriz de covarianza $\boldsymbol{\Sigma}$, estudiada en la Sección 1.2.1, $\boldsymbol{\eta}_2 = \mathrm{vech}\left(\boldsymbol{\Sigma}^{-1}\right)$ la cual se supone constante para cualquier valor de $\mathbf{X}$. Más generalmente,

$$\begin{aligned} \boldsymbol{\eta}_1 &= \overline{\boldsymbol{\eta}}_1 + \mathbf{D}\mathbf{x} = \boldsymbol{\Gamma}_1\mathbf{f} = (\mathbf{f}^T \otimes \mathbf{I}_{k_1})\mathrm{vec}(\boldsymbol{\Gamma}_1) \\ \boldsymbol{\eta}_2 &= \overline{\boldsymbol{\eta}}_2 = \boldsymbol{\Gamma}_2 = (1 \otimes \mathbf{I}_{k_2})\mathrm{vec}(\boldsymbol{\Gamma}_2) \end{aligned} \tag{1.11}$$

donde $\mathbf{f}^T = (1, \mathbf{x})^T : 1 \times (p+1)$, $\boldsymbol{\Gamma}_1 = (\overline{\boldsymbol{\eta}}_1, \mathbf{D}) : k_1 \times (p+1)$ y $\boldsymbol{\Gamma}_2 = (\overline{\boldsymbol{\eta}}_2) : k_2 \times 1$. Luego, un modelo lineal generalizado multivariado en forma matricial es el siguiente:

$$\boldsymbol{\eta}_{\mathbf{x}} = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{pmatrix} = \begin{pmatrix} (\mathbf{f}^T \otimes \mathbf{I}_{k_1}) & 0 \\ 0 & \mathbf{I}_{k_2} \end{pmatrix} \begin{pmatrix} \mathrm{vec}(\boldsymbol{\Gamma}_1) \\ \mathrm{vec}(\boldsymbol{\Gamma}_2) \end{pmatrix} = \mathbf{F}\boldsymbol{\Gamma} \tag{1.12}$$

donde $\mathbf{F} = \begin{pmatrix} (\mathbf{f}^T \otimes \mathbf{I}_{k_1}) & 0 \\ 0 & \mathbf{I}_{k_2} \end{pmatrix} : k \times (k_1(p+1)+k_2)$ es la matriz de diseño y $\boldsymbol{\Gamma} = \begin{pmatrix} \mathrm{vec}(\boldsymbol{\Gamma}_1) \\ \mathrm{vec}(\boldsymbol{\Gamma}_2) \end{pmatrix}$ son los coeficientes a estimar con los datos.

1.3.2. Estimación

Bajo las suposiciones establecidas para los errores al formular el modelo (1.10), los parámetros desconocidos pueden ser estimados mediante el método basado en máxima verosimilitud. Aunque en algunos casos especiales, se pueden hallar expresiones explícitas para estos estimadores (como en el caso de errores gaussianos), usualmente son necesarios métodos numéricos. Típicamente estos métodos son iterativos y están basados en el algoritmo de Newton-Raphson.

Para obtener el estimador de máxima verosimilitud

$$\widehat{\mathbf{\Gamma}} = \begin{pmatrix} \text{vec}\left(\widehat{\mathbf{\Gamma}}_1\right) \\ \text{vec}\left(\widehat{\mathbf{\Gamma}}_2\right) \end{pmatrix},$$

contamos con n observaciones independientes $\mathbf{y}_1, \ldots, \mathbf{y}_n$ del vector $\mathbf{Y}$ y $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sus respectivos valores de $\mathbf{X}$. Estos datos siguen el modelo (1.10) junto con (1.12):

$$\boldsymbol{\eta}_i = \begin{pmatrix} \boldsymbol{\eta}_{1i} \\ \boldsymbol{\eta}_{2i} \end{pmatrix} = \begin{pmatrix} (\mathbf{f}_i^T \otimes \mathbf{I}_{k_1}) & 0 \\ 0 & \mathbf{I}_{k_2} \end{pmatrix} \begin{pmatrix} \text{vec}(\mathbf{\Gamma}_1) \\ \text{vec}(\mathbf{\Gamma}_2) \end{pmatrix} \doteq \mathbf{F}_i \mathbf{\Gamma}$$

donde $\mathbf{f}_i^T = (1, \mathbf{x}_i)^T : 1 \times (p+1)$ con $i = 1, \ldots, n$. El logaritmo de la función de verosimilitud para estos datos se puede expresar de la siguiente forma:

$$\mathcal{L}_n(\mathbf{\Gamma}) = \sum_{i=1}^{n} (\boldsymbol{\eta}_i^T \mathbf{T}(\mathbf{y}_i) - \psi(\boldsymbol{\eta}_i)) = \sum_{i=1}^{n} ((\mathbf{F}_i \mathbf{\Gamma})^T \mathbf{T}(\mathbf{y}_i) - \psi(\mathbf{F}_i \mathbf{\Gamma}) \doteq \sum_{i=1}^{n} l_i. \tag{1.13}$$

La ecuación de máxima verosimilitud $U(\mathcal{L}_n) = \dfrac{\partial \mathcal{L}_n}{\partial \mathbf{\Gamma}} = 0$ es en este caso

$$U(\mathcal{L}_n) = \sum_{i=1}^{n} \mathbf{F}_i^T \mathbf{T}(\mathbf{y}_i) - \mathbf{F}_i^T \nabla \psi(\mathbf{F}_i \mathbf{\Gamma}) = 0,$$

donde $\nabla \psi(\cdot) : \mathbb{R}^k \to \mathbb{R}^k$ es el operador gradiente.

Es claro que, si la función ψ no tiene una forma sencilla, no es posible obtener una forma explícita para la estimación $\widehat{\mathbf{\Gamma}}$. Es por ello que, en los modelos lineales generalizados los estimadores de máxima verosimilitud, cuando existen, se obtienen mediante los método iterativos IRLS o de Newton-Rawson [**Green, 1984**]. Habiendo obtenido en el paso t, $\mathbf{\Gamma}^{(t)}$ y con ello $\boldsymbol{\eta}_i^{(t)} = \mathbf{F}_i \mathbf{\Gamma}^{(t)}$ para $i = 1, \ldots, n$, la iteración $(t+1)$ obtenida para el vector de coeficientes $\mathbf{\Gamma}$ es

$$\mathbf{\Gamma}^{(t+1)} = \left(\sum_{i=1}^{n} \mathbf{F}_i^T \mathbf{W}_i^{(t)} \mathbf{F}_i \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{F}_i^T \mathbf{W}_i^{(t)} \mathbf{z}_i \right), \tag{1.14}$$

donde para cada $i = 1, \ldots, n$

$$\mathbf{z}_i^{(t)} = \mathbf{F}_i \mathbf{\Gamma}^{(t)} + (\mathbf{W}_i^{(t)})^{-1} \mathbf{d}_i^{(t)} \text{ con} \tag{1.15}$$

$$\mathbf{d}_i^{(t)} = \frac{\partial l_i}{\partial \boldsymbol{\eta}_i} = \mathbf{T}(\mathbf{y}_i) - \nabla \psi(\boldsymbol{\eta}_i^{(t)}) = \mathbf{T}(\mathbf{y}_i) - \nabla \psi(\mathbf{F}_i \mathbf{\Gamma}^{(t)}) \tag{1.16}$$

$$\mathbf{W}_i^{(t)} = -\frac{\partial^2 l_i}{\partial \boldsymbol{\eta}_i \partial \boldsymbol{\eta}_i} = \mathbf{H}\psi(\boldsymbol{\eta}_i^{(t)}) = \mathbf{H}\psi(\mathbf{F}_i \mathbf{\Gamma}^{(t)}) \tag{1.17}$$

y definimos $\mathbf{H}\psi(\cdot) : \mathbb{R}^k \to \mathbb{R}^{k\times k}$ como la matriz Hessiana de la función ψ. Los vectores $\mathbf{z}_i^{(t)}$ de la (t)-esima iteración son denominados vectores de seudo-respuestas. Esto se debe a que la iteración $\mathbf{\Gamma}^{(t+1)}$ (1.14) es la solución del problema de mínimos cuadrados pesados:

$$\arg\min_{\mathbf{\Gamma}} \sum_{i=1}^{n} \left(\mathbf{z}_i^{(t)} - \mathbf{F}_i\mathbf{\Gamma}\right)^T \mathbf{W}_i^{(t)} \left(\mathbf{z}_i^{(t)} - \mathbf{F}_i\mathbf{\Gamma}\right).$$

Es decir, si suprimimos el superíndice de la iteración, cada iteración IRLS involucra ajustar el modelo de regresión lineal multivariado

$$\mathbf{Z} = \mathbf{F}\mathbf{\Gamma} + \mathbf{U}$$

para el cual contamos con los datos $\mathbf{z}_1, \ldots, \mathbf{z}_n$ para la respuesta y $\mathbf{F}_1, \ldots, \mathbf{F}_n$ para los predictores y además $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)^T$ son los errores independientes con media cero y matrices de covariazas desigual $\mathbf{W}_i^{-1}$.

1.3.3. Distribución asintótica del estimador de máxima verosimilitud

Dependiendo si los predictores son considerados fijos o aleatorios, siempre que la matriz de información observada $\mathcal{J}_n(\mathbf{\Gamma}) = \sum_{i=1}^{n} \mathbf{F}_i^T \mathbf{H}\psi(\mathbf{F}_i\mathbf{\Gamma})\mathbf{F}_i$ o la esperanza $\mathbf{E}(\mathbf{F}^T\mathbf{H}\psi(\mathbf{F}\mathbf{\Gamma})\mathbf{F})$ sean definidas positivas, respectivamente, el estimador de máxima verosimilitud $\widehat{\mathbf{\Gamma}}$ del modelo (1.12) existe y es único (ver Lema 3.5.1 de **Spokoiny and Dickhaus, 2015**). Para obtener la consistencia y la distribución asintótica del estimador $\widehat{\mathbf{\Gamma}}$ se requieren las condiciones usuales de regularidad, las cuales son satisfechas para los casos más conocidos de familias de exponenciales (Ver **Fahrmeir and Kaufmann, 1986** para modelos con respuesta binomial, multinomial y Poisson, **Wedderburn, 1976** para modeos con respuesta normal, Poisson y Gamma, entre otros). Son condiciones que aseguran la intercambiabilidad de la diferenciación e integración (así la matriz de Fisher está bien definida y es igual a la varianza del estadístico suficiente) y que la derivada tercera del logaritmo de la verosimilitud $\mathcal{L}$ correspondiente a una observación esté acotada por una función integrable (independiente de los parámetros) en un entorno del verdadero parámetro. Más detalles pueden verse en la Sección 3.4.1 de **Fahrmeir and Tutz, 2001**. Un estudio cuidadoso sobre consistencia y distribución asintótica se encuentra en **Fahrmeir and Kaufmann, 1985** y **Haberman, 1977**. Por otro lado, como veremos para el caso de regresión de rango reducido, también puede aplicarse el Teorema 5.1 de **Lehmann and Casella, 1998** para obtener existencia, consistenia y distribución asintótica del estimador de máxma verosimilitud. Además, para el caso condicionado a los predictores, es decir cuando $\mathbf{X}$ no es aleatorio, se necesitan condiciones en la matriz de diseño $\mathbf{F}$

para poder utilizar el teorema de Lindeberg-Feller y así obtener la distribución asintótica en este caso.

La siguiente proposición presenta la distribución asintótica del estimador de máxima verosimilitud para los casos condicionados o no en $\mathbf{X}$. Damos aquí un bozquejo de la prueba que hace uso de las condiciones habituales de regularidad.

Proposición 1.2. *Sea el vector aleatorio* $\mathbf{Y}|\mathbf{X}$ *cuya distribución pertenece a una familia exponencial a* k *parámetros naturales de la forma (1.10) y sea* $\mathbf{H}\psi$ *la matriz Hessiana de la función* ψ *correspondiente a cada familia. Teniendo en cuenta la separación de los* k *parámetros naturales en* k_1 *y* k_2*, sea la partición de* $\mathbf{H}\psi$*:*

$$\mathbf{H}\psi = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{pmatrix}$$

con

$$\mathbf{H}_{11} = \frac{\partial^2 \psi}{\partial \boldsymbol{\eta}_1^T \partial \boldsymbol{\eta}_1} : k_1 \times k_1 \qquad \mathbf{H}_{12} = \frac{\partial^2 \psi}{\partial \boldsymbol{\eta}_2^T \partial \boldsymbol{\eta}_1} : k_1 \times k_2$$

$$\mathbf{H}_{21} = \mathbf{H}_{12}^T : k_2 \times k_1 \qquad \mathbf{H}_{22} = \frac{\partial^2 \psi}{\partial \boldsymbol{\eta}_2^T \partial \boldsymbol{\eta}_2} : k_2 \times k_2.$$

El estimador de máxima verosimilitud del modelo (1.12) satisface

$$\sqrt{n}\left(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0\right) \to \mathcal{N}(0, \mathbf{V}_{\boldsymbol{\Gamma}_0}) \tag{1.18}$$

donde $\boldsymbol{\Gamma}_0$ *es el valor poblacional verdadero para* $\boldsymbol{\Gamma}$ *y* $\mathbf{V}_{\boldsymbol{\Gamma}_0} = \mathcal{J}(\boldsymbol{\Gamma}_0)^{-1}$*, con*

$$\mathcal{J}(\boldsymbol{\Gamma}_0) = \mathbf{E}\begin{pmatrix} (\mathbf{f}\mathbf{f}^T \otimes \mathbf{H}_{11}(\mathbf{F}\boldsymbol{\Gamma}_0)) & (\mathbf{f} \otimes \mathbf{H}_{12}(\mathbf{F}\boldsymbol{\Gamma}_0)) \\ (\mathbf{f}^T \otimes \mathbf{H}_{21}(\mathbf{F}\boldsymbol{\Gamma}_0)) & \mathbf{H}_{22}(\mathbf{F}\boldsymbol{\Gamma}_0) \end{pmatrix}$$

en el caso que $\mathbf{X}$ *sea aleatorio. Si se condiciona en* $\mathbf{X}$*, considerar*

$$\mathcal{J}(\boldsymbol{\Gamma}_0) = \lim_{n\to\infty} \mathcal{J}_n(\boldsymbol{\Gamma}_0) \doteq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} (\mathbf{f}_i\mathbf{f}_i^T \otimes \mathbf{H}_{11}(\mathbf{F}_i\boldsymbol{\Gamma}_0)) & (\mathbf{f}_i \otimes \mathbf{H}_{12}(\mathbf{F}_i\boldsymbol{\Gamma}_0)) \\ (\mathbf{f}_i^T \otimes \mathbf{H}_{21}(\mathbf{F}_i\boldsymbol{\Gamma}_0)) & \mathbf{H}_{22}(\mathbf{F}_i\boldsymbol{\Gamma}_0) \end{pmatrix},$$

donde se asume que dicho límite existe.

Demostración. La demostración de esta proposición se basa en el Teorema central del límite (para el caso de $\mathbf{X}$ aleatorio) y en el Teorema de Lindeberg-Feller (para el caso de $\mathbf{X}$

fijo). Como $\frac{\partial \mathcal{L}_n}{\partial \boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Gamma}}) = 0$, donde $\mathcal{L}_n(\boldsymbol{\Gamma})$ está dado por la ecuación (1.13) si expandimos $\frac{\partial \mathcal{L}_n}{\partial \boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Gamma}})$ en series de Taylor alrededor de $\boldsymbol{\Gamma}_0$ tenemos, bajo condiciones de regularidad, que

$$\begin{aligned}
\sqrt{n}\left(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0\right) &= -\sqrt{n}\left(\frac{\partial^2 \mathcal{L}_n}{\partial^2 \boldsymbol{\Gamma}^2}(\boldsymbol{\Gamma}_0) + \frac{1}{2}\frac{\partial^3 \mathcal{L}_n}{\partial^3 \boldsymbol{\Gamma}^3}(\widetilde{\boldsymbol{\Gamma}})(\mathbf{I} \otimes (\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0))\right)^{-1} \frac{\partial \mathcal{L}_n}{\partial \boldsymbol{\Gamma}}(\boldsymbol{\Gamma}_0) \\
&= \sqrt{n}\left(\sum_{i=1}^{n} \mathbf{F}_i^T \mathbf{H}\psi(\mathbf{F}_i\boldsymbol{\Gamma}_0)\mathbf{F}_i + \sum_{i=1}^{n} \mathbf{F}_i^T \mathbf{G}\psi(\mathbf{F}_i\widetilde{\boldsymbol{\Gamma}})(\mathbf{F}_i \otimes \mathbf{F}_i)(\mathbf{I} \otimes (\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0))\right)^{-1} \\
&\quad \left(\sum_{i=1}^{n} \mathbf{F}_i^T \mathbf{T}(\mathbf{y}_i) - \mathbf{F}_i^T \nabla\psi(\mathbf{F}_i\boldsymbol{\Gamma}_0)\right)
\end{aligned}$$

donde $\widetilde{\boldsymbol{\Gamma}} = \frac{t}{n}\boldsymbol{\Gamma}_0 + \left(1 - \frac{t}{n}\right)\widehat{\boldsymbol{\Gamma}}$ para algún $t \in (0,1)$ y $\mathbf{G}\psi(\cdot)$ es la derivada tercera de dimensiones $k^2 \times k$ de la función ψ. Dicha derivada de tercer orden es acotada en probabilidad, por las condiciones de regularidad supuestas para la derivada tercera del logaritmo de la función de verosimiltitud para una muestra $\mathcal{L}$. Además, bajo condiciones de regularidad, $\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0 = o_p(1)$. Por lo que,

$$\sqrt{n}\left(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0\right) \simeq \sqrt{n}\left(\sum_{i=1}^{n} \mathbf{F}_i^T \mathbf{H}\psi(\mathbf{F}_i\boldsymbol{\Gamma}_0)\mathbf{F}_i\right)^{-1}\left(\sum_{i=1}^{n} \mathbf{F}_i^T \mathbf{T}(\mathbf{y}_i) - \mathbf{F}_i^T \nabla\psi(\mathbf{F}_i\boldsymbol{\Gamma}_0)\right),$$

donde el símbolo $\simeq$ indica que ambos miembros son asintóticamente iguales en distribución cuando $n \to \infty$.

Definimos las variables aleatorias $\mathbf{Z}_i = \mathbf{F}_i^T \mathbf{T}(\mathbf{y}_i) - \mathbf{F}_i^T \nabla\psi(\mathbf{F}_i\boldsymbol{\Gamma}_0)$. Luego, en el caso en que $\mathbf{X}$ sea un vector aleatorio, $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ son independientes e idénticamente distribuidas. La condición de independencia e igualmente distribuidas se sigue del hecho de que $(\mathbf{x}_i, \mathbf{y}_i)$ es una muestra i.i.d. y además

$$\begin{aligned}
\mathbf{E}_{(\mathbf{X},\mathbf{Y})}(\mathbf{Z}_i) &= \mathbf{E}_{\mathbf{X}}\left(\mathbf{E}_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{F}_i^T \mathbf{T}(\mathbf{y}_i) - \mathbf{F}_i^T \nabla\psi(\mathbf{F}_i\boldsymbol{\Gamma}_0)\right)\right) \\
&= \mathbf{E}_{\mathbf{X}}\left(\mathbf{F}_i^T \mathbf{E}_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{T}(\mathbf{y}_i) - \nabla\psi(\mathbf{F}_i\boldsymbol{\Gamma}_0)\right)\right) = 0
\end{aligned}$$

$$\begin{aligned}
\mathrm{var}_{(\mathbf{X},\mathbf{Y})}(\mathbf{Z}_i) &= \mathrm{var}_{\mathbf{X}}\left(\mathbf{E}_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{F}_i^T \mathbf{T}(\mathbf{y}_i) - \mathbf{F}_i^T \nabla\psi(\mathbf{F}_i\boldsymbol{\Gamma}_0)\right)\right) \\
&\quad + \mathbf{E}_{\mathbf{X}}\left(\mathrm{var}_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{F}_i^T \mathbf{T}(\mathbf{y}_i) - \mathbf{F}_i^T \nabla\psi(\mathbf{F}_i\boldsymbol{\Gamma}_0)\right)\right) \\
&= 0 + \mathbf{E}_{\mathbf{X}}\left(\mathbf{F}_i^T \mathrm{var}_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{T}(\mathbf{y}_i)\right)\mathbf{F}_i\right) \\
&= \mathbf{E}_{\mathbf{X}}\left(\mathbf{F}_i^T \mathbf{H}\psi(\mathbf{F}_i\boldsymbol{\Gamma}_0)\mathbf{F}_i\right) \\
&= \mathbf{E}_{\mathbf{X}}\left(\mathbf{F}^T \mathbf{H}\psi(\mathbf{F}\boldsymbol{\Gamma}_0)\mathbf{F}\right) \\
&= \mathcal{J}(\boldsymbol{\Gamma}_0) = \mathbf{V}_{\boldsymbol{\Gamma}_0}^{-1}
\end{aligned}$$

donde en las últimas igualdades hemos usado el resultado de la Proposición 1.1. Luego, por el Teorema central del límite, tenemos que

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\right) \to \mathcal{N}(0, \mathbf{V}_{\mathbf{\Gamma}_0}^{-1}). \tag{1.19}$$

Por otro lado, como $\frac{1}{n}\sum_{i=1}^{n}\mathbf{F}_i^T\mathbf{H}\psi(\mathbf{F}_i\mathbf{\Gamma}_0)\mathbf{F}_i$ converge en probabilidad a $\mathbf{E}_{\mathbf{X}}\left(\mathbf{F}\mathbf{H}\psi(\mathbf{F}\mathbf{\Gamma}_0)\mathbf{F}\right) = \mathbf{V}_{\mathbf{\Gamma}_0}^{-1}$, aplicando el Teorema de Slutsky tenemos que

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{F}_i^T\mathbf{H}\psi(\mathbf{F}_i\mathbf{\Gamma}_0)\mathbf{F}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\right) \to \mathcal{N}(0, \mathbf{V}_{\mathbf{\Gamma}}).$$

Por lo tanto, $\sqrt{n}\left(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}_0\right)$ también converge en distribución a una normal con media 0 y varianza $\mathbf{V}_{\mathbf{\Gamma}_0}$.

En el caso en que $\mathbf{X}$ no sea aleatorio, las variables aleatorias $\mathbf{Z}_i = \mathbf{F}_i^T\mathbf{T}(\mathbf{y}_i) - \mathbf{F}_i^T\nabla\psi(\mathbf{F}_i\mathbf{\Gamma}_0)$ son independientes con media 0 pero distinta varianza:

$$\begin{aligned}\text{var}_{\mathbf{Y}}\left(\mathbf{Z}_i\right) = \text{var}_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{Z}_i\right) &= \text{var}_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{F}_i^T\mathbf{T}(\mathbf{y}_i) - \mathbf{F}_i^T\nabla\psi(\mathbf{F}_i\mathbf{\Gamma}_0)\right)\\ &= \mathbf{F}_i^T\text{var}_{\mathbf{Y}|\mathbf{X}}\left(\mathbf{T}(\mathbf{y}_i)\right)\mathbf{F}_i = \mathbf{F}_i^T\mathbf{H}\psi(\mathbf{F}_i\mathbf{\Gamma}_0)\mathbf{F}_i.\end{aligned}$$

Sean las variables aleatorias $\mathbf{Y}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{Z}_i$ y $B_n = \text{var}(\mathbf{Y}_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{F}_i^T\mathbf{H}\psi(\mathbf{F}_i\mathbf{\Gamma}_0)\mathbf{F}_i$. Asumiendo que existe el límite $\mathcal{J}(\mathbf{\Gamma}_0) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbf{F}_i^T\mathbf{H}\psi(\mathbf{F}_i\mathbf{\Gamma}_0)\mathbf{F}_i$, podemos aplicar el Teorema de Lindeberg-Feller bajo ciertas condiciones en la matriz $\mathbf{F}$ (ver Sección 2.8 de **van der Vaart, 1998**). Entonces tenemos que $\mathbf{Y}_n \to \mathcal{N}(0, \mathcal{J}(\mathbf{\Gamma}_0))$.

Luego, aplicando el Teorema de Slutsky se obtiene el resultado:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{F}_i^T\mathbf{H}\psi(\mathbf{F}_i\mathbf{\Gamma}_0)\mathbf{F}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{F}_i^T\mathbf{T}(\mathbf{y}_i) - \mathbf{F}_i^T\nabla\psi(\mathbf{F}_i\mathbf{\Gamma}_0)\right) \to \mathcal{N}(0, \mathcal{J}(\mathbf{\Gamma}_0)^{-1}).$$

□

1.3.4. Normal multivariada como familia exponencial

Supongamos que la distribución de $\mathbf{Y}|\mathbf{X}$ es $N_r(\boldsymbol{\mu}, \mathbf{\Sigma})$, luego de acuerdo a (1.5) y (1.7), el modelo centrado modelando los parámetros naturales será:

$$\begin{aligned}\boldsymbol{\eta}_1 &= \mathbf{\Sigma}^{-1}\boldsymbol{\mu} = \mathbf{\Gamma}_1\mathbf{x}\\ \boldsymbol{\eta}_2 &= \text{vech}(\mathbf{\Sigma}^{-1}) = \mathbf{\Gamma}_2.\end{aligned} \tag{1.20}$$

Es decir que $k_1 = r$ y $k_2 = r(r+1)/2$.

Observar que en este caso el enfoque de los GLMM, donde modelamos linealmente los parámetros naturales de la familia exponencial a la cual pertenece la distribución, se diferencia del modelo de regresión lineal multivariado pues no se modela la media $\boldsymbol{\mu}$ en función de $\mathbf{x}$ sino que se modela $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Sin embargo, es clara la relación entre estos modelos pues de acuerdo a (1.20) y (1.1), $\boldsymbol{\Gamma}_1 = \boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}$. Por lo tanto, el estimador de máxima verosimilitud $\widehat{\boldsymbol{\Gamma}}_1$ obtenido via GLMM y el producto de los estimadores de máxima verosimilitud $\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}$ usando el modelo de regresión lineal son los mismos, como así también sus varianzas asintóticas.

La matriz de covarianza asintótica para $\boldsymbol{\Gamma}_1$ utilizando la Proposición 1.2 está dada por

$$\mathbf{V}_{\boldsymbol{\Gamma}_1} = \left(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right) + \left(\boldsymbol{\Gamma}_1^T \boldsymbol{\Sigma} \boldsymbol{\Gamma}_1 \otimes \boldsymbol{\Sigma}^{-1}\right) + \left(\boldsymbol{\Gamma}_1^T \otimes \boldsymbol{\Gamma}_1\right) \mathbf{K}_{rp}. \tag{1.21}$$

Si queremos obtener la varianza asintótica de $\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\beta}}$ a partir (1.3), debemos usar la regla de Cramer. Así, $\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\beta}}$ es asintóticamente normal con media $\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}$ y varianza dada por

$$\mathbf{V}_{\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}} = \left(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right) + \left(\boldsymbol{\beta}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta} \otimes \boldsymbol{\Sigma}^{-1}\right) + \left(\boldsymbol{\beta}^T \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right) \mathbf{K}_{rp}. \tag{1.22}$$

Teniendo en cuenta que $\boldsymbol{\Gamma}_1 = \boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}$, (1.21) y (1.22) son las mismas.

1.4. Demostraciones del Cápitulo 1

Prueba de (1.21): Vamos a calcular la varianza asintótica del estimador de máxima verosimilitud de $\mathbf{\Gamma}_1$ usando la Proposición 1.2. En primer lugar, veamos quien es $\mathcal{J}_n$ en este caso. De acuerdo a (1.5) y (1.6), obtenemos las siguientes derivadas de segundo orden que constituyen la matriz $\mathbf{H}\psi$.

$$\begin{aligned}
\mathbf{H}_{11i} &= \mathbf{\Sigma} \\
\mathbf{H}_{12i} &= -(\boldsymbol{\eta}_1^T\mathbf{\Sigma}\otimes\mathbf{\Sigma})\mathbf{D}_r = -(\mathbf{x}_i^T\mathbf{\Gamma}_1^T\mathbf{\Sigma}\otimes\mathbf{\Sigma})\mathbf{D}_r = -(\mathbf{x}_i^T\boldsymbol{\beta}^T\otimes\mathbf{\Sigma})\mathbf{D}_r \\
\mathbf{H}_{22i} &= \mathbf{D}_r^T((\mathbf{\Sigma}\boldsymbol{\eta}_1\boldsymbol{\eta}_1^T\mathbf{\Sigma}+\frac{1}{2}\mathbf{\Sigma})\otimes\mathbf{\Sigma})\mathbf{D}_r = \mathbf{D}_r^T((\mathbf{\Sigma}\mathbf{\Gamma}_1\mathbf{x}_i\mathbf{x}_i^T\mathbf{\Gamma}_1^T\mathbf{\Sigma}+\frac{1}{2}\mathbf{\Sigma})\otimes\mathbf{\Sigma})\mathbf{D}_r \\
&= \mathbf{D}_r^T((\boldsymbol{\beta}\mathbf{x}_i\mathbf{x}_i^T\boldsymbol{\beta}^T+\frac{1}{2}\mathbf{\Sigma})\otimes\mathbf{\Sigma})\mathbf{D}_r.
\end{aligned}$$

Luego,

$$\mathcal{J}_n = \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} (\mathbf{x}_i\mathbf{x}_i^T\otimes\mathbf{\Sigma}) & (\mathbf{x}_i\otimes(\mathbf{x}_i^T\boldsymbol{\beta}^T\otimes\mathbf{\Sigma})\mathbf{D}_r) \\ (\mathbf{x}_i^T\otimes(\mathbf{D}_r^T(\boldsymbol{\beta}\mathbf{x}_i\otimes\mathbf{\Sigma})) & \mathbf{D}_r^T\left(\left(\boldsymbol{\beta}\mathbf{x}_i\mathbf{x}_i^T\boldsymbol{\beta}^T+\frac{1}{2}\mathbf{\Sigma}\right)\otimes\mathbf{\Sigma}\right)\mathbf{D}_r \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n(\mathbf{x}_i\mathbf{x}_i^T)\otimes\mathbf{\Sigma} & \left(\frac{1}{n}\sum_{i=1}^n(\mathbf{x}_i\mathbf{x}_i^T)\boldsymbol{\beta}^T\otimes\mathbf{\Sigma}\right)\mathbf{D}_r \\ \mathbf{D}_r^T\left(\boldsymbol{\beta}\frac{1}{n}\sum_{i=1}^n(\mathbf{x}_i\mathbf{x}_i^T)\otimes\mathbf{\Sigma}\right) & \mathbf{D}_r^T\left(\left(\boldsymbol{\beta}\frac{1}{n}\sum_{i=1}^n(\mathbf{x}_i\mathbf{x}_i^T)\boldsymbol{\beta}^T+\frac{1}{2}\mathbf{\Sigma}\right)\otimes\mathbf{\Sigma}\right)\mathbf{D}_r \end{pmatrix}.$$

Así, obtenemos que la varianza asintótica de $(\widehat{\mathbf{\Gamma}}_1,\widehat{\mathbf{\Gamma}}_2)$ es

$$\mathbf{V}_{\mathbf{\Gamma}} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{X}}\otimes\mathbf{\Sigma} & \left(\mathbf{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\otimes\mathbf{\Sigma}\right)\mathbf{D}_r) \\ \mathbf{D}_r^T\left(\beta\mathbf{\Sigma}_{\mathbf{X}}\otimes\mathbf{\Sigma}\right)) & \mathbf{D}_r^T\left(\left(\boldsymbol{\beta}\mathbf{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T+\frac{1}{2}\mathbf{\Sigma}\right)\otimes\mathbf{\Sigma}\right)\mathbf{D}_r \end{pmatrix}^{-1}$$

donde $\mathbf{\Sigma}_{\mathbf{X}} = \text{var}(\mathbf{X})$ en el caso que $\mathbf{X}$ sea aleatorio, y en el caso que se condicione en $\mathbf{X}$ o sea fijo, $\mathbf{\Sigma}_{\mathbf{X}} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^T$. Por lo tanto, la distribución asintótica de $\widehat{\mathbf{\Gamma}}_1$ está dado por

$$\mathbf{V}_{\mathbf{\Gamma}_1} = \left(\mathbf{\Sigma}_{\mathbf{X}}\otimes\mathbf{\Sigma} - \left(\mathbf{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\otimes\mathbf{\Sigma}\right)\mathbf{D}_r\left(\mathbf{D}_r^T\left(\left(\boldsymbol{\beta}\mathbf{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T+\frac{1}{2}\mathbf{\Sigma}\right)\otimes\mathbf{\Sigma}\right)\mathbf{D}_r\right)^{-1}\mathbf{D}_r^T\left(\boldsymbol{\beta}^T\mathbf{\Sigma}_{\mathbf{X}}\otimes\mathbf{\Sigma}\right)\right)^{-1}.$$

Para obtener una forma explícita de la inversa de la ecuación anterior, aplicamos la identidad de Woodbury (ver Apéndice A). Así,

$$\begin{aligned}
\mathbf{V}_{\boldsymbol{\Gamma}_1} &= \left(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right) - \left(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right)\left(\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T \otimes \boldsymbol{\Sigma}\right)\mathbf{D}_r \\
&\quad \left[-\mathbf{D}_r^T\left(\left(\boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T + \frac{1}{2}\boldsymbol{\Sigma}\right)\otimes\boldsymbol{\Sigma}\right)\mathbf{D}_r + \mathbf{D}_r^T\left(\boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\boldsymbol{\Sigma}\right)\left(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\left(\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\otimes\boldsymbol{\Sigma}\right)\mathbf{D}_r\right]^{-1} \\
&\quad \mathbf{D}_r^T\left(\boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\boldsymbol{\Sigma}\right)\left(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right) \\
&= \left(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right) - \left(\boldsymbol{\beta}^T\otimes\mathbf{I}_r\right)\mathbf{D}_r\left[\mathbf{D}_r^T\left(-\frac{1}{2}\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}\right)\mathbf{D}_r\right]^{-1}\mathbf{D}_r^T\left(\boldsymbol{\beta}\otimes\mathbf{I}_r\right) \\
&= \left(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right) + 2\left(\boldsymbol{\beta}^T\otimes\mathbf{I}_{r^2}\right)\mathbf{D}_r\mathbf{E}_r\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathbf{E}_r^T\mathbf{D}_r^T\left(\boldsymbol{\beta}\otimes\mathbf{I}_r\right) \\
&= \left(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right) + 2\left(\boldsymbol{\beta}^T\otimes\mathbf{I}_r\right)\frac{1}{2}(\mathbf{I}_r + \mathbf{K}_{rr})\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\left(\boldsymbol{\beta}\otimes\mathbf{I}_r\right) \\
&= \left(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right) + \left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\otimes\boldsymbol{\Sigma}^{-1}\right) + \left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right)\mathbf{K}_{rp}.
\end{aligned}$$

En las últimas tres igualdades hemos usado las Propiedades A.3 y A.4 de las matrices de expansión, duplicación y conmutación que se encuentran en el Apéndice A.

□

Prueba de (1.22): Calculamos ahora la varianza asintótica de $\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\beta}}$ usando que $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})$ tienen la distribución asintótica dada en (1.3) y aplicando la regla de Cramer. Teniendo en cuenta las expresiones $\text{vec}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}) = (\boldsymbol{\beta}^T \otimes \mathbf{I}_r)\text{vec}(\boldsymbol{\Sigma}^{-1}) = (\mathbf{I}_p \otimes \boldsymbol{\Sigma}^{-1})\text{vec}(\boldsymbol{\beta})$, obtenemos las siguientes derivadas

$$\frac{\partial \text{vec}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta})}{\partial \text{vec}^T\boldsymbol{\beta}} = \mathbf{I}_p \otimes \boldsymbol{\Sigma}^{-1} \qquad \frac{\partial \text{vec}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta})}{\partial \text{vech}^T\boldsymbol{\Sigma}} = -(\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\mathbf{D}_r.$$

Por lo tanto, $\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\beta}}$ es asintóticamente normal con media $\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}$ y varianza dada por

$$\begin{aligned}
\mathbf{V}_{\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}} &= \left(\mathbf{I}_p\otimes\boldsymbol{\Sigma}^{-1}, -(\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\mathbf{D}_r\right)\begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma} & 0\\ 0 & 2\mathbf{E}_r(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\mathbf{E}_r^T\end{pmatrix} \\
&\quad \begin{pmatrix}\mathbf{I}_p\otimes\boldsymbol{\Sigma}^{-1}\\ -\mathbf{D}_r^T(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\otimes\boldsymbol{\Sigma}^{-1})\end{pmatrix} \\
&= \left(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right) + 2\left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathbf{D}_r\mathbf{E}_r\left(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}\right)\mathbf{E}_r^T\mathbf{D}_r^T\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\otimes\boldsymbol{\Sigma}^{-1}\right) \\
&= \left(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right) + 2\left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\frac{1}{2}(\mathbf{I}_{r^2}+\mathbf{K}_{rr})\left(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}\right)\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\otimes\boldsymbol{\Sigma}^{-1}\right) \\
&= \left(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right) + \left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\left(\boldsymbol{\beta}\otimes\mathbf{I}_r\right) + \left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathbf{K}_{rr}(\boldsymbol{\beta}\otimes\mathbf{I}_r) \\
&= \left(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right) + \left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\otimes\boldsymbol{\Sigma}^{-1}\right) + \left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right)\mathbf{K}_{rp}.
\end{aligned}$$

□

CAPÍTULO 2

REGRESIÓN LINEAL MULTIVARIADA DE RANGO REDUCIDO

En el modelo usual de regresión lineal multivariada, que relaciona un conjunto de r respuestas con un conjunto de p variables predictoras, se asume implícitamente que la matriz de coeficientes $r \times p$ es de rango completo. El estimador de mínimos cuadrados o máxima verosimilitud de la matriz de coeficientes es el mismo que si se realiza cada una de las r regresiones individuales. Por lo tanto, aunque posiblemente las variables respuestas estén correlacionadas, esta información no está contemplada en la estimación de los coeficientes de regresión. Hay dos cuestiones prácticas con respecto a este modelo de regresión multivariado. En primer lugar, la estimación precisa de todos los coeficientes de regresión puede requerir un número relativamente grande de observaciones y en segundo lugar, incluso si se dispone de los datos, la interpretación simultánea de los coeficientes de regresión puede llegar a ser difícil de manejar. Una forma de abordar estas cuestiones es a través de la suposición de que el rango de matriz de coeficientes pueda ser deficiente. Tales modelos se denominan modelos de regresión de rango reducido y han sido un aporte importante en la teoría clásica del análisis multivariado. La principal diferencia con el modelo de regesión lineal multivariado (1.1) es la restricción adicional en el rango de la matriz de coeficientes $\boldsymbol{\beta}$, de esta forma es posible obtener estimadores sin requerir un gran tamaño de muestra y son más eficientes si el modelo es cierto. El trabajo [**Anderson, 1951**] fue el primero en considerar en detalle el problema de regresión con rango reducido. Luego, en [**Izenman, 1975**] se introduce el término *Reduced rank regression* (RRR, de las siglas en inglés) y se examina este modelo en detalle. En este capítulo se repasan los resultados más importantes de este modelo, se proponen otros estimadores de rango reducido y se comparan sus eficiencias

asintóticas de forma análitica. En el Capítulo 4 se mostrarán estos resultados por medio de simulaciones.

2.1. Estructura del modelo de regresión de rango reducido

Los estimadores de máxima verosimilitud o mínimos cuadrados obtenidos del modelo clásico de regresión lineal multivariada (1.1), $\widehat{\boldsymbol{\alpha}}_{OLS}$, $\widehat{\boldsymbol{\beta}}_{OLS}$ y $\widehat{\boldsymbol{\Sigma}}_{OLS}$, no se ven modificados si las variables respuestas $\mathbf{Y}$ están correlacionadas o no. El número de parámetros en la matriz de coeficientes de regresión puede ser muy grande aún cuando es moderado el número de variables cuyas relaciones se quieren estudiar. De esta manera, en situaciones prácticas es necesario reducir el número de parámetros en el modelo. Existe la idea general de que la estimación de $\boldsymbol{\beta}$ puede optimizarse reduciendo la dimensión de $\mathbf{X}$ e $\mathbf{Y}$ y el modelo de rango reducido es uno de los métodos más conocido para esto. Dicho modelo asume que $\text{rank}(\boldsymbol{\beta}) = d \leq \min(p, r)$, pudiendo expresar $\boldsymbol{\beta}$ como

$$\boldsymbol{\beta} = \mathbf{AB} \text{ con } \mathbf{A} : r \times d \text{ y } \mathbf{B} : d \times p, \tag{2.1}$$

donde $\mathbf{A}$ y $\mathbf{B}$ son ambas de rango completo d. Los estimadores de máxima verosimilitud para este modelo fueron estudiados por **Anderson, 1951**, **Izenman, 1975**, **Reinsel and Velu, 1998** bajo diferentes restricciones sobre $\mathbf{A}$ y $\mathbf{B}$ para obtener identificabilidad, ya que la descomposición $\boldsymbol{\beta} = \mathbf{AB}$ no es única. Sin embargo, notemos que el parámetro de interés $\boldsymbol{\beta}$ es identificable en el sentido que $\text{span}(\mathbf{A}) = \text{span}(\boldsymbol{\beta})$ y $\text{span}(\mathbf{B}^T) = \text{span}(\boldsymbol{\beta}^T)$ son únicos.

Supongamos que contamos con la muestra $(\mathbf{y}_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ que siguen el modelo (1.1) con $\text{rank}(\boldsymbol{\beta}) = d \leq \text{mín}(p, r)$. Luego, la función objetivo a maximizar es el logaritmo de verosimilitud que para estos datos se puede expresar de la siguiente forma:

$$\mathcal{L}_n(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \simeq -\frac{n}{2}\left\{\log|\boldsymbol{\Sigma}| + \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\alpha} - \boldsymbol{\beta}\mathbf{x}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\alpha} - \boldsymbol{\beta}\mathbf{x}_i)\right\} \tag{2.2}$$

que es maximizada bajo la restricción $\text{rank}(\boldsymbol{\beta}) = d$ para obtener los estimadores de máxima verosimilitud (MLE) $\widehat{\boldsymbol{\alpha}}_{RR}$, $\widehat{\boldsymbol{\beta}}_{RR}$ y $\widehat{\boldsymbol{\Sigma}}_{RR}$ del modelo de rango reducido. Dichos estimadores se presentan en el siguiente resultado (**Anderson, 1999** y **Cook et al., 2015**).

Proposición 2.1. *Los estimadores de máxima verosimilitud* $\widehat{\boldsymbol{\alpha}}_{RR}$, $\widehat{\boldsymbol{\beta}}_{RR}$ *y* $\widehat{\boldsymbol{\Sigma}}_{RR}$ *para* $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ *y* $\boldsymbol{\Sigma}$ *respectivamente, a partir de maximizar la función de verosimilitud* L_n *en (2.2) están dados por*

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{RR} &= \overline{\mathbf{y}} \in \mathbb{R}^r \\ \widehat{\boldsymbol{\beta}}_{RR} &= S_{\mathbf{y}}^{1/2} C_{\mathbf{yx}}^{(d)} S_{\mathbf{x}}^{-1/2} \in \mathbb{R}^{r\times p} \\ \widehat{\boldsymbol{\Sigma}}_{RR} &= S_{\mathbf{y}} - \widehat{\boldsymbol{\beta}}_{RR} S_{\mathbf{xy}} \in \mathbb{R}^{r\times r}\end{aligned}$$

donde

$$\begin{aligned}C_{\mathbf{yx}} &= S_{\mathbf{y}}^{-1/2} S_{\mathbf{yx}} S_{\mathbf{x}}^{-1/2} = S_{\mathbf{y}}^{-1/2} \widehat{\boldsymbol{\beta}}_{OLS} S_{\mathbf{x}}^{1/2} \\ C_{\mathbf{yx}}^{(d)} &= C_{\mathbf{yx}} \mathbf{V}_{(d)} \mathbf{V}_{(d)}^T\end{aligned}$$

y $\mathbf{V}_{(d)}$ *es la matriz que contiene los primeros* d *autovectores de* $S^T S$ *donde*

$$S = S_{\mathbf{y}|\mathbf{x}}^{-1/2} S_{\mathbf{yx}} S_{\mathbf{x}}^{-1/2} = S_{\mathbf{y}|\mathbf{x}}^{-1/2} \widehat{\boldsymbol{\beta}}_{OLS} S_{\mathbf{x}}^{1/2}.$$

Observación 2.2. Si $C_{\mathbf{yx}} = \mathbf{U}^T \boldsymbol{\Lambda} \mathbf{R}$ expresa su descomposición en valores singulares (SVD, de ahora en adelante), las filas de la matriz $\mathbf{R}$ contiene los autovectores de $C_{\mathbf{yx}}^T C_{\mathbf{yx}}$. En el Lema 2.13, que se encuentra al final del capítulo, mostramos que dichos autovectores coinciden con los autovectores de $S^T S$, es decir que $\mathbf{V}_{(d)} = \mathbf{R}_{(d)}$. Entonces si descomponemos las matrices $\mathbf{U}$ y $\boldsymbol{\Lambda}$ de la forma

$$\mathbf{U}^T = \begin{pmatrix} \mathbf{U}_{(d)} & \mathbf{U}_{(r-d)} \end{pmatrix} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_1 & 0 \\ 0 & \boldsymbol{\Lambda}_0 \end{pmatrix}$$

con $\mathbf{U}_{(d)} : r \times d$, $\mathbf{U}_{(r-d)} : r \times (r-d)$, $\boldsymbol{\Lambda}_1 : d \times d$ y $\boldsymbol{\Lambda}_0 : (r-d) \times (p-d)$ tenemos que

$$\begin{aligned}C_{\mathbf{yx}}^{(d)} &= \mathbf{U}^T \boldsymbol{\Lambda} \mathbf{R} \mathbf{V}_{(d)} \mathbf{V}_{(d)}^T = \mathbf{U}^T \boldsymbol{\Lambda} \mathbf{R} \mathbf{R}_{(d)} \mathbf{R}_{(d)}^T \\ &= \mathbf{U}^T \boldsymbol{\Lambda} \begin{pmatrix} \mathbf{I}_d \\ \mathbf{0} \end{pmatrix} \mathbf{R}_{(d)}^T = \mathbf{U}^T \boldsymbol{\Lambda} \begin{pmatrix} \mathbf{R}_{(d)}^T \\ \mathbf{0} \end{pmatrix} \\ &= \mathbf{U}_{(d)} \boldsymbol{\Lambda}_1 \mathbf{R}_{(d)}^T. \end{aligned} \tag{2.3}$$

Es decir que, de acuerdo a la Proposición 2.1, el estimador de máxima verosimilitud de $\boldsymbol{\beta}$ bajo el modelo de rango reducido consiste en:

1. Premultiplicar y posmultiplicar el estimador OLS por $S_{\mathbf{y}}^{-1/2}$ y $S_{\mathbf{x}}^{1/2}$ respectivamente, lo que nos da $C_{\mathbf{yx}}$.
2. Obtener la SVD de $C_{\mathbf{yx}}$ y truncarla a los primeros d vectores y valores singulares como en (2.3). Denotamos a esta operación como $C_{\mathbf{yx}}^{(d)}$.

3. Volver a la escala original premultiplicando y posmultiplicando $C_{\mathbf{yx}}^{(d)}$ por $S_{\mathbf{y}}^{1/2}$ y $S_{\mathbf{x}}^{-1/2}$.

Observación 2.3. Si $d = min(p, r)$ se tiene que $C_{\mathbf{yx}}^{(d)} = C_{\mathbf{yx}}$, de donde se obtienen los estimadores de máxima verosimilitud usuales para $\boldsymbol{\beta}$ y $\boldsymbol{\Sigma}$ dados en (1.2):

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{OLS} &= S_{\mathbf{y}}^{1/2} C_{\mathbf{yx}}^{(d)} S_{\mathbf{x}}^{-1/2} = S_{\mathbf{y}}^{1/2} C_{\mathbf{yx}} S_{\mathbf{x}}^{-1/2} = S_{\mathbf{y}}^{1/2} S_{\mathbf{y}}^{-1/2} S_{\mathbf{yx}} S_{\mathbf{x}}^{-1/2} S_{\mathbf{x}}^{-1/2} = S_{\mathbf{yx}} S_{\mathbf{x}}^{-1} \\ \widehat{\boldsymbol{\Sigma}}_{OLS} &= S_{\mathbf{y}} - \widehat{\boldsymbol{\beta}}_{OLS} S_{\mathbf{xy}} = S_{\mathbf{y}} - S_{\mathbf{yx}} S_{\mathbf{x}}^{-1} S_{\mathbf{xy}} = S_{\mathbf{y}|\mathbf{x}}.\end{aligned}$$

2.2. Distribución asintótica del estimador de máxima verosimilitud

La distribución asintótica de estos estimadores fueron estudiados por **[Anderson, 1999]**, **[Reinsel and Velu, 1998]**, **[Stoica and Viberg, 1996]** y **[Cook et al., 2015]**. Repasaremos acá la prueba presentada en este último trabajo ya que es independiente de la parametrización usada, es decir no depende de qué manera se descompuso $\boldsymbol{\beta}$ en $\mathbf{A}$ y $\mathbf{B}$. Además, usaremos estos resultados cuando probemos la distribución asintótica en el caso de GLMM de rango reducido en el Capítulo 3.

Para ello, definimos los parámetros $\mathbf{g}$ del modelo estandar (1.1) sin restricción y $\boldsymbol{\theta}$ para el modelo de rango reducido (1.1) con (2.1),

$$\mathbf{g} = \begin{pmatrix} \text{vec}(\boldsymbol{\beta}) \\ \text{vech}(\boldsymbol{\Sigma}) \end{pmatrix} \qquad \boldsymbol{\theta} = \begin{pmatrix} \text{vec}(\mathbf{A}) \\ \text{vec}(\mathbf{B}) \\ \text{vech}(\boldsymbol{\Sigma}) \end{pmatrix}.$$

El parámetro $\boldsymbol{\alpha}$ es omitido pues su estimador toma la misma forma en ambos modelos: $\widehat{\boldsymbol{\alpha}}_{RR/OLS} = \overline{\mathbf{y}}$. Bajo el modelo de regresión de rango reducido tenemos $\mathbf{g} = \mathbf{g}(\boldsymbol{\theta})$ y por lo tanto su correspondiente estimador de máxima verosimilitud es $\widehat{\mathbf{g}}_{RR} = \mathbf{g}(\widehat{\boldsymbol{\theta}})$. Además, denotamos a los estimadores de máxima verosimilitud del modelo de rango completo como $\widehat{\mathbf{g}}_{OLS} = (\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS}), \text{vech}(\widehat{\boldsymbol{\Sigma}}_{OLS}))^T$.

La matriz de información de Fisher para $\mathbf{g}$ es:

$$J_{\mathbf{g}} = \begin{pmatrix} J_{\boldsymbol{\beta}} & 0 \\ 0 & J_{\boldsymbol{\Sigma}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}^{-1} & 0 \\ 0 & \frac{1}{2}\mathbf{D}_r^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{D}_r \end{pmatrix}$$

y definimos $\boldsymbol{\Delta}$ como el gradiente $\boldsymbol{\Delta} = \dfrac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. Luego las varianzas asintóticas para los estimadores $\widehat{\mathbf{g}}_{OLS}$ y $\widehat{\mathbf{g}}_{RR}$, y en particular para $\widehat{\boldsymbol{\beta}}_{OLS}$ y $\widehat{\boldsymbol{\beta}}_{RR}$, son resumidos en la siguiente proposición:

Proposición 2.4. *Asumiendo que* $\boldsymbol{\varepsilon}|\mathbf{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ *y que* $\text{rank}(\boldsymbol{\beta}) = d$,

$$\text{avar}(\sqrt{n}\widehat{\mathbf{g}}_{OLS}) = J_{\mathbf{g}}^{-1}$$

$$\text{avar}(\sqrt{n}\widehat{\mathbf{g}}_{RR}) = \boldsymbol{\Delta}(\boldsymbol{\Delta}^T J_{\mathbf{g}} \boldsymbol{\Delta})^{\ddagger} \boldsymbol{\Delta}^T$$

donde el supraíndice ‡ *indica una inversa generalizada de la matriz. En particular,* $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})$ *y* $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{RR} - \boldsymbol{\beta})$ *son ambos asintóticamente normales con media cero y las siguientes matrices de covarianzas*

$$\text{avar}\{\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS})\} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma} \tag{2.4}$$

$$\text{avar}\{\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{RR})\} = (\mathbf{P}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})} \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}) + (\mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})} \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})} \boldsymbol{\Sigma}). \tag{2.5}$$

Observación 2.5. Notar que entre las matrices (2.4) y (2.5) existe la siguiente relación:

$$(\mathbf{P}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})} \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}) + (\mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{X}})} \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})} \boldsymbol{\Sigma}) \leq \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}.$$

Este orden en las matrices de covarianza expresa que si estamos bajo las hipótesis del modelo (1.1) con (2.1), el estimador $\widehat{\boldsymbol{\beta}}_{RR}$ es más eficiente asintóticamente que $\widehat{\boldsymbol{\beta}}_{OLS}$. Además, notar que la matriz de covarianza asinótica (2.5) depende de $\mathbf{A}$ y $\mathbf{B}$ solo a través de proyecciones. Es decir que no depende de la descomposición de $\boldsymbol{\beta}$ en $\mathbf{A}$ y $\mathbf{B}$ que se ha considerado, teniendo en cuenta que esta descomposición no es única.

La demostración de la Proposición 2.4 se encuentra en el trabajo [**Cook et al., 2015**] y se basa en un teorema de [**Shapiro, 1986**], que reproducimos en el Teorema 2.7 ya que lo utilizaremos en el capítulo siguiente. Previamente consideramos la siguiente definición:

Definición 2.6. Sea $\boldsymbol{\Theta} \subset \mathbb{R}^q$ un espacio de parámetros abierto y $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$. Sea $\boldsymbol{\Xi}$ el conjunto imagen de $\boldsymbol{\Theta}$ a través de la función $\mathbf{g}$, es decir $\boldsymbol{\Xi} = \{\boldsymbol{\xi} : \boldsymbol{\xi} = \mathbf{g}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. El punto $\boldsymbol{\theta}_0$ es regular si

1. la matriz jacobiana $\boldsymbol{\Delta} = \partial \mathbf{g}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^T$ tiene el mismo rango que $\boldsymbol{\Delta}_0 = \partial \mathbf{g}(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}^T$ para todo $\boldsymbol{\theta}$ en un entorno de $\boldsymbol{\theta}_0$.
2. Existen los entornos $\mathcal{U}$ y $\mathcal{V}$ de $\boldsymbol{\theta}_0$ y $\boldsymbol{\xi}_0 = g(\boldsymbol{\theta}_0)$ respectivamente, tal que $\boldsymbol{\Xi} \cap \mathcal{V} = \mathbf{g}(\mathcal{U})$.

Teorema 2.7. *Supongamos que* $\boldsymbol{\theta}$ *es un vector de parámetros de dimensión* q, $\boldsymbol{\theta} \in \Theta$ *con* Θ *un conjunto abierto y conexo de* $\mathbb{R}^q$. *Sea* $\boldsymbol{\theta}_0$ *el valor poblacional de* $\boldsymbol{\theta}$. *Definimos* $\boldsymbol{\xi} = \mathbf{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \ldots, g_m(\boldsymbol{\theta}))^T : \Theta \to \mathbb{R}^m$, *donde* $g_i(\boldsymbol{\theta})$ *es dos veces continuamente diferenciable en* Θ,

$i = 1, \ldots, m$. Consideramos la matriz jacobiana de dimensiones $m \times q$, $\boldsymbol{\Delta} = \dfrac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ que no es necesariamente de rango completo. Además suponemos que:

1. *$\boldsymbol{\tau}_n$ es un estimador del valor poblacional $\mathbf{g}(\boldsymbol{\theta}_0) \in \mathbb{R}^m$ asintóticamente normal. Es decir, $\sqrt{n}(\boldsymbol{\tau}_n - \mathbf{g}(\boldsymbol{\theta}_0)) \to \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma})$*
2. *contamos con la función $H(\mathbf{a}, \mathbf{b})$ que para $\mathbf{a} = \boldsymbol{\tau}_n$ satisface*
   a) *$H(\boldsymbol{\tau}_n, \boldsymbol{\xi}) \geqslant 0$ para todo $\boldsymbol{\xi} \in \mathbf{g}(\Theta)$*
   b) *$H(\boldsymbol{\tau}_n, \boldsymbol{\xi}) = 0$ si y solo si $\boldsymbol{\tau}_n = \boldsymbol{\xi}$*
   c) *$H(\boldsymbol{\tau}_n, \boldsymbol{\xi})$ es al menos dos veces continuamente diferenciable en la segunda variable $\boldsymbol{\xi}$*
   d) *existen $\epsilon > 0$ y $\delta > 0$ tal que si $\|\boldsymbol{\tau}_n - \boldsymbol{\xi}\| > \delta$, se tiene que $H(\boldsymbol{\tau}_n, \boldsymbol{\xi}) \geqslant \epsilon$, donde $\|\cdot\|$ es la norma Euclídea.*

   Llamamos $\mathbf{V} = \dfrac{1}{2}\dfrac{\partial^2 H}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^T}(\mathbf{g}(\boldsymbol{\theta}_0), \mathbf{g}(\boldsymbol{\theta}_0))$.
3. *el punto $\boldsymbol{\theta}_0$ es regular*
4. $\operatorname{rank}(\boldsymbol{\Delta}) = \operatorname{rank}(\boldsymbol{\Delta}^T \mathbf{V} \boldsymbol{\Delta})$.

Luego, el estimador $\mathbf{g}(\widehat{\boldsymbol{\theta}})$ que minimiza la función de discrepancia H sobre Θ, es un estimador consistente de $\mathbf{g}(\boldsymbol{\theta}_0)$ y $\sqrt{n}(\mathbf{g}(\widehat{\boldsymbol{\theta}}) - \mathbf{g}(\boldsymbol{\theta}_0))$ es asintóticamente normal con media $\boldsymbol{0}$ y matriz de covarianza $\mathbf{P}_{\Delta(\mathbf{V})} \boldsymbol{\Gamma} \mathbf{P}_{\Delta(\mathbf{V})}^T$.

Demostración de la Proposición 2.4: Por propiedades de los estimadores de máxima verosimilitud la varianza asintótica de $\widehat{\mathbf{g}}_{OLS}$ es $J_{\mathbf{g}}^{-1}$. En particular, como $J_{\mathbf{g}}$ es diagonal en bloque, la matriz de coeficientes $\widehat{\boldsymbol{\beta}}_{OLS}$ tiene como varianza asintótica $J_{\boldsymbol{\beta}}^{-1} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}$.

Por otro lado, vamos a aplicar el Teorema 2.7 para obtener la distribución asintótica de los estimadores del modelo de rango reducido. Observemos que en este contexto tenemos el vector de parámetros $\boldsymbol{\theta} = (\operatorname{vec}(\mathbf{A}), \operatorname{vec}(\mathbf{B}), \operatorname{vech}(\boldsymbol{\Sigma})) : q \times 1$, con $q = (p + r)d + r(r + 1)/2$ que vive en el subespacio abierto $\Theta \subset \mathbb{R}^q$ definido a partir de los conjuntos abiertos $\Pi^{m \times n}$ y S_+^m tal que $\Theta = \operatorname{vec}(\Pi^{r \times d}) \times \operatorname{vec}(\Pi^{d \times p}) \times S_+^r$. El valor poblacional $\boldsymbol{\theta}_0$ es el valor poblacional de $(\operatorname{vec}(\mathbf{A})^T, \operatorname{vec}(\mathbf{B})^T, \operatorname{vec}(\boldsymbol{\Sigma})^T)^T$, el cual denotamos de la forma $(\operatorname{vec}(\mathbf{A}_0)^T, \operatorname{vec}(\mathbf{B}_0)^T, \operatorname{vec}(\boldsymbol{\Sigma}_0)^T)^T$.

La función $\mathbf{g} : \Theta \to \mathbb{R}^{pr + r(r+1)/2}$ es tal que

$$\mathbf{g}(\boldsymbol{\theta}) = \mathbf{g}\begin{pmatrix} \operatorname{vec}(\mathbf{A}) \\ \operatorname{vec}(\mathbf{B}) \\ \operatorname{vech}(\boldsymbol{\Sigma}) \end{pmatrix} = \begin{pmatrix} \operatorname{vec}(\mathbf{AB}) \\ \operatorname{vech}(\boldsymbol{\Sigma}) \end{pmatrix} = \begin{pmatrix} \operatorname{vec}(\boldsymbol{\beta}) \\ \operatorname{vech}(\boldsymbol{\Sigma}) \end{pmatrix}.$$

La matriz $\boldsymbol{\Delta}$ de dimensiones $pr + r(r+1)/2 \times (p+r)d + r(r+1)/2$ es

$$\boldsymbol{\Delta} = \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \begin{pmatrix} \mathbf{B}_0^T \otimes \mathbf{I}_r & \mathbf{I}_p \otimes \mathbf{A}_0 & 0 \\ 0 & 0 & \mathbf{I}_{r(r+1)/2} \end{pmatrix}.$$

En este contexto, τ_n es el estimador del modelo completo $\tau_n = (\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS})^T, \text{vech}(\widehat{\boldsymbol{\Sigma}}_{OLS})^T)^T$ del cual conocemos su distribución asintótica, es decir que $\boldsymbol{\Gamma} = J_{\mathbf{g}}^{-1}$.

Definimos la siguiente función:

$$F(\mathbf{g}, \widehat{\mathbf{g}}_{OLS}) = \frac{2}{n}\left\{\mathcal{L}_n(\widehat{\boldsymbol{\beta}}_{OLS}, \widehat{\boldsymbol{\Sigma}}_{OLS}) - \mathcal{L}_n(\boldsymbol{\beta}, \boldsymbol{\Sigma})\right\},$$

donde $\mathcal{L}_n$ es la función objetivo de máxima verosimilitud de $\mathbf{Y}$ dado $\mathbf{X}$ en el modelo (1.1) bajo normalidad de los errores. $F(\mathbf{g}, \widehat{\mathbf{g}}_{OLS})$ satisface los puntos $a - d$ del enunciado del teorema y en este caso $\mathbf{V} = J_{\mathbf{g}}$. La condición 4 de la igualdad de los rangos se satisface pues $\text{rank}(\boldsymbol{\Delta}) = d(p+r) - d^2 + (r+1)r/2$ y al ser $\mathbf{V} = J_{\mathbf{g}}$ de rango completo, tenemos que $\text{rank}(\boldsymbol{\Delta}^T J_{\mathbf{g}} \boldsymbol{\Delta}) = d(p+r) - d^2 + (r+1)r/2$.

Es claro que el estimador $\mathbf{g}(\widehat{\boldsymbol{\theta}})$ que minimiza la función $F(\mathbf{g}, \widehat{\mathbf{g}}_{OLS})$ es el estimador de máxima verosimilitud bajo el modelo de rango reducido y de acuerdo al Teorema 2.7 la varianza asintótica de $\mathbf{g}(\widehat{\theta}) = \widehat{\mathbf{g}}_{RR}$ es

$$\begin{aligned} \mathbf{P}_{\Delta(\mathbf{V})} \boldsymbol{\Gamma} \mathbf{P}_{\Delta(\mathbf{V})}^T &= \boldsymbol{\Delta}(\boldsymbol{\Delta}^T \mathbf{V} \boldsymbol{\Delta})^\dagger \boldsymbol{\Delta}^T \mathbf{V} \boldsymbol{\Gamma} \mathbf{V} \boldsymbol{\Delta} (\boldsymbol{\Delta}^T \mathbf{V} \boldsymbol{\Delta})^\dagger \boldsymbol{\Delta}^T \\ &= \boldsymbol{\Delta}(\boldsymbol{\Delta}^T J_{\mathbf{g}} \boldsymbol{\Delta})^\dagger \boldsymbol{\Delta}^T J_{\mathbf{g}} J_{\mathbf{g}}^{-1} J_{\mathbf{g}} \boldsymbol{\Delta} (\boldsymbol{\Delta}^T J_{\mathbf{g}} \boldsymbol{\Delta})^\dagger \boldsymbol{\Delta}^T \\ &= \boldsymbol{\Delta}(\boldsymbol{\Delta}^T J_{\mathbf{g}} \boldsymbol{\Delta})^\dagger \boldsymbol{\Delta}^T. \end{aligned} \tag{2.6}$$

En particular, estamos interesados en obtener la parte de la matriz de covarianza asintótica (2.6) que corresponde al estimador de máxima verosimilitud de $\boldsymbol{\beta}$. Para ello, en primer lugar vamos a aplicar el Lema A.10 del Apéndice que nos dice que para cualquier matriz $\mathbf{M}$ y para cualquier inversa generalizada vale que

$$\mathbf{M}(\mathbf{M}^T\mathbf{M})^\dagger \mathbf{M}^T = \mathbf{M}(\mathbf{M}^T\mathbf{M})^\ddagger \mathbf{M}^T,$$

de donde se deduce fácilmente que también es cierto para cualquier matriz $\mathbf{W} > 0$,

$$\mathbf{M}(\mathbf{M}^T\mathbf{W}\mathbf{M})^\dagger \mathbf{M}^T = \mathbf{M}(\mathbf{M}^T\mathbf{W}\mathbf{M})^\ddagger \mathbf{M}^T.$$

Es decir, que podemos usar cualquier inversa generalizada en (2.6). En las cuentas que siguen, suprimimos el subindice 0 en los valores poblacionales pero es claro que estamos tratando con

los valores poblacionales. El factor central de (2.6) es

$$\begin{aligned}\boldsymbol{\Delta}^T J_{\mathbf{g}} \boldsymbol{\Delta} &= \begin{pmatrix} \mathbf{B}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{B}^T \otimes \boldsymbol{\Sigma}^{-1} & \mathbf{B}\boldsymbol{\Sigma}_{\mathbf{x}} \otimes \boldsymbol{\Sigma}^{-1}\mathbf{A} & 0 \\ \boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{B}^T \otimes \mathbf{A}^T\boldsymbol{\Sigma}^{-1} & \boldsymbol{\Sigma}_{\mathbf{x}} \otimes \mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} & 0 \\ 0 & 0 & \frac{1}{2}\mathbf{D}_r^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{D}_r \end{pmatrix} \\ &\doteq \begin{pmatrix} \Delta_1 & 0 \\ 0 & \Delta_2 \end{pmatrix},\end{aligned}$$

donde en la última igualdad introducimos notación para cada bloque diagonal de la matriz. Es decir,

$$\Delta_1 = \begin{pmatrix} \mathbf{B}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{B}^T \otimes \boldsymbol{\Sigma}^{-1} & \mathbf{B}\boldsymbol{\Sigma}_{\mathbf{x}} \otimes \boldsymbol{\Sigma}^{-1}\mathbf{A} \\ \boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{B}^T \otimes \mathbf{A}^T\boldsymbol{\Sigma}^{-1} & \boldsymbol{\Sigma}_{\mathbf{x}} \otimes \mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A} \end{pmatrix} \text{ y } \Delta_2 = \mathbf{D}_r^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{D}_r.$$

Luego, observemos que si $\Delta_1^\ddagger$ y $\Delta_2^\ddagger$ indican una inversa generalizada de Δ_1 y Δ_2 entonces

$$(\boldsymbol{\Delta}^T J_{\mathbf{g}} \boldsymbol{\Delta})^\ddagger = \begin{pmatrix} \Delta_1 & 0 \\ 0 & \Delta_2 \end{pmatrix}^\ddagger = \begin{pmatrix} \Delta_1^\ddagger & 0 \\ 0 & \Delta_2^\ddagger \end{pmatrix}$$

es una inversa generalizada de $\boldsymbol{\Delta}^T J_{\mathbf{g}} \boldsymbol{\Delta}$. Como estamos interesados en obtener el bloque diagonal superior de la expresión (2.6) y por la estructura diagonal en bloque de $\boldsymbol{\Delta}$ y de $\boldsymbol{\Delta}^T J_{\mathbf{g}} \boldsymbol{\Delta}$ nos queda que

$$\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{RR})) = \begin{pmatrix} \mathbf{B}^T \otimes \mathbf{I} & \mathbf{I} \otimes \mathbf{A} \end{pmatrix} \Delta_1^\ddagger \begin{pmatrix} \mathbf{B} \otimes \mathbf{I} \\ \mathbf{I} \otimes \mathbf{A}^T \end{pmatrix}.$$

Basándonos en el Lema A.8, proponemos la siguiente inversa generalizada de Δ_1

$$\Delta_1^\ddagger = \begin{pmatrix} (\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{B}^T)^{-1} \otimes (\boldsymbol{\Sigma} - \mathbf{A}(\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}^T) & 0 \\ 0 & \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1} \end{pmatrix}$$

la cual verifica que $\Delta_1\Delta_1^\ddagger\Delta_1 = \Delta_1$.

Por lo tanto,

$$\begin{aligned}
\text{avar}(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{RR})) &= (\mathbf{B}^T \otimes \mathbf{I})((\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{B}^T)^{-1}) \otimes (\boldsymbol{\Sigma} - \mathbf{A}(\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}^T))(\mathbf{B} \otimes \mathbf{I}) \\
&\quad +(\mathbf{I} \otimes \mathbf{A})(\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1})(\mathbf{I} \otimes \mathbf{A}^T) \\
&= (\mathbf{B}^T(\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{B}^T)^{-1}\mathbf{B} \otimes (\boldsymbol{\Sigma} - \mathbf{A}(\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}^T)) \\
&\quad +(\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{A}(\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}^T) \\
&= (\mathbf{P}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \boldsymbol{\Sigma}) + (\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})}\boldsymbol{\Sigma}) - (\mathbf{P}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})}\boldsymbol{\Sigma}) \\
&= (\mathbf{P}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \boldsymbol{\Sigma}) + ((\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} - \mathbf{P}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}) \otimes \mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})}\boldsymbol{\Sigma}) \\
&= (\mathbf{P}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \boldsymbol{\Sigma}) + (\mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})}\boldsymbol{\Sigma}).
\end{aligned}$$

□

2.3. Otros estimadores de rango reducido

2.3.1. Estimador basado en el Teorema de minimización cuadrática

En el trabajo **Cook and Ni, 2005** se presenta un resultado que establece las condiciones para extender los resultados del Teorema 2.7 a ciertos estimadores hallados vía minimización cuadrática. Utilizando este resultado podemos obtener un nuevo estimador de la matriz de coeficientes para el modelo de regresión lineal de rango reducido que tenga la misma varianza asintótica que el estimador MLE. El siguiente corolario es consecuencia directa del Lema A.3 y A.4 de **Cook and Ni, 2005**.

Corolario 2.8. *Supongamos que se verifican las condiciones del Teorema 2.7 pero* $H(\boldsymbol{\tau}_n, \mathbf{g}(\theta))$ *es de la forma*

$$\left(\boldsymbol{\tau}_n - \mathbf{g}(\boldsymbol{\theta})\right)^T \mathbf{V}_n \left(\boldsymbol{\tau}_n - \mathbf{g}(\boldsymbol{\theta})\right),$$

donde $\{\mathbf{V}_n > 0\}$ *es una sucesión de matrices aleatorias que convergen a* $\mathbf{V} > 0$ *en probabilidad. Entonces la tesis del Teorema 2.7 sigue siendo válida.*

Basándonos en estos resultados, es posible definir otro estimador del modelo de rango reducido a través de la definición de una nueva función de discrepancia. Vamos a considerar la siguiente función

$$F(\mathbf{g}, \widehat{\mathbf{g}}_{OLS}) = (\mathbf{g} - \widehat{\mathbf{g}}_{OLS})^T \widehat{J}_{\mathbf{g}} (\mathbf{g} - \widehat{\mathbf{g}}_{OLS}), \tag{2.7}$$

la cual podemos expresar también de la siguiente forma:

$$\begin{aligned} F_1(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}_{OLS}) + F_2(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_{OLS}) &= (\text{vec}(\boldsymbol{\beta}) - \text{vec}(\widehat{\boldsymbol{\beta}}_{OLS}))^T \left(S_{\mathbf{x}} \otimes S_{\mathbf{y}|\mathbf{x}}^{-1}\right) (\text{vec}(\boldsymbol{\beta}) - \text{vec}(\widehat{\boldsymbol{\beta}}_{OLS})) \\ &\quad + (\text{vech}(\boldsymbol{\Sigma}) - \text{vech}(S_{\mathbf{y}|\mathbf{x}}^{-1}))^T \mathbf{D}_r^T \left(S_{\mathbf{y}|\mathbf{x}}^{-1} \otimes S_{\mathbf{y}|\mathbf{x}}^{-1}\right) \mathbf{D}_r \\ &\quad (\text{vech}(\boldsymbol{\Sigma}) - \text{vech}(S_{\mathbf{y}|\mathbf{x}}^{-1})). \end{aligned}$$

Sea $\mathbf{g}_{RR2} = (\widehat{\boldsymbol{\beta}}_{RR2}, \widehat{\boldsymbol{\Sigma}}_{RR2})$, donde $\widehat{\boldsymbol{\beta}}_{RR2}$ es el valor de $\boldsymbol{\beta}$ que minimiza el primer término de F dentro el conjunto de matrices $r \times p$ de rango $d \leq \text{mín}(p, r)$ y $\widehat{\boldsymbol{\Sigma}}_{RR2}$ es la matriz simétrica y definida positiva que minimiza el segundo término de F. Luego, es posible obtener la distribución asintótica de este nuevo estimador basándonos en el resultado anterior. De esta forma tenemos la siguiente proposición:

Proposición 2.9. *Asumiendo que* $\epsilon|\mathbf{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ *y que* $\text{rank}(\boldsymbol{\beta}) = d$*, luego*

$$\text{avar}(\sqrt{n}\widehat{\mathbf{g}}_{RR2}) = \boldsymbol{\Delta}(\boldsymbol{\Delta}^T J_{\mathbf{g}} \boldsymbol{\Delta})^{\ddagger} \boldsymbol{\Delta}^T$$

donde

$$\boldsymbol{\Delta} = \begin{pmatrix} \mathbf{B}^T \otimes \mathbf{I}_r & \mathbf{I}_p \otimes \mathbf{A} & 0 \\ 0 & 0 & \mathbf{I}_{r(r+1)/2} \end{pmatrix}.$$

En particular, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{RR2} - \boldsymbol{\beta})$ *es asintóticamente normal con media cero y las siguiente matriz de covarianza*

$$\text{avar}\{\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{RR2})\} = (\mathbf{P}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})} \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \boldsymbol{\Sigma}) + (\mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})} \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})} \boldsymbol{\Sigma}). \tag{2.8}$$

Demostración. El resultado es consecuencia directa del Corolario 2.8 pues $S_{\mathbf{x}}$ y $S_{\mathbf{y}|\mathbf{x}}$ son estimadores consistentes de $\boldsymbol{\Sigma}_{\mathbf{X}}$ y $\boldsymbol{\Sigma}$ respectivamente. □

En conclusión, es posible obtener un estimador de rango reducido de forma explícita, tan eficiente como el estimador MLE, a partir de la minimización de una forma cuadrática que involucra el estimador de rango completo $\widehat{\boldsymbol{\beta}}_{OLS}$. Daremos a continuación la forma explícita de este estimador.

Proposición 2.10. *Asumiendo que* $\epsilon|\mathbf{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ *y que* $\text{rank}(\boldsymbol{\beta}) = d$*, el estimador* $\widehat{\boldsymbol{\beta}}_{RR2}$ *es*

$$\widehat{\boldsymbol{\beta}}_{RR2} = S_{\mathbf{y}|\mathbf{x}}^{1/2} \left(S_{\mathbf{y}|\mathbf{x}}^{-1/2} \widehat{\boldsymbol{\beta}}_{OLS} S_{\mathbf{x}}^{1/2}\right)^{(d)} S_{\mathbf{x}}^{-1/2},$$

mientras que $\widehat{\boldsymbol{\Sigma}}_{RR2}$ *coincide con el estimador de máxima verosimilitud de* $\boldsymbol{\Sigma}$ *del modelo de rango completo.*

Demostración. Comenzemos minimizando (2.7) en $\mathbf{B}$, luego

$$\begin{aligned} F_1(\mathbf{A},\mathbf{B},\widehat{\boldsymbol{\beta}}_{OLS}) &= (\text{vec}(\mathbf{AB})-\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS}))^T(S_{\mathbf{x}}\otimes S_{\mathbf{y}|\mathbf{x}}^{-1})(\text{vec}(\mathbf{AB})-\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS})) \\ &= \text{vec}^T(\mathbf{B})(S_{\mathbf{x}}\otimes \mathbf{A}^T S_{\mathbf{y}|\mathbf{x}}^{-1}\mathbf{A})\text{vec}(\mathbf{B}) - 2\text{vec}^T(\mathbf{B})(S_{\mathbf{x}}\otimes \mathbf{A}^T S_{\mathbf{y}|\mathbf{x}}^{-1})\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS}) \\ &\quad +\text{vec}^T(\widehat{\boldsymbol{\beta}}_{OLS})(S_{\mathbf{x}}\otimes S_{\mathbf{y}|\mathbf{x}}^{-1})\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS}). \end{aligned}$$

Como F_1 es una forma cuadrática definida positiva en $\mathbf{B}$ para $\mathbf{A}$ fijo, para obtener el valor donde se minimiza, derivamos con respecto a vec($\mathbf{B}$) y obtenemos que

$$\frac{\partial F(\boldsymbol{\beta},\widehat{\boldsymbol{\beta}}_{OLS})}{\partial \text{vec}(\mathbf{B})} = 2(S_{\mathbf{x}}\otimes \mathbf{A}^T S_{\mathbf{y}|\mathbf{x}}^{-1}\mathbf{A})\text{vec}(\mathbf{B}) - 2(S_{\mathbf{x}}\otimes \mathbf{A}^T S_{\mathbf{y}|\mathbf{x}}^{-1})\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS}).$$

Entonces,

$$\begin{aligned} \text{vec}(\widehat{\mathbf{B}}) &= (S_{\mathbf{x}}\otimes \mathbf{A}^T S_{\mathbf{y}|\mathbf{x}}^{-1}\mathbf{A})^{-1}(S_{\mathbf{x}}\otimes \mathbf{A}^T S_{\mathbf{y}|\mathbf{x}}^{-1})\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS}) \\ &= \text{vec}((\mathbf{A}^T S_{\mathbf{y}|\mathbf{x}}^{-1}\mathbf{A})^{-1}\mathbf{A}^T S_{\mathbf{y}|\mathbf{x}}^{-1}\widehat{\boldsymbol{\beta}}_{OLS}). \end{aligned}$$

Luego, $\mathbf{A}\widehat{\mathbf{B}} = \mathbf{P}_{\mathbf{A}(S_{\mathbf{y}|\mathbf{x}}^{-1})}\widehat{\boldsymbol{\beta}}_{OLS}$. Ahora remplazamos $\widehat{\mathbf{B}}$ en (2.7) y obtenemos F_1 como función de $\mathbf{A}$

$$\begin{aligned} F_1(\mathbf{A},\widehat{\mathbf{B}},\widehat{\boldsymbol{\beta}}_{OLS}) &= (\text{vec}(\mathbf{P}_{\mathbf{A}(S_{\mathbf{y}|\mathbf{x}}^{-1})}\widehat{\boldsymbol{\beta}}_{OLS})-\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS}))^T(S_{\mathbf{x}}\otimes S_{\mathbf{y}|\mathbf{x}}^{-1})(\text{vec}(\mathbf{P}_{\mathbf{A}(S_{\mathbf{y}|\mathbf{x}}^{-1})}\widehat{\boldsymbol{\beta}}_{OLS})-\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS})) \\ &= \text{vec}^T(\widehat{\boldsymbol{\beta}}_{OLS})[(S_{\mathbf{x}}\otimes \mathbf{P}_{\mathbf{A}(S_{\mathbf{y}|\mathbf{x}}^{-1})}^T S_{\mathbf{y}|\mathbf{x}}^{-1}\mathbf{P}_{\mathbf{A}(S_{\mathbf{y}|\mathbf{x}}^{-1})}) - (S_{\mathbf{x}}\otimes \mathbf{P}_{\mathbf{A}(S_{\mathbf{y}|\mathbf{x}}^{-1})}^T S_{\mathbf{y}|\mathbf{x}}^{-1}) \\ &\quad -(S_{\mathbf{x}}\otimes S_{\mathbf{y}|\mathbf{x}}^{-1}\mathbf{P}_{\mathbf{A}(S_{\mathbf{y}|\mathbf{x}}^{-1})}) + (S_{\mathbf{x}}\otimes S_{\mathbf{y}|\mathbf{x}}^{-1})]\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS}) \\ &= \text{vec}^T(\widehat{\boldsymbol{\beta}}_{OLS})[-(S_{\mathbf{x}}\otimes \mathbf{P}_{\mathbf{A}(S_{\mathbf{y}|\mathbf{x}}^{-1})}^T S_{\mathbf{y}|\mathbf{x}}^{-1}) + (S_{\mathbf{x}}\otimes S_{\mathbf{y}|\mathbf{x}}^{-1})]\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS}) \end{aligned}$$

donde en la última igualdad hemos usado que $\mathbf{P}_{\mathbf{A}(S_{\mathbf{y}|\mathbf{x}}^{-1})}^T S_{\mathbf{y}|\mathbf{x}}^{-1}\mathbf{P}_{\mathbf{A}(S_{\mathbf{y}|\mathbf{x}}^{-1})} = S_{\mathbf{y}|\mathbf{x}}^{-1}\mathbf{P}_{\mathbf{A}(S_{\mathbf{y}|\mathbf{x}}^{-1})}$. Entonces, el problema es equivalente a maximizar en $\mathbf{A}$ la siguiente función

$$\begin{aligned} \text{vec}^T(\widehat{\boldsymbol{\beta}}_{OLS})(S_{\mathbf{x}}\otimes \mathbf{P}_{\mathbf{A}(S_{\mathbf{y}|\mathbf{x}}^{-1})}^T S_{\mathbf{y}|\mathbf{x}}^{-1})\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS}) &= \text{vec}^T(\widehat{\boldsymbol{\beta}}_{OLS})\text{vec}(\mathbf{P}_{\mathbf{A}(S_{\mathbf{y}|\mathbf{x}}^{-1})}^T S_{\mathbf{y}|\mathbf{x}}^{-1}\widehat{\boldsymbol{\beta}}_{OLS}S_{\mathbf{x}}) \\ &= \text{tr}(\widehat{\boldsymbol{\beta}}_{OLS}^T S_{\mathbf{y}|\mathbf{x}}^{-1}\mathbf{A}(\mathbf{A}^T S_{\mathbf{y}|\mathbf{x}}^{-1}\mathbf{A})^{-1}\mathbf{A}^T S_{\mathbf{y}|\mathbf{x}}^{-1}\widehat{\boldsymbol{\beta}}_{OLS}S_{\mathbf{x}}) \\ &= \text{tr}(\mathbf{P}_{(\mathbf{A}S_{\mathbf{y}|\mathbf{x}}^{-1/2})}S_{\mathbf{y}|\mathbf{x}}^{-1/2}\widehat{\boldsymbol{\beta}}_{OLS}S_{\mathbf{x}}\widehat{\boldsymbol{\beta}}_{OLS}^T S_{\mathbf{y}|\mathbf{x}}^{-1/2}). \end{aligned}$$

Sea $M = S_{\mathbf{y}|\mathbf{x}}^{-1/2}\widehat{\boldsymbol{\beta}}_{OLS}S_{\mathbf{x}}^{1/2}$, luego

$$\begin{aligned} MM^T &= \left(S_{\mathbf{y}|\mathbf{x}}^{-1/2}\widehat{\boldsymbol{\beta}}_{OLS}S_{\mathbf{x}}^{1/2}\right)\left(S_{\mathbf{y}|\mathbf{x}}^{-1/2}\widehat{\boldsymbol{\beta}}_{OLS}^T S_{\mathbf{x}}^{1/2}\right)^T \\ &= S_{\mathbf{y}|\mathbf{x}}^{-1/2}\widehat{\boldsymbol{\beta}}_{OLS}S_{\mathbf{x}}\widehat{\boldsymbol{\beta}}_{OLS}^T S_{\mathbf{y}|\mathbf{x}}^{-1/2}. \end{aligned}$$

Si $\nu_1, \ldots, \nu_d$ son los primeros autovectores de MM^T y $\mathbf{V}_d = [\nu_1, \ldots, \nu_d] : r \times d$, se tiene que para cualquier matriz $\mathbf{H} : d \times d$ invertible (ver Lema 2.1 de [**Reinsel and Velu, 1998**])

$$\begin{aligned}
\widehat{\mathbf{A}} &= S_{\mathbf{y}|\mathbf{x}}^{1/2}\mathbf{V}_d\mathbf{H}, \\
\widehat{\mathbf{B}} &= (\widehat{\mathbf{A}}^T S_{\mathbf{y}|\mathbf{x}}^{-1}\widehat{\mathbf{A}})^{-1}\widehat{\mathbf{A}}^T S_{\mathbf{y}|\mathbf{x}}^{-1}\widehat{\boldsymbol{\beta}}_{OLS} \\
&= \mathbf{H}^{-1}\mathbf{V}_d^T S_{\mathbf{y}|\mathbf{x}}^{-1/2}\widehat{\boldsymbol{\beta}}_{OLS}.
\end{aligned}$$

Por lo tanto,

$$\begin{aligned}
\widehat{\mathbf{A}}\widehat{\mathbf{B}} &= S_{\mathbf{y}|\mathbf{x}}^{1/2}\mathbf{V}_d\mathbf{V}_d^T S_{\mathbf{y}|\mathbf{x}}^{-1/2}\widehat{\boldsymbol{\beta}}_{OLS} \\
&= S_{\mathbf{y}|\mathbf{x}}^{1/2}\mathbf{V}_d\mathbf{V}_d^T M S_{\mathbf{x}}^{-1} \\
&= S_{\mathbf{y}|\mathbf{x}}^{1/2}\left(S_{\mathbf{y}|\mathbf{x}}^{-1/2}\widehat{\boldsymbol{\beta}}_{OLS}S_{\mathbf{x}}^{1/2}\right)^{(d)} S_{\mathbf{x}}^{-1}.
\end{aligned}$$

Por otro lado, es claro que el valor de $\boldsymbol{\Sigma}$ que minimiza F_2 coincide con $\widehat{\boldsymbol{\Sigma}}_{OLS}$. □

Llamamos a este estimador **de minimización cuadrática** o **sub-d estandarizado**. Esta segunda denominación se debe a que $\widehat{\boldsymbol{\beta}}_{RR2}$ consiste en truncar la descomposición en valores singulares del estimador del modelo de rango completo estandarizado por su matriz de covarianza asintótica estimada y luego de esto, retomar a la escala original multiplicando nuevamente por la raíz cuadrada de la matriz de covarianza asintótica estimada. Los siguientes pasos detallan esta idea:

1. Calcular $(S_{\mathbf{x}}^{1/2} \otimes S_{\mathbf{y}|\mathbf{x}}^{-1/2})\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS})$ y luego reconstruir por columna la matriz de dimensiones $r \times p$ a partir de este vector obtenido. Denominamos a esta matriz como $\widehat{\boldsymbol{\Gamma}}$.
2. Truncar la descomposición en valores singulares la matriz obtenida en el paso anterior. Es decir obtener $\widehat{\boldsymbol{\Gamma}}^{(d)}$.
3. Calcular $(S_{\mathbf{x}}^{-1/2} \otimes S_{\mathbf{y}|\mathbf{x}}^{1/2})\text{vec}(\widehat{\boldsymbol{\Gamma}}^{(d)})$ y luego reconstruir por columna la matriz de dimensiones $r \times p$ a partir de este vector obtenido.

Notar las semejanzas y diferencias entre los estimadores de máxima verosimilitud y sub-d estandarizado del parámetro $\boldsymbol{\beta}$. Ambos tienen la misma eficiencia asintótica y se obtienen siguiendo los pasos 1., 2. y 3. Pero en el primer caso, se estandariza $\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS})$ mediante la matriz $S_{\mathbf{x}}^{1/2} \otimes S_{\mathbf{y}}^{-1/2}$ y en el segundo caso, esto se lleva a cabo mediante la matriz $S_{\mathbf{x}}^{1/2} \otimes S_{\mathbf{y}|\mathbf{x}}^{-1/2}$ que es exactamente estandarizar por su varianza asintótica estimada.

2.3.2. Otro estimador propuesto

Es posible definir otro estimador para $\boldsymbol{\beta}$ del modelo de rango reducido (1.1) con (2.1) teniendo en cuenta siempre que el rango del estimador sea d. Los objetivos generales de esta propuesta, serán extenderlo a GLMM de rango reducido en el próximo capítulo (que en dicho contexto serán más simples de obtener) y también aplicarlos en el último capítulo de la tesis.

Consideremos la siguiente descomposición en valores singulares de $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = \mathbf{U}^T \begin{pmatrix} \boldsymbol{\Lambda} & 0 \\ 0 & 0 \end{pmatrix} \mathbf{R}$$

donde $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ es la matriz diagonal que contiene los valores singulares de $\boldsymbol{\beta}$, $\lambda_1 \geq \ldots \geq \lambda_d > 0$. La matriz ortogonal $\mathbf{U}^T = (\mathbf{U}_1, \mathbf{U}_0)$ es de orden $r \times r$ con $\mathbf{U}_1 : r \times d$ y $\mathbf{U}_0 : r \times (r-d)$, cumple que $\mathbf{U}_1\mathbf{U}_1^T + \mathbf{U}_0\mathbf{U}_0^T = \mathbf{I}_r$, $\mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I}_d$, $\mathbf{U}_0^T\mathbf{U}_0 = \mathbf{I}_{r-d}$ y $\mathbf{U}_0^T\mathbf{U}_1 = \mathbf{0}$. La matriz ortogonal $\mathbf{R}^T = (\mathbf{R}_1, \mathbf{R}_0)$ de orden $p \times p$ con $\mathbf{R}_1 : p \times d$ y $\mathbf{R}_0 : p \times (p-d)$ cumple, al igual que $\mathbf{U}$, con $\mathbf{R}_1\mathbf{R}_1^T + \mathbf{R}_0\mathbf{R}_0^T = \mathbf{I}_p$, $\mathbf{R}_1^T\mathbf{R}_1 = \mathbf{I}_d$, $\mathbf{R}_0^T\mathbf{R}_0 = \mathbf{I}_{p-d}$ y $\mathbf{R}_0^T\mathbf{R}_1 = \mathbf{0}$.
Los d vectores singulares a izquierda $\mathbf{U}_1 = (U_1^{(1)}, \ldots, U_d^{(1)})$ de $\boldsymbol{\beta}$ que corresponde a los d valores singulares no nulos, son una base para span($\boldsymbol{\beta}$). Esto es porque podemos expresar a $\boldsymbol{\beta}$ como $\boldsymbol{\beta} = \mathbf{U}_1\boldsymbol{\Lambda}\mathbf{R}_1^T$ y $\boldsymbol{\Lambda}\mathbf{R}_1^T$ es de rango completo d. Análogamente, para $\widehat{\boldsymbol{\beta}}_{OLS}$, la SVD es:

$$\widehat{\boldsymbol{\beta}}_{OLS} = \widehat{\mathbf{U}}^T \begin{pmatrix} \widehat{\boldsymbol{\Lambda}}_1 & 0 \\ 0 & \widehat{\boldsymbol{\Lambda}}_0 \end{pmatrix} \widehat{\mathbf{R}}$$

con $\widehat{\mathbf{U}}^T = (\widehat{\mathbf{U}}_1, \widehat{\mathbf{U}}_0)$ y $\widehat{\mathbf{R}}^T = (\widehat{\mathbf{R}}_1, \widehat{\mathbf{R}}_0)$ donde las particiones son las mismas que la SVD de $\boldsymbol{\beta}$. Las matrices $\widehat{\boldsymbol{\Lambda}}_1 = \text{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_d)$ y $\widehat{\boldsymbol{\Lambda}}_0$ de dimensiones $d \times d$ y $(r-d) \times (p-d)$ respectivamente contienen en su diagonal principal los valores singulares $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{\min(r,p)}$ de $\widehat{\boldsymbol{\beta}}_{OLS}$ en orden decreciente. Observar que $\widehat{\boldsymbol{\Lambda}}_0$ tiene la forma:

$$\widehat{\boldsymbol{\Lambda}}_0 = \begin{pmatrix} \widehat{\lambda}_{d+1} & & \\ & \cdots & \\ & & \widehat{\lambda}_p \\ & \mathbf{0} & \end{pmatrix}, \text{ si } r > p \quad \text{ó} \quad \widehat{\boldsymbol{\Lambda}}_0 = \begin{pmatrix} \widehat{\lambda}_{d+1} & & \\ & \cdots & \mathbf{0} \\ & & \widehat{\lambda}_r \end{pmatrix}, \text{ si } r \leq p.$$

Además, como $\widehat{\boldsymbol{\beta}}_{OLS}$ converge en probabilidad a $\boldsymbol{\beta}$ y $\boldsymbol{\beta}$ es de rango d, se tiene que $\widehat{\lambda}_{d+1} < \widehat{\lambda}_d$ con probabilidad 1. Luego, proponemos el siguiente estimador:

Estimador Sub-d $\widehat{\boldsymbol{\beta}}_{OLS}^{(d)}$: Consiste en truncar la descomposición en valores singulares del estimador del modelo de rango completo. Es decir, que este estimador es obtenido haciendo

$\widehat{\mathbf{\Lambda}}_0 = 0$. Luego

$$\widehat{\boldsymbol{\beta}}_{OLS}^{(d)} = \widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T.$$

Es importante resaltar la simplicidad de este estimador ya que no requiere cálculos complejos más que la descomposición SVD de matrices. A continuación presentamos la distribución asintótica del estimador $\widehat{\boldsymbol{\beta}}_{OLS}^{(d)}$.

Proposición 2.11. *Asumiendo que* $\boldsymbol{\varepsilon}|\mathbf{X} \sim \mathcal{N}(0, \mathbf{\Sigma})$ *y que* $\text{rank}(\boldsymbol{\beta}) = d$,

$$\text{avar}\left(\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{OLS}^{(d)})\right) = \mathbf{V}_{\boldsymbol{\beta}_{OLS}^{(d)}},$$

donde

$$\begin{aligned}\mathbf{V}_{\boldsymbol{\beta}_{OLS}^{(d)}} &= \mathbf{P}_{\mathbf{B}^T}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{P}_{\mathbf{B}^T} \otimes \mathbf{\Sigma} + \mathbf{Q}_{\mathbf{B}^T}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{Q}_{\mathbf{B}^T} \otimes \mathbf{P}_{\mathbf{A}}\mathbf{\Sigma}\mathbf{P}_{\mathbf{A}} \\ &\quad + \mathbf{P}_{\mathbf{B}^T}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{Q}_{\mathbf{B}^T} \otimes \mathbf{\Sigma}\mathbf{P}_{\mathbf{A}} + \mathbf{Q}_{\mathbf{B}^T}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{P}_{\mathbf{B}^T} \otimes \mathbf{P}_{\mathbf{A}}\mathbf{\Sigma}. \end{aligned} \tag{2.9}$$

Demostración. Como $\widehat{\boldsymbol{\beta}}_{OLS} = \widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T + \widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T$ y sabemos que vale (2.4), tenemos que

$$\sqrt{n}\text{vec}(\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T + \widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T) \to \mathcal{N}(0, \mathbf{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{\Sigma}).$$

Entonces,

$$\begin{aligned}
&\sqrt{n}\left[\begin{pmatrix}\mathbf{P}_{\mathbf{R}_1}\\ \mathbf{P}_{\mathbf{R}_0}\end{pmatrix} \otimes \begin{pmatrix}\mathbf{P}_{\mathbf{U}_1}\\ \mathbf{P}_{\mathbf{U}_0}\end{pmatrix}\right]\text{vec}(\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T + \widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T) \\
= \;& \sqrt{n}\text{vec}\left(\begin{pmatrix}\mathbf{P}_{\mathbf{U}_1}\\ \mathbf{P}_{\mathbf{U}_0}\end{pmatrix}(\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T + \widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T)\begin{pmatrix}\mathbf{P}_{\mathbf{R}_1} & \mathbf{P}_{\mathbf{R}_0}\end{pmatrix}\right) \\
= \;& \sqrt{n}\text{vec}\begin{pmatrix}\mathbf{P}_{\mathbf{U}_1}\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T\mathbf{P}_{\mathbf{R}_1} + \mathbf{P}_{\mathbf{U}_1}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_1} - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T & \mathbf{P}_{\mathbf{U}_1}\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T\mathbf{P}_{\mathbf{R}_0} + \mathbf{P}_{\mathbf{U}_1}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_0} \\ \mathbf{P}_{\mathbf{U}_0}\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T\mathbf{P}_{\mathbf{R}_1} + \mathbf{P}_{\mathbf{U}_0}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_1} & \mathbf{P}_{\mathbf{U}_0}\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T\mathbf{P}_{\mathbf{R}_0} + \mathbf{P}_{\mathbf{U}_0}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_0}\end{pmatrix} \\
\doteq \;& \sqrt{n}\text{vec}\begin{pmatrix}\mathbf{A}_n - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T & \mathbf{C}_n \\ \mathbf{B}_n & \mathbf{D}_n\end{pmatrix} = \sqrt{n}\begin{pmatrix}\text{vec}\begin{pmatrix}\mathbf{A}_n - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T\\ \mathbf{B}_n\end{pmatrix} \\ \text{vec}\begin{pmatrix}\mathbf{C}_n\\ \mathbf{D}_n\end{pmatrix}\end{pmatrix}
\end{aligned} \tag{2.10}$$

donde en la última línea, introducimos notación para referirnos a cada bloque de la matriz (2.10). Luego, (2.10) es asintóticamente normal con la siguiente matriz de covarianza:

$$\begin{aligned}\mathbf{G} &= \left[\begin{pmatrix}\mathbf{P}_{\mathbf{R}_1}\\ \mathbf{P}_{\mathbf{R}_0}\end{pmatrix}\otimes\begin{pmatrix}\mathbf{P}_{\mathbf{U}_1}\\ \mathbf{P}_{\mathbf{U}_0}\end{pmatrix}\right](\mathbf{\Sigma}_{\mathbf{x}}^{-1}\otimes\mathbf{\Sigma})\left[\begin{pmatrix}\mathbf{P}_{\mathbf{R}_1}\\ \mathbf{P}_{\mathbf{R}_0}\end{pmatrix}\otimes\begin{pmatrix}\mathbf{P}_{\mathbf{U}_1}\\ \mathbf{P}_{\mathbf{U}_0}\end{pmatrix}\right]^T\\ &= \begin{bmatrix}\mathbf{P}_{\mathbf{R}_1}\otimes\begin{pmatrix}\mathbf{P}_{\mathbf{U}_1}\\ \mathbf{P}_{\mathbf{U}_0}\end{pmatrix}\\ \mathbf{P}_{\mathbf{R}_0}\otimes\begin{pmatrix}\mathbf{P}_{\mathbf{U}_1}\\ \mathbf{P}_{\mathbf{U}_0}\end{pmatrix}\end{bmatrix}(\mathbf{\Sigma}_{\mathbf{x}}^{-1}\otimes\mathbf{\Sigma})\begin{bmatrix}\mathbf{P}_{\mathbf{R}_1}\otimes\begin{pmatrix}\mathbf{P}_{\mathbf{U}_1} & \mathbf{P}_{\mathbf{U}_0}\end{pmatrix} & \mathbf{P}_{\mathbf{R}_0}\otimes\begin{pmatrix}\mathbf{P}_{\mathbf{U}_1} & \mathbf{P}_{\mathbf{U}_0}\end{pmatrix}\end{bmatrix}.\end{aligned}$$

Ahora veamos que algunos términos de (2.10) convergen a cero en probabilidad. Para ellos vamos a usar fundamentalemente que $\sqrt{n}\widehat{\mathbf{\Lambda}}_0 = O_p(1)$, $\sqrt{n}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T = O_p(1)$, $\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T = O_p(1)$, $\mathbf{P}_{\mathbf{U}_1} = \mathbf{P}_{\widehat{\mathbf{U}}_1} + O_p(n^{-1/2})$, $\mathbf{P}_{\mathbf{U}_0} = \mathbf{P}_{\widehat{\mathbf{U}}_0} + O_p(n^{-1/2})$, $\mathbf{P}_{\mathbf{R}_1} = \mathbf{P}_{\widehat{\mathbf{R}}_1} + O_p(n^{-1/2})$ y $\mathbf{P}_{\mathbf{R}_0} = \mathbf{P}_{\widehat{\mathbf{R}}_0} + O_p(n^{-1/2})$.

$$\begin{aligned}\sqrt{n}\mathbf{P}_{\mathbf{U}_1}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_1} &= \sqrt{n}(\mathbf{P}_{\widehat{\mathbf{U}}_1} + O_p(n^{-1/2}))\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T(\mathbf{P}_{\widehat{\mathbf{R}}_1} + O_p(n^{-1/2})) \qquad (2.11)\\ &= \sqrt{n}O_p(n^{-1/2})\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T O_p(n^{-1/2}) = O_p(n^{-1})\end{aligned}$$

$$\begin{aligned}\sqrt{n}\mathbf{P}_{\mathbf{U}_0}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_1} &= \sqrt{n}(\mathbf{P}_{\widehat{\mathbf{U}}_0} + O_p(n^{-1/2}))\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T(\mathbf{P}_{\widehat{\mathbf{R}}_1} + O_p(n^{-1/2})) \qquad (2.12)\\ &= \sqrt{n}O_p(n^{-1})\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T + \sqrt{n}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T O_p(n^{-1/2})\\ &= O_p(n^{-1}) + O_p(n^{-1/2}) = O_p(n^{-1/2})\end{aligned}$$

$$\begin{aligned}\sqrt{n}\mathbf{P}_{\mathbf{U}_1}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_0} &= \sqrt{n}(\mathbf{P}_{\widehat{\mathbf{U}}_1} + O_p(n^{-1/2}))\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T(\mathbf{P}_{\widehat{\mathbf{R}}_0} + O_p(n^{-1/2})) \qquad (2.13)\\ &= \sqrt{n}O_p(n^{-1})\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T + \sqrt{n}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T O_p(n^{-1/2})\\ &= O_p(n^{-1}) + O_p(n^{-1/2}) = O_p(n^{-1/2})\end{aligned}$$

$$\begin{aligned}\sqrt{n}\mathbf{P}_{\mathbf{U}_0}\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T\mathbf{P}_{\mathbf{R}_0} &= \sqrt{n}(\mathbf{P}_{\widehat{\mathbf{U}}_0} + O_p(n^{-1/2}))\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T(\mathbf{P}_{\widehat{\mathbf{R}}_0} + O_p(n^{-1/2})) \qquad (2.14)\\ &= \sqrt{n}O_p(n^{-1})\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T = O_p(n^{-1/2})\end{aligned}$$

Por otro lado, observemos que

$$\begin{aligned}
\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T &= (\mathbf{P}_{\mathbf{U}_1} + \mathbf{P}_{\mathbf{U}_0})(\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T)(\mathbf{P}_{\mathbf{R}_1} + \mathbf{P}_{\mathbf{R}_0}) \\
&= \mathbf{P}_{\mathbf{U}_1}\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T\mathbf{P}_{\mathbf{R}_1} + \mathbf{P}_{\mathbf{U}_1}\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T\mathbf{P}_{\mathbf{R}_0} + \mathbf{P}_{\mathbf{U}_0}\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T\mathbf{P}_{\mathbf{R}_1} + \\
&\quad \mathbf{P}_{\mathbf{U}_0}\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T\mathbf{P}_{\mathbf{R}_0} - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T \\
&= \mathbf{A}_n + \mathbf{B}_n + \mathbf{C}_n + \mathbf{P}_{\mathbf{U}_0}\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T\mathbf{P}_{\mathbf{R}_0} - \mathbf{P}_{\mathbf{U}_1}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_1} - \\
&\quad \mathbf{P}_{\mathbf{U}_0}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_1} - \mathbf{P}_{\mathbf{U}_1}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_0} - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T
\end{aligned}$$

$$\begin{aligned}
\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T &= (\mathbf{P}_{\mathbf{U}_1} + \mathbf{P}_{\mathbf{U}_0})(\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T)(\mathbf{P}_{\mathbf{R}_1} + \mathbf{P}_{\mathbf{R}_0}) \\
&= \mathbf{P}_{\mathbf{U}_1}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_1} + \mathbf{P}_{\mathbf{U}_1}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_0} + \mathbf{P}_{\mathbf{U}_0}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_1} + \\
&\quad \mathbf{P}_{\mathbf{U}_0}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_0} \\
&= \mathbf{D}_n + \mathbf{P}_{\mathbf{U}_1}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_1} + \mathbf{P}_{\mathbf{U}_1}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_0} + \\
&\quad \mathbf{P}_{\mathbf{U}_0}\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{P}_{\mathbf{R}_1} - \mathbf{P}_{\mathbf{U}_0}\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T\mathbf{P}_{\mathbf{R}_0}
\end{aligned}$$

y además por (2.11), (2.12), (2.13) y (2.14) tenemos que

$$\operatorname{avar}\left(\sqrt{n}\begin{pmatrix}\operatorname{vec}(\widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T)\\ \operatorname{vec}(\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T)\end{pmatrix}\right) = \operatorname{avar}\left(\sqrt{n}\begin{pmatrix}\operatorname{vec}(\mathbf{A}_n + \mathbf{B}_n + \mathbf{C}_n - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T)\\ \operatorname{vec}(\mathbf{D}_n)\end{pmatrix}\right).$$

Como

$$\begin{aligned}
&\sqrt{n}\begin{pmatrix}\operatorname{vec}(\mathbf{A}_n + \mathbf{B}_n + \mathbf{C}_n - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T)\\ \operatorname{vec}(\mathbf{D}_n)\end{pmatrix} \\
= \sqrt{n}&\begin{pmatrix}\mathbf{I}_p \otimes \begin{pmatrix}\mathbf{I}_r & \mathbf{I}_r\end{pmatrix} & \mathbf{I}_p \otimes \begin{pmatrix}\mathbf{I}_r & 0\end{pmatrix}\\ 0 & \mathbf{I}_p \otimes \begin{pmatrix}0 & \mathbf{I}_r\end{pmatrix}\end{pmatrix}\begin{pmatrix}\operatorname{vec}\begin{pmatrix}\mathbf{A}_n - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T\\ \mathbf{B}_n\end{pmatrix}\\ \operatorname{vec}\begin{pmatrix}\mathbf{C}_n\\ \mathbf{D}_n\end{pmatrix}\end{pmatrix}
\end{aligned}$$

obtenemos que

$$\operatorname{avar}\left(\sqrt{n}\begin{pmatrix}\operatorname{vec}(\mathbf{A}_n + \mathbf{B}_n + \mathbf{C}_n - \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T)\\ \operatorname{vec}(\mathbf{D}_n)\end{pmatrix}\right) = \mathbf{H}$$

donde

$$\begin{aligned}
\mathbf{H} &= \begin{pmatrix} \mathbf{I}_p \otimes \begin{pmatrix} \mathbf{I}_r & \mathbf{I}_r \end{pmatrix} & \mathbf{I}_p \otimes \begin{pmatrix} \mathbf{I}_r & 0 \end{pmatrix} \\ 0 & \mathbf{I}_p \otimes \begin{pmatrix} 0 & \mathbf{I}_r \end{pmatrix} \end{pmatrix} \mathbf{G} \begin{pmatrix} \mathbf{I}_p \otimes \begin{pmatrix} \mathbf{I}_r \\ \mathbf{I}_r \end{pmatrix} & 0 \\ \mathbf{I}_p \otimes \begin{pmatrix} \mathbf{I}_r \\ 0 \end{pmatrix} & \mathbf{I}_p \otimes \begin{pmatrix} 0 \\ \mathbf{I}_r \end{pmatrix} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{P}_{\mathbf{R}_1} \otimes \mathbf{I}_r + \mathbf{P}_{\mathbf{R}_0} \otimes \mathbf{P}_{\mathbf{U}_1} \\ \mathbf{P}_{\mathbf{R}_0} \otimes \mathbf{P}_{\mathbf{U}_0} \end{pmatrix} (\mathbf{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{\Sigma}) \begin{pmatrix} \mathbf{P}_{\mathbf{R}_1} \otimes \mathbf{I}_r + \mathbf{P}_{\mathbf{R}_0} \otimes \mathbf{P}_{\mathbf{U}_1} & \mathbf{P}_{\mathbf{R}_0} \otimes \mathbf{P}_{\mathbf{U}_0} \end{pmatrix} \\
&\doteq \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{12}^T & \mathbf{M}_{22} \end{pmatrix}
\end{aligned}$$

con

$$\begin{aligned}
\mathbf{M}_{11} &= \mathbf{P}_{\mathbf{R}_1}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{P}_{\mathbf{R}_1} \otimes \mathbf{\Sigma} + \mathbf{P}_{\mathbf{R}_0}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{P}_{\mathbf{R}_0} \otimes \mathbf{P}_{\mathbf{U}_1}\mathbf{\Sigma}\mathbf{P}_{\mathbf{U}_1} + \mathbf{P}_{\mathbf{R}_1}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{P}_{\mathbf{R}_0} \otimes \mathbf{\Sigma}\mathbf{P}_{\mathbf{U}_1} + \\
&\quad \mathbf{P}_{\mathbf{R}_0}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{P}_{\mathbf{R}_1} \otimes \mathbf{P}_{\mathbf{U}_1}\mathbf{\Sigma} \\
\mathbf{M}_{22} &= \mathbf{P}_{\mathbf{R}_0}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{P}_{\mathbf{R}_0} \otimes \mathbf{P}_{\mathbf{U}_0}\mathbf{\Sigma}\mathbf{P}_{\mathbf{U}_0} \\
\mathbf{M}_{12} &= \mathbf{P}_{\mathbf{R}_1}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{P}_{\mathbf{R}_0} \otimes \mathbf{\Sigma}\mathbf{P}_{\mathbf{U}_0} + \mathbf{P}_{\mathbf{R}_0}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{P}_{\mathbf{R}_0} \otimes \mathbf{P}_{\mathbf{U}_1}\mathbf{\Sigma}\mathbf{P}_{\mathbf{U}_0}
\end{aligned}$$

Por lo tanto, $\text{avar}(\widehat{\boldsymbol{\beta}}_{OLS}^{(d)}) = \mathbf{M}_{11}$.

Como $\boldsymbol{\beta} = \mathbf{AB} = \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{R}_1^T$, podemos suponer que $\mathbf{A} = \mathbf{U}_1\mathbf{\Lambda}_1$ y $\mathbf{B} = \mathbf{R}_1^T$. Luego, $\mathbf{P}_{\mathbf{R}_1} = \mathbf{P}_{\mathbf{B}^T}$ y además

$$\mathbf{P}_{\mathbf{A}} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A} = \mathbf{U}_1\mathbf{\Lambda}_1(\mathbf{\Lambda}_1\mathbf{U}_1^T\mathbf{U}_1\mathbf{\Lambda}_1)^{-1}\mathbf{\Lambda}\mathbf{U}_1^T = \mathbf{U}_1\mathbf{U}_1^T = \mathbf{P}_{\mathbf{U}_1}.$$

Remplazando las expresiones correspondientes llegamos a (2.9). □

Observación 2.12. Puede probarse que entre las matrices (2.8) y (2.9) existe la siguiente relación:

$$(\mathbf{P}_{\mathbf{B}^T(\mathbf{\Sigma}_\mathbf{x})} \otimes \mathbf{I}_r + \mathbf{Q}_{\mathbf{B}^T(\mathbf{\Sigma}_\mathbf{x})} \otimes \mathbf{P}_{\mathbf{A}(\mathbf{\Sigma}^{-1})})(\mathbf{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{\Sigma})(\mathbf{P}_{\mathbf{B}^T(\mathbf{\Sigma}_\mathbf{x})} \otimes \mathbf{I}_r + \mathbf{Q}_{\mathbf{B}^T(\mathbf{\Sigma}_\mathbf{x})} \otimes \mathbf{P}_{\mathbf{A}(\mathbf{\Sigma}^{-1})})^T \geq$$
$$(\mathbf{P}_{\mathbf{B}^T} \otimes \mathbf{I}_r + \mathbf{Q}_{\mathbf{B}^T} \otimes \mathbf{P}_{\mathbf{A}})(\mathbf{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{\Sigma})(\mathbf{P}_{\mathbf{B}^T} \otimes \mathbf{I}_r + \mathbf{Q}_{\mathbf{B}^T} \otimes \mathbf{P}_{\mathbf{A}}).$$

Este orden en las matrices de covarianza expresa que si bien el estimador $\widehat{\boldsymbol{\beta}}_{OLS}^{(d)}$ es sencillo de obtener a partir del estimador OLS, es menos eficiente que el estimador de máxima verosimilitud, como era de esperarse.

Por otro lado, no siempre es cierto que $\mathbf{V}_{\boldsymbol{\beta}_{OLS}^{(d)}} \leqslant \mathbf{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{\Sigma}$. Es decir que no siempre el estimador sub-d resulta ser más asintóticamente eficiente que el estimador de rango completo.

Un ejemplo donde no se verifica que $\mathbf{V}_{\boldsymbol{\beta}_{OLS}^{(d)}} \leqslant \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}$ es considerar $p = r = 2$, $d = 1$, y las matrices

$$\boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{y } \mathbf{B} = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

con $\rho = 0.5$. Sin embargo, en las simulaciones que se presentan en el Capítulo 4, el estimador Sub-d estandarizado arroja errores de estimación mas pequeños que el estimador OLS.

El objetivo del próximo capítulo será extender los estimadores y resultados asintóticos a los modelos lineales generalizados de rango reducido. Además, en el Capítulo 4 se llevarán a cabo simulaciones para comparar la eficiencia de los estimadores presentados en este capítulo donde la distribución normal es un ejemplo en particular de familias exponenciales multivariadas.

2.4. Demostraciones del Cápitulo 2

Lema 2.13. *Las matrices $C_{\mathbf{yx}}^T C_{\mathbf{yx}}$ y $S^T S$ poseen los mismos autovectores.*

Demostración. De acuerdo a las definiciones de $C_{\mathbf{yx}}$ y S,

$$\begin{aligned} C_{\mathbf{yx}}^T C_{\mathbf{yx}} &= S_{\mathbf{x}}^{-1/2} S_{\mathbf{xy}} S_{\mathbf{y}}^{-1} S_{\mathbf{yx}} S_{\mathbf{x}}^{-1/2} \\ S^T S &= S_{\mathbf{x}}^{-1/2} S_{\mathbf{xy}} S_{\mathbf{y|x}}^{-1} S_{\mathbf{yx}} S_{\mathbf{x}}^{-1/2}. \end{aligned}$$

Luego, si aplicamos la Identidad de Woodbury a la igualdad $S_{\mathbf{y}} = S_{\mathbf{y|x}} + S_{\mathbf{yx}} S_{\mathbf{x}}^{-1} S_{\mathbf{xy}}$, tenemos que

$$S_{\mathbf{y}}^{-1} = S_{\mathbf{y|x}} - S_{\mathbf{y|x}}^{-1} S_{\mathbf{yx}} (S_{\mathbf{x}} + S_{\mathbf{xy}} S_{\mathbf{y|x}}^{-1} S_{\mathbf{yx}})^{-1} S_{\mathbf{xy}} S_{\mathbf{y|x}}^{-1}$$

por lo que

$$C_{\mathbf{yx}}^T C_{\mathbf{yx}} = S^T S - S_{\mathbf{x}}^{-1/2} S_{\mathbf{xy}} S_{\mathbf{y|x}}^{-1} S_{\mathbf{yx}} (S_{\mathbf{x}} + S_{\mathbf{xy}} S_{\mathbf{y|x}}^{-1} S_{\mathbf{yx}})^{-1} S_{\mathbf{xy}} S_{\mathbf{y|x}}^{-1} S_{\mathbf{yx}} S_{\mathbf{x}}^{-1/2}.$$

Sea $\mathbf{v}$ un autovector de $S^T S$ y λ su correspondiente autovalor, entonces

$$\begin{aligned} C_{\mathbf{yx}}^T C_{\mathbf{yx}} \mathbf{v} &= \lambda \mathbf{v} - S_{\mathbf{x}}^{-1/2} S_{\mathbf{xy}} S_{\mathbf{y|x}}^{-1} S_{\mathbf{yx}} (S_{\mathbf{x}} + S_{\mathbf{xy}} S_{\mathbf{y|x}}^{-1} S_{\mathbf{yx}})^{-1} S_{\mathbf{xy}} S_{\mathbf{y|x}}^{-1} S_{\mathbf{yx}} S_{\mathbf{x}}^{-1/2} \mathbf{v} \\ &= \lambda \mathbf{v} - S_{\mathbf{x}}^{-1/2} S_{\mathbf{xy}} S_{\mathbf{y|x}}^{-1} S_{\mathbf{yx}} (S_{\mathbf{x}} + S_{\mathbf{xy}} S_{\mathbf{y|x}}^{-1} S_{\mathbf{yx}})^{-1} S_{\mathbf{x}}^{1/2} \lambda \mathbf{v} \\ &= \lambda \mathbf{v} - \lambda S_{\mathbf{x}}^{-1/2} S_{\mathbf{xy}} S_{\mathbf{y|x}}^{-1} S_{\mathbf{yx}} S_{\mathbf{x}}^{-1/2} S_{\mathbf{x}}^{1/2} (S_{\mathbf{x}} + S_{\mathbf{xy}} S_{\mathbf{y|x}}^{-1} S_{\mathbf{yx}})^{-1} S_{\mathbf{x}}^{1/2} \mathbf{v} \\ &= \lambda \mathbf{v} - \lambda S S^T (\mathbf{I}_p + S S^T)^{-1} \mathbf{v} \\ &= (\lambda - \lambda^3/(1+\lambda)) \mathbf{v}. \end{aligned}$$

Por lo tanto, también es autovector de $C_{\mathbf{yx}}^T C_{\mathbf{yx}}$. □

CAPÍTULO 3

MODELOS LINEALES GENERALIZADOS DE RANGO REDUCIDO

El estudio del modelo clásico de regresión con rango reducido, estudiado en detalle en el Capítulo 2, ha sido limitado a las familias gaussianas, es decir que sus principales aplicaciones son con datos donde la variable respuesta es de tipo continua. Una forma de eliminar esta limitación es considerar los modelos lineales generalizados multivariados. En el Capítulo 1 se ha visto que esta clase de modelos abarca una amplia gama de tipos de respuesta multivariadas. La idea de aplicar regresión de rango reducido a otros modelos ha aparecido en la literatura, un ejemplo es [**Anderson, 1984**] que aplica esta idea a los modelos logísticos multinomiales. Más tarde, en [**Yee and Hastie, 2003**] se presenta esta teoría de forma más general cuando $\mathbf{Y}|\mathbf{X}$ verifica (1.10) y (1.11), es decir la respuesta dada los predictores provienen de una familia exponencial. Este trabajo, permite que los potenciales beneficios de la regresión de rango reducido puedan ser transportados a una amplia clase de modelos. Bajo este enfoque, ellos obtienen estimadores de máxima versosimilitud pero no presentan resultados acerca de la distribución asintótica de los mismos. En este capítulo se estudia estos modelos en detalle, se obtiene la distribución asintótica del estimador de máxima verosimilitud, se presentan otros estimadores de rango reducido, sus varianzas asintóticas y la relación entre ellas. Este resultado se utilizará en el Capítulo 4 para complementar los ejemplos de [**Yee and Hastie, 2003**] calculando intervalos de confianza asintóticos para los parámetros de interés. Además se aplicarán estos modelos y resultados en el Capítulo 6 donde se estudiará reducción suficiente de dimensiones.

3.1. Estructura del modelo

Suponemos que en el modelo (1.10) con (1.11), el rango de la matriz de coeficientes $\mathbf{D}$ es $d < \min\{k_1, p\}$. Luego, de la misma forma que antes, suponemos que $\mathbf{D} = \mathbf{AB}$ con $\mathbf{A} : k_1 \times d$ y $\mathbf{B} : d \times p$ ambas de rango completo. Entonces el modelo GLMM de rango reducido lo expresamos como

$$\begin{aligned} \boldsymbol{\eta}_1 &= \overline{\boldsymbol{\eta}}_1 + \mathbf{ABx} = \boldsymbol{\Gamma}_1 \mathbf{f} = (\mathbf{f}^t \otimes \mathbf{I}_{k_1}) \mathrm{vec}(\boldsymbol{\Gamma}_1) \\ \boldsymbol{\eta}_2 &= \overline{\boldsymbol{\eta}}_2 = \boldsymbol{\Gamma}_2 = (1 \otimes \mathbf{I}_{k_2}) \mathrm{vec}(\boldsymbol{\Gamma}_2) \end{aligned}$$

donde $\mathbf{f}^T = (1, \mathbf{x}^T) : 1 \times (p+1)$, $\boldsymbol{\Gamma}_1 = (\overline{\boldsymbol{\eta}}_1, \mathbf{AB}) : k_1 \times (p+1)$ y $\boldsymbol{\Gamma}_2 = (\overline{\boldsymbol{\eta}}_2) : k_2 \times 1$. En forma matricial, lo podemos sintetizar de la siguiente forma

$$\boldsymbol{\eta}_{\mathbf{x}} = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} (\mathbf{f}^T \otimes \mathbf{I}_{k_1}) & 0 \\ 0 & \mathbf{I}_{k_2} \end{pmatrix} \begin{pmatrix} \mathrm{vec}(\boldsymbol{\Gamma}_1) \\ \mathrm{vec}(\boldsymbol{\Gamma}_2) \end{pmatrix} = \mathbf{F}\boldsymbol{\Gamma},$$

donde $\mathbf{F} = \begin{pmatrix} (\mathbf{f}^T \otimes \mathbf{I}_{k_1}) & 0 \\ 0 & \mathbf{I}_{k_2} \end{pmatrix} : k \times (k_1(p+1)+k_2)$ es la matriz de diseño y $\boldsymbol{\Gamma} = \begin{pmatrix} \mathrm{vec}(\boldsymbol{\Gamma}_1) \\ \mathrm{vec}(\boldsymbol{\Gamma}_2) \end{pmatrix}$ son los coeficientes a estimar con los datos.

Sea la partición de $\mathbf{x}$ en $(\mathbf{x}_1^T, \mathbf{x}_2^T)^T$ y de acuedo a esta partición, sea $\mathbf{D} = (\mathbf{D}_1, \mathbf{D}_2)$. En **[Anderson, 1951]** se sugiere una extensión del modelo clásico de regresión con rango reducido donde aplicamos esta idea solamente a un subconjunto de los predictores. De la misma manera, podemos considerar esta modificación a los modelos lineales generalizados:

$$\begin{aligned} \boldsymbol{\eta}_1 &= \overline{\boldsymbol{\eta}}_1 + \mathbf{Cx}_1 + \mathbf{ABx}_2 = \boldsymbol{\Gamma}_1 \mathbf{f} = (\mathbf{f}^T \otimes \mathbf{I}_{k_1}) \mathrm{vec}(\boldsymbol{\Gamma}_1) \\ \boldsymbol{\eta}_2 &= \overline{\boldsymbol{\eta}}_2 = \boldsymbol{\Gamma}_2 = (1 \otimes \mathbf{I}_{k_2}) \mathrm{vec}(\boldsymbol{\Gamma}_2) \end{aligned} \tag{3.1}$$

donde $\mathbf{f}^T = (1, \mathbf{x}_1^T, \mathbf{x}_2^T) : 1 \times (p+1)$, $\boldsymbol{\Gamma}_2 = (\overline{\boldsymbol{\eta}}_2) : k_2 \times 1$ y $\boldsymbol{\Gamma}_1 = (\overline{\boldsymbol{\eta}}_1, \mathbf{C}, \mathbf{AB}) : k_1 \times (1+p)$. Es decir que $\mathbf{D}_1 = \mathbf{C}$ de rango completo y $\mathbf{D}_2 = \mathbf{AB}$. Esta clase de modelos se denomina modelos generalizados lineales reducido parciales (**[Yee and Hastie, 2003]**).

3.2. Estimadores de máxima verosimilitud

Para el caso de GLMM sin restrición, estudiado en la Sección 1.3.2, hemos visto que la solución de la ecuaciones de máxima verosimilitud es equivalente a las solución obtenida mediante el algoritmo iterativo IRLS. Sin embargo, no contamos con una forma explícita de los estimadores sino que estos se obtienen de manera iterativa. En el trabajo **[Yee and Hastie, 2003]** proponen

una extensión del algoritmo IRLS estándar para obtener estimadores de máxima verosimilitud de los parámetros de interés de este modelo. Dicha metodología se presenta a continuación y notar que en el caso que $d = \text{mín}(k_1, p_2)$, el algoritmo alternativo se reduce al usual. Los pasos son los siguientes:

Algoritmo IRLS adaptado para GLMM reducido

Sea $(\mathbf{y}_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ una muestra del modelo (3.1).

Paso 1: Dado los estimadores actuales de los coeficientes $\mathbf{A}^{(t)}, \mathbf{B}^{(t)}, \mathbf{C}^{(t)}, \bar{\boldsymbol{\eta}}^{(t)} = (\bar{\boldsymbol{\eta}}_1^{(t)}, \bar{\boldsymbol{\eta}}_2^{(t)})$ y con ello los valores actuales de $\mathbf{W}_i^{(t)}, \mathbf{d}_i^{(t)}$ y $\mathbf{z}_i^{(t)}$ $i = 1, \ldots, n$, actualizamos

$$\begin{aligned}\text{vec}\left(\mathbf{B}^{(t+1)}\right) &= \left(\sum_{i=1}^{n} \mathbf{x}_{i2}\mathbf{x}_{i2}^T \otimes (\mathbf{A}^{(t)})^T \mathbf{W}_{i11}^{(t)} \mathbf{A}^{(t)}\right)^{-1} \\ &\quad \sum_{i=1}^{n} \mathbf{x}_{i2} \otimes (\mathbf{A}^{(t)})^T \left(\mathbf{W}_{i11}^{(t)}, \mathbf{W}_{i12}^{(t)}\right) \left(\left(\mathbf{z}_{i1}^{(t)} - \bar{\boldsymbol{\eta}}_1^{(t)} - \mathbf{C}^{(t)}\mathbf{x}_{i1}\right)^T, \left(\mathbf{z}_{i2}^{(t)} - \bar{\boldsymbol{\eta}}_2^{(t)}\right)^T\right)^T\end{aligned}$$

donde

- $\mathbf{W}_i^{(t)}$ y $\mathbf{z}_i^{(t)}$ están definidas en (1.16), (1.17) y (1.15) pero en este caso

$$\boldsymbol{\Gamma}^{(t)} = \begin{pmatrix} \text{vec}\left((\overline{\boldsymbol{\eta}}_1^{(t)}, \mathbf{C}^{(t)}, \mathbf{A}^{(t)}\mathbf{B}^{(t)})\right) \\ \text{vec}\left(\overline{\boldsymbol{\eta}}_2^{(t)}\right) \end{pmatrix},$$

- además, hemos particionado las matrices $\mathbf{W}_i^{(t)}$ y los vectores $\mathbf{z}_i^{(t)}$ conforme a las dimensiones de $\bar{\boldsymbol{\eta}}_1^{(t)}, \bar{\boldsymbol{\eta}}_2^{(t)}$. Es decir

$$\mathbf{z}_i^{(t)} = \begin{pmatrix} \mathbf{z}_{i1}^{(t)} \\ \mathbf{z}_{i2}^{(t)} \end{pmatrix} \quad \text{y} \quad \mathbf{W}_i^{(t)} = \begin{pmatrix} \mathbf{W}_{i11}^{(t)} & \mathbf{W}_{i12}^{(t)} \\ \mathbf{W}_{i21}^{(t)} & \mathbf{W}_{i22}^{(t)} \end{pmatrix}$$

con $\mathbf{z}_{i1}^{(t)} : k_1 \times 1$ y $\mathbf{z}_{i2}^{(t)} : k_2 \times 1$; $\mathbf{W}_{i11}^{(t)} : k_1 \times k_1$, $\mathbf{W}_{i12}^{(t)} : k_1 \times k_2$, $\mathbf{W}_{i21}^{(t)} : k_2 \times k_1$ y $\mathbf{W}_{i22}^{(t)} : k_2 \times k_2$.

Paso 2: Sea $\mathbf{h}_i^{(t)} = \left(1, \mathbf{x}_{i1}, \mathbf{B}^{(t+1)}\mathbf{x}_{i2}\right)$ y $\mathbf{H}_i = \begin{pmatrix} \left(\mathbf{h}_i^{(t)} \otimes \mathbf{I}_{k_1}\right) & 0 \\ 0 & \mathbf{I}_{k_2} \end{pmatrix}$. Luego,

$$\begin{pmatrix} \text{vec}\left(\bar{\boldsymbol{\eta}}_1^{(t+1)}, \mathbf{C}^{(t+1)}, \mathbf{A}^{(t+1)}\right) \\ \text{vec}\left(\bar{\boldsymbol{\eta}}_2^{(t+1)}\right) \end{pmatrix} = \left(\sum_{i=1}^{n} \mathbf{H}_i^{(t)} \mathbf{W}_i^{(t)} \mathbf{H}_i^{(t)T}\right)^{-1} \sum_{i=1}^{n} \mathbf{H}_i^{(t)} \mathbf{W}_i^{(t)} \mathbf{b}_i^{(t)}$$

$$\text{donde } \mathbf{b}_i^{(t)} = \mathbf{H}_i^{(t)} \begin{pmatrix} \text{vec}\left(\bar{\boldsymbol{\eta}}_1^{(t)}, \mathbf{C}^{(t)}, \mathbf{A}^{(t)}\right) \\ \text{vec}\left(\bar{\boldsymbol{\eta}}_2^{(t)}\right) \end{pmatrix} + \left(\mathbf{W}_i^{(t)}\right)^{-1} \mathbf{d}_i^{(t)}.$$

Estos pasos se deducen teniendo en cuenta el problema de mínimos cuadrados pesados, el cual se resuelve en cada iteración del algoritmo IRLS estándar. Dicho problema adaptado al

caso reducido tiene la siguiente forma:

$$\arg\min_{\mathbf{\Gamma}} \sum_{i=1}^{n} \left(\mathbf{z}_i^{(t)} - \mathbf{F}_i\mathbf{\Gamma}\right)^T \mathbf{W}_i^{(t)} \left(\mathbf{z}_i^{(t)} - \mathbf{F}_i\mathbf{\Gamma}\right) \tag{3.2}$$

con

$$\left(\mathbf{z}_i^{(t)} - \mathbf{F}_i\mathbf{\Gamma}\right) = \begin{pmatrix} \mathbf{z}_{i1}^{(t)} - \bar{\boldsymbol{\eta}}_1^{(t)} - \mathbf{C}^{(t)}\mathbf{x}_{i1} - (\mathbf{x}_{i2}^T \otimes \mathbf{A}^{(t)})\text{vec}(\mathbf{B}^{(t)}) \\ \mathbf{z}_{i2}^{(t)} - \bar{\boldsymbol{\eta}}_2^{(t)} \end{pmatrix}.$$

Entonces un estimador de $\mathbf{B}$, el cual expresamos como $\mathbf{B}^{(t+1)}$, puede obtenerse derivando e igualando a cero la expresión (3.2). Cada término tiene la forma:

$$\begin{aligned} &\left(\mathbf{x}_{i2}^T \otimes \mathbf{A}^{(t)}\right) \mathbf{W}_{i11}^{(t)} \left(\mathbf{z}_{i1}^{(t)} - \bar{\boldsymbol{\eta}}_1^{(t)} - \mathbf{C}^{(t)}\mathbf{x}_{i1} - \left(\mathbf{x}_{i2}^T \otimes \mathbf{A}^{(t)}\right) \text{vec}(\mathbf{B}^{(t)})\right) \\ &+ \left(\mathbf{x}_{i2}^T \otimes \mathbf{A}^{(t)}\right) \mathbf{W}_{i12}^{(t)} \left(\mathbf{z}_{i2}^{(t)} - \bar{\boldsymbol{\eta}}_2^{(t)}\right) = 0 \end{aligned}$$

de donde se obtiene la expresión para $\mathbf{B}^{(t+1)}$ del paso 1. Entonces, una vez actualizado $\mathbf{B}$, el siguiente paso es actualizar los demás coeficientes. Para esto, simplemente se considera el nuevo conjunto de predictores $\mathbf{B}^{(t+1)}\mathbf{x}_i$ con $i = 1, \ldots, n$ y se procede como el caso del modelo GLMM con rango completo. Los estimadores de máxima verosimilitud obtenidos mediante este método iterativo lo denotamos como $\widehat{\mathbf{\Gamma}}_{RR} = (\widehat{\mathbf{\Gamma}}_{1RR}, \widehat{\mathbf{\Gamma}}_{2RR})$ y en particular a la matriz de coeficientes que se ha reducido el rango como $\widehat{\mathbf{D}}_{2RR}$

3.3. Distribución asintótica del estimador de máxima verosimilitud

En esta sección presentamos los resultados acerca de la distribución asintótica del estimador de máxima verosimilitud del modelo (3.1). Además mostramos que en el caso particular de la distribución normal multivariada, cuando se consideran sus parámetros naturales, estos resultados coinciden con los ya conocidos y expuestos en la Proposición 2.4.

Teorema 3.1. *Asumiendo que* $\mathbf{Y}|\mathbf{X}$ *sigue el modelo (1.10) con (3.1) y que las condiciones usuales de regularidad se cumplen para el modelo de rango completo (1.11),*

$$\text{avar}(\widehat{\mathbf{\Gamma}}_{RR}) = \mathbf{\Delta}(\mathbf{\Delta}^T \mathbf{V}_{\mathbf{\Gamma}_0}^{-1} \mathbf{\Delta})^{\ddagger} \mathbf{\Delta}^T, \tag{3.3}$$

donde $\mathbf{\Gamma}_0$ *es el valor poblacional verdadero para* $\mathbf{\Gamma}$*,* $\mathbf{V}_{\mathbf{\Gamma}_0}$ *es la matriz de covarianza asintótica del estimador de máxima verosimilitud de GLMM de rango completo (1.11) y*

$$\mathbf{\Delta} = \begin{pmatrix} \mathbf{I}_{k_1(p_1+1)} & 0 & 0 & 0 \\ 0 & \mathbf{B}^T \otimes \mathbf{I}_{k_1} & \mathbf{I}_{p_2} \otimes \mathbf{A} & 0 \\ 0 & 0 & 0 & \mathbf{I}_{k_2} \end{pmatrix}.$$

Demostración. La distribución asintótica del estimador $\widehat{\mathbf{\Gamma}}_{RR}$ es consecuencia de una extensión del resultado [Cook and Ni, 2005] que a su vez es una extensión del resultado de [Shapiro, 1986]. Definimos la función de discrepancia:

$$F_n(\mathbf{\Gamma}, \widehat{\mathbf{\Gamma}}_{RC}) = \frac{1}{n}\left\{\mathcal{L}_n(\widehat{\mathbf{\Gamma}}_{RC}) - \mathcal{L}_n(\mathbf{\Gamma})\right\}$$

donde $\mathcal{L}_n$ es la función objetivo de máxima verosimilitud de $\mathbf{Y}$ dado $\mathbf{X}$ de acuerdo al modelo (1.10) junto con (3.1). Luego, el primer paso de esta demostración consiste en probar de forma análoga el Teorema 1 de [Shapiro, 1985] que podemos escribir F_n de la forma

$$F_n(\mathbf{\Gamma}, \widehat{\mathbf{\Gamma}}_{RC}) = \left(\widehat{\mathbf{\Gamma}}_{RC} - \mathbf{\Gamma}\right)^T \mathbf{V}_n(\widehat{\mathbf{\Gamma}}_{RC}, \mathbf{\Gamma}) \left(\widehat{\mathbf{\Gamma}}_{RC} - \mathbf{\Gamma}\right) \tag{3.4}$$

con $\mathbf{V}_n(\widehat{\mathbf{\Gamma}}_{RC}, \mathbf{\Gamma}) = \displaystyle\int_0^1 \int_0^1 t\frac{1}{n}\sum_{i=1}^{n} \mathbf{F}_i\mathbf{H}\psi(\mathbf{F}_i\widehat{\mathbf{\Gamma}}_{RC} + ut\mathbf{F}_i(\mathbf{\Gamma} - \widehat{\mathbf{\Gamma}}_{RC}))\mathbf{F}_i du dt.$

En el caso de [Shapiro, 1986] la función $\mathbf{V}_n$ solo puede depender de n a travez de $\widehat{\mathbf{\Gamma}}_{RC}$ mientras que la extensión en [Cook and Ni, 2005], $\mathbf{V}_n$ puede depender de la muestra completa pero no de $\mathbf{\Gamma}$. En [Cook and Ni, 2005] prueban que la distribución asintótica del mínimo de nF_n es la misma si se reemplaza $\mathbf{V}_n$ por su límite en probabilidad. Este mismo resultado puede extenderse a nuestro caso si mayores modificaciones si $\mathbf{V}_n(\widehat{\mathbf{\Gamma}}_{RC}, \mathbf{\Gamma})$ converge en probabilidad a $\mathbf{V}(\mathbf{\Gamma}_0, \mathbf{\Gamma})$ uniformemente en $\mathbf{\Gamma}$ en un entorno de $\mathbf{\Gamma}_0$ y si el estimador de máxima verosimilitud del modelo GLMM de rango reducido es consistente.

Observemos que en este contexto tenemos el subespacio abierto $\Theta = \text{vec}(\Pi^{k_1\times(p_1+1)}) \times \text{vec}(\Pi^{k_1\times d}) \times \text{vec}(\Pi^{p_2\times d}) \times \mathbb{R}^{k_2} \subset \mathbb{R}^{k_1(p_1+1)+(k_1+p_2)d+k_2}$, el vector de parámetros $\boldsymbol{\theta}^T = (\overline{\boldsymbol{\eta}}_1, \text{vec}(\mathbf{C})^T, \text{vec}(\mathbf{A})^T, \text{vec}(\mathbf{B})^T, \text{vec}(\mathbf{\Gamma}_2)^T) \in \Theta$ y la función $\mathbf{g} : \Theta \to \mathbb{R}^{k_1(p+1)+k_2}$, tal que

$$\mathbf{g}(\boldsymbol{\theta}) = \mathbf{g}(\overline{\boldsymbol{\eta}}_1, \text{vec}(\mathbf{C}), \text{vec}(\mathbf{A}), \text{vec}(\mathbf{B}), \text{vec}(\mathbf{\Gamma}_2)^T) = (\overline{\boldsymbol{\eta}}_1, \text{vec}(\mathbf{C}), \text{vec}(\mathbf{AB}), \text{vec}(\mathbf{\Gamma}_2)^T).$$

Por lo que, en este caso, la matriz $\boldsymbol{\Delta} = \dfrac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ de dimensiones $(k_1p + k_2) \times (k_1p_1 + (k_1 + p_2)d + k_2)$ es la que se definió en el enunciado del teorema. □

Observación 3.2. No es posible obtener la consistencia a priori del estimador de máxima verosimilitud $\widehat{\mathbf{\Gamma}}_{RR}$ a partir de esta metodología. Sin embargo, probamos que para una parametrización de $\boldsymbol{\theta}$ en particular, las condiciones de regularidad clásicas impuestas para el modelo GLMM de rango completo implican condiciones de regularidad en el modelo GLMM de rango reducido y así se obtiene la consistencia del estimador $\widehat{\mathbf{\Gamma}}_{RR}$. A posteriori, como la distribución asintótica es independiente de la parametrización, vía los resultados de [Shapiro, 1986] se obtiene (3.3). Esta prueba se encuentra al final del capítulo.

3.3.1. Distribución normal multivariada como familia exponencial

En la Sección 1.3.4 analizamos el caso donde $\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y suponíamos el modelo lineal generalizado de regresión centrado (1.20). Ahora, estamos interesados en el siguiente modelo GLMM de rango reducido

$$\begin{aligned} \boldsymbol{\eta}_1 &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \boldsymbol{\Gamma}_1\mathbf{x} = \mathbf{ABx} \\ \boldsymbol{\eta}_2 &= \text{vech}(\boldsymbol{\Sigma}^{-1}) = \boldsymbol{\Gamma}_2 \end{aligned} \tag{3.5}$$

donde $\boldsymbol{\Gamma}_1 = \mathbf{AB}$ con $\mathbf{A} : r \times d$ y $\mathbf{B} : d \times p$, ambas de rango completo d. Aplicaremos los resultados asintóticos del Teorema 3.1, para estudiar la varianza asintótica de los estimadores de máxima verosimilitud, $\widehat{\mathbf{A}}\widehat{\mathbf{B}}$ y $\widehat{\boldsymbol{\Sigma}}^{-1}$, del modelo (3.5). En este caso la matriz $\boldsymbol{\Delta}$ del Teorema 3.1 tiene la forma

$$\boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{\Delta}_1 & 0 \\ 0 & \mathbf{I}_{r(r+1)/2} \end{pmatrix} = \begin{pmatrix} \mathbf{B}^T \otimes \mathbf{I}_r & \mathbf{I}_p \otimes \mathbf{A} & 0 \\ 0 & 0 & \mathbf{I}_{r(r+1)/2} \end{pmatrix},$$

donde $\boldsymbol{\Delta}_1 = \begin{pmatrix} \mathbf{B}^T \otimes \mathbf{I}_r & \mathbf{I}_p \otimes \mathbf{A} \end{pmatrix}$. Por lo tanto,

$$\text{avar}\begin{pmatrix} \text{vec}(\widehat{\mathbf{A}}\widehat{\mathbf{B}}) \\ \text{vech}(\widehat{\boldsymbol{\Sigma}}^{-1}) \end{pmatrix} = \boldsymbol{\Delta}(\boldsymbol{\Delta}^T\mathbf{V}_{\boldsymbol{\Gamma}}^{-1}\boldsymbol{\Delta})^{\ddagger}\boldsymbol{\Delta}^T \tag{3.6}$$

donde $\mathbf{V}_{\boldsymbol{\Gamma}}$ es la varianza asintótica de los estimadores de $(\text{vec}(\widehat{\boldsymbol{\Gamma}}_1)^T, \text{vech}(\widehat{\boldsymbol{\Sigma}}^{-1})^T)$ bajo el modelo de rango completo (1.20). En la Sección 1.3.4 habiamos obtenido que

$$\mathbf{V}_{\boldsymbol{\Gamma}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma} & \left(\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma}_1^T\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}\right)\mathbf{D}_r \\ \mathbf{D}_r^T\left(\boldsymbol{\Sigma}\boldsymbol{\Gamma}_1\boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}\right) & \mathbf{D}_r^T\left(\left(\boldsymbol{\Sigma}\boldsymbol{\Gamma}_1\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\Gamma}_1^T\boldsymbol{\Sigma} + \frac{1}{2}\boldsymbol{\Sigma}\right) \otimes \boldsymbol{\Sigma}\right)\mathbf{D}_r \end{pmatrix}^{-1}. \tag{3.7}$$

Si desarrollamos la expresión (3.6) en forma detallada obtenemos que la varianza asintótica de $\widehat{\boldsymbol{\Gamma}}_1$ está dada por

$$\begin{aligned} \text{avar}(\text{vec}(\widehat{\boldsymbol{\Gamma}}_1)) &= \mathbf{P}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \boldsymbol{\Sigma}^{-1} + \mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma})}\boldsymbol{\Sigma}^{-1} \\ &\quad + (\boldsymbol{\Gamma}_1^T\boldsymbol{\Sigma}\boldsymbol{\Gamma}_1 \otimes \boldsymbol{\Sigma}^{-1}) + (\boldsymbol{\Gamma}_1^T \otimes \boldsymbol{\Gamma}_1)\mathbf{K}_{pr}. \end{aligned} \tag{3.8}$$

Las cuentas para obtener la expresión (3.8) se encuentran al final de este capítulo. De la misma forma que vimos en la Sección 1.3.4, podemos establecer la relación que entre el modelo GLMM de rango reducido (3.5) y el modelo de regresión lineal clásico de rango reducido (2.1) estudiado en detalle en el Capítulo 2. Este último modelo, asume que $\boldsymbol{\beta} = \mathbf{A}_{\boldsymbol{\beta}}\mathbf{B}_{\boldsymbol{\beta}}$ con $\mathbf{A}_{\boldsymbol{\beta}} : r \times d$ y $\mathbf{B}_{\boldsymbol{\beta}} : d \times p$ ambas de rango completo d (lo denotamos con el subíndice $\boldsymbol{\beta}$ para diferenciarlos de

$\mathbf{A}$ y $\mathbf{B}$ del modelo (3.5)). La relación claramente es que $\mathbf{\Gamma}_1 = \mathbf{\Sigma}^{-1}\mathbf{A}_{\boldsymbol{\beta}}\mathbf{B}_{\boldsymbol{\beta}}$. Como los estimadores de máxima verosimilitud de $\mathbf{\Gamma}_1$ y de $\mathbf{\Sigma}^{-1}\mathbf{A}_{\boldsymbol{\beta}}\mathbf{B}_{\boldsymbol{\beta}}$ coinciden, también lo harán sus distribuciones asintóticas.

Utilizando el resultado de la Proposición 2.4 y la regla de Cramer obtenemos que $\text{vec}(\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{A}}_{\boldsymbol{\beta}}\widehat{\mathbf{B}}_{\boldsymbol{\beta}})$ es asintóticamente normal con la siguiente matriz de covarianza:

$$\begin{aligned}\text{avar}(\text{vec}(\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{A}}_{\boldsymbol{\beta}}\widehat{\mathbf{B}}_{\boldsymbol{\beta}})) &= (\mathbf{P}_{\mathbf{B}^T_{\boldsymbol{\beta}}(\mathbf{\Sigma_x})}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\otimes\mathbf{\Sigma}^{-1}) + (\mathbf{Q}_{\mathbf{B}^T_{\boldsymbol{\beta}}(\mathbf{\Sigma_x})}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\otimes\mathbf{\Sigma}^{-1}\mathbf{P}_{\mathbf{A}_{\boldsymbol{\beta}}(\mathbf{\Sigma}^{-1})}) \\ &\quad +(\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1}\boldsymbol{\beta}\otimes\mathbf{\Sigma}^{-1}) + (\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1}\otimes\mathbf{\Sigma}^{-1}\boldsymbol{\beta})\mathbf{K}_{pr}. \qquad (3.9)\end{aligned}$$

Ya que tenemos las siguientes igualdades

$$\mathbf{\Gamma}_1 = \mathbf{AB} = \mathbf{\Sigma}^{-1}\boldsymbol{\beta} = \mathbf{\Sigma}^{-1}\mathbf{A}_{\boldsymbol{\beta}}\mathbf{B}_{\boldsymbol{\beta}},$$

podemos suponer, sin pérdida de generalidad, que $\mathbf{A} = \mathbf{\Sigma}^{-1}\mathbf{A}_{\boldsymbol{\beta}}$ y $\mathbf{B} = \mathbf{B}_{\boldsymbol{\beta}}$. Por lo tanto,

$$\begin{aligned}\text{avar}(\text{vec}(\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{A}}_{\boldsymbol{\beta}}\widehat{\mathbf{B}}_{\boldsymbol{\beta}})) &= \mathbf{P}_{\mathbf{B}^T(\mathbf{\Sigma_x})}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\otimes\mathbf{\Sigma}^{-1} + \mathbf{Q}_{\mathbf{B}^T(\mathbf{\Sigma_x})}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\otimes\mathbf{P}_{\mathbf{A}(\mathbf{\Sigma})}\mathbf{\Sigma}^{-1} \\ &\quad +(\mathbf{\Gamma}_1^T\mathbf{\Sigma}\mathbf{\Gamma}_1\otimes\mathbf{\Sigma}^{-1}) + (\mathbf{\Gamma}_1^T\otimes\mathbf{\Gamma}_1)\mathbf{K}_{pr}\end{aligned}$$

que es igual a la expresión obtenida en (3.8).

3.4. Otros estimadores de rango reducido propuestos

En esta sección presentamos los estimadores **Sub-d** y de **minimización cuadrática** para los modelos lineales generalizados de rango reducido. Ambos, están motivados en aquellos estimadores que se propusieron en el Capítulo 2, con el objetivo de obtener estimadores que sean fáciles de calcular en esta clase más amplia de modelos de GLMM de rango reducido. Además presentamos sus distribuciones asintóticas.

3.4.1. Estimador Sub-d

En el modelo (3.1) hemos supuesto que la matriz $\mathbf{D}_2$ es de rango $d < \min(k_1, p_2)$. Luego, la siguiente descomposición en valores singulares de $\mathbf{D}_2$ tendrá la forma:

$$\mathbf{D}_2 = \mathbf{U}^T\begin{pmatrix}\mathbf{\Lambda} & 0\\ 0 & 0\end{pmatrix}\mathbf{R}$$

donde $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ es la matriz diagonal que contiene los valores singulares de $\mathbf{D}_2$, $\lambda_1 \geq \ldots \geq \lambda_d > 0$. La matriz ortogonal $\mathbf{U}^T = (\mathbf{U}_1, \mathbf{U}_0)$ es de orden $k_1 \times k_1$ con $\mathbf{U}_1 : k_1 \times d$ y $\mathbf{U}_0 : k_1 \times (k_1 - d)$, cumple que $\mathbf{U}_1\mathbf{U}_1^T + \mathbf{U}_0\mathbf{U}_0^T = \mathbf{I}_{k_1}$, $\mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I}_d$, $\mathbf{U}_0^T\mathbf{U}_0 = \mathbf{I}_{k_1-d}$ y $\mathbf{U}_0^T\mathbf{U}_1 = \mathbf{0}$.

La matriz ortogonal $\mathbf{R}^T = (\mathbf{R}_1, \mathbf{R}_0)$ de orden $p_2 \times p_2$ con $\mathbf{R}_1 : p_2 \times d$ y $\mathbf{R}_0 : p_2 \times (p_2 - d)$ cumple, al igual que $\mathbf{U}$, con $\mathbf{R}_1\mathbf{R}_1^T + \mathbf{R}_0\mathbf{R}_0^T = \mathbf{I}_{p_2}$, $\mathbf{R}_1^T\mathbf{R}_1 = \mathbf{I}_d$, $\mathbf{R}_0^T\mathbf{R}_0 = \mathbf{I}_{p_2-d}$ y $\mathbf{R}_0^T\mathbf{R}_1 = \mathbf{0}$. Luego $\mathbf{D}_2 = \mathbf{U}_1\mathbf{\Lambda}\mathbf{R}_1^T$.

Sea $\widehat{\mathbf{D}}_{2RC}$ el estimador de máxima verosimilitud de rango completo y consideremos su SVD:

$$\widehat{\mathbf{D}}_{2RC} = \widehat{\mathbf{U}}^T \begin{pmatrix} \widehat{\mathbf{\Lambda}}_1 & 0 \\ 0 & \widehat{\mathbf{\Lambda}}_0 \end{pmatrix} \widehat{\mathbf{R}}$$

con $\widehat{\mathbf{U}}^T = (\widehat{\mathbf{U}}_1, \widehat{\mathbf{U}}_0)$ y $\widehat{\mathbf{R}}^T = (\widehat{\mathbf{R}}_1, \widehat{\mathbf{R}}_0)$, donde las particiones son las mismas que la SVD de $\mathbf{D}_2$. Las matrices $\widehat{\mathbf{\Lambda}}_1 = \text{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_d)$ y $\widehat{\mathbf{\Lambda}}_0$ de dimensiones $d \times d$ y $(k_1 - d) \times (p_2 - d)$ respectivamente contienen en su diagonal principal los valores singulares $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{\min(k_1,p_2)}$ de $\widehat{\mathbf{D}}_{2RC}$ en orden decreciente. Luego, proponemos el siguiente estimador de rango reducido:

- **Estimador Sub-d $\widehat{\mathbf{D}}_{2RC}^{(d)}$:** Consiste en truncar la descomposición en valores singulares del estimador del modelo de rango completo. Es decir, que este estimador es obtenido haciendo $\widehat{\mathbf{D}}_0 = 0$. Luego

$$\widehat{\mathbf{D}}_{2RC}^{(d)} = \widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T.$$

La siguiente proposición presenta la distribución asintótica del estimador Sub-d $\widehat{\mathbf{D}}_{2RC}^{(d)}$.

Proposición 3.3. *Asumiendo que* $\mathbf{Y}|(\mathbf{X} = \mathbf{x}) \sim \mathcal{F}_{\boldsymbol{\eta}_\mathbf{x}, \mathbf{T}, \psi}$ *y que los parámetros naturales* $\boldsymbol{\eta}$ *siguen el modelo (3.1),*

$$\text{avar}\left(\sqrt{n}\text{vec}(\widehat{\mathbf{D}}_{2RC}^{(d)})\right) = \mathbf{V}_{\widehat{\mathbf{D}}_{2RC}^{(d)}},$$

donde

$$\mathbf{V}_{\widehat{\mathbf{D}}_{2RC}^{(d)}} = (\mathbf{P}_{\mathbf{B}^T} \otimes \mathbf{I}_{k_1} + \mathbf{Q}_{\mathbf{B}^T} \otimes \mathbf{P}_\mathbf{A})\, \mathbf{V}_{\mathbf{D}_2}\, (\mathbf{P}_{\mathbf{B}^T} \otimes \mathbf{I}_{k_1} + \mathbf{Q}_{\mathbf{B}^T} \otimes \mathbf{P}_\mathbf{A})$$

y $\mathbf{V}_{\mathbf{D}_2}$ *es la matriz de covarianza asintótica correspondiente al estimador* $\widehat{\mathbf{D}}_{2RC}$.

La demostración de este resultado es análoga a la demostración de la Proposición 2.11.

3.4.2. Estimador basado en el teorema de minimización cuadrática

En el Capítulo 2, donde estudiamos el modelo de regresión lineal de rango reducido, vimos que el estimador de minimización cuadrática es asintóticamente tan eficiente como el estimador de máxima verosimilitud. Esto pudo demostrarse aplicando el Corolario 2.8 donde se define una nueva función de discrepancia cuadrática. Para el caso de los GLMM de rango reducido, no es posible reproducir la misma idea. Dicha minimización cuadrática implica que si expresamos $\mathbf{D}_2 = \mathbf{AB}$ con $\mathbf{A} : k_1 \times d$ y $\mathbf{B} : d \times p_2$ ambas matrices de rango completo d y $\widehat{\mathbf{V}}_{\mathbf{D}_2}$ es un estimador

consistente de la matriz de covarianza asintótica $\mathbf{V}_{\mathbf{D}_2}$, el objetivo es obtener los valores de $\mathbf{A}$ y $\mathbf{B}$ que minimicen

$$\text{vec}^T\left(\widehat{\mathbf{D}}_{2RC} - \mathbf{AB}\right)\widehat{\mathbf{V}}_{\mathbf{D}_2}^{-1}\text{vec}\left(\widehat{\mathbf{D}}_{2RC} - \mathbf{AB}\right). \tag{3.10}$$

Los valores que se obtienen para $\mathbf{A}$ y $\mathbf{B}$ son:

- $\mathbf{B} = \left((\mathbf{I}_{p_2} \otimes \mathbf{A}^T)\widehat{\mathbf{V}}_{\mathbf{D}_2}^{-1}(\mathbf{I}_{p_2} \otimes \mathbf{A})\right)^{-1}(\mathbf{I}_{p_2} \otimes \mathbf{A}^T)\widehat{\mathbf{V}}_{\mathbf{D}_2}^{-1}\text{vec}(\widehat{\mathbf{D}}_{2RC})$
- $\mathbf{A} = \text{argmax}_{\mathbf{A}\in\Pi^{k_1\times d}}\left(\text{tr}\left(\mathbf{P}_{(\widehat{\mathbf{V}}_{\mathbf{D}_2}^{-1/2}(\mathbf{I}_{p_2}\otimes\mathbf{A}))}\widehat{\mathbf{V}}_{\mathbf{D}_2}^{-1/2}\text{vec}\left(\widehat{\mathbf{D}}_{2RC}\right)\text{vec}^T\left(\widehat{\mathbf{D}}_{2RC}\right)\widehat{\mathbf{V}}_{\mathbf{D}_2}^{-1/2}\right)\right)$

y no es posible obtener una forma explícita para $\mathbf{A}$. No obstante, esta limitación no impide poder generalizar el estimador de minimización cuadrática al contexto de este capítulo.

Teniendo en cuenta las expresiones de las derivadas parciales con respecto a vec($\mathbf{A}$) y vec($\mathbf{B}$) de la expresión cuadrática (3.10), proponemos el siguiente algoritmo iterativo para obtener los valores de vec($\mathbf{A}$) y vec($\mathbf{B}$) que minimizan (3.10):

Algoritmo iterativo de minimización cuadrático adaptado para GLMM reducido

1. Obtener los estimadores de la matriz de coeficientes del modelo de rango completo $\widehat{\mathbf{D}}_{2RC}$ y de su respectiva matriz de covarianza asintótica $\widehat{\mathbf{V}}_{\mathbf{D}_2}$.
2. Proponer valores iniciales de $\widehat{\mathbf{A}}^{(t)}$ y $\widehat{\mathbf{B}}^{(t)}$. Por ejemplo, puede considerarse como valores iniciales el estimador Sub-d $\widehat{\mathbf{D}}_{2RC}^{(d)}$.
3. Repetir los siguientes pasos
   a) Dado el valor actual de $\widehat{\mathbf{A}}^{(t)}$, actualizar $\widehat{\mathbf{B}}^{(t)}$ de la siguiente forma:

   $$\widehat{\mathbf{B}}^{(t+1)} = \left((\mathbf{I}_{p_2} \otimes \widehat{\mathbf{A}}^{(t)})^T\widehat{\mathbf{V}}_{\mathbf{D}_2}^{-1}(\mathbf{I}_{p_2} \otimes \widehat{\mathbf{A}}^{(t)})\right)^{-1}(\mathbf{I}_{p_2} \otimes \widehat{\mathbf{A}}^{(t)})^T\widehat{\mathbf{V}}_{\mathbf{D}_2}^{-1}\text{vec}(\widehat{\mathbf{D}}_{2RC}).$$

   b) Dado el valor actual de $\widehat{\mathbf{B}}^{(t+1)}$ obtenido en el paso anterior, actualizar $\widehat{\mathbf{A}}^{(t)}$ de la siguiente forma:

   $$\widehat{\mathbf{A}}^{(t+1)} = \left((\widehat{\mathbf{B}}^{(t+1)} \otimes \mathbf{I}_r)\widehat{\mathbf{V}}_{\mathbf{D}_2}^{-1}(\widehat{\mathbf{B}}^{(t+1)} \otimes \mathbf{I}_r)^T\right)^{-1}(\widehat{\mathbf{B}}^{(t+1)} \otimes \mathbf{I}_r)\widehat{\mathbf{V}}_{\mathbf{D}_2}^{-1}\text{vec}(\widehat{\mathbf{D}}_{2RC}).$$

Denotamos al estimador obtenido mediante este algoritmo como $\widehat{\mathbf{D}}_{2RR2}$ y lo llamamos **estimador de minimización cuadrática**. La siguiente proposición, cuya demostración es análoga a la demostración de la Proposición 2.9 y está basada en el Corolario 2.8, expone la distribución asintótica del estimador $\widehat{\mathbf{D}}_{RR2}$.

Proposición 3.4. *Asumiendo que* $\mathbf{Y}|(\mathbf{X} = \mathbf{x}) \sim \mathcal{F}_{\boldsymbol{\eta}_\mathbf{x},\mathbf{T},\psi}$ *y que los parámetros naturales* $\boldsymbol{\eta}$ *siguen el modelo (3.1),*

$$\text{avar}\left(\sqrt{n}\text{vec}(\widehat{\mathbf{D}}_{2RR2})\right) = \text{avar}\left(\sqrt{n}\text{vec}(\widehat{\mathbf{D}}_{2RR})\right),$$

donde $\widehat{\mathbf{D}}_{2RR}$ *es el estimador de máxima verosimilitud obtenido mediante el algoritmo IRLS adaptado para GLMM reducido.*

En conclusión, es posible obtener un estimador de rango reducido tan eficiente como el estimador MLE, a partir de la minimización de una forma cuadrática que involucra el estimador de rango completo $\widehat{\mathbf{D}}_{2RC}$ y el estimador de su varianza asintótica $\widehat{\mathbf{V}}_{\mathbf{D}_2}$. Si bien, no se ha podido obtenerse una forma explícita de dicho estimador, el algoritmo propuesto es simple y de rápida implemetación y ejecución.

3.5. Tests asintóticos para la dimensión d

En esta sección presentamos tests asintóticos para estimar el rango d de la matriz de coeficientes $\mathbf{D}$ basados en los test propuestos en [**Bura and Yang, 2011**]. Estos test requieren que el estimador propuesto para el modelo de rango completo sea asintóticamente normal y el conocimiento de su varianza asintótica. Como ya hemos expresado, el algoritmo IRLS estándar proporciona estimadores de máxima verosimilitud para los GLMM. Es decir que, en nuestro contexto, las propiedades asintóticas de $\widehat{\mathbf{D}}$ se deducen a través de las propiedades que gozan los estimadores de máxima verosimilitud.

En la Sección 1.3.3 hemos estudiado las propiedades asintóticas necesarias del modelo de rango completo. De acuerdo a la Proposición 1.2, el modelo de rango completo

$$\begin{aligned}\boldsymbol{\eta}_1 &= \overline{\boldsymbol{\eta}}_1 + \mathbf{D}\mathbf{x} = \boldsymbol{\Gamma}_1\mathbf{f} = (\mathbf{f}^T \otimes \mathbf{I}_{k_1})\text{vec}(\boldsymbol{\Gamma}_1)\\ \boldsymbol{\eta}_2 &= \overline{\boldsymbol{\eta}}_2 = \boldsymbol{\Gamma}_2 = (1 \otimes \mathbf{I}_{k_2})\text{vec}(\boldsymbol{\Gamma}_2),\end{aligned}$$

donde $\mathbf{f}^T = (1, \mathbf{x})^T : 1 \times (p+1)$, $\boldsymbol{\Gamma}_1 = (\overline{\boldsymbol{\eta}}_1, \mathbf{D}) : k_1 \times (p+1)$ y $\boldsymbol{\Gamma}_2 = (\overline{\boldsymbol{\eta}}_2) : k_2 \times 1$ tiene la distribución asintótica

$$\sqrt{n}\left(\begin{pmatrix}\text{vec}(\widehat{\boldsymbol{\Gamma}}_1)\\ \text{vec}(\widehat{\boldsymbol{\Gamma}}_2)\end{pmatrix} - \begin{pmatrix}\text{vec}(\boldsymbol{\Gamma}_1)\\ \text{vec}(\boldsymbol{\Gamma}_2)\end{pmatrix}\right) \to \mathcal{N}(0, \mathbf{V}_{\boldsymbol{\Gamma}}).$$

Sea la matriz $\mathbf{M} = (0, \mathbf{I}_p)$ de dimensiones $p \times (p+1)$, luego podemos expresar

$$\text{vec}(\mathbf{D}) = (\mathbf{M} \otimes \mathbf{I}_{k_1})\begin{pmatrix}\mathbf{I}_{k_1(p+1)} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\text{vec}(\boldsymbol{\Gamma}_1)\\ \text{vec}(\boldsymbol{\Gamma}_2)\end{pmatrix}.$$

En consecuencia tenemos que $\sqrt{n}\text{vec}(\widehat{\mathbf{D}} - \mathbf{D}) \to \mathcal{N}(\mathbf{0}, \mathbf{V}_{\mathbf{D}})$ con

$$\mathbf{V}_{\mathbf{D}} = (\mathbf{M} \otimes \mathbf{I}_{k_1})\begin{pmatrix}\mathbf{I}_{k_1(p+1)} & \mathbf{0}\end{pmatrix}\mathbf{V}_{\boldsymbol{\Gamma}}\begin{pmatrix}\mathbf{I}_{k_1(p+1)}\\ \mathbf{0}\end{pmatrix}(\mathbf{M}^T \otimes \mathbf{I}_{k_1}).$$

Por lo tanto, $\widehat{\mathbf{D}}$ es asintóticamente normal y podemos aplicar el test chi-cuadrado ponderado asintótico o el test chi-cuadrado asintótico de Wald basados en los valores singulares más chicos de $\widehat{\mathbf{D}}$ desarrollados en [**Bura and Yang, 2011**].

Sea $\widehat{\mathbf{D}} = \widehat{\mathbf{U}}^T\widehat{\mathbf{\Lambda}}\widehat{\mathbf{R}}$ la descomposición en valores singulares de $\widehat{\mathbf{D}}$ con $\widehat{\mathbf{\Lambda}} = \text{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{\text{mín}(k_1,p)})$, donde $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_{\text{mín}(r,k)}$ son los valores singulares de $\widehat{\mathbf{D}}$, $\widehat{\mathbf{U}}$ es la matriz $k_1 \times k_1$ de vectores singulares a izquierda y $\widehat{\mathbf{R}}$ es la matriz $p \times p$ de vectores singulares a derecha. Para $m = 0, \ldots, \text{mín}(k_1, p)$, sea $\widehat{\mathbf{U}}_0$ la sub-matriz $k_1 \times (k_1 - m)$ de las últimas $k_1 - m$ columnas de $\widehat{\mathbf{U}}^T$ y $\widehat{\mathbf{R}}_0$ la sub-matriz $r \times (p-m)$ de las últimas $p-m$ columnas de $\widehat{\mathbf{R}}^T$. Además, sea $\widehat{\mathbf{Q}} = (\widehat{\mathbf{R}}_0^T \otimes \widehat{\mathbf{U}}_0^T)\widehat{\mathbf{V}}_{\mathbf{D}}(\widehat{\mathbf{R}}_0 \otimes \widehat{\mathbf{U}}_0)$ y $\widehat{\mathbf{D}}_0 = \widehat{\mathbf{U}}_0^T\widehat{\mathbf{D}}\widehat{\mathbf{R}}_0$. De acuerdo a [**Bura and Yang, 2011**], cuando $m = \text{rank}(\mathbf{D})$, el estadístico

$$\Lambda_1(m) = n \sum_{i=m+1}^{\text{mín}(k_1,p)} \widehat{\lambda}_i^2$$

para $m = 0, \ldots, \text{mín}(k_1, p)$ tiene distribución asintótica dada por

$$\sum_{i=1}^{s} \omega \chi_i^2,$$

donde $s = \text{mín}(\text{rank}(\mathbf{V}_{\mathbf{D}}), (k_1 - m)(p - m))$ y las χ_i^2 son chi-cuadradas independientes con 1 grado de libertad y pesos ω_i dados por los autovalores ordenados de $\mathbf{Q}$, el valor poblacional de $\widehat{\mathbf{Q}}$. El estadístico del segundo test estudiado en [**Bura and Yang, 2011**] es

$$\Lambda_2(m) = n\text{vec}(\widehat{\mathbf{D}}_0^T)\widehat{\mathbf{Q}}^{\dagger}\text{vec}(\widehat{\mathbf{D}}_0)$$

con $\widehat{\mathbf{Q}}^{\dagger}$ la inversa de Moore-Penrose de $\widehat{\mathbf{Q}}$. Bajo la suposición de que $m = \text{rank}(\mathbf{D})$, $\Lambda_2(m)$ es asintóticamente chi-cuadrado con $\text{mín}(\text{rank}(\text{var}(\widehat{\mathbf{D}})), (k_1 - m)(p - m))$ grados de libertad.

La dimensión es estimada como el primer valor de m para el cual la hipótesis $d = m$ no se puede rechazar a un pre-especificado valor de α cuando se llevan a cabo los test secuenciales de $d = m$ versus $d > m$ para $m = 0, \ldots, \text{mín}(k_1, p)$.

Las hipótesis en los test secuenciales están jerárquicamente ordenados en el sentido de que para probar $d = m + 1$, uno tiene que rechazar primero $d = m$. Dicha aplicación secuencial de tests no ajustados que requieren $\text{mín}(k_1, p)$ pasos para testear $\text{mín}(k_1, p)$ hipótesis nulas se llama *gatekeeping serial* y es análoga a los test de unión-intersección ([**Berger, 1982**]). Una característica importante de este proceso es que cada hipótesis nula se testea secuencialmente a nivel α general y es controlado para el test secuencial completo (ver [**Dmitrienko et al., 2010**], Sección 5.3; [**Westfall and Krishen, 2001**]).

3.6. Demostraciones del Cápitulo 3

Consistencia del estimador MLE para una parametrización en particular

Sea $\mathbb{R}_d^{k\times p} = \{\mathbf{M} \in \mathbb{R}^{k\times p} \text{ de rango } d \leq \text{mín}(k,p)\}$. Supongamos que el verdadero parámetro de interés $\boldsymbol{\Gamma}_0 = \text{vec}(\mathbf{D}_0)$ es tal que $\mathbf{D}_0 \in \mathbb{R}_d^{k\times p}$. El objetivo es estudiar la existencia, consistencia y distribución asintótica del estimador MLE en el espacio $\mathbb{R}_d^{k\times p}$, el cual hemos denotado con $\widehat{\boldsymbol{\Gamma}}_{RR}$. $\widehat{\boldsymbol{\Gamma}}_{RC}$ representa el MLE bajo el modelo de rango completo en el espacio $\mathbb{R}_r^{k\times p}$.

Sin pérdida de generalidad, asumimos que $\mathbf{D}_0 = \begin{pmatrix} \mathbf{I}_d \\ \mathbf{A}_0 \end{pmatrix} \mathbf{B}_0$, con $\mathbf{B}_0 \in \mathbb{R}_d^{d\times p}$ y $\mathbf{A}_0 \in \mathbb{R}^{(k-d)\times d}$. Luego, definimos el siguiente conjunto

$$\Theta = \{\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) : \boldsymbol{\theta}_1 \in \mathbb{R}^{(k-d)\times d} \text{ y } \boldsymbol{\theta}_2 \in \mathbb{R}^{dp} : \boldsymbol{\theta}_2 = \text{vec}(T), T \in \mathbb{R}_d^{d\times p}\} = \{\Theta_1 \times \Theta_2\}.$$

El conjunto Θ es abierto como lo establece el siguiente lema.

Lema 3.5. *El conjunto*

$$\Theta = \{(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) : \boldsymbol{\theta}_1 \in \mathbb{R}^{(k-d)\times d} \text{ y } \boldsymbol{\theta}_2 \in \mathbb{R}^{dp} : \boldsymbol{\theta}_2 = \text{vec}(\mathbf{T}), \mathbf{T} \in \mathbb{R}_d^{d\times p}\} = \{\Theta_1 \times \Theta_2\}$$

es abierto.

Demostración. El espacio $\Theta_1 = \mathbb{R}^{(k-d)\times d}$ es abierto. Así, necesitamos probar que

$$\Theta_2 = \{\boldsymbol{\theta}_2 \in \mathbb{R}^{dp} : \boldsymbol{\theta}_2 = \text{vec}(\mathbf{T}), \mathbf{T} \in \mathbb{R}_d^{d\times p}\} \tag{3.11}$$

es un conjunto abierto. Sea Θ_2^c el complemento de Θ_2. Ya que Θ_2^c es cerrado si y solo si Θ_2 es abierto, debemos ver que Θ_2^c es cerrado en $\mathbb{R}^{dp}$. Consideremos la sucesión convergente $\{\boldsymbol{\theta}_n\} \subset \Theta_2^c$ con $\boldsymbol{\theta}_n \to \boldsymbol{\theta}$ (esta sucesión existe ya que $\mathbb{R}^{dr}$ es un espacio métrico). Luego, debemos ver que $\boldsymbol{\theta} \in \Theta_2^c$. De (3.11), $\boldsymbol{\theta}_n \in \Theta_2^c$ implica que $\boldsymbol{\theta}_n = \text{vec}(\mathbf{T}_n)$ con $\mathbf{T}_n \in \mathbb{R}_d^{d\times p}$ de rango estrictamente menor que d, luego $|\mathbf{T}_n\mathbf{T}_n^T| = 0$. Ya que $\text{vec}(\mathbf{T}_n) \to \boldsymbol{\theta} = \text{vec}(\mathbf{T})$ y la traspuesta y el determinante son funciones continuas, $|\mathbf{T}\mathbf{T}^T| = 0$. Esto es, $\mathbf{T}$ es de rango menor que d y asi $\boldsymbol{\theta} \in \Theta_2^c$. □

Además, definimos

$$\Omega = \{\mathbf{m} \in \mathbb{R}^{kp} : \mathbf{m} = \text{vec}(\mathbf{M}), \mathbf{M} \in \mathbb{R}_d^{k\times p} \text{ con } \mathbf{M} = \begin{pmatrix} \mathbf{I}_d \\ \mathbf{A} \end{pmatrix} \mathbf{B}, \text{ donde } \mathbf{A} \in \mathbb{R}^{(k-d)\times d} \text{ y } \mathbf{B} \in \mathbb{R}_d^{d\times p}\}$$

y la función $\mathbf{p} : \Theta \to \Omega$ por $\mathbf{p}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \boldsymbol{\Gamma}$ tal que

$$\boldsymbol{\Gamma} = \text{vec}(\mathbf{D}) = \text{vec}\left(\begin{pmatrix} \mathbf{I}_d \\ \mathbf{A} \end{pmatrix} \mathbf{B}\right) \text{ con } \text{vec}(\mathbf{A}) = \boldsymbol{\theta}_1 \text{ y } \text{vec}(\mathbf{B}) = \boldsymbol{\theta}_2.$$

Puede probarse que $\mathbf{p}$ es uno a uno, bicontinua y dos veces continuamente diferenciable. Esto es consecuencia del hecho que el Jacobiano de la función $\mathbf{p}$, que denotamos con $\boldsymbol{\Delta}_{\mathbf{p}}(\boldsymbol{\theta})$, es de rango completo. Con la anterior notación indicamos con $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0)$ al valor del parámetro verdadero. Supongamos que las condiciones de regularidad del Teorema 5.1 de **[Lehmann and Casella, 1998]** son verificables. Estas son:

- Existe un subconjunto abierto W de Θ que contiene el valor verdadero $\boldsymbol{\theta}_0$ tal que para casi todo $\mathbf{F}$, el logaritmo de la densidad $f(\mathbf{Y}|\mathbf{F}, \boldsymbol{\theta})$ admite tercera derivada $(\partial^3/\partial^3\theta^3) \log f(\mathbf{Y}|\mathbf{F}, \boldsymbol{\theta})$ para todo $\boldsymbol{\theta} \in \Theta$. Esta es la misma condición de regularidad pedida para el modelo de rango completo ya que la tercera derivada con respecto a $\boldsymbol{\theta}$ es combinación de las derivadas de orden 3 de f y $\mathbf{p}$.
- Necesitamos que la primera y la segunda derivada de $\log f$ verifique las ecuaciones $\mathbf{E}\left(\dfrac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{Y}|\mathbf{F}, \boldsymbol{\theta})\right) = 0$ y

$$\mathbf{E}_{\boldsymbol{\theta}}\left(-\frac{\partial^2}{\partial^2 \boldsymbol{\theta}^2} \log f(\mathbf{Y}|\mathbf{F}, \boldsymbol{\theta})\right) = \mathbf{I}(\boldsymbol{\theta}). \tag{3.12}$$

  Tenemos que

$$\begin{aligned}
\frac{\partial \log f(\mathbf{Y}|\mathbf{F}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \boldsymbol{\Delta}_{\mathbf{p}}^T(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\left(\mathbf{F}^T(\mathbf{T}(\mathbf{Y}) - \nabla\psi(\mathbf{F}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)))\right) \\
\frac{\partial^2 \log f(\mathbf{Y}|\mathbf{F}, \boldsymbol{\theta})}{\partial^2 \boldsymbol{\theta}^2}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= -\boldsymbol{\Delta}_{\mathbf{p}}^T(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\mathbf{F}^T\nabla^2\psi(\mathbf{F}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2))\mathbf{F}\boldsymbol{\Delta}_{\mathbf{p}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\
&\quad + \left[\mathbf{T}(\mathbf{Y}) - \nabla\psi(\mathbf{F}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2))^T\mathbf{F} \otimes \mathbf{I}_{(k-d)d+dp}\right]\nabla\boldsymbol{\Delta}_{\mathbf{p}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)
\end{aligned}$$

  Y usando el hecho que

$$\mathbf{E}_{(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}\left[\mathbf{T}(\mathbf{Y}) - \nabla\psi(\mathbf{F}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2))\right] = 0 \tag{3.13}$$

  obtenemos la primera ecuación directamente y nuevamente si (3.13) se verifica en el modelo de rango completo, también se verifica aquí.
- Necesitamos que $\mathbf{I}(\boldsymbol{\theta})$ sea finita y definida positiva para todo $\boldsymbol{\theta} \in \Theta$. Aquí $\mathbf{I}(\boldsymbol{\theta}) = \boldsymbol{\Delta}_{\mathbf{p}}^T(\boldsymbol{\theta})\mathbf{V}^{-1}\boldsymbol{\Delta}_{\mathbf{p}}(\boldsymbol{\theta})$ donde $\mathbf{V}^{-1} = \mathbf{E}\left(\mathbf{F}\nabla^2\psi(\mathbf{F}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2))\mathbf{F}\right)$ es finita y definida positiva. Luego, esto es cierto si se verifica para el modelo de rango completo ya que $\boldsymbol{\Delta}_{\mathbf{p}}(\boldsymbol{\theta})$ es de rango completo. Por lo tanto los estadísticos

$$\frac{\partial}{\partial \boldsymbol{\theta}_1} \log f(\mathbf{Y}|\mathbf{F}, \boldsymbol{\theta}), \ldots, \frac{\partial}{\partial \boldsymbol{\theta}_s} \log f(\mathbf{Y}|\mathbf{F}, \boldsymbol{\theta})$$

  son independientes con probabilidad 1.

- Suponemos que existe una función M tal que

$$\left|\frac{\partial^3}{\partial^3 \boldsymbol{\theta}^3} \log f(\mathbf{Y}|\mathbf{F}, \boldsymbol{\theta})\right| \leq M(\mathbf{F}) \text{ para todo } \boldsymbol{\theta} \in W$$

  donde $\mathbf{E}(M(\mathbf{F})) < \infty$.
- La distribución $f(\mathbf{Y}|\mathbf{F}, \boldsymbol{\theta})$ de las observaciones son distintas (en otro caso, $\boldsymbol{\theta}$ no puede ser estimado consistentemente). Esta es la misma condición para el modelo de rango completo y si esta es cierta, lo será para el caso de rango reducido ya que $\mathbf{p}$ es biyectiva.
- Las distribuciones $f(\mathbf{Y}|\mathbf{F}, \boldsymbol{\theta})$ tiene soporte común. Esto es cierto pues estamos en el contexto de distribuciones que pertenecen a familia de exponenciales.

Luego, existe la solución $(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n})$ de la ecuación de verosimilitud que son consistentes para $(\boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0)$ y asintóticamente normal con

$$\sqrt{n}((\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2) - (\boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0)) \to \mathcal{N}(0, (\boldsymbol{\Delta}_{\mathbf{p}}(\boldsymbol{\theta}_0)^T \mathbf{V}_{\boldsymbol{\Gamma}_0}^{-1} \boldsymbol{\Delta}_{\mathbf{p}}(\boldsymbol{\theta}_0))^{-1}), \tag{3.14}$$

donde $\mathbf{V}_{\boldsymbol{\Gamma}_0}^{-1} = \mathbf{E}_{\boldsymbol{\theta}_0}\left(\mathbf{F}\nabla^2\psi(\mathbf{F}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2))\mathbf{F}\right)$. Usando el Delta método,

$$\sqrt{n}(\widehat{\boldsymbol{\Gamma}}_{RR} - \mathbf{p}(\boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0)) \to \mathcal{N}(0, \boldsymbol{\Delta}_{\mathbf{p}}(\boldsymbol{\theta}_0)(\boldsymbol{\Delta}_{\mathbf{p}}(\boldsymbol{\theta}_0)^T \mathbf{V}_{\boldsymbol{\Gamma}_0}^{-1} \boldsymbol{\Delta}_{\mathbf{p}}(\boldsymbol{\theta}_0))^{-1} \boldsymbol{\Delta}_{\mathbf{p}}(\boldsymbol{\theta}_0)^T) \tag{3.15}$$

y expresamos (3.15) como

$$\sqrt{n}(\widehat{\boldsymbol{\Gamma}}_{RR} - \boldsymbol{\Gamma}_0) \to \mathcal{N}(0, \boldsymbol{\Delta}_{\mathbf{p}}(\boldsymbol{\theta}_0)(\boldsymbol{\Delta}_{\mathbf{p}}(\boldsymbol{\theta}_0)^T \mathbf{V}_{\boldsymbol{\Gamma}_0}^{-1} \boldsymbol{\Delta}_{\mathbf{p}}(\boldsymbol{\theta}_0))^{-1} \boldsymbol{\Delta}_{\mathbf{p}}(\boldsymbol{\theta}_0)^T). \tag{3.16}$$

Luego, teniendo en cuenta que la re-parametrización de $\boldsymbol{\theta}$ no debe afectar las propiedades asintóticas de $\mathbf{p}(\boldsymbol{\theta})$, a través de los resultados de **[Shapiro, 1986]** se obtiene (3.3).

Observar que este caso se ha considerado que los predictores son aleatorios, el caso en que los predictores $\mathbf{X}$ son considerados valores fijos, puede estudiarse por medio de condiciones adicionales en la matriz de diseño $\mathbf{F}$.

Prueba de (3.8): En la Sección 1.3.4 mostramos que la varianza asintótica de los estimadores de $(\text{vec}(\boldsymbol{\Gamma}_1)^T, \text{vech}(\boldsymbol{\Sigma}^{-1})^T)$ bajo el modelo GLMM de rango completo (1.20) es $\mathbf{V}_{\boldsymbol{\Gamma}}$, la cual está expresada en (3.7).

Además, como $\boldsymbol{\Gamma}_1 = \boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}$ con $\boldsymbol{\beta}$ la matriz de coeficientes del modelo de regresión lineal clásico (1.1), mostramos que al aplicar la regla de Cramer a la varianza asintótica (1.3), el

resultado que se obtiene coincide con $\mathbf{V_\Gamma}$. Es decir que si llamamos

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ 0 & \mathbf{S}_{22} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial \text{vec}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta})}{\partial \text{vec}^T(\boldsymbol{\beta})} & \dfrac{\partial \text{vec}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta})}{\partial \text{vech}^T(\boldsymbol{\Sigma})} \\ \dfrac{\partial \text{vech}(\boldsymbol{\Sigma}^{-1})}{\partial \text{vec}^T(\boldsymbol{\beta})} & \dfrac{\partial \text{vech}(\boldsymbol{\Sigma}^{-1})}{\partial \text{vech}^T(\boldsymbol{\Sigma})} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_p \otimes \boldsymbol{\Sigma}^{-1} & -(\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{D}_r \\ 0 & -\mathbf{E}_r(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{D}_r \end{pmatrix},$$

tenemos que

$$\mathbf{V_\Gamma} = \mathbf{S} \begin{pmatrix} J_{\boldsymbol{\beta}}^{-1} & 0 \\ 0 & J_{\boldsymbol{\Sigma}}^{-1} \end{pmatrix} \mathbf{S}^T = \mathbf{S} \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma} & 0 \\ 0 & 2\mathbf{E}_r\left(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}\right)\mathbf{E}_r^T \end{pmatrix} \mathbf{S}^T.$$

Luego, podemos expresar $\mathbf{V}_\mathbf{\Gamma}^{-1}$ de la siguiente forma:

$$\begin{aligned} \mathbf{V}_\mathbf{\Gamma}^{-1} &= \begin{pmatrix} \mathbf{S}_{11} & 0 \\ \mathbf{S}_{12}^T & \mathbf{S}_{22}^T \end{pmatrix}^{-1} \begin{pmatrix} J_{\boldsymbol{\beta}} & 0 \\ 0 & J_{\boldsymbol{\Sigma}} \end{pmatrix} \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ 0 & \mathbf{S}_{22} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \mathbf{S}_{11}^{-1} & 0 \\ -(\mathbf{S}_{22}^T)^{-1}\mathbf{S}_{12}^T\mathbf{S}_{11}^{-1} & (\mathbf{S}_{22}^T) \end{pmatrix} \begin{pmatrix} J_{\boldsymbol{\beta}} & 0 \\ 0 & J_{\boldsymbol{\Sigma}} \end{pmatrix} \begin{pmatrix} \mathbf{S}_{11}^{-1} & -\mathbf{S}_{11}^{-1}\mathbf{S}_{12}^T(\mathbf{S}_{22}^T)^{-1} \\ 0 & (\mathbf{S}_{22}^T)^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{S}_{11}^{-1}J_{\boldsymbol{\beta}}\mathbf{S}_{11}^{-1} & -\mathbf{S}_{11}^{-1}J_{\boldsymbol{\beta}}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}^T(\mathbf{S}_{22}^T)^{-1} \\ -(\mathbf{S}_{22}^T)^{-1}\mathbf{S}_{12}^T\mathbf{S}_{11}^{-1}J_{\boldsymbol{\beta}}\mathbf{S}_{11}^{-1} & (\mathbf{S}_{22}^T)^{-1}J_{\boldsymbol{\Sigma}}\mathbf{S}_{22}^{-1} + (\mathbf{S}_{22}^T)^{-1}\mathbf{S}_{21}^T\mathbf{S}_{11}^{-1}J_{\boldsymbol{\beta}}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1} \end{pmatrix}. \end{aligned}$$

Ahora nos ocuparemos de desarrollar la expresión (3.6) detalladamente. En primer lugar,

$$(\boldsymbol{\Delta}^T\mathbf{V}_\mathbf{\Gamma}^{-1}\boldsymbol{\Delta})^\ddagger = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}^\ddagger \doteq \begin{pmatrix} \mathbf{M}^{11} & \mathbf{M}^{12} \\ \mathbf{M}^{21} & \mathbf{M}^{22} \end{pmatrix},$$

donde el supraíndice ‡ indica una inversa generalizada de la matriz,

$$\begin{aligned} \mathbf{M}_{11} &= \boldsymbol{\Delta}_1^T\mathbf{S}_{11}^{-1}J_{\boldsymbol{\beta}}\mathbf{S}_{11}^{-1}\boldsymbol{\Delta}_1 \\ \mathbf{M}_{12} &= -\boldsymbol{\Delta}_1^T\mathbf{S}_{11}^{-1}J_{\boldsymbol{\beta}}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}(\mathbf{S}_{22}^T)^{-1} \\ \mathbf{M}_{21} &= \mathbf{M}_{12}^T \\ \mathbf{M}_{22} &= (\mathbf{S}_{22}^T)^{-1}J_{\boldsymbol{\Sigma}}\mathbf{S}_{22}^{-1} + (\mathbf{S}_{22}^T)^{-1}\mathbf{S}_{21}^T\mathbf{S}_{11}^{-1}J_{\boldsymbol{\beta}}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1} \end{aligned}$$

y $\mathbf{M}^{11}, \mathbf{M}^{12}, \mathbf{M}^{21}$ y $\mathbf{M}^{22}$ se corresponden con los bloques de la inversa generalizada. Finalmente tenemos la siguiente expresión

$$\begin{aligned} \boldsymbol{\Delta}(\boldsymbol{\Delta}^T\mathbf{V}_\mathbf{\Gamma}^{-1}\boldsymbol{\Delta})^\ddagger\boldsymbol{\Delta}^T &= \begin{pmatrix} \boldsymbol{\Delta}_1 & 0 \\ 0 & \mathbf{I}_{r(r+1)/2} \end{pmatrix} \begin{pmatrix} \mathbf{M}^{11} & \mathbf{M}^{12} \\ \mathbf{M}^{21} & \mathbf{M}^{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Delta}_1^T & 0 \\ 0 & \mathbf{I}_{r(r+1)/2} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Delta}_1\mathbf{M}^{11}\boldsymbol{\Delta}_1^T & \boldsymbol{\Delta}_1\mathbf{M}^{12} \\ \mathbf{M}^{21}\boldsymbol{\Delta}_1^T & \mathbf{M}^{22} \end{pmatrix}. \end{aligned}$$

Primero veamos que $\mathbf{M}^{22} = \mathbf{S}_{22} J_{\boldsymbol{\Sigma}}^{-1} \mathbf{S}_{22}^T$, pues esta expresión corresponde a la varianza asintótica del estimador de máxima verosimilitud de $\boldsymbol{\Gamma}_2$ bajo el modelo (3.5). Para ello tengamos en cuenta que

$$\mathbf{S}_{11}^{-1} J_{\boldsymbol{\beta}} \mathbf{S}_{11}^{-1} = \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}$$

$$\mathbf{S}_{11}^{-1} J_{\boldsymbol{\beta}} \mathbf{S}_{11}^{-1} \mathbf{S}_{12} = -(\boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\beta} \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_r) \mathbf{D}_r$$

$$\mathbf{S}_{12}^T \mathbf{S}_{11}^{-1} J_{\boldsymbol{\beta}} \mathbf{S}_{11}^{-1} \mathbf{S}_{12} = \mathbf{D}_r^T (\boldsymbol{\Sigma}^{-1} \boldsymbol{\beta} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\beta}^T \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) \mathbf{D}_r.$$

Luego de acuerdo al Lema A.8 del Apéndice,

$$\begin{aligned}
\mathbf{M}^{22} &= (\mathbf{M}_{22} - \mathbf{M}_{21} \mathbf{M}_{11}^{\ddagger} \mathbf{M}_{21})^{\ddagger} \\
&= ((\mathbf{S}_{22}^T)^{-1} J_{\boldsymbol{\Sigma}} \mathbf{S}_{22}^{-1} + (\mathbf{S}_{22}^T)^{-1} \mathbf{S}_{12}^T \mathbf{S}_{11}^{-1} J_{\boldsymbol{\beta}} \mathbf{S}_{11}^{-1} \mathbf{S}_{12} \mathbf{S}_{22}^{-1} \\
&\quad -(\mathbf{S}_{22}^T)^{-1} \mathbf{S}_{12}^T \mathbf{S}_{11}^{-1} J_{\boldsymbol{\beta}} \mathbf{S}_{11}^{-1} \boldsymbol{\Delta}_1 (\boldsymbol{\Delta}_1^T \mathbf{S}_{11}^{-1} J_{\boldsymbol{\beta}} \mathbf{S}_{11}^{-1} \boldsymbol{\Delta}_1)^{\ddagger} \boldsymbol{\Delta}_1^T \mathbf{S}_{11}^{-1} J_{\boldsymbol{\beta}} \mathbf{S}_{11}^{-1} \mathbf{S}_{12}^T (\mathbf{S}_{22}^T)^{-1})^{\ddagger}.
\end{aligned}$$

Veremos ahora que los últimos dos términos son iguales. En efecto, aplicando la Propiedad A.4 tenemos por un lado que

$$\begin{aligned}
(\mathbf{S}_{22}^T)^{-1} \mathbf{S}_{12}^T \mathbf{S}_{11}^{-1} J_{\boldsymbol{\beta}} \mathbf{S}_{11}^{-1} \mathbf{S}_{12} \mathbf{S}_{22}^{-1} &= \mathbf{D}_r^T (\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) \mathbf{E}_r^T \mathbf{D}_r^T (\boldsymbol{\Sigma}^{-1} \boldsymbol{\beta} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\beta}^T \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) \mathbf{D}_r \mathbf{E}_r (\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) \mathbf{D}_r \\
&= \mathbf{D}_r^T (\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) \frac{\mathbf{I}_{r^2} + \mathbf{K}_{rr}}{2} (\boldsymbol{\Sigma}^{-1} \boldsymbol{\beta} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\beta}^T \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) \\
&\quad \frac{\mathbf{I}_{r^2} + \mathbf{K}_{rr}}{2} (\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) \mathbf{D}_r \\
&= \mathbf{D}_r^T (\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) (\boldsymbol{\Sigma}^{-1} \boldsymbol{\beta} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\beta}^T \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) (\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) \mathbf{D}_r \\
&= \mathbf{D}_r^T (\boldsymbol{\beta} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\beta} \otimes \boldsymbol{\Sigma}) \mathbf{D}_r.
\end{aligned}$$

Por otro lado,

$$(\boldsymbol{\beta} \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \mathbf{I}_r) \boldsymbol{\Delta}_1 (\boldsymbol{\Delta}_1^T \mathbf{S}_{11}^{-1} J_{\boldsymbol{\beta}} \mathbf{S}_{11}^{-1} \boldsymbol{\Delta}_1)^{\ddagger} \boldsymbol{\Delta}_1^T (\boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\beta}^T \otimes \mathbf{I}_r) =$$

$$(\boldsymbol{\beta} \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \mathbf{I}_r) \boldsymbol{\Delta}_1 (\boldsymbol{\Delta}_1^T (\boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}) \boldsymbol{\Delta}_1)^{\ddagger} \boldsymbol{\Delta}_1^T (\boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\beta}^T \otimes \mathbf{I}_r) =$$

$$(\boldsymbol{\beta} \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \mathbf{I}_r) (\mathbf{P}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})} \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \boldsymbol{\Sigma}^{-1} + \mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})} \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma})} \boldsymbol{\Sigma}^{-1}) (\boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\beta}^T \otimes \mathbf{I}_r) =$$

$$(\boldsymbol{\beta} \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{B}^T (\mathbf{B} \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{B}^T)^{-1} \mathbf{B} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\beta}^T \otimes \boldsymbol{\Sigma}^{-1})$$

$$+\boldsymbol{\beta} \boldsymbol{\Sigma}_{\mathbf{X}} (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} - \mathbf{B}^T (\mathbf{B} \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{B}^T)^{-1} \mathbf{B} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}) \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\beta}^T \otimes \mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma}^{-1})} \boldsymbol{\Sigma}^{-1} =$$

$$(\boldsymbol{\beta} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\beta}^T \otimes \boldsymbol{\Sigma}^{-1})$$

donde hemos usado en la última igualdad que $\boldsymbol{\beta} = \boldsymbol{\Sigma} \boldsymbol{\Gamma}_1 = \boldsymbol{\Sigma} \mathbf{A} \mathbf{B}$ y en la segunda igualdad que

$$\boldsymbol{\Delta}_1 (\boldsymbol{\Delta}_1^T (\boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}) \boldsymbol{\Delta}_1)^{\ddagger} \boldsymbol{\Delta}_1^T = \mathbf{P}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})} \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \boldsymbol{\Sigma}^{-1} + \mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})} \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma})} \boldsymbol{\Sigma}^{-1}. \tag{3.17}$$

La igualdad (3.17) se deduce de forma análoga a la demostración del Proposición 2.4 del Capítulo 2. Luego, el tercer término de la expresión de $\mathbf{M}^{22}$ es

$$
\begin{aligned}
&(\mathbf{S}_{22}^T)^{-1}\mathbf{S}_{12}^T\mathbf{S}_{11}^{-1}J_{\boldsymbol{\beta}}\mathbf{S}_{11}^{-1}\boldsymbol{\Delta}_1(\boldsymbol{\Delta}_1^T\mathbf{S}_{11}^{-1}J_{\boldsymbol{\beta}}\mathbf{S}_{11}^{-1}\boldsymbol{\Delta}_1)^{\ddagger}\boldsymbol{\Delta}_1^T\mathbf{S}_{11}^{-1}J_{\boldsymbol{\beta}}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}^T(\mathbf{S}_{22}^T)^{-1} = \\
&\mathbf{D}_r^T(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\mathbf{E}_r^T\mathbf{D}_r^T(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_r)(\boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\otimes\boldsymbol{\Sigma}^{-1})(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_r)\mathbf{D}_r^T\mathbf{E}_r^T(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\mathbf{D}_r = \\
&\mathbf{D}_r^T(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\frac{\mathbf{I}_{r^2}+\mathbf{K}_{rr}}{2}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\frac{\mathbf{I}_{r^2}+\mathbf{K}_{rr}}{2}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\mathbf{D}_r = \\
&\mathbf{D}_r^T(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\mathbf{D}_r = \\
&\mathbf{D}_r^T(\boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\otimes\boldsymbol{\Sigma})\mathbf{D}_r.
\end{aligned}
$$

Por lo tanto el segundo y tercer término se cancelan y

$$\mathbf{M}^{22} = ((\mathbf{S}_{22}^T)^{-1}J_{\boldsymbol{\Sigma}}\mathbf{S}_{22}^{-1})^{\ddagger} = \mathbf{S}_{22}^T J_{\boldsymbol{\Sigma}}^{-1}\mathbf{S}_{22}.$$

Ahora vamos a obtener una expresión explícita de $\boldsymbol{\Delta}_1\mathbf{M}^{11}\boldsymbol{\Delta}_1^T$ que corresponde a la varianza asintótica del estimador de máxima verosimilitud de $\boldsymbol{\Gamma}_1$ bajo el modelo (3.5). Nuevamente, usando el Lema A.8 tenemos

$$
\begin{aligned}
\mathbf{M}^{11} &= \mathbf{M}_{11}^{\ddagger} + \mathbf{M}_{11}^{\ddagger}\mathbf{M}_{12}\mathbf{S}_{22}J_{\boldsymbol{\Sigma}}^{-1}\mathbf{S}_{22}^T\mathbf{M}_{12}^T\mathbf{M}_{11}^{\ddagger} \\
&= (\boldsymbol{\Delta}_1^T(\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\boldsymbol{\Sigma})\boldsymbol{\Delta}_1)^{\ddagger} + (\boldsymbol{\Delta}_1^T(\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\boldsymbol{\Sigma})\boldsymbol{\Delta}_1)^{\ddagger}\boldsymbol{\Delta}_1^T 2(\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_r) \\
&\quad \mathbf{D}_r\mathbf{E}_r(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\mathbf{E}_r^T\mathbf{D}_r^T(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\mathbf{I}_r)\boldsymbol{\Delta}_1(\boldsymbol{\Delta}_1^T(\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\boldsymbol{\Sigma})\boldsymbol{\Delta}_1)^{\ddagger}
\end{aligned}
$$

y a partir de la Propiedad A.3 y de la Propiedad A.4 del Apéndice podemos expresar

$$
\begin{aligned}
&2(\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_r)\mathbf{D}_r\mathbf{E}_r(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\mathbf{E}_r^T\mathbf{D}_r^T(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\mathbf{I}_r) = \\
&2(\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_r)\frac{\mathbf{I}_{r^2}+\mathbf{K}_{rr}}{2}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\frac{\mathbf{I}_{r^2}+\mathbf{K}_{rr}}{2}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\mathbf{I}_r) = \\
&(\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_r)((\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})+(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\mathbf{K}_{rr})(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\mathbf{I}_r) = \\
&(\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\beta\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\boldsymbol{\Sigma})+(\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\otimes\boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}})\mathbf{K}_{pr}.
\end{aligned}
$$

Por lo tanto,

$$
\begin{aligned}
\boldsymbol{\Delta}_1\mathbf{M}^{11}\boldsymbol{\Delta}_1^T &= \boldsymbol{\Delta}_1(\boldsymbol{\Delta}_1^T(\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\boldsymbol{\Sigma})\boldsymbol{\Delta}_1)^{\ddagger}\boldsymbol{\Delta}_1^T + \boldsymbol{\Delta}_1(\boldsymbol{\Delta}_1^T(\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\boldsymbol{\Sigma})\boldsymbol{\Delta}_1)^{\ddagger}\boldsymbol{\Delta}_1^T \\
&\quad [(\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\beta\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\boldsymbol{\Sigma})+(\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\otimes\boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}})\mathbf{K}_{pr}]\boldsymbol{\Delta}_1(\boldsymbol{\Delta}_1^T(\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\boldsymbol{\Sigma})\boldsymbol{\Delta}_1)^{\ddagger}\boldsymbol{\Delta}_1^T.
\end{aligned}
$$

Utilizando la igualdad (3.17) para $\boldsymbol{\Delta}_1(\boldsymbol{\Delta}_1^T(\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\boldsymbol{\Sigma})\boldsymbol{\Delta}_1)^{\ddagger}\boldsymbol{\Delta}_1^T$ y que $\boldsymbol{\beta} = \boldsymbol{\Sigma}\mathbf{A}\mathbf{B}$, tenemos que

$$(\mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\otimes\mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma})})(\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\beta\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\boldsymbol{\Sigma}) = 0 \tag{3.18}$$

$$(\mathbf{Q}_{\mathbf{B}^T(\boldsymbol{\Sigma}_{\mathbf{x}})}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\otimes\mathbf{P}_{\mathbf{A}(\boldsymbol{\Sigma})})(\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta}^T\otimes\boldsymbol{\beta}\boldsymbol{\Sigma}_{\mathbf{X}}) = 0. \tag{3.19}$$

Usando (3.18) y (3.19), el segundo término de $\mathbf{\Delta}_1\mathbf{M}^{11}\mathbf{\Delta}_1^T$ resulta

$$(\mathbf{P}_{\mathbf{B}^T(\mathbf{\Sigma_x})}\mathbf{\Sigma}_\mathbf{x}^{-1}\otimes\mathbf{\Sigma}^{-1})[(\mathbf{\Sigma_X}\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1}\beta\mathbf{\Sigma_X}\otimes\mathbf{\Sigma})+(\mathbf{\Sigma_X}\boldsymbol{\beta}^T\otimes\boldsymbol{\beta}\mathbf{\Sigma_X})\mathbf{K}_{pr}](\mathbf{P}_{\mathbf{B}^T(\mathbf{\Sigma_x})}\mathbf{\Sigma}_\mathbf{x}^{-1}\otimes\mathbf{\Sigma}^{-1})=$$
$$(\mathbf{B}^T\mathbf{A}^T\mathbf{\Sigma AB})+(\mathbf{B}^T\mathbf{A}^T\otimes\mathbf{AB})\mathbf{K}_{pr}=$$
$$(\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1}\boldsymbol{\beta}\otimes\mathbf{\Sigma}^{-1})+(\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1}\otimes\mathbf{\Sigma}^{-1}\boldsymbol{\beta})\mathbf{K}_{pr}.$$

Por lo tanto

$$\begin{aligned}\mathbf{\Delta}_1\mathbf{M}^{11}\mathbf{\Delta}_1^T &= \mathbf{P}_{\mathbf{B}^T(\mathbf{\Sigma_x})}\mathbf{\Sigma}_\mathbf{x}^{-1}\otimes\mathbf{\Sigma}^{-1}+\mathbf{Q}_{\mathbf{B}^T(\mathbf{\Sigma_x})}\mathbf{\Sigma}_\mathbf{x}^{-1}\otimes\mathbf{P}_{\mathbf{A}(\mathbf{\Sigma})}\mathbf{\Sigma}^{-1}\\ &\quad+(\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1}\boldsymbol{\beta}\otimes\mathbf{\Sigma}^{-1})+(\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1}\otimes\mathbf{\Sigma}^{-1}\boldsymbol{\beta})\mathbf{K}_{pr}\end{aligned}$$

o lo que es lo mismo,

$$\begin{aligned}\mathbf{\Delta}_1\mathbf{M}^{11}\mathbf{\Delta}_1^T &= \mathbf{P}_{\mathbf{B}^T(\mathbf{\Sigma_x})}\mathbf{\Sigma}_\mathbf{x}^{-1}\otimes\mathbf{\Sigma}^{-1}+\mathbf{Q}_{\mathbf{B}^T(\mathbf{\Sigma_x})}\mathbf{\Sigma}_\mathbf{x}^{-1}\otimes\mathbf{P}_{\mathbf{A}(\mathbf{\Sigma})}\mathbf{\Sigma}^{-1}\\ &\quad+(\mathbf{\Gamma}_1^T\mathbf{\Sigma}\mathbf{\Gamma}_1\otimes\mathbf{\Sigma}^{-1})+(\mathbf{\Gamma}_1^T\otimes\mathbf{\Gamma}_1)\mathbf{K}_{pr}\end{aligned}$$

que es lo que queríamos probar.

□

Prueba de (3.9): Por otro lado, de acuerdo a la Proposición 2.4, la distribución asintótica de los estimadores de máxima verosimilitud del modelo de regresión lineal de rango reducido es

$$\text{avar}\left(\sqrt{n}\begin{pmatrix}\text{vec}(\widehat{\mathbf{A}}_{\boldsymbol{\beta}}\widehat{\mathbf{B}}_{\boldsymbol{\beta}})\\ \text{vech}(\widehat{\mathbf{\Sigma}})\end{pmatrix}\right)=$$
$$\begin{pmatrix}(\mathbf{P}_{\mathbf{B}_{\boldsymbol{\beta}}^T(\mathbf{\Sigma_x})}\mathbf{\Sigma}_\mathbf{x}^{-1}\otimes\mathbf{\Sigma})+(\mathbf{Q}_{\mathbf{B}_{\boldsymbol{\beta}}^T(\mathbf{\Sigma_x})}\mathbf{\Sigma}_\mathbf{x}^{-1}\otimes\mathbf{P}_{\mathbf{A}_{\boldsymbol{\beta}}(\mathbf{\Sigma}^{-1})}\mathbf{\Sigma}) & 0\\ 0 & 2\mathbf{E}_r(\mathbf{\Sigma}\otimes\mathbf{\Sigma})\mathbf{E}_r^T\end{pmatrix}.$$

Luego, si queremos la distribución asintótica de $\text{vec}(\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{A}}_{\boldsymbol{\beta}}\widehat{\mathbf{B}}_{\boldsymbol{\beta}})$, utilizamos las siguientes derivadas

$$\begin{aligned}\frac{\partial\text{vec}(\mathbf{\Sigma}^{-1}\boldsymbol{\beta})}{\partial\text{vec}^T\boldsymbol{\beta}} &= \mathbf{I}_p\otimes\mathbf{\Sigma}^{-1}\\ \frac{\partial\text{vec}(\mathbf{\Sigma}^{-1}\boldsymbol{\beta})}{\partial\text{vech}^T\mathbf{\Sigma}} &= -(\boldsymbol{\beta}^T\otimes\mathbf{I}_r)(\mathbf{\Sigma}^{-1}\otimes\mathbf{\Sigma}^{-1})\mathbf{D}_r\end{aligned}$$

y aplicamos la regla de Cramer. De esta forma, $\text{vec}(\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{A}}_{\boldsymbol{\beta}}\widehat{\mathbf{B}}_{\boldsymbol{\beta}})$ es asintóticamente normal con la siguiente matriz de covarianza:

$$\begin{aligned}
\mathbf{V} &= \left(\mathbf{I}_p \otimes \mathbf{\Sigma}^{-1} \quad -(\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma}^{-1})\mathbf{D}_r \right) \text{avar} \begin{pmatrix} \text{vec}(\widehat{\mathbf{A}}_{\boldsymbol{\beta}}\widehat{\mathbf{B}}_{\boldsymbol{\beta}}) \\ \text{vech}(\widehat{\mathbf{\Sigma}}) \end{pmatrix} \begin{pmatrix} \mathbf{I}_p \otimes \mathbf{\Sigma}^{-1} \\ -\mathbf{D}_r^T(\mathbf{\Sigma}^{-1}\boldsymbol{\beta} \otimes \mathbf{\Sigma}^{-1}) \end{pmatrix} \\
&= (\mathbf{P}_{\mathbf{B}_{\boldsymbol{\beta}}^T(\mathbf{\Sigma_x})}\mathbf{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{\Sigma}^{-1}) + (\mathbf{Q}_{\mathbf{B}_{\boldsymbol{\beta}}^T(\mathbf{\Sigma_x})}\mathbf{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{\Sigma}^{-1}\mathbf{P}_{\mathbf{A}_{\boldsymbol{\beta}}(\mathbf{\Sigma}^{-1})}) \\
&\quad +2(\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma}^{-1})\mathbf{D}_r\mathbf{E}_r(\mathbf{\Sigma} \otimes \mathbf{\Sigma})\mathbf{E}_r^T\mathbf{D}_r^T(\mathbf{\Sigma}^{-1}\boldsymbol{\beta} \otimes \mathbf{\Sigma}^{-1}).
\end{aligned}$$

El segundo término de $\mathbf{V}$ puede simplificarse de la siguiente forma

$$\begin{aligned}
&2(\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma}^{-1})\mathbf{D}_r\mathbf{E}_r(\mathbf{\Sigma} \otimes \mathbf{\Sigma})\mathbf{E}_r^T\mathbf{D}_r^T(\mathbf{\Sigma}^{-1}\boldsymbol{\beta} \otimes \mathbf{\Sigma}^{-1}) = \\
&2(\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma}^{-1})\frac{\mathbf{I}_{r^2} + \mathbf{K}_{rr}}{2}(\mathbf{\Sigma} \otimes \mathbf{\Sigma})\frac{\mathbf{I}_{r^2} + \mathbf{K}_{rr}}{2}(\mathbf{\Sigma}^{-1}\boldsymbol{\beta} \otimes \mathbf{\Sigma}^{-1}) = \\
&(\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma}^{-1})((\mathbf{\Sigma} \otimes \mathbf{\Sigma}) + (\mathbf{\Sigma} \otimes \mathbf{\Sigma})\mathbf{K}_{rr})(\mathbf{\Sigma}^{-1}\boldsymbol{\beta} \otimes \mathbf{\Sigma}^{-1}) = \\
&(\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1}\boldsymbol{\beta} \otimes \mathbf{\Sigma}^{-1}) + (\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma}^{-1}\boldsymbol{\beta})\mathbf{K}_{pr}.
\end{aligned}$$

Por lo tanto,

$$\begin{aligned}
\mathbf{V} &= (\mathbf{P}_{\mathbf{B}_{\boldsymbol{\beta}}^T(\mathbf{\Sigma_x})}\mathbf{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{\Sigma}^{-1}) + (\mathbf{Q}_{\mathbf{B}_{\boldsymbol{\beta}}^T(\mathbf{\Sigma_x})}\mathbf{\Sigma}_{\mathbf{x}}^{-1} \otimes \mathbf{\Sigma}^{-1}\mathbf{P}_{\mathbf{A}_{\boldsymbol{\beta}}(\mathbf{\Sigma}^{-1})}) \\
&\quad +(\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1}\boldsymbol{\beta} \otimes \mathbf{\Sigma}^{-1}) + (\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1} \otimes \mathbf{\Sigma}^{-1}\boldsymbol{\beta})\mathbf{K}_{pr}.
\end{aligned}$$

□

CAPÍTULO 4

SIMULACIONES Y EJEMPLO CON DATOS REALES

En este capítulo se presentan simulaciones para analizar el comportamiento de los diferentes estimadores propuestos para los GLMM de rango reducido y para verificar los resultados del Teorema 3.1 y la Proposición 3.3 de forma empírica para diferentes tamaños de muestra y varios valores del rango de la matriz de coeficientes de la regresión d. Además en la Sección 4.3, se utilizan los resultados del Capítulo 3 para completar el ejemplo de [**Yee and Hastie, 2003**].

4.1. Simulaciones

Para las simulaciones consideramos los modelos lineales genealizados (4.1), (4.2), (4.3) y (4.4) que se detallan a continuación, para las familias exponenciales multinomial, Bernoulli multivariado y normal, respectivamente.

En todas las simulaciones primero fijamos la matriz de coeficientes $\mathbf{\Gamma}$. Las matrices $\mathbf{A}$ y $\mathbf{B}$ se obtuvieron a través de números aleatorios uniformes en el intervalo [0,1] y estandarizamos $\mathbf{\Gamma} = \mathbf{AB}$ tal que $\|\mathbf{\Gamma}\|_F = 1$ donde $\|\cdot\|_F$ denota la norma de Frobenius de las matrices. Definimos el error de estimación como $\|\mathbf{\Gamma} - \widehat{\mathbf{\Gamma}}\|_F$, donde $\widehat{\mathbf{\Gamma}}$ será el estimador correspondiente. En todos los modelos calculamos el estimador de rango completo y los estimadores de rango reducido: de máxima verosimilitud, sub-d y de minimización cuadrática. Para la distribución de $\mathbf{Y}|\mathbf{X}$, los predictores $\mathbf{X}$ fueron simulados con distribución $\mathcal{N}(0, \mathbf{I}_p)$. Todas las figuras que se presentan fueron generadas en base al promedio de los resultados obtenidos en 200 replicas de conjuntos de datos independientes.

Hemos considerado dos situaciones para los valores de la dimensión de la matriz $\mathbf{\Gamma} : k \times p$ y su rango. En un primer caso $(a) - (d, k, p) = (1, 10, 20)$, es decir que la reducción del rango es

muy clara y el segundo caso $(b) - (d, k, p) = (3, 6, 20)$, es decir que la matriz $\boldsymbol{\Gamma}$ es casi de rango completo.

En un primer caso, $\mathbf{Y}|\mathbf{X}$ es multinomial. Obtuvimos la muestra de $\mathbf{Y}$ teniendo en cuenta el siguiente modelo para $p_{1\mathbf{x}}, \ldots, p_{r\mathbf{x}}$, las probabilidades de que ocurra cada uno de los r posibles resultados:

$$\boldsymbol{\eta}_{\mathbf{x}} = \begin{pmatrix} \log(p_{1\mathbf{x}}/p_{r\mathbf{x}}) \\ \log(p_{2\mathbf{x}}/p_{r\mathbf{x}}) \\ \cdots \\ \log(p_{(r-1)\mathbf{x}}/p_{r\mathbf{x}}) \end{pmatrix} = \boldsymbol{\Gamma}\mathbf{x}. \tag{4.1}$$

En el segundo caso se considera que $\mathbf{Y}|\mathbf{X}$ tiene distribución Bernoulli multivariada, donde las componentes $Y_j|\mathbf{X}$, $j = 1, \ldots, r$ son independientes condicionadas en $\mathbf{X}$. Obtuvimos la muestra de $\mathbf{Y}$ teniendo en cuenta el siguiente modelo para los parámetros naturales de esta familia exponencial:

$$\boldsymbol{\eta}_{\mathbf{x}} = \begin{pmatrix} \log(p_{1\mathbf{x}}/(1-p_{1\mathbf{x}})) \\ \log(p_{2\mathbf{x}}/(1-p_{2\mathbf{x}})) \\ \cdots \\ \log(p_{r\mathbf{x}}/(1-p_{r\mathbf{x}})) \end{pmatrix} = \boldsymbol{\Gamma}\mathbf{x}. \tag{4.2}$$

En el tercer modelo debilitamos la suposición del modelo anterior de la independencia condicional dado los predictores, de las componentes del vector respuesta $\mathbf{Y}$. En su lugar, asumimos que la presencia de correlación de grado dos entre dichas componentes. Específicamente, suponemos que solamente están presentes las correlaciones entre Y_1 e Y_2, Y_3 e Y_4, ..., Y_{r-1} e Y_r. Así, el modelo para los parámetros naturales será en este caso

$$\boldsymbol{\eta}_{\mathbf{x}} = \begin{pmatrix} \eta^1 & \ldots & \eta^r & \eta^{12} & \eta^{34} & \ldots & \eta^{(r-1)r} \end{pmatrix}^T = \boldsymbol{\Gamma}\mathbf{x}, \tag{4.3}$$

donde dichos parámetros fueron definidos en la Sección 1.2.2 del Capítulo 1. El último modelo considerado asume que $\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma})$. La matriz de covarianza se fijó como en **Cook et al., 2015**. Tal como se vio en el Capítulo 1 y 3, los parámetros naturales de la distribución normal son $\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathbf{x}}, \text{vech}(\boldsymbol{\Sigma}^{-1})\right)$. Es por esto, que la muestra de $\mathbf{Y}$ se obtuvo teniendo en cuenta el siguiendo modelo para $\boldsymbol{\mu}_{\mathbf{x}}$

$$\boldsymbol{\eta}_{\mathbf{x}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathbf{x}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\mathbf{x} = \boldsymbol{\Gamma}\mathbf{x}. \tag{4.4}$$

La figuras 1, 2, 3 y 4 fueron construidas sobre el promedio de los errores de estimación $\|\boldsymbol{\Gamma} - \widehat{\boldsymbol{\Gamma}}\|_F$ para los modelos (4.1), (4.2), (4.3) y (4.4) respectivamente, versus el tamaño muestra en escala logarítmica. El tamaño muestral va desde 160 a 1000 para todos los modelos.

Observar que en todos los casos los estimadores de rango reducido presentan un error de estimación menor que el estimador de rango completo y esta diferencia está más acentuada cuando d es mucho más chico que el número de filas y columnas de $\mathbf{\Gamma}$.

Con respecto a los estimadores de rango reducido, como es de esperar, el estimador de máxima verosimilitud y de minimización cuadrática presentaron mejores resultados con respecto al estimador Sub-d. Además, para los modelos (4.1), (4.2) y (4.4) el estimador de minimización cuadrática presenta similares, y en algunos casos casi idénticos, errores de estimación que el estimador ML. En el modelo (4.3) también se observa un desempeño satifactorio del estimador de minimización cuadrática, pero los errores de estimación difieren de los obtenidos por máxima verosimilitud.

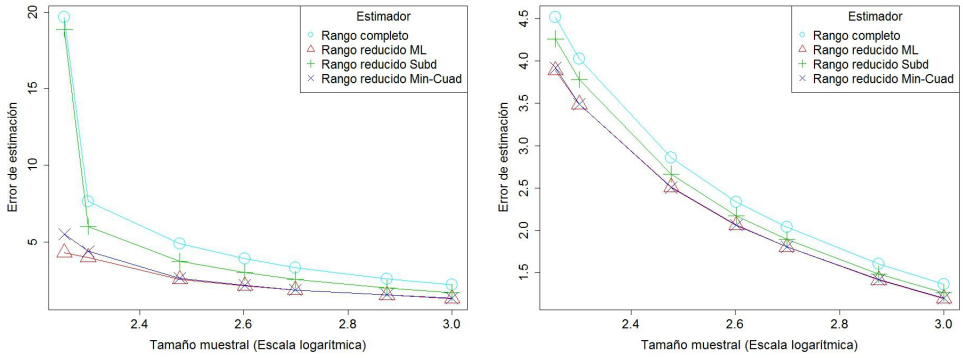


FIGURA 1. Gráficos del error de estimación para el caso multinomial. El lado derecho corresponde al caso (a) y el lado izquierdo al caso (b).

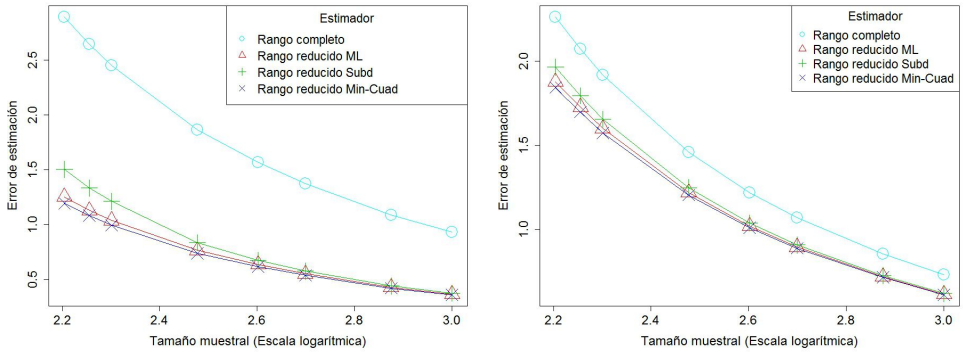


FIGURA 2. Gráficos del error de estimación para el caso Bernoulli independientes. El lado derecho corresponde al caso (a) y el lado izquierdo al caso (b).

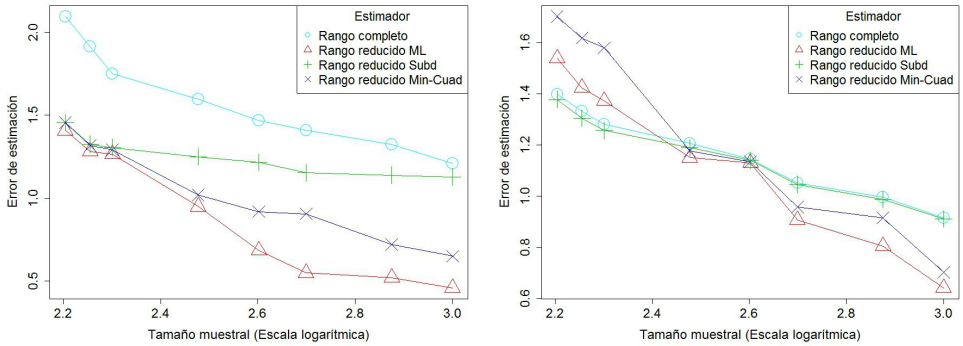


FIGURA 3. Gráficos del error de estimación para el caso Bernoulli con correlación. El lado derecho corresponde al caso (a) y el lado izquierdo al caso (b).

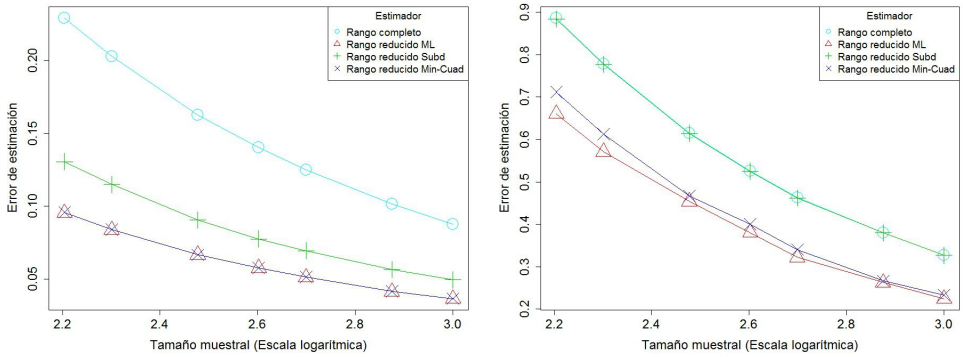


FIGURA 4. Gráficos de los errores de estimación para el caso normal. El lado derecho corresponde al caso (a) y el lado izquierdo al caso (b).

Para corroborar los resultados asintóticos obtenidos en el Teorema 3.1 y en la Proposición 3.3, los cuales presentan las distribuciones asintóticas de los estimadores de máxima verosimilitud y sub-d para GLMM de rango reducido, simulamos datos del modelo (4.2) para los casos $(c) - (d, k, p) = (1, 10, 10)$ y $(d) - (d, k, p) = (3, 6, 10)$. En estas simulaciones, consideramos que la matriz $\mathbf{A} : k \times d$ tiene la forma $\mathbf{A} = (1, \ldots, 1)/\sqrt{k}$ para el caso (c) y $\mathbf{A}^T = (\mathbf{I}_d, 0)/\sqrt{d}$ para el caso (d).

Se han realizado 1000 replicas y el tamaño de muestra n varió desde 80 hasta 3000. Para cada tamaño de muestra y cada replica obtuvimos los estimadores de rango reducido de máxima verosimilitud y Sub-d. Luego para cada n, las varianzas estimadas a partir de la muestra de estimadores de rango reducido se compararon con su respectivas varianzas asintóticas poblacionales. Esto se hizo calculando la norma de Frobenius, es decir $\|\widehat{\mathbf{W}}_{\mathbf{\Gamma}} - \mathbf{W}_{\mathbf{\Gamma}}\|_F$.

Los gráficos del tamaño de muestra versus la norma de la diferencia obtenida para cada estimador se encuentran en la Figura 5, donde se observa que a medida que n crece dichas

normas tienden al valor cero. Estos resultados experimentales, ratifican los resultados teóricos que se obtuvieron en el Capítulo 3 para estos modelos.

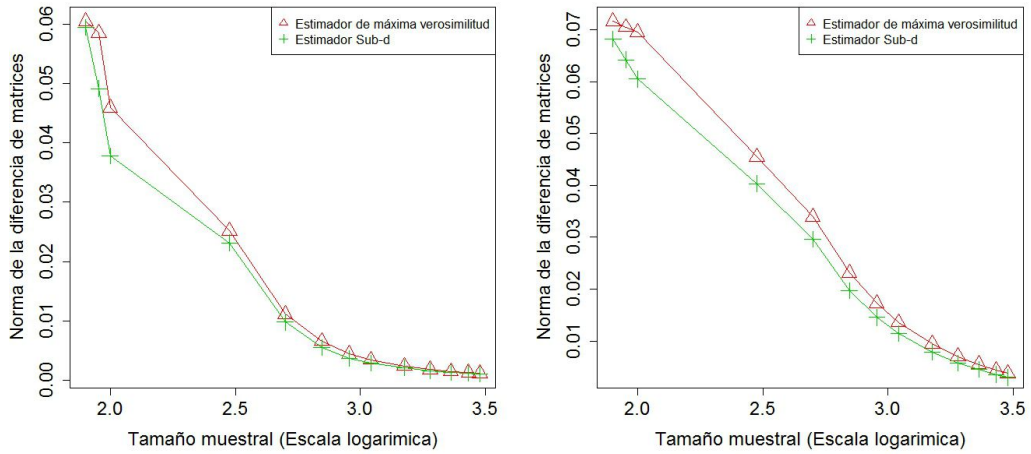


FIGURA 5. La primera figura corresponde a los resultados para el caso (c) y la segunda figura corresponde al caso (d).

4.2. Comparación con la metodología propuesta en [Yee and Hastie, 2003]

En [Yee and Hastie, 2003] no cuentan con resultados asintóticos para los estimadores de máxima verosimilitud de rango reducido propuestos para el modelo (3.1). Es por esto que proponen remplazar esta varianza asintótica por una matriz particionada que se construye a partir de resultados asintóticos conocidos para modelos más simples que se obtienen si se suponen fijas y conocidas la matriz de coeficientes $\mathbf{A}$ por un lado, o la matriz de coeficientes $\mathbf{B}$ por otro lado. Es decir, si la matriz $\mathbf{B}$ es conocida, el modelo (3.1) se reduce en estimar la matriz de rango completo $(\mathbf{C}\ \mathbf{A})$ con predictores $(\mathbf{X}_1\ ,\ \mathbf{B}\mathbf{X}_2)$. Y si la matriz $\mathbf{A}$ está fija, el modelo (3.1) consiste en estimar $\mathbf{C}$ y $\mathbf{B}$ donde los predictores $\mathbf{X}_2$ están sujetos a restriciones lineales conocidas, determinadas por la matriz $\mathbf{A}$ (ver Sección 2.1 de [Yee and Hastie, 2003]).

Sea $\boldsymbol{\theta} = (\theta_1^T, \theta_2^T, \theta_3^T)^T$ donde $\theta_1 = \text{vec}(\mathbf{A})$, $\theta_2 = \text{vec}(\mathbf{C})$ y $\theta_3 = \text{vec}(\mathbf{B})$, y la partición

$$-\frac{\partial^2 l}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} = -\ddot{l} = -\begin{pmatrix} \ddot{l}_{11} & \ddot{l}_{12} & \ddot{l}_{13} \\ \ddot{l}_{21} & \ddot{l}_{22} & \ddot{l}_{23} \\ \ddot{l}_{31} & \ddot{l}_{32} & \ddot{l}_{33} \end{pmatrix}$$

donde $\ddot{l}_{jk} = \partial^2 l/(\partial\theta_j\partial\theta_k^T)$ y l es la función loglikelihood que corresponde a cada familia exponencial. Luego, en [Yee and Hastie, 2003] proponen obtener los errores estándar de $\widehat{\boldsymbol{\theta}}$ a

través de calculo de la matriz completa $-\partial^2 l/(\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T)$, evaluada en $\widehat{\boldsymbol{\theta}}$ e invirtiendo esta. Las matrices bloques

$$\begin{pmatrix} \ddot{l}_{11} & \ddot{l}_{12} \\ \ddot{l}_{21} & \ddot{l}_{22} \end{pmatrix} \quad \text{y} \quad \begin{pmatrix} \ddot{l}_{22} & \ddot{l}_{23} \\ \ddot{l}_{32} & \ddot{l}_{33} \end{pmatrix}$$

pueden ser obtenidas fijando θ_3 y θ_1 respectivamente. Y para el bloque $-\ddot{l}_{13}$ proponen usar la igualdad $-\partial^2 l/(\partial\theta_1\partial\theta_3^T) = -\dfrac{\partial\theta_3^T(\theta_1)}{\partial\theta_1}(\partial^2 l/(\partial\theta_3\partial\theta_3^T))$.

4.3. Datos: Estado civil

Para aplicar los resultados presentados a lo largo del Capítulo 3, vamos a analizar el conjunto de datos *Estado civil de los hombres de Nueva Zelanda* que fue estudiado en detalle en el trabajo [**Yee and Hastie, 2003**]. Los datos se obtuvieron por medio de un cuestionario realizado a una gran número de empleados de la población de Nueva Zelanda. Para lograr homogeneidad, el análisis se restringió a un subconjunto de $n = 4105$ datos correspondientes a los resultados de sexo masculino que no presentaran datos perdidos en las variables utilizadas. Para más detalles acerca de estos datos ver [**Yee and Wild, 1996**]. El objetivo del análisis es explorar si ciertos estilos de vida y variables psicológicas están asociadas con el estado civil, especialmente separado o divorciado. La variable respuesta Y es estado civil y consiste en 4 posibles opciones identificadas como 1 = soltero, 2 = separado o divorciado, 3 = viudo y 4 = casado o en viviendo en pareja. Las variables predictoras están dadas en el Cuadro 1 de [**Yee and Hastie, 2003**]. Hay 12 predictores que son binarios (1/0 para presencia/ausencia, respectivamente) y que han sido dispuestos, según los autores, de forma tal que la presencia de alguno de ellos es indicador de un mal estilo de vida o características psicológicas negativas. El objetivo del estudio es investigar si estas 12 variables indicadoras, más 2 variables continuas que son la edad y los años de estudio, están relacionadas con Y y de que forma.

Si identificamos la respuesta Y con el vector $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$ que tiene en su j-ésima posición un 1 si $Y = j$, $j = 1, 2, 3, 4$ y cero en las restantes, es claro que podemos atribuir a $\mathbf{Y}|\mathbf{X}$ una distribución multinomial siendo $\mathbf{X}$ los 14 predictores mencionados anteriormente. Esta distribución fue presentada en la Sección 1.2.3 como familia exponencial. Para este conjunto de datos $r = 4$, $m = 1$ y p_1, p_2, p_3 y p_4 indican las probabilidades de que Y tome los valores 1, 2, 3 y 4 respectivamente. El vector de parámetros naturales es $\boldsymbol{\eta}_{\mathbf{x}} = (\eta_1, \eta_2, \eta_3) = (\log(p_1/p_4), \log(p_2/p_4), \log(p_3/p_4))$ y el estadístico suficiente es

$\mathbf{T}(\mathbf{Y}) = (Y_1, Y_2, Y_3)$. El modelo de regresión de rango completo que asumimos para los parámetros naturales es

$$\boldsymbol{\eta}_{\mathbf{x}} = \begin{pmatrix} \log(p_1/p_4) \\ \log(p_2/p_4) \\ \log(p_3/p_4) \end{pmatrix} = \overline{\boldsymbol{\eta}} + \mathbf{D}\mathbf{x} \tag{4.5}$$

donde $\overline{\boldsymbol{\eta}} \in \mathbb{R}^3$ es el intercepto y $\mathbf{D} \in \mathbb{R}^{3\times 14}$ es la matriz de coeficientes. Los estimadores de máxima verosimilitud para este modelo pueden obtenerse a través del algoritmo IRLS estandar presentado en la Sección 1.3.2. Tener en cuenta que en este caso la función

$$\psi(\boldsymbol{\eta}_{\mathbf{x}}) = \log\left(1 + \sum_{j=1}^{3} e^{\eta_j}\right)$$

por lo que en dicho algoritmo IRLS de estimación tendremos que

$$\mathbf{d}_i = (y_{1i}, y_{2i}, y_{3i}) - (p_{1i}, p_{2i}, p_{3i})$$

$$\mathbf{W}_i = \begin{pmatrix} p_{1i}(1-p_{1i}) & -p_{1i}p_{2i} & -p_{1i}p_{3i} \\ -p_{1i}p_2 & p_{2i}(1-p_{2i}) & -p_{2i}p_{3i} \\ -p_{1i}p_3 & -p_{2i}p_{3i} & p_{3i}(1-p_{3i}) \end{pmatrix}.$$

Además las matrices $\mathbf{W}_i$, $i = 1, \ldots, n$ serán necesarias para el cálculo computacional de las varianzas asintóticas de los estimadores de máxima verosimilitud de $\mathbf{D}$. No reproducimos aquí los estimadores obtenidos para $\widehat{\mathbf{D}}$ y $\widehat{\overline{\boldsymbol{\eta}}}$, ya que estos pueden verse en **[Yee and Hastie, 2003]**. Sin embargo, resaltamos que solamente el intercepto y 5 coeficientes de la matriz de coeficientes 3×14 resultaron significativos.

Notar que entre ciertas variables binarias es esperable que exista un alto grado de correlación, por ejemplo entre *nerves* y *nervous* ó *worry* y *worried* y contrariamente esta presencia de alta correlación no es tan esperable entre las variables continuas. De acuerdo a esto, proponemos un modelo de regresión de rango reducido parcial para este conjunto de datos. Por un lado, $\mathbf{x}_1$ representan las variables continuas (*age30 y logedu1*), las cuales no son sujetas a restricción de rango, y $\mathbf{x}_2$ representan las 12 variables binarias:

$$\boldsymbol{\eta}_{\mathbf{x}} = \begin{pmatrix} \log(p_1/p_4) \\ \log(p_2/p_4) \\ \log(p_3/p_4) \end{pmatrix} = \overline{\boldsymbol{\eta}} + \mathbf{D}\mathbf{x} = \overline{\boldsymbol{\eta}} + \mathbf{C}\mathbf{x}_1 + \mathbf{A}\mathbf{B}\mathbf{x}_2, \tag{4.6}$$

donde $\mathbf{C} : 3 \times 2$, $\mathbf{A} : 3 \times d$ y $\mathbf{B} : d \times 12$. El valor d fue estimado por medio de los test de dimensión presentados en la Sección 3.5, para los cuales fue necesario el cálculo de la varianza

asintótica del estimador de máxima verosimilitud de rango completo $\widehat{\mathbf{D}}$. Los p-valores fueron los siguientes:

	$H_0 : d = 0$ vs $H_1 : d > 0$	$H_0 : d = 1$ vs $H_1 : d > 1$
Test 1: Chi-cuadrado ponderado asintótico	0.079	0.458
Test 2: Chi-cuadrado asintótico de Wald	$2.22 * 10^{-8}$	0.071

CUADRO 1. p-valores obtenidos en los test de dimensión.

Teniendo en cuenta estos resultados y que el primer test es muy conservador, tenemos evidencia suficiente para suponer que d es 1. En **[Yee and Hastie, 2003]**, basandose en lo sugerido en **[Reinsel and Velu, 1998]**, proponen aplicar el criterio AIC para determinar el valor d. Los resultados que obtuvieron concuerdan con lo obtenido en los test del Cuadro 1, pues el valor de d que minimiza el criterio AIC se da en $d = 1$.

Los estimadores de máxima verosimilitud para el modelo (4.6) se obtuvieron implementando la función rrvglm del paquete VGAM del software [R], utilizando la restricción *corner* en el grupo de separados/divorciados. Los Cuadros 2 y 3 muestran estos resultados con sus respectivos errores estandar entre paréntesis.

Variable	$\log(p_1/p_4)$	$\log(p_2/p_4)$	$\log(p_3/p_4)$
Intercept	-1.762 *	-2.699 *	-6.711 *
	(0.377)	(0.383)	(1.047)
age30	-0.191 *	0.012	0.086 *
	(0.008)	(0.008)	(0.023)
logedu1	0.338	-0.365	-0.089
	(0.225)	(0.213)	(0.559)

CUADRO 2. Estimadores de $\overline{\boldsymbol{\eta}}$ y $\mathbf{C}$ del modelo (4.6).

Los errores estándar para estos estimadores fueron obtenidos de la versión muestral del resultado del Teorema 3.1. Es decir,

$$\widehat{\text{avar}}\begin{pmatrix} \widehat{\overline{\boldsymbol{\eta}}} \\ \text{vec}(\widehat{\mathbf{C}}) \\ \text{vec}(\widehat{\mathbf{A}}\widehat{\mathbf{B}}) \end{pmatrix} = \widehat{\boldsymbol{\Delta}}(\widehat{\boldsymbol{\Delta}}^T\widehat{\mathbf{V}}^{-1}\widehat{\boldsymbol{\Delta}})^{\ddagger}\widehat{\boldsymbol{\Delta}}^T \tag{4.7}$$

Variable	$\log(p_1/p_4)$	$\log(p_2/p_4)$	$\log(p_3/p_4)$
binge	0.569 *	0.786 *	1.114 *
	(0.125)	(0.196)	(0.409)
smokenow	0.222 *	0.306 *	0.434 *
	(0.088)	(0.119)	(0.208)
sun	0.011	0.015	0.021
	(0.084)	(0.116)	(0.164)
nerves	-0.054	-0.074	-0.105
	(0.101)	(0.139)	(0.197)
nervous	0.312 *	0.430 *	0.609 *
	(0.124)	(0.168)	(0.291)
hurt	0.180	0.248	0.352 *
	(0.089)	(0.122)	(0.199)
tense	0.302 *	0.416 *	0.590 *
	(0.122)	(0.163)	(0.284)
miserable	0.019	0.0268	0.038
	(0.094)	(0.129)	(0.185)
fedup	0.117	0.161	0.229
	(0.094)	(0.129)	(0.185)
worry	0.003	0.004	0.005
	(0.106)	(0.146)	(0.207)
worrier	-0.116	-0.160	-0.227
	(0.092)	(0.128)	(0.193)
mood	-0.037	-0.052	-0.073
	(0.087)	(0.120)	(0.172)

CUADRO 3. Estimadores de **AB** del modelo (4.6).

donde en esta caso

$$\widehat{\mathbf{\Delta}} = \begin{pmatrix} \mathbf{I}_9 & 0 & 0 \\ 0 & \widehat{\mathbf{B}}^T \otimes \mathbf{I}_3 & \mathbf{I}_{12} \otimes \widehat{\mathbf{A}} \end{pmatrix} \quad y \quad \widehat{\mathbf{V}}^{-1} = \frac{1}{n} \sum_{i=1}^{n} \left((1 \ \mathbf{x})^T (1 \ \mathbf{x}) \otimes \widehat{\mathbf{W}}_i \right)$$

es la inversa de la varianza asintótica muestral de los estimadores de máxima verosimilitud del modelo de rango completo (4.5) con $\widehat{\mathbf{W}}_i$ dado por

$$\widehat{\mathbf{W}}_i = \begin{pmatrix} \widehat{p}_{1i}(1-\widehat{p}_{1i}) & -\widehat{p}_{1i}\widehat{p}_{2i} & -\widehat{p}_{1i}\widehat{p}_{3i} \\ -\widehat{p}_{1i}\widehat{p}_{2} & \widehat{p}_{2i}(1-\widehat{p}_{2i}) & -\widehat{p}_{2i}\widehat{p}_{3i} \\ -\widehat{p}_{1i}\widehat{p}_{3} & -\widehat{p}_{2i}\widehat{p}_{3i} & \widehat{p}_{3i}(1-\widehat{p}_{3i}) \end{pmatrix}.$$

A partir de estos errores estándar se calcularon intervalos de confianza e identificamos con asterisco (*) aquellos coeficientes que resultaron significativos.

Yee and Hastie, 2003 calculan los errores estándar para $\mathbf{A}$ y $\mathbf{B}$ como se presentó en la Sección 4.2. Por esta razón, concluyen acerca de la significancia de las componentes de la matriz $\mathbf{B}$ y de las componentes de la matriz $\mathbf{A}$ en forma separada y lo mismo sucede con la interpretación de los resultados. Es decir, primero analizan $\boldsymbol{\nu} = \mathbf{B}\mathbf{x}_1$ y luego $\mathbf{A}\boldsymbol{\nu}$, concluyendo primero sobre cuales son los predictores significativos en el problema y luego de qué forma estas combinaciones $\boldsymbol{\nu} = \mathbf{B}\mathbf{x}_1$ influyen en cada variable respuesta. Sin embargo, no es sencillo analizar el producto de los errores estándar y concluir acerca de la significancia del producto de las componentes de la matriz de coeficientes $\mathbf{AB}$.

A partir de los resultados que se obtuvieron en el Teorema 3.1, es posible obtener de forma sencilla los errores estándar de cada componente de la matriz de coeficientes $\mathbf{AB}$ y por lo tanto es posible analizar la significancia de cada variable predictora sobre cada variable respuesta. Por ejemplo, de acuerdo a los resultados en **Yee and Hastie, 2003** para este conjunto de datos, la variable predictora *hurt* es considerada significativa para los tres grupos (solteros, separados, viudos) pero no es así según los resultados presentados en el Cuadro 3, donde el predictor *hurt* solo es importante en el grupo de viudos y no en los grupos de casados y solteros.

En conclusión, a partir de los resultados obtenidos, podemos interpretar la matriz completa $\mathbf{AB}$ del Cuadro 3 de la siguiente forma:

- Notar que todos los coeficientes significativos son positivos. Estos corresponden a las variables: *binge, smokenow, nervous, tense* y *hurt* solo para viudos. Esto sugiere que si las variables binarias indican un mal estilo de vida y características psicológicas negativas de la persona, según los autores, las chances de ser soltero, divociado o viudo aumentarán con respecto a la de ser casado, manteniendo la variable edad fija.
- Notar que los coeficientes que corresponden a la respuesta $\log(p_3/p_4)$ son dos veces más grandes que los de $\log(p_1/p_4)$ sugiriendo la importancia relativa de los predictores en cada grupo.

CAPÍTULO 5

REDUCCIÓN SUFICIENTE DE DIMENSIONES

El objetivo general del análisis de regresión, del cual se ha estudiado diferentes modelos en los capítulos anteriores, es estudiar la distribución de una variable respuesta $\mathbf{Y}$ en $\mathbb{R}^r$ dado un vector de predictores $\mathbf{X}$ en $\mathbb{R}^p$. En cualquier análisis estadístico, el primer paso consiste en elaborar gráficos con el conjunto de datos disponibles. Especialmente en los casos en que p es chico, estos gráficos permiten visualizar un posible modelo que se ajuste a la muestra dada. Sin embargo, cuando el número de predictores es grande, elaborar un modelo paramétrico adecuado comienza a ser complejo y las herramientas visuales ya no son tan informativas. Además, aunque intuitivamente parezca favorable contar con una gran número predictores que puedan explicar la variable respuesta, no siempre ello implica poder obtener más información. Para soslayar este tipo de dificultades, emergen las ideas de reducción suficente de dimensiones en regresión (SDR), donde el objetivo principal es reducir la dimensión de $\mathbf{X}$ sin perder información acerca de la respuesta $\mathbf{Y}$ y sin requerir un modelo previo para $\mathbf{Y}|\mathbf{X}$. Esto permite la elaboración de gráficos que contienen toda la información que $\mathbf{X}$ tiene sobre $\mathbf{Y}$, en una dimensión menor a la dimensión donde vive $\mathbf{X}$. En este capítulo se repasan conceptos básicos y se presenta una breve introducción de los principales métodos propuestos en SDR que servirán de base para presentar los resultados nuevos en el próximo capítulo.

5.1. Definiciones e ideas básicas

Para una variable respuesta $\mathbf{Y} \in \mathbb{R}^r$ y un conjunto de predictores $\mathbf{X} \in \mathbb{R}^p$, estamos interesados en hallar una función $R(\mathbf{X})$ que retenga toda la información que $\mathbf{X}$ tiene sobre $\mathbf{Y}$. La siguiente definición formaliza la noción de reducción suficiente de dimensiones en regresión [**Cook, 2007**].

Definición 5.1. Una reducción $R : \mathbb{R}^p \to \mathbb{R}^d$, con $d \leq p$ es suficiente para la regresión de $\mathbf{Y}|\mathbf{X}$ si satisface la siguiente condición

$$\mathbf{Y}|\mathbf{X} \sim \mathbf{Y}|R(\mathbf{X}) \tag{5.1}$$

El siguiente lema establece las condiciones que son equivalentes a la condición (5.1) de la Definición 5.1. La demostración se encuentra al final del capítulo.

Lema 5.2. *Si* $\mathbf{X}$ *e* $\mathbf{Y}$ *son vectores aleatorios y* $R(\mathbf{X})$ *es una función medible en el vector de predictores, entonces los siguientes puntos son equivalentes:*

(i) $\mathbf{Y}|\mathbf{X} \sim \mathbf{Y}|R(\mathbf{X})$

(ii) $\mathbf{X}|(\mathbf{Y}, R(\mathbf{X})) \sim \mathbf{X}|R(\mathbf{X})$

(iii) $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|R(\mathbf{X})$

Debido a este resultado, aún cuando $\mathbf{X}$ no sea aleatorio, diremos que una reducción es suficiente para estudiar $\mathbf{Y}$ en función de $\mathbf{X}$, si cumple con alguno de los estamentos establecidos por el Lema 5.2.

La mayoría de las metodologías en SDR están basadas en la regresión inversa de $\mathbf{X}$ en $\mathbf{Y}$. La justificación de este enfoque deriva en la equivalencia **(ii)** con **(i)**, que muestra que una reducción suficiente para $\mathbf{Y}$ puede ser determinada a partir de la regresión $\mathbf{X}|\mathbf{Y}$ y luego pasar a la regresión directa $\mathbf{Y}|\mathbf{X}$ sin especificar la distribución marginal de $\mathbf{Y}$ o la distribución condicional de $\mathbf{Y}|\mathbf{X}$. El atractivo de regresión inversa es que en la mayoría de los problemas de regresión la respuesta es de una dimensión, mientras que el número de predictores puede ser grande haciendo que la regresión directa de $\mathbf{Y}$ sobre $\mathbf{X}$ sea muy difícil de modelar. En contraste, la regresión inversa de $\mathbf{X}$ en $\mathbf{Y}$ consta de p regresiones unidimensionales, que son mucho más simples de visualizar y permite deducir un posible modelo. Esta misma ventaja es la que se tiene después de obtener una reducción $R(\mathbf{X})$, poder graficar $\mathbf{Y}$ versus $R(\mathbf{X})$ sabiendo que no se ha perdido información y pensar a partir de este gráfico un posible modelo.

Una de las características que definen las mayoría de las reducciones suficientes estudiadas en la literatura estadística es que ellas son lineales, es decir que la función R tiene la forma $R(\mathbf{X}) = \boldsymbol{\alpha}^T\mathbf{X}$ con $\boldsymbol{\alpha} \in \mathbb{R}^{p\times d}$. Observar que en este contexto, si $\boldsymbol{\alpha}^T\mathbf{X}$ es una reducción suficiente y $\boldsymbol{\rho}$ es una matriz $d \times d$ no singular, luego $\boldsymbol{\rho}\boldsymbol{\alpha}^T\mathbf{X}$ es también una reducción suficiente. Así, $\boldsymbol{\alpha}$ no es única y lo que realmente interesa identificar es el subespacio generado por las columnas de $\boldsymbol{\alpha}$, es decir, nuestro parámetro de interés no es $\boldsymbol{\alpha}$ por sí mismo, sino span($\boldsymbol{\alpha}$) o $\mathbf{P}_{\boldsymbol{\alpha}}$ con $\mathbf{P}_{\boldsymbol{\alpha}}$ la matriz de proyección sobre el espacio generado por las columnas de $\boldsymbol{\alpha}$. Este subespacio lo denotamos como $\mathcal{S}_\alpha \equiv \text{span}(\boldsymbol{\alpha})$ y lo llamamos subespacio de reducción suficiente (DRS).

Además, en el sentido estricto de la definición, $R(\mathbf{X}) = \mathbf{X}$ es una reducción suficiente y por lo tanto $\mathcal{S}_{\mathbf{I}_p}$ es un subespacio de redución suficiente, por lo que la tarea escencial de SDR es caracterizar y estimar la reducción suficiente más pequeña. Específicamente, un subespacio $\mathcal{S}$ se dice que es un subespacio de reducción suficiente minimal, si $\mathcal{S}$ es un DSR y $\dim(\mathcal{S}) \leq \dim(\mathcal{S}_{drs})$ para todos los DRS $\mathcal{S}_{drs}$. Otra especificación de subespacio más pequeño, siempre en el contexto de reducciones lineales, es aquel DSR $\mathcal{S} \subset \mathcal{S}_{drs}$ para todo los DRS $\mathcal{S}_{drs}$. Este DRS, cuando existe, recibe el nombre de *subespacio central*, se denota $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ y es el único DRS minimal. Este subespacio goza de varias propiedades que son útiles en la práctica. Entre ellas, si $\mathbf{A}$ es una matriz $p \times p$ de rango completo y $\mathbf{b}$ un vector en $\mathbb{R}^p$, luego $\mathcal{S}_{\mathbf{Y}|\mathbf{A}^T\mathbf{X}+\mathbf{b}} = \mathbf{A}^{-1}\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. Es decir, que no hay pérdida de información bajo transformaciones afines de rango completo de $\mathbf{X}$. Un estudio completo y detallado sobre $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ puede encontrarse en el Capítulo 6 de [**Cook, 1998**].

Cuando nos restringimos a las reducciones lineales, la noción de minimalidad se ha establecido a través de subespacios. Sin embargo, en un contexto más amplio donde la reducción puede ser lineal o no, una definición más apropiada de reducción suficiente minimal es la siguiente.

Definición 5.3. Una reducción suficiente $R(\mathbf{X})$, es minimal para la regresión $\mathbf{Y}|\mathbf{X}$ si para cualquier otra reducción suficiente $S(\mathbf{X})$, $R(\mathbf{X})$ es función de $S(\mathbf{X})$.

Observemos que esta definición está conectada a la noción de estadístico suficiente minimal, si vemos a la variable observable $\mathbf{Y}$ como un parámetro.

5.2. Reducción suficiente basada en momentos

La estimación de reducciones suficientes lineales fue originariamente basado en momentos o funciones de los momentos de la distribución condicional $\mathbf{X}|Y$ (Sliced Inverse Regression [**Li, 1991**], Slice Average Variance Estimation [**Cook and Weisberg, 1991**], principal Hessian direction [**Li, 1992**], Parametric Inverse Regression [**Bura and Cook, 2001**], Minimum Average Variance Estimation [**Xia et al., 2002**], [**Li et al., 2005**], [**Cook and Ni, 2005**], [**Zhu and Zeng, 2006**], [**Cook and Li, 2002**], Directional Regression [**Li and Wang, 2007**]). Estos métodos requieren que los predictores sean de tipo continuo, suponen condiciones sobre estos momentos, y en general logran capturar sólo una parte de la reducción.

Indicando con $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ la media y la covarianza de $\mathbf{X}$ y suponiendo que $\boldsymbol{\Sigma}$ es definida positiva, usualmente es común en estos métodos trabajar con la estandarización de los predictores $\mathbf{Z} =$

$\boldsymbol{\Sigma}^{-1/2}(\mathbf{X}-\boldsymbol{\mu})$ y el correspondiente subespacio $\mathcal{S}_{Y|\mathbf{Z}}$ para la respuesta $Y \in \mathbb{R}$. Luego, el subespacio central en la escala original de los predictores está dado por $\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{Y|\mathbf{Z}}$.

El método *Sliced Inverse Regression* (SIR, de ahora en adelante) está basado en el siguiente resultado fundamental de [**Li, 1991**].

Proposición 5.4. *Sea* $\boldsymbol{\alpha} \in \mathbb{R}^{p\times d}$ *una base para el subespacio central* $\mathcal{S}_{Y|\mathbf{Z}}$ *y* $\mathbf{P}_{\boldsymbol{\alpha}}$ *la matriz de proyección sobre el* $\text{span}(\boldsymbol{\alpha})$. *Si la esperanza condicional* $\mathbf{E}(\mathbf{Z}|\boldsymbol{\alpha}^T\mathbf{Z})$ *es lineal en* $\boldsymbol{\alpha}^T\mathbf{Z}$*; esto es*

$$\mathbf{E}(\mathbf{Z}|\boldsymbol{\alpha}^T\mathbf{Z}) = \mathbf{P}_{\boldsymbol{\alpha}}\mathbf{Z} \tag{5.2}$$

entonces $\mathbf{E}(\mathbf{Z}|Y) \in \mathcal{S}_{Y|\mathbf{Z}}$.

Esta proposición establece que es posible identificar una parte del subespacio $\mathcal{S}_{Y|\mathbf{Z}}$ a través de la esperanza condicional $\mathbf{E}(\mathbf{Z}|Y)$. Para obtener una forma relativamente sencilla de estimar el vector $\mathbf{E}(\mathbf{Z}|Y)$, la respuesta es remplazada con una versión discreta $\widetilde{Y}$, construida a través de una partición del rango de Y en H intervalos. Bajo la condición (5.2), conocida como condición de linealidad, se tiene que

$$\mathbf{E}(\mathbf{Z}|\widetilde{Y}) \in \mathcal{S}_{\widetilde{Y}|\mathbf{Z}} \subseteq \mathcal{S}_{Y|\mathbf{Z}}$$

de modo que el espacio generado por las columnas de

$$\boldsymbol{\Theta} = \text{var}(\mathbf{E}(\mathbf{Z}|\widetilde{Y})) = \mathbf{E}(\mathbf{E}(\mathbf{Z}|\widetilde{Y})\mathbf{E}(\mathbf{Z}|\widetilde{Y})^T)$$

está contenido en $\mathcal{S}_{Y|\mathbf{Z}}$, es decir $\text{span}(\boldsymbol{\Theta}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. Por lo tanto, una parte del subespacio central puede ser construida como

$$\text{span}\{\mathbf{t}_1, \ldots, \mathbf{t}_d\} = \text{span}(\boldsymbol{\Theta}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$$

donde $d = \text{rank}(\boldsymbol{\Theta})$ y $\mathbf{t}_j$ con $j = 1, \ldots, d$ son los autovectores correspondientes a los autovalores no nulos $\lambda_1 > \ldots > \lambda_d$ en la descomposición espectral $\boldsymbol{\Theta} = \sum_{j=1}^{d} \lambda_j \mathbf{t}_j \mathbf{t}_j^T$. En la práctica, $\mathcal{S}_{Y|\mathbf{Z}}$ es estimado como el espacio generado por los autovectores de $\widehat{\boldsymbol{\Theta}}$, la versión muestral de $\boldsymbol{\Theta}$.

Siguiendo la misma linea, el método *Slice Average Variance Estimation* (SAVE, de ahora en adelante) considera una segunda condición, además de la condición de linealidad, en los momentos de $\mathbf{X}$. Resumimos estas ideas en la siguiente proposición [**Cook and Weisberg, 1991**]:

Proposición 5.5. *Sea* $\boldsymbol{\alpha} \in \mathbb{R}^{p\times d}$ *una base para el subespacio central* $\mathcal{S}_{Y|\mathbf{Z}}$. *Bajo las condiciones*

- $\mathbf{E}(\mathbf{Z}|\boldsymbol{\alpha}^T\mathbf{Z}) = \mathbf{P}_{\boldsymbol{\alpha}}\mathbf{Z}$
- $\text{var}(\mathbf{Z}|\boldsymbol{\alpha}^T\mathbf{Z}) = \mathbf{I}_p - \mathbf{P}_{\boldsymbol{\alpha}}$

se tiene que $\text{span}(\mathbf{E}(\mathbf{I}_p - \text{var}(\mathbf{Z}|Y))^2) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$.

La segunda condición de este resultado es conocida como condición de varianza constante. La Proposición 5.5 establece la base del método SAVE. Sea λ_j y $\mathbf{v}_j$ los autovalores y autovectores de $\mathbf{E}(\mathbf{I}_p - \text{var}(\mathbf{Z}|Y))^2$, $j = 1, \ldots, p$ respectivamente y sea $d \leq p$ el número de autovalores no nulos. De acuerdo a la Proposición 5.5, estos autovectores correspondientes a los d autovalores no nulos están en $\mathcal{S}_{Y|\mathbf{Z}}$ y así, permiten construir una parte del subespacio central. Los predictores poblacionales SAVE son definidos como $\mathbf{v}_j^T\mathbf{Z}$, $j = 1, \ldots, d$. Para su uso en aplicaciones, nuevamente consideramos la versión discreta $\widetilde{Y}$ como en SIR. Estimamos la varianza $\text{var}(\mathbf{Z}|\widetilde{Y})$ en cada intervalo y la denotamos con $\mathbf{V}_s$, $s = 1, \ldots, H$. Luego, sea $\mathbf{M} = \sum_{s=1}^{H} f_s(\mathbf{I} - \mathbf{V}_s)^2$ donde f_s denota la fracción de observaciones en el intervalo s. El j-ésimo predictor muestral SAVE es $\widehat{\mathbf{b}}_j^T\mathbf{Z}$ $j = 1, \ldots, d$ donde $\widehat{\mathbf{b}}_j$ es el j-ésimo autovector de $\mathbf{M}$ correspondiente a los d autovalores no nulos ordenados $\widehat{\lambda}_1 > \ldots > \widehat{\lambda}_d$.

El primer paso para extender las ideas de reducción suficiente de dimensiones a regresiones con predictores continuos y categóricos fue propuesto en el trabajo [**Chiaromonte et al., 2002**], donde se introduce *Sir Parcial*. De acuerdo a este enfoque, además del vector de predictores continuos $\mathbf{X} \in \mathbb{R}^p$, se cuenta con una variable explicativa cualitativa W que representa una o más categorías con valores $w = 1, \ldots, C$ que se indentifican con C subpoblaciones. El objetivo es hallar y estimar reducciones suficientes lineales de la forma $(\widetilde{\boldsymbol{\zeta}}^T\mathbf{X}, W)$ para la regresión de Y en $(\mathbf{X}, W)$, es decir $\widetilde{\boldsymbol{\zeta}}$ tal que

$$Y \perp\!\!\!\perp \mathbf{X}|(\widetilde{\boldsymbol{\zeta}}^T\mathbf{X}, W).$$

La siguiente proposición conecta el $\text{span}(\widetilde{\boldsymbol{\zeta}})$ con los subespacios $\text{span}(\widetilde{\boldsymbol{\zeta}}_w)$ $w = 1, \ldots, C$ donde $\widetilde{\boldsymbol{\zeta}}_w$ indica una base para el subespacio de reducción suficiente para la regresión $Y_w|\mathbf{X}_w$ con Y_w y $\mathbf{X}_w$ la restricción de las variables en cada subpoblación $w = 1, \ldots, C$; $\bigoplus$ indica la suma directa entre dos subespacios $(V_1 \oplus V_2) = \{v_1 + v_2; v_1 \in V_1, v_2 \in V_2\}$.

Proposición 5.6. *Sea* $\mathbf{X}_w$ *la restricción de las variables continuas en cada subpoblación* $w = 1, \ldots, C$, *y sea* $\widetilde{\boldsymbol{\zeta}}_w$ *una base para el subespacio de reducción suficiente para la regresión* $Y_w|\mathbf{X}_w$. *Luego,*

$$\text{span}(\widetilde{\boldsymbol{\zeta}}) = \text{span}\left(\bigoplus_{w=1}^{C} \widetilde{\boldsymbol{\zeta}}_w\right).$$

Basándose en esta proposición, la idea de esta metodología es que $\widetilde{\boldsymbol{\zeta}}$ puede ser estimado combinando las estimaciones de las reducciones suficientes dentro de cada subpoblación y para

ello adaptan el método SIR para la estimación de $\widetilde{\boldsymbol{\zeta}}$. Denominan a este metodo Sir Parcial. El algoritmo Sir Parcial es una forma efectiva de considerar ambos predictores, categóricos y continuos, utilizando metodología SDR conocida, que sólo pueden aplicarse a predictores continuos. Con este enfoque, se identifica una reducción suficiente para la regresión de Y en $(\mathbf{X}, W)$. Las deventajas que presenta este enfoque es que no permite mezclar las variables continuas y categórica en la reducción, ya que cada subpoblación se considera por separado. Además si C es grande, el tamaño de la muestra en cada $Y_w|\ X_w$ puede ser muy pequeña, lo que lleva a una mayor imprecisión en la estimación.

5.3. Reducción suficiente basada en modelos

Los métodos SIR y SAVE, proporcionan estimadores $\sqrt{n}$ consistentes de una parte del subespacio central $\mathcal{S}_{Y|\mathbf{X}}$ bajo ciertas condiciones de regularidad y han sido ampliamente utilizados en aplicaciones, no obstante presentan conocidas limitaciones y desventajas en la práctica. En particular, el subespacio de reducción suficiente estimada por SIR es un subconjunto propio del subespacio central por lo que no proporciona una exhautiva estimación (esta desventaja también es heredada por Sir Parcial). Además, posee dificultades para detectar direcciones que están asociadas a direcciones no lineales en la función media $\mathbf{E}(Y|\mathbf{X})$. El método SAVE, a pesar que se propuso en respuesta a estas limitaciones y que captura, en general, un parte mayor del subespacio central, es poco eficiente para encontrar direcciones lineales en la función media $\mathbf{E}(Y|\mathbf{X})$.

Con el fin de evitar estas limitaciones, las reducciones suficientes basadas en modelos fueron introducidas en [**Cook, 2007**]. Las principales ventajas que presenta este enfoque es la identificación de reducciones suficientes de los predictores en forma exhautivas para la regresión de Y en $\mathbf{X}$; esto es, que contienen toda la información relevante para Y que hay en $\mathbf{X}$. Además ya que se cuenta con un modelo, es posible obtener estimadores de máxima verosimilitud de las reducciones suficientes lo cuales son óptimos en términos de eficiencia cuando el modelo es cierto.

En los trabajos [**Cook, 2007**] y [**Cook and Forzani, 2008**] se considera el caso $\mathbf{X}|Y$ con distribución normal y desarrollan la metodología *Principal Fitted Component* (PFC). Bajo este enfoque, se asume que $\mathbf{X}_y = \mathbf{X}|(Y = y)$ es normalmente distribuido con media $\boldsymbol{\mu}_y$ y matriz de covarianza constante $\boldsymbol{\Delta} > 0$. Sea $\overline{\boldsymbol{\mu}} = \mathbf{E}(\mathbf{X})$ y sea $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$ una matriz semiortogonal cuyas columnas forman una base para el subespacio d dimensional $S_{\boldsymbol{\Gamma}} = \text{span}\{\boldsymbol{\mu}_y - \overline{\boldsymbol{\mu}} : y \in \Omega_Y\}$,

donde Ω_Y es el espacio muestral de Y. Luego, podemos escribir

$$\mathbf{X}_y = \overline{\boldsymbol{\mu}} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y + \boldsymbol{\varepsilon}, \tag{5.3}$$

donde $\boldsymbol{\varepsilon}$ es independiente de Y y distribuido normalmente con media 0 y matriz de covarianza $\boldsymbol{\Delta}$ y $\boldsymbol{\nu}_y = (\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T(\boldsymbol{\mu}_y - \overline{\boldsymbol{\mu}}) \in \mathbb{R}^d$. La matriz $\boldsymbol{\Gamma}$ no es identificable en este modelo pues para cualquier matriz de rango completo $\mathbf{A} : d \times d$, podemos obtener una parametrización equivalente $\boldsymbol{\Gamma}\boldsymbol{\nu}_y = (\boldsymbol{\Gamma}\mathbf{A}^{-1})(\mathbf{A})\boldsymbol{\nu}_Y$. Sin embargo, $\mathcal{S}_{\boldsymbol{\Gamma}}$ es identificable y estimable. Bajo el modelo (5.3), **Cook and Forzani, 2008** demuestran que $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$ es el subespacio de reducción suficiente minimal, más aun $R(\mathbf{X}) = \boldsymbol{\Gamma}\boldsymbol{\Delta}^{-1}\mathbf{X}$ es la reducción suficiente minimal y en este caso es lineal.

La formulación de PFC propone que los vectores $\boldsymbol{\nu}_y$ sean modelados de la forma $\boldsymbol{\nu}_y = \mathbf{B}(\mathbf{f}_y - \mathbf{E}(\mathbf{f}_Y))$, donde $\mathbf{f}_y \in \mathbb{R}^r$ es un vector de valores conocidos que depende de y y $\mathbf{B}$ es una matriz $d \times r$ sin restricciones de rango $d \leqslant \text{mín}(r, p)$. Bajo este modelo para $\boldsymbol{\nu}_y$ cada coordenada X_{yj} $j = 1, \ldots, p$ de $\mathbf{X}_y$ sigue un modelo lineal con vector predictor $\mathbf{f}_y$. En consecuencia, podemos utilizar los gráficos de X_{yj} versus y, $j = 1, \ldots, p$ para obtener información acerca de como elegir a $\mathbf{f}_y$ (Ver **Cook, 1998**, Capítulo 10). Por ejemplo en el trabajo **Bura and Cook, 2001** la Figura 1b, que se reproduce en la Figura 1, presenta un gráfico de dispersión de 5 variables de las cuales M es la respuesta y las 4 restantes son predictores. De acuerdo a este gráfico, parece razonable ajustar la relación de cada predictor con la variable respuesta a través de la función $\log(\cdot)$, indicando que en ese ejemplo $\mathbf{f}_y = \log(y)$ puede ser adecuado. Observar que esta herramienta visual no es posible aplicarla en la regresión directa de Y en $\mathbf{X}$, pues consiste en una función de regresión más compleja como hemos explicado con anterioridad. En algunos casos, puede haber una elección natural de $\mathbf{f}_y$. Supongamos por ejemplo que Y es categórica con $C = r + 1$ categorías H_w, $w = 1, \ldots, C$. Luego, definimos la w coordenada del vactor f_{yw} de $\mathbf{f}_y$ como

$$f_{yw} = J(y \in H_w) - n_w/n \ , \ w = 1, \ldots, r,$$

donde J es la función indicadora y n_w es el número de observaciones que pertenecen a la clase H_w. Cuando Y es una variable continua, puede considerarse un razonable y flexible conjunto de funciones bases, como los polinomios, que pueder muy útil cuando no es práctico aplicar los métodos gráficos a todos los predictores. Otra opción consiste en particionar los valores de Y en C intervalos (categorías) y construir $\mathbf{f}_y$ como para el caso de Y categórica. Se puede ver más de este tema en **Adragni and Cook, 2009**.

A partir del modelo (5.3), **Cook and Forzani, 2008** obtienen estimadores de máxima verosimilitud explícitos del subespacio central. Además, presentan resultados y simulaciones que

argumentan acerca de la robustez de los métodos SDR basados en modelos normales si se asume que en dicho modelo los errores son independientes de Y con momentos finitos pero posiblemente no normales y si la especificación de $\mathbf{f}_y$ es errónea. Debilitando estas hipótesis del modelo, establecen características robustas acerca del subespacio de reducción suficiente poblacional $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$ y acerca de los estimadores de máxima verosimilitud de la reducción suficiente $R(\mathbf{X}) = \boldsymbol{\Gamma}\boldsymbol{\Delta}^{-1}\mathbf{X}$. Ellos establecen que el subespacio de reducción suficiente $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$ está contenido en $\mathcal{S}_{Y|\mathbf{X}}$, es decir que aún se captura una parte del subespacio central. Además, demuestran que dicho estimador de $R(\mathbf{X}) = \boldsymbol{\Gamma}\boldsymbol{\Delta}^{-1}\mathbf{X}$ es todavía $\sqrt{n}$-consistente si la matriz $d \times r$ de correlaciones de los elementos de $\boldsymbol{\nu}_y$ y $\mathbf{f}_y$ tiene rango d.

El caso donde la matriz de covarianza condicional constante no es satisfecha, es estudiado en detalle en el trabajo **[Cook and Forzani, 2009]**. Ellos demuestran que una reducción suficiente lineal no es necesariamente minimal. Volveremos a este tema en la Sección 6.2.1.

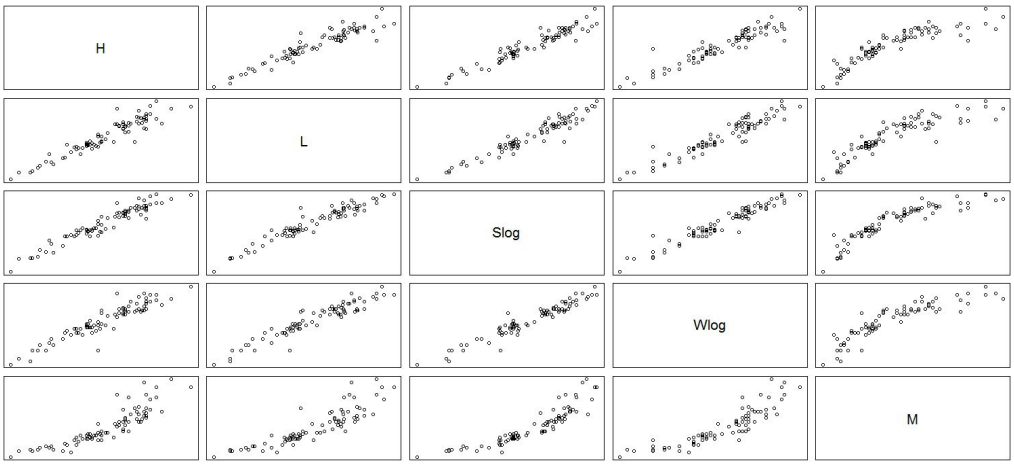


FIGURA 1. Gráfico de dispersión de las variable respuesta M y los predictores $H, L, Slog$ y $Wlog$ del conjunto de datos *Horse mussels* del Paquete dr del Software [R].

Hasta aquí, solamente modelos con predictores continuos han sido considerados en esta sección. El trabajo de **[Cook and Li, 2009]** fue el primer paso hacia el desarrollo de metodologías SDR donde predictores continuos y categóricos se consideren en forma conjunta bajo el enfoque de regresión inversa basada en modelos. Ellos proponen que las componentes del vector aleatorio $\mathbf{X}_y = (X_{yj})\,,\ j = 1, \ldots, p$ sean condicionalmente independientes dada la respuesta Y y que cada X_{yj} de $\mathbf{X}_y$ pertenezca a una familia exponencial a un parámetro con densidad o función de

probabilidad puntual de la forma:

$$f_j(x|\eta_{yj}, Y = y) = \exp(x\eta_{yj} - \psi_j(\eta_{yj}))h_j(x), \tag{5.4}$$

donde los parámetros naturales η_{yj} $j = 1, \ldots, p$ son funciones de y como indica el subíndice. Sea $\boldsymbol{\eta}_y \in \mathbb{R}^p$ que contiene los elementos η_{yj} con $j = 1, \ldots, p$, $\overline{\boldsymbol{\eta}} = \mathbf{E}(\boldsymbol{\eta}_Y)$ y $\mathcal{S}_{\boldsymbol{\Gamma}} = \text{span}\{\boldsymbol{\eta}_y - \overline{\boldsymbol{\eta}} : y \in \Omega_Y\}$. Luego,

$$\begin{aligned} \boldsymbol{\eta}_y &= \overline{\boldsymbol{\eta}} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y \\ \eta_{yj} &= \overline{\eta}_j + \boldsymbol{\gamma}_j^T \boldsymbol{\nu}_y \text{ con } j = 1, \ldots, p \end{aligned} \tag{5.5}$$

donde $\boldsymbol{\nu}_y = \boldsymbol{\Gamma}^T\boldsymbol{\eta}_y - \overline{\boldsymbol{\eta}}$, y $\boldsymbol{\gamma}_j^T$ es la j-ésima fila de $\boldsymbol{\Gamma}$. Siguiendo la formulación de PFC, ellos modelan $\boldsymbol{\nu}_y$ como $\boldsymbol{\nu}_y = \boldsymbol{\beta}(\mathbf{f}_y - \mathbf{E}(\mathbf{f}_Y))$, donde $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$ tiene rango $d \leq \text{mín}(p, r)$ y $\mathbf{f}_y \in \mathbb{R}^r$ son vectores conocidos que dependen de y. Es decir, que cada coordenada X_{yj}, $j = 1, \ldots, p$ de $\mathbf{X}_y$ sigue un modelo lineal generalizado en $\mathbf{f}_y$. Sustituyendo $\boldsymbol{\beta}\mathbf{f}_y$ por $\boldsymbol{\nu}_y$ en (5.5) obtenemos que:

$$\boldsymbol{\eta}_y = \boldsymbol{\zeta} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y, \tag{5.6}$$

donde $\boldsymbol{\zeta} = \overline{\boldsymbol{\eta}} - \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{E}(\mathbf{f}_Y)$. Ellos denominan este modelo como *Generalized Principal Fitted Components* (GPFC). La matriz $\boldsymbol{\Gamma}$ no es identificable en el modelo pero $\mathcal{S}_{\boldsymbol{\Gamma}}$ si es identificable y estimable. El espacio de parámetros al cual pertenece $S_{\boldsymbol{\Gamma}}$ se denomina *Grassmann* $\mathcal{G}_{d,p}$ de dimensión d en $\mathbb{R}^p$. $\mathcal{G}_{d,p}$ es de dimensión $d(p-d)$ [**Chikuse, 2002**], que es el número de parámetros necesitados para especificar un único subespacio. En el caso que $d = \text{mín}(p, r)$, la matriz de coeficientes $\boldsymbol{\Gamma}\boldsymbol{\beta}$ es una matriz sin restricciones $\boldsymbol{\Phi} \in \mathbb{R}^{p\times r}$ y ellos se refieren al modelo resultante $\boldsymbol{\eta}_y = \boldsymbol{\zeta} + \boldsymbol{\Phi}\mathbf{f}_y$ como el modelo completo, el cual sirve como punto de referencia para evaluar modelos con $d < \text{mín}(p, r)$.

Bajo este modelo obtienen que la reducción suficiente minimal es $\boldsymbol{\Gamma}^T\mathbf{X}$, la cual es lineal en los predictores. Sin embargo, la hipótesis de independencia condicional de los predictores es muy restrictiva y limita las aplicaciones en conjunto de datos reales. Además, como las familias exponenciales deben ser a un parámetro, el caso donde los predictores son condicionalmente independiente con distribución normal con media $\boldsymbol{\mu}_y$ y varianza $\sigma_y^2\mathbf{I}$ queda excluido.

La estimación de la reducción es hecha vía optimización sobre *Grassmann*, un algoritmo de optimización específico y complejo que requiere el conocimiento de la función de máxima verosimilitud completa y sus derivadas parciales y valores iniciales muy precisos (ver por ejemplo

[**Edelman et al., 1998**], [**Adragni et al., 2012**]). En el próximo capítulo, donde extenderemos esta teoría, presentaremos otro algoritmo que es más habitual en la literatura estadística y que puede aplicarse al modelo GPFC obteniendo los mismos estimadores.

5.4. Reducción suficiente basada en métodos kernel

Una característica que define la mayoría de los metodos SDR, y en particular los discutidos anteriormente, es que estos obtienen reducciones suficientes que son proyecciones del vector de predictores en un subespacio de dimensión inferior. Es decir, estiman sólo reducciones lineales perdiendo información y eficiencia en aquellos casos donde la reducción suficiente minimal no es lineal en $\mathbf{X}$. En la búsqueda de superar la naturaleza lineal de la mayoría de los métodos RDS, varios trabajos recientes han combinado reducción suficiente de dimensiones con la teoría de espacios reproducidos en espacios de Hilbert (RKHS) ([**Akaho, 2001**], [**Bach and Jordan, 2002**], [**Fukumizu et al., 0304**], [**Fukumizu et al., 2009**], [**Fukumizu et al., 2007**], [**Wu, 2008**],[**Wu et al., 2008**], [**Hsing and Ren, 2009**], [**Yeh et al., 2009**], [**Zhu and Li, 2011**], [**Kim and Pavlovic, 2013**], entre otros).

La idea de los métodos kernel es trabajar en un espacio transformado de las variables $\mathbf{X}$ e Y y de esta forma obtener más información acerca de las variables y/o obtener reducciones no necesariamente lineales en $\mathbf{X}$, aplicando métodos SDR tradicionales en el espacio transformado. Por lo tanto, estos métodos trabajan en los siguientes espacios transformados,

$$\begin{aligned} \mathbf{x} &\rightarrow \phi_{\mathbf{X}}(\mathbf{x}) \\ \mathbf{y} &\rightarrow \phi_{\mathbf{Y}}(\mathbf{y}) \end{aligned}$$

donde $\phi_{\mathbf{Y}}$ y $\phi_{\mathbf{X}}$ son funciones que pertenecen a los espacios de Hilbert $\mathcal{H}_{k_{\mathbf{y}}}$ y $\mathcal{H}_{k_{\mathbf{x}}}$ inducidos por las funciones kernel $k_{\mathbf{x}}(\cdot,\cdot)$ y $k_{\mathbf{y}}(\cdot,\cdot)$ definidas en el espacio de las $\mathbf{x}$ e $\mathbf{y}$, respectivamente. La idea propuesta en [**Fukumizu et al., 0304**], el método *Kernel Dimension Reduction* (KDR), es elegir una matriz $\mathbf{B} \in \mathbb{R}^{p\times d}$ que minimice un criterio de dependencia que cuantifica la noción de dependencia condicional: $\mathbf{y} \perp\!\!\!\perp \mathbf{x}|\mathbf{z} = \mathbf{B}^T\mathbf{x}$ a través de un orden en los operadores de covarianza definidos positivos. Este enfoque se formaliza en el siguiente teorema:

Teorema 5.7. $E[\mathrm{var}(\phi_{\mathbf{Y}}(\mathbf{y}))|\mathbf{x}] \preccurlyeq E[\mathrm{var}(\phi_{\mathbf{Y}}(\mathbf{y}))|\mathbf{B}^T\mathbf{x}]$ *donde la igualdad vale si y solo si* $\mathbf{y} \perp\!\!\!\perp \mathbf{x}|\mathbf{z} = \mathbf{B}^T\mathbf{x}$.

Este teorema trasladado a la versión muestral, es equivalente a encontrar $\widehat{\mathbf{B}}$ tal que

$$\min_{\mathbf{B}} \text{tr}[\mathbf{K}_y(\mathbf{K}_{\mathbf{B}^T\mathbf{x}} + n\epsilon\mathbf{I}_n)^{-1}] \tag{5.7}$$

sujeto a $\mathbf{B}^T\mathbf{B} = \mathbf{I}_q$, donde $\mathbf{K_y}$ y $\mathbf{K}_{\mathbf{B}^T\mathbf{x}}$ son matrices de dimensiones $n \times n$ con $\mathbf{K}_{\mathbf{B}^T\mathbf{x}}(i,j) = k_{\mathbf{x}}(\mathbf{B}^T\mathbf{x}_i, \mathbf{B}^T\mathbf{x}_j)$ y $\mathbf{K_y}(i,j) = k_{\mathbf{y}}(\mathbf{y}_i, \mathbf{y}_j)$ calculadas sobre la muestra $\{\mathbf{y}_i\}$ y $\{\mathbf{B}^T\mathbf{x}_i\}$, $i = 1, \ldots, n$ respectivamente y ϵ es un parámetro de regularización para que sea posible la inversión en (5.7).

KDR no asume ninguna restricción en la distribución de $(\mathbf{Y}, \mathbf{X})$ debido a que heredan la generalidad del Teorema 5.7 en que se basan. Además, a pesar de su formulación en RKHS, la reducción suficiente obtenida es lineal en el espacio original de las $\mathbf{X}$. Como contrapartida, la optimización (5.7) es no convexa la cual implica un alto costo computacional.

El método *Kernel Sliced Invere Regression* (KSIR), propuesto en [**Wu, 2008**], aplica básicamente las ideas de SIR a las variables transformadas $\boldsymbol{\phi}_{\mathbf{X}}(\mathbf{x})$ en lugar de $\mathbf{x}$ y obtiene reducciones lineales en $\boldsymbol{\phi}_{\mathbf{X}}(\mathbf{x})$. En general, no se tiene un conocimiento explícito de $\boldsymbol{\phi}_{\mathbf{X}}(\mathbf{x})$, si no que se cuenta con la información de $\langle \boldsymbol{\phi}_{\mathbf{X}}(\mathbf{x}_i), \boldsymbol{\phi}_{\mathbf{X}}(\mathbf{x}_j)\rangle_{\mathcal{H}} = k_{\mathbf{x}}(\mathbf{x}_i, \mathbf{x}_j)$. Por lo tanto, en este caso no es posible dar una solución explícita en la escala original de $\mathbf{X}$. KSIR estima $\text{var}(\mathbf{E}[\boldsymbol{\phi}(\mathbf{x})|y])$ particionando la respuesta (similar a SIR) en H intervalos $\{S_j\}_{j=1}^{H}$. Sea $\boldsymbol{\phi}_{\mathbf{X}} = [\boldsymbol{\phi}_{\mathbf{X}}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}_{\mathbf{X}}(\mathbf{x}_n)]$ y $\mathbf{C} : n \times H$ la matriz que contiene en su i-ésima fila cero en todas las posiciones excepto en la j-ésima posición, donde $y_i \in S_j$ que contiene un 1. Además, sea $\mathbf{P} : H \times H$ sea la matriz diagonal cuya j-ésima entrada es $p_j = \mathbf{P}(j,j) = |S_j|/n$. La estimación del subespacio central se obtiene siguiendo los pasos:

1. Particionar $y_1, \ldots, y_n$ en H intervalos. Calcular los promedios por intervalos, $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_H] = \dfrac{1}{n}\boldsymbol{\phi}_{\mathbf{X}}\mathbf{C}\mathbf{P}^{-1}$.
2. Estimar la covarianza muestral, es decir $\mathbf{V} = \mathbf{M}\mathbf{P}\mathbf{M}^T$. Sus primeras d autofunciones son denotadas con $\mathbf{v}_1, \ldots, \mathbf{v}_d$.
3. Las direcciones obtenidas son $\mathbf{b}_l = \Sigma_{\mathbf{x}}^{-1}\mathbf{v}_l$ para $l = 1, \ldots, d$.

Aunque el cálculo anterior no puede, en general, llevarse a cabo explícitamente en el espacio transformado, se puede realizar de la siguiente forma. En el paso 2, se resuelve el autosistema $\mathbf{V}\mathbf{v} = \lambda\mathbf{v}$ y luego representamos $\mathbf{v}$ como combinación lineal de $\{\boldsymbol{\phi}_{\mathbf{X}}(\mathbf{x}_i)\}_{i=1}^{n}$, es decir, $\mathbf{v} = \sum \alpha_i \boldsymbol{\phi}_{\mathbf{X}}(\mathbf{x}_i) = \boldsymbol{\phi}_{\mathbf{X}}\boldsymbol{\alpha}$, donde $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^T$. Premultiplicando el autosistema por $\boldsymbol{\phi}_{\mathbf{X}}^T$ da el siguiente sistema dual:

$$\frac{1}{n^2}\mathbf{K}_{\mathbf{x}}\mathbf{C}\mathbf{P}^{-1}\mathbf{C}^T\mathbf{K}_{\mathbf{x}}\boldsymbol{\alpha} = \lambda\mathbf{K}_{\mathbf{x}}\boldsymbol{\alpha}, \tag{5.8}$$

donde $\mathbf{K_x}$ es la matriz $n \times n$ definida con anterioridad. La ecuación (5.8) es el problema de autovalores generalizado, de donde debemos encontrar los primeros d autovectores $\boldsymbol{\alpha}$.

Las direcciones $\mathbf{b}$ del paso 3, pueden ser obtenidas de $\mathbf{v}$ aplicando un desarrollo similar. Para resolver $\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{b}$, remplazamos el operador de covarianza $\boldsymbol{\Sigma}_{\mathbf{x}}$ por su estimador $\frac{1}{n}\boldsymbol{\phi}_{\mathbf{X}}\boldsymbol{\phi}_{\mathbf{X}}^T$. Expresando $\mathbf{b} = \sum_{i=1}^{n} \beta_i \boldsymbol{\phi}_{\mathbf{X}}(\mathbf{x}_i) = \boldsymbol{\phi}_{\mathbf{X}}\boldsymbol{\beta}$, con $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_n]^T$ y premultiplicando el sistema por $\boldsymbol{\phi}_{\mathbf{X}}^T$, obtenemos:

$$\boldsymbol{\beta} = n\mathbf{K}_{\mathbf{x}}^{-1}\boldsymbol{\alpha}. \tag{5.9}$$

La reducción suficiente estimada (que resultará no lineal en el espacio original) es representado por las d funciones bases $\{\mathbf{b}_l\}_{l=1}^{d}$ utilizando $\{\boldsymbol{\beta}^l\}_{l=1}^{d}$, las d soluciones de (5.9).

Covariance Operator Inverse Regression (COIR) propuesto en **Kim and Pavlovic, 2013** está basado en la descomposición en autovectores de la matriz de covarianza de la regesión inversa. Sin embargo, a diferencia de KSIR, COIR no necesita particionar el rango de Y ya que utiliza los operadores de covarianza de la variable $\mathbf{X}$ y la respuesta Y. Esto hace que sea más adecuado para sus aplicaciones con la respuestas multivariantes. En el caso que la respuesta fuera univariada, COIR es una modificación de KSIR donde un modelo no paramétrico se utiliza en la regresión inversa de $\mathbf{X}|Y$ en lugar de utilizar una partición de Y.

Los métodos kernel dan un paso importante hacia la identificación de reducciones suficientes no lineales y tiene un carácter general al poder aplicarse a cualquier tipo de datos, pero en la mayoría de los casos no es posible obtener una forma explícita de la reducción y pueden ser menos eficientes que aquellos métodos donde puede establecerse un modelo parámetrico. Esto será corroborado a través de ejemplos en el siguiente capítulo.

Las metodologías presentadas hasta aquí son los métodos vigentes en SDR y los usaremos como base de comparación de los nuevos resultados presentados en el próximo capítulo.

5.5. Demostraciones del Cápitulo 5

Demostración del Lema 5.2: Vamos a suponer, sin pérdida de generalidad, que existe la densidad conjunta de $(Y, \mathbf{X})$. Luego, en la mayoría de los pasos de la demostración podremos utilizar la definición de densidad condicional. El punto **(iii)** es cierto si y solo si la función densidad de $(Y, \mathbf{X})|R(\mathbf{X})$ se factoriza, es decir:

$$f_{(Y,\mathbf{X})|R(\mathbf{X})} = f_{Y|R(\mathbf{X})} \cdot f_{\mathbf{X}|R(\mathbf{X})}$$

Como además $f_{(Y,\mathbf{X})|R(\mathbf{X})} = \dfrac{f_{(Y,\mathbf{X},R(\mathbf{X}))}}{f_{R(\mathbf{X})}}$ tenemos que

$$\begin{aligned} f_{\mathbf{X}|R(\mathbf{X})} &= \frac{f_{(Y,\mathbf{X},R(\mathbf{X}))}}{f_{Y|R(\mathbf{X})} \cdot f_{R(\mathbf{X})}} \\ &= \frac{f_{(Y,\mathbf{X},R(\mathbf{X}))}}{f_{(Y,R(\mathbf{X}))}} \\ &= f_{\mathbf{X}|Y,(R(\mathbf{X}))} \end{aligned}$$

Es decir, es cierto el punto **(ii)**. Con las mismas igualdades podemos ver que el punto **(ii)** implica **(iii)**. Si el punto **(ii)** es cierto tenemos que $f_{\mathbf{X}|(Y,R(\mathbf{X}))} = f_{\mathbf{X}|R(\mathbf{X})}$, luego:

$$\begin{aligned} \frac{f_{(Y,\mathbf{X},R(X))}}{f_{(Y,R(\mathbf{X}))}} &= \frac{f_{(\mathbf{X},R(\mathbf{X}))}}{f_{R(\mathbf{X})}} \\ &\Updownarrow \\ \frac{f_{(Y,\mathbf{X},R(\mathbf{X}))}}{f_{(\mathbf{X},R(\mathbf{X}))}} &= \frac{f_{(Y,R(\mathbf{X}))}}{f_{R(\mathbf{X})}} \\ &\Updownarrow \\ f_{(Y|\mathbf{X},R(\mathbf{X}))} &= f_{Y|R(\mathbf{X})} \end{aligned}$$

Como $f_{Y|\mathbf{X}} = f_{Y|(\mathbf{X},R(\mathbf{X}))}$, se tiene que **(i)** es equivalente a **(ii)**.

□

CAPÍTULO 6

REDUCCIÓN SUFICIENTE DE DIMENSIONES PARA FAMILIAS EXPONENCIALES

En este capítulo se considera el problema de identificar reducciones suficientes en regresiones con respuesta Y en $\mathbb{R}$ y predictores que pueden ser todos continuos, todos categóricos o una mezcla de variables continuas o categóricas. Basados en el enfoque SDR de proponer un modelo para la regresión inversa, se asume que la distribución de $\mathbf{X}|Y$ pertenece a una familia exponencial, es decir $\mathbf{X}|Y \sim \mathcal{F}_{\boldsymbol{\eta},\mathbf{T},\psi}$ pero sin imponer la condición de independencia condicional de los predictores como en el caso de [**Cook and Li, 2009**]. Se identifica la reducción suficiente minimal para la regresión de Y dado $\mathbf{X}$ y se muestra que esta no es necesariamente lineal en los predictores pero sí en $\mathbf{T}(\mathbf{X})$, siendo $\mathbf{T}(\mathbf{X})$ un estadístico suficiente de la familia exponencial. Dependiendo de la forma de $\mathbf{T}(\mathbf{X})$, la reducción suficiente minimal podrá ser lineal o no lineal en los predictores, como es el caso de la distribución Bernoulli multivariada que se estudiará en detalle. La estimación de la reducción suficiente minimal se obtiene ajustando modelos GLMM, propios de las familias exponenciales y se propone aplicar el algoritmo IRLS para obtener estimadores de máxima verosimilitud. Se denotará al modelo y a los estimadores obtenidos con abreviatura EF-DR.

6.1. Estructura del modelo

Asumimos que el vector aleatorio $\mathbf{X}_y = \mathbf{X}|(Y = y) \in \mathbb{R}^p$ tiene densidad o función de probabilidad puntual dada por

$$f(\mathbf{x}|\boldsymbol{\eta}_y, Y = y) = e^{\boldsymbol{\eta}_y^T \mathbf{T}(\mathbf{X}) - \psi(\boldsymbol{\eta}_y)} h(\mathbf{x}), \tag{6.1}$$

es decir, pertenece a una familia exponencial donde los parámetros naturales η_{yj}, $j = 1 \ldots, k$ son funciones de y como indica el primer subíndice , $\boldsymbol{\eta}_y = (\eta_{y1}, \ldots, \eta_{yk})^T$ y $\mathbf{T}(\mathbf{X})$ es el estadístico suficiente de la familia. La dimensión del vector de parámetros naturales satisface en general $k \geq p$, por ejemplo si $\mathbf{X}_y \sim \mathcal{N}_p(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$, la dimensión de $\boldsymbol{\eta}_y$ es $k = p + p(p+1)/2$.

Siguiendo la teoría de GLMM estudiado en los capítulos 1 y 3, asumimos que los parámetros naturales son función lineal de $\boldsymbol{\nu}_Y$. Luego

$$\boldsymbol{\eta}_Y = \overline{\boldsymbol{\eta}} + \mathbf{A}\boldsymbol{\nu}_Y,$$

donde $\overline{\boldsymbol{\eta}} = \mathbf{E}(\boldsymbol{\eta}_Y)$, $\mathbf{A} \in \mathbb{R}^{k \times d}$ es una matriz de rango completo semiortogonal tal que sus columnas forman una base para $\mathcal{S}_{\mathbf{A}} = \text{span}\{\boldsymbol{\eta}_y - \overline{\boldsymbol{\eta}} : y \in \Omega_Y\}$ con Ω_Y el espacio muestral de Y y $\mathbf{E}(\boldsymbol{\nu}_Y) = 0$. Con el objetivo de explicitar la dependencia de $\boldsymbol{\nu}_Y$ en Y, asumimos que el vector de coordenadas sigue el modelo $\boldsymbol{\nu}_Y = \mathbf{B}(\mathbf{f}_Y - \mathbf{E}(\mathbf{f}_Y))$, donde $\mathbf{f}_Y \in \mathbb{R}^r$ es un vector de valores conocidos que depende de Y y $\mathbf{B}^T \in \mathbb{R}^{r \times d}$ es una matriz sin restricciones de rango $d \leq \text{mín}(k, r)$. Esta formulación es similar a la de PFC ([**Cook, 2007**], [**Cook and Forzani, 2008**]) y a regresión inversa paramétrica (PIR) [**Bura and Cook, 2001**] para predictores continuos. Bajo este modelo, cada coordenada η_{Yj}, $j = 1, \ldots, k$ sigue un modelo lineal generalizado con $\mathbf{f}_Y$ como vector predictor. Por lo tanto, podemos utilizar las herramientas desarrolladas en la Sección 5.3 para obtener información acerca de como elegir y construir $\mathbf{f}_Y$. Sea $\mathbf{D} = \mathbf{AB}$, con $\mathbf{A} : k \times d$, $\mathbf{B}^T : r \times d$ y $\mathbf{f}_Y \in \mathbb{R}^r$ funciones conocidas que dependen de Y de modo que

$$\boldsymbol{\eta}_Y = \overline{\boldsymbol{\eta}} + \mathbf{D}(\mathbf{f}_Y - \overline{\mathbf{f}}) \tag{6.2}$$

$$= \overline{\boldsymbol{\eta}} + \mathbf{AB}(\mathbf{f}_Y - \overline{\mathbf{f}}). \tag{6.3}$$

Las ecuaciones (6.2) y (6.3) expresan modelos lineales generalizados sin restricción y de rango reducido respectivamente. Si bien en el Capítulo 1 y 3 diferenciabamos los k parámetros naturales en k_1 y k_2 (para poder generalizar e incluir casos como el de la normal multivariada), en este capítulo no haremos dicha distinción para lograr mayor compresión y simplificar notaciones. Sin embargo, todos los resultados pueden adaptarse a esos casos.

6.2. Reducción suficiente minimal

En esta sección identificamos la reducción minimal suficiente para la regresión $Y|\mathbf{X}$ cuando $\mathbf{X}|Y$ sigue el modelo (6.1). Además analizamos algunos resultados previos que son ejemplos de familias exponenciales y en especial estudiamos su conexión con el trabajo [**Cook and Li, 2009**].

Teorema 6.1. *Si* $\mathbf{X}|Y$ *tiene densidad o función de probabilidad puntual en una familia exponencial dada por (6.1), entonces la reducción suficiente minimal para la regresión de* $Y|\mathbf{X}$ *es*

$$R(\mathbf{X}) = \boldsymbol{\alpha}^T \left(\mathbf{T}(\mathbf{X}) - \mathbf{E}(\mathbf{T}(\mathbf{X}))\right),$$

donde $\boldsymbol{\alpha} \in \mathbb{R}^{k\times d}$ *es tal que* $\mathcal{S}_{\boldsymbol{\alpha}} = \text{span}\{\boldsymbol{\eta}_Y - \overline{\boldsymbol{\eta}} : Y \in \Omega_Y\}$ *y* $\mathbf{T}(\mathbf{X})$ *es un estadístico suficiente de la familia para los parámetros naturales de la familia exponencial definida en (6.1).*

Demostración. De acuerdo al Teorema de Lehmann-Scheffé (ver Teorema 6.2.13 de **[Casella and Berger, 1990]**), la afirmación de que $\boldsymbol{\alpha}^T \left(\mathbf{T}(\mathbf{X}) - \mathbf{E}(\mathbf{T}(\mathbf{X}))\right)$ sea una reducción suficiente minimal deriva de mostrar que son equivalentes los siguientes puntos: (i) $\log\left(f_{\mathbf{X}|Y}(\mathbf{x})/f_{\mathbf{X}|Y}(\mathbf{z})\right)$ es independiente de Y y (ii) $\boldsymbol{\alpha}^T \left(\mathbf{T}(\mathbf{x}) - \mathbf{E}(\mathbf{T}(\mathbf{x}))\right) = \boldsymbol{\alpha}^T \left(\mathbf{T}(\mathbf{z}) - \mathbf{E}(\mathbf{T}(\mathbf{z}))\right)$ o equivalentemente $\boldsymbol{\alpha}^T\ \mathbf{T}(\mathbf{x}) = \boldsymbol{\alpha}^T\ \mathbf{T}(\mathbf{z})$. De acuerdo a (6.1), la independencia de $\log\left(f_{\mathbf{X}|Y}(\mathbf{x})/f_{\mathbf{X}|Y}(\mathbf{z})\right)$ de Y es equivalente a

$$\log\frac{h(\mathbf{x})}{h(\mathbf{z})} + (\mathbf{T}(\mathbf{x}) - \mathbf{T}(\mathbf{z}))^T \boldsymbol{\eta}_Y = c, \tag{6.4}$$

donde c no depende de Y. Tomando esperanza con respecto a Y obtenemos que (6.4) es equivalente a

$$(\mathbf{T}(\mathbf{x}) - \mathbf{T}(\mathbf{z}))^T(\boldsymbol{\eta}_Y - \overline{\boldsymbol{\eta}}) = 0 \text{ para todo } y \in \Omega_Y. \tag{6.5}$$

Sea $\boldsymbol{\alpha} \in \mathbb{R}^{k\times d}$ una matriz cuyas columnas generen el subespacio $\text{span}\{\boldsymbol{\eta}_Y - \overline{\boldsymbol{\eta}} : Y \in \Omega_Y\}$. Luego, $\boldsymbol{\eta}_Y - \overline{\boldsymbol{\eta}} = \boldsymbol{\alpha}\boldsymbol{\nu}_Y$ para algún $\boldsymbol{\nu}_Y \in \mathbb{R}^d$ y (6.5) es equivalente a que $(\mathbf{T}(\mathbf{x}) - \mathbf{T}(\mathbf{z}))^T\boldsymbol{\alpha}\boldsymbol{\nu}_Y = 0$ para todo $y \in \Omega_Y$, es decir que $\boldsymbol{\alpha}^T\mathbf{T}(\mathbf{x}) = \boldsymbol{\alpha}^T\mathbf{T}(\mathbf{z})$. Por lo tanto, $\boldsymbol{\alpha}^T \left(\mathbf{T}(\mathbf{X}) - \mathbf{E}(\mathbf{T}(\mathbf{X}))\right)$ es una reducción suficiente minimal para la regresión de $Y|\mathbf{X}$. □

Consideremos una subclase de familias exponenciales, que fue especificamente discutida en **[Cook and Li, 2009]**, denominadas familias exponenciales cuadráticas cuya densidad depende de sus primeros dos momentos:

Definición 6.2. La distribución de $\mathbf{X}|Y$ pertenece a una familia exponencial cuadrática si su densidad (o función de probabilidad puntual) está dada por

$$\begin{aligned} f(\mathbf{x}|\boldsymbol{\eta}_y, Y = y) &= e^{\boldsymbol{\eta}_y^T\mathbf{T}(\mathbf{X}) - \psi(\boldsymbol{\eta}_y)}h(\mathbf{x}) \\ &= e^{\boldsymbol{\eta}_{y,1}^T\mathbf{X} + \boldsymbol{\eta}_{y,2}^T\text{vec}(\mathbf{X}\mathbf{X}^T) - \psi(\boldsymbol{\eta}_y)}h(\mathbf{x}). \end{aligned} \tag{6.6}$$

Corolario 6.3. *Si la distribución de* $\mathbf{X}|Y$ *pertenece a una familia exponencial cuadrática con densidad o función de probabilidad puntual (6.6), la reducción minimal suficiente para la regresión de* $Y|\mathbf{X}$ *está dada por*

$$R(\mathbf{X}) = \left(\boldsymbol{\alpha}_1^T(\mathbf{X} - \mathbf{E}(\mathbf{X})) + \boldsymbol{\alpha}_2^T(\text{vec}(\mathbf{X}\mathbf{X}^T) - \mathbf{E}(\text{vec}(\mathbf{X}\mathbf{X}^T)))\right),$$

donde $\text{span}(\boldsymbol{\alpha}_1) = \text{span}\{\boldsymbol{\eta}_{Y,1} - \mathbf{E}(\boldsymbol{\eta}_{Y,1}), Y \in \Omega_Y\}$ *y* $\text{span}(\boldsymbol{\alpha}_2) = \text{span}\{\boldsymbol{\eta}_{Y,2} - \mathbf{E}(\boldsymbol{\eta}_{Y,2}), Y \in \Omega_Y\}$.

La demostración del corolario se sigue directamente del Teorema 6.1 y del hecho que el estadístico suficiente para la familia exponencial es $\mathbf{T}(\mathbf{X}) = (\mathbf{X}, \text{vec}^T(\mathbf{X}\mathbf{X}^T))$.

Observación 6.4. Si consideramos la siguiente reducción suficiente

$$R(\mathbf{X}) = \begin{pmatrix} \boldsymbol{\alpha}_1^T(\mathbf{X} - \mathbf{E}(\mathbf{X})) \\ \boldsymbol{\alpha}_2^T(\text{vec}(\mathbf{X}\mathbf{X}^T) - \mathbf{E}(\text{vec}(\mathbf{X}\mathbf{X}^T))) \end{pmatrix},$$

el subespacio de reducción suficiente correspondiente es $\mathcal{S} = \text{span}((\boldsymbol{\alpha}_1^T, 0)^T) \oplus \text{span}((0, \boldsymbol{\alpha}_2^T)^T)$, cuya dimensión es más grande que la dimensión del subesapcio de reducción suficiente minimal definido en el Corolario 6.3.

6.2.1. Algunos ejemplos conocidos

Aquí describimos algunos resultados previos que son ejemplos de familias exponenciales cuadráticas y se incluyen en el marco de nuestros resultados.

a. **Normal con varianza constante:** $\mathbf{X}|Y \sim \mathcal{N}(\boldsymbol{\mu}_Y, \boldsymbol{\Delta})$

   Ya que $\mathbf{X}|Y$ posee densidad perteneciente a una familia exponencial cuadrática, aplicamos el Corolario 6.3 y obtenemos que $\boldsymbol{\alpha}_1$ es tal que sus columnas son una base para el subespacio $\boldsymbol{\Delta}^{-1}\text{span}\{\boldsymbol{\mu}_Y - \boldsymbol{\mu}, Y \in \Omega_Y\}$ y $\boldsymbol{\alpha}_2 = 0$. Es decir, $\boldsymbol{\alpha}_1^T(\mathbf{X} - \mathbf{E}(\mathbf{X}))$ es la reducción minimal suficiente, que es el resultado obtenido en **[Cook, 2007]** y **[Cook and Forzani, 2008]**.

b. **Normal con varianza que depende de** Y **a través de un efecto multiplicativo:** $\mathbf{X}|Y \sim \mathcal{N}(\boldsymbol{\mu}_Y, c_Y\boldsymbol{\Delta})$

   En este caso, escribimos la densidad de $\mathbf{X}|Y$ de la siguiente forma:

$$f_{\mathbf{X}|Y=y}(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\mu}_y - \boldsymbol{\mu})^T c_y\boldsymbol{\Delta}(\boldsymbol{\mu}_y - \boldsymbol{\mu})\right)}{2\pi^{p/2}|c_y\boldsymbol{\Delta}|^{1/2}} \times$$
$$\exp\left(-\frac{1}{2}\text{vec}((\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T)^T\text{vec}(c_y^{-1}\boldsymbol{\Delta}^{-1}) + (\mathbf{x} - \boldsymbol{\mu})^T c_y^{-1}\boldsymbol{\Delta}^{-1}(\boldsymbol{\mu}_y - \boldsymbol{\mu})\right).$$

Luego, de acuerdo a la Observación 6.4, $\boldsymbol{\alpha}_1 = \boldsymbol{\Delta}^{-1}\text{span}\{(\boldsymbol{\mu}_Y - \boldsymbol{\mu}), Y \in \Omega_Y\}$ y $\boldsymbol{\alpha}_2 = \text{vec}(\boldsymbol{\Delta}^{-1})$. Por lo tanto, una reducción suficiente es $\{\boldsymbol{\alpha}_1^T(\mathbf{X}-\boldsymbol{\mu}), (\mathbf{X}-\boldsymbol{\mu})^T\boldsymbol{\Delta}^{-1}(\mathbf{X}-\boldsymbol{\mu})\}$, como en **Bura and Forzani, 2015**.

c. **Normal con varianza no constante:** $\mathbf{X}|Y \sim \mathcal{N}(\boldsymbol{\mu}_Y, \boldsymbol{\Delta}_Y)$

Similarmente al caso b., obtenemos que el estadístico suficiente es $\mathbf{T}(\mathbf{X}) = (\mathbf{X} - \boldsymbol{\mu}, \text{vec}((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T))$ y $\boldsymbol{\eta}_Y = (\boldsymbol{\Delta}_Y^{-1}(\boldsymbol{\mu}_Y - \boldsymbol{\mu}), -\text{vec}(\boldsymbol{\Delta}_Y^{-1})/2)$. Sea $\boldsymbol{\alpha}_1 = \text{span}(\boldsymbol{\Delta}_Y^{-1}(\boldsymbol{\mu}_Y - \boldsymbol{\mu}))$ y consideremos la matriz $\mathbf{S} = (\mathbf{S}_j)_{j=1}^q$ de dimensiones $p^2 \times q$, tal que $\text{span}(\mathbf{S}) = \text{span}(\text{vec}(\boldsymbol{\Delta}_Y^{-1} - \mathbf{E}(\boldsymbol{\Delta}_Y^{-1})))$. Reordenando las columnas de $\mathbf{S}$ en matrices de dimensiones $p \times p$, podemos construir las matrices simétricas $\mathbf{S}_j$, $j = 1, \ldots, q$. Luego, una reducción suficiente es $\{\boldsymbol{\alpha}_1^T\mathbf{X}, (\mathbf{X}-\boldsymbol{\mu})^T\mathbf{S}_1(\mathbf{X}-\boldsymbol{\mu}), \ldots, (\mathbf{X}-\boldsymbol{\mu})^T\mathbf{S}_q(\mathbf{X}-\boldsymbol{\mu})\}$, resultado obtenido en **Forzani, 2007**.

Cook and Forzani, 2009 también consideran este caso, es decir $\mathbf{X}|Y \sim \mathcal{N}(\boldsymbol{\mu}_Y, \boldsymbol{\Delta}_Y)$. Ellos obtienen una reducción lineal suficiente de la forma $\boldsymbol{\alpha}^T\mathbf{X}$ donde $\boldsymbol{\alpha}$ debe sastifacer que

$$\text{span}(\boldsymbol{\alpha}) \subset \text{span}\{\boldsymbol{\Delta}^{-1}(\boldsymbol{\mu}_y - \boldsymbol{\mu}), \boldsymbol{\Delta}_y^{-1} - \boldsymbol{\Delta}^{-1} : y \in \Omega_Y\}. \tag{6.7}$$

El Corolario 6.3 muestra que la reducción lineal (6.7) no es necesariamente minimal. Por ejemplo, en el caso que $\boldsymbol{\Delta}_Y = c_Y\boldsymbol{\Delta}$ la reducción suficiente minimal *lineal* obtenida en (6.7) es $\mathbb{R}^p$, mientras que la reducción minimal es $\left(\boldsymbol{\alpha}_1^T(\mathbf{X}-\boldsymbol{\mu}) + (\mathbf{X}-\boldsymbol{\mu})^T\boldsymbol{\Delta}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right)$ con $\boldsymbol{\alpha}_1 = \text{span}\{\boldsymbol{\Delta}^{-1}(\boldsymbol{\mu}_y - \boldsymbol{\mu}) : y \in \Omega_Y\}$.

6.2.2. Relación con el modelo GPFC del trabajo Cook and Li, 2009

El modelo GPFC propuesto en **Cook and Li, 2009** y presentado en la Sección 5.3 es un caso particular del modelo propuesto en este capítulo. GPFC asume que cada X_{yj} de $\mathbf{X}_y$ pertenece a una familia exponencial a un parámetro con densidad dada por (5.4) y que las p componentes del vector aleatorio $\mathbf{X}_y = (X_{yj})$ son condicionalmente independientes dada la respuesta Y. La densidad (6.1), en este caso, tiene la forma

$$f(\mathbf{x}|\boldsymbol{\eta}_y, Y = y) = \prod_{j=1}^{p} e^{\eta_{yj}x_j - \psi_j(\eta_{yj})}h_j(x_j) = e^{\boldsymbol{\eta}_y^T\mathbf{X} - \psi(\boldsymbol{\eta}_y)}h(\mathbf{x}), \tag{6.8}$$

es decir que p es el número de parámetros naturales, los cuales se disponen en el vector de parámetros naturales $\boldsymbol{\eta}_y = (\eta_{y1}, \ldots, \eta_{yp})$, $\mathbf{T}(\mathbf{X}) = \mathbf{X}$ es el estadístico suficiente, $\psi(\boldsymbol{\eta}_y) = \sum_{j=1}^{p} \psi_j(\eta_{yj})$ y $h(\mathbf{x}) = h_1(x_1)\ldots h_p(x_p)$. Este modelo, si bien es sencillo, no permite incluir algunos ejemplos estudiados en la literatura de SDR. Por ejemplo, en mencionado trabajo consideran el caso especial de normales independientes $X_j|Y$ con media μ_{jY} y varianza constante σ_j^2,

$j = 1, \ldots, p$. Sin embargo, el caso de independencia condicional de los predictores con varianza que dependen de Y no puede incluirse en el enfoque del trabajo de **Cook and Li, 2009**. Para ver esto, observemos que si $X_j|Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, $j = 1, \ldots, p$ y X_j es independiente de X_k dado Y para $j \neq k$,

$$f(\mathbf{x}|y) = \prod_{j=1}^{p} f(x_j|y) = \frac{1}{(2\pi)^{p/2}} \frac{1}{\prod_{j=1}^{p} \sigma_{jy}} \exp\left(-\frac{1}{2}\sum_{j=1}^{p} \frac{\mu_{jj}^2}{\sigma_{jy}^2}\right) \exp\left(-\frac{1}{2}\sum_{j=1}^{p} \frac{x_j^2}{\sigma_{jy}^2} + \sum_{j=1}^{p} \frac{x_j \mu_{jy}}{\sigma_{jy}^2}\right).$$

Es decir que en este caso el vector de parámetros naturales es de longitud $2p$, conformado por $\boldsymbol{\eta}_y = (\mu_{1y}/\sigma_{y1}^2, \ldots, \mu_{py}/\sigma_{yp}^2, \sigma_{y1}^{-2}/2, \ldots, \sigma_{yp}^{-2}/2)$ y con estádistico suficiente $\mathbf{T}(\mathbf{X}) = (x_1, \ldots, x_p, x_1^2, \ldots, x_p^2)$. Por lo tanto, este modelo no puede adecuarse a (6.8).

Con el objetivo de demostrar que la metodología presentada por ellos puede ser todavía útil cuando los predictores son condicionalmente dependientes, consideran las familias exponenciales cuadráticas. Sin embargo consideran este caso con $\boldsymbol{\eta}_{Y,2} = \boldsymbol{\eta}_2$ en la Definición 6.2. Es decir que de acuerdo al Corolario 6.3, $R(\mathbf{X})$ se reduce solamente a $\boldsymbol{\alpha}_1^T\mathbf{X}$ y por lo tanto la reducción suficiente no contiene la parte cuadrática.

Para este modelo de familia exponencial cuadrática, proponen un estimador de $\boldsymbol{\alpha}_1$ basado en momentos. Ellos consideran las versiones muestrales de las varianzas $\boldsymbol{\Sigma}_s = \text{var}(\mathbf{X}|Y \in H_s)$ y $\mathbf{W}_s = \text{var}(\mathbf{f}_Y|Y \in H_s)$ donde el rango de Y se ha dividido en h slice H_s, $s = 1, \ldots, h$ y el estimador de máxima verosimilitud de $\boldsymbol{\Phi} = \boldsymbol{\Gamma}\boldsymbol{\beta}$ del modelo que asume la independencia condicional de los predictores (5.6). Luego, si h es grande, $\mathcal{S}_{\boldsymbol{\alpha}_1}$ puede aproximarse en la población por el span de los autovectores correspondientes a los autovalores no nulos de la matriz $\mathbf{M} = \mathbf{E}\left(\boldsymbol{\Sigma}_s^{-1}\mathbf{G}_s^{-1}\boldsymbol{\Phi}\mathbf{W}_s\boldsymbol{\Phi}^T\mathbf{G}_s^{-1}\boldsymbol{\Sigma}_s^{-1}\right)$. Luego, una estimador de $\mathcal{S}_{\boldsymbol{\alpha}_1}$ es el span de los primeros d autovectores de $\widehat{\mathbf{M}}$, la versión muestral consistente de $\mathbf{M}$. Un ejemplo de gran interés de familia exponencial cuadrática con $\boldsymbol{\eta}_{Y,2} = \boldsymbol{\eta}_2$, es la distribución normal multivariada con matriz de covarianza constante $\text{var}(\mathbf{X}|Y = y) = \boldsymbol{\Delta}$ independiente de y. Dicho modelo fue estudiado en detalle en **Cook and Forzani, 2008** donde deducen estimadores de máxima verosimilitud de la reducción suficiente minimal, por lo que esta metodología de momentos propuesta resulta obsoleta en este ejemplo.

El método propuesto en **Cook and Li, 2009** para obtener los estimadores de máxima verosimilitud del modelo (5.4) es el algoritmo de optimización sobre Grassmann, que se diferencia claramente de la metodología que presentaremos a continuación la cual resulta sencilla y de rápida aplicación. Además, esta metodología puede aplicarse para obtener estimadores de máxima verosimilitud para el modelo de familia exponencial cuadrática, para el cual **Cook and Li, 2009** solo proponen estimadores de momentos.

6.3. Estimaciones, test de dimensiones y propiedades asintóticas

6.3.1. Estimadores propuestos cuando d es conocido

Supongamos que contamos con una muestra aleatoria de tamaño n de $\mathbf{X} = (X_1, \ldots, X_p)^T$ e Y. Asumiendo que la muestra sigue el modelo (6.1) con $\boldsymbol{\eta}_{y_i}$ que satisface (6.3) y/o (6.2), el objetivo principal es estimar $\text{span}\{\boldsymbol{\eta}_Y - \overline{\boldsymbol{\eta}}, Y \in \Omega_Y\}$. Para esto proponemos dos enfoques, por un lado el propósito es estimar $\text{span}(\mathbf{A})$ y por otro lado, el $\text{span}(\mathbf{D})$ será requerido. En ambas situaciones, estimadores de máxima verosimilitud son usados. En el primer caso, podemos estimar los parámetros de interés usando el método iterativo propuesto en **Yee and Hastie, 2003** que extiende el algoritmo IRSL estándar o bien el algoritmo de minimización cuadrática que sabemos que proporciona estimadores tan eficientes como los de máxima verosimilitud. Ambos fueron presentados en el Capítulo 3. En el segundo caso, donde no se imponen restricciones en el rango de la matriz $\mathbf{D}$, el algoritmo IRLS usal presentado en el Capítulo 1 puede ser utilizado.

Si utilizamos el primer enfoque, necesitamos estimar $\overline{\boldsymbol{\eta}}$ y $\mathbf{AB}$. Así, una estimación de la base del subespacio $\text{span}\{\boldsymbol{\eta}_Y - \overline{\boldsymbol{\eta}}, Y \in \Omega_Y\}$ será $\widehat{\boldsymbol{\alpha}}_1 = \widehat{\mathbf{A}}$. En el segundo enfoque, luego de obtener los estimadores de máxima verosimilitud de $\overline{\boldsymbol{\eta}}$ y $\mathbf{D}$, los primeros d vectores singulares a izquierda de $\widehat{\mathbf{D}}$ o lo que es lo mismo, los primeros d autovectores de $\widehat{\mathbf{D}}\widehat{\mathbf{D}}^T$, pueden ser utilizados como una estimación de la base de $\{\boldsymbol{\eta}_Y - \overline{\boldsymbol{\eta}}, Y \in \Omega_Y\}$. Sea $\widehat{\boldsymbol{\alpha}}_2$ la matriz que contiene en sus columnas dichos vectores y denotamos como $\widehat{\boldsymbol{\alpha}}_2 = \left(\widehat{\mathbf{D}}\right)_{(d)}$.

Los algoritmos de estimación propuestos son rápidos, pues ambos son problemas de optimización cuadrática en cada iteración y son fáciles de implementar ya que solo requiere las derivadas de la función de máxima verosimilitud, que en el caso de familia exponenciales pueden ser calculadas sin calcular la función de máxima verosimilitud. Este puede ser aplicado en lugar de la optimización sobre *Grassmann* propuesta en **Cook and Li, 2009**. Por ejemplo, en las simulaciones presentadas en la Sección 6.4 hemos reproducidos de forma rápida los estimadores propuestos en **Cook and Li, 2009**, bajo la suposición de independencia condicional de los predictores.

6.3.2. Distribuciones asintóticas de los estimadores propuestos

Como nuestro interés no está centrado en $\widehat{\boldsymbol{\alpha}}_1 = \widehat{\mathbf{A}}$ y en $\widehat{\boldsymbol{\alpha}}_2 = \left(\widehat{\mathbf{D}}\right)_{(d)}$ sino en el subespacio que generan, estudiaremos las distribuciones asintóticas de las matrices de proyección $\mathbf{P}_{\widehat{\boldsymbol{\alpha}}_1}$ y $\mathbf{P}_{\widehat{\boldsymbol{\alpha}}_2}$.

Para ello, consideremos la siguiente descomposición en valores singulares de la matriz $\mathbf{D}$:

$$\mathbf{D} = \mathbf{U}^T \begin{pmatrix} \mathbf{\Lambda} & 0 \\ 0 & 0 \end{pmatrix} \mathbf{R}, \tag{6.9}$$

donde $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ es la matriz diagonal que contiene los valores singulares de $\mathbf{D}$, $\lambda_1 \geq \ldots \geq \lambda_d > 0$. La matriz ortogonal $\mathbf{U}^T = (\mathbf{U}_1, \mathbf{U}_0)$ es de orden $k \times k$ con $\mathbf{U}_1 : k \times d$ y $\mathbf{U}_0 : k \times (k-d)$, cumple que $\mathbf{U}_1\mathbf{U}_1^T + \mathbf{U}_0\mathbf{U}_0^T = \mathbf{I}_k$, $\mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I}_d$, $\mathbf{U}_0^T\mathbf{U}_0 = \mathbf{I}_{k-d}$ y $\mathbf{U}_0^T\mathbf{U}_1 = \mathbf{0}$. La matriz ortogonal $\mathbf{R}^T = (\mathbf{R}_1, \mathbf{R}_0)$ de orden $r \times r$ con $\mathbf{R}_1 : r \times d$ y $\mathbf{R}_0 : r \times (r-d)$ cumple, al igual que $\mathbf{U}$, con $\mathbf{R}_1\mathbf{R}_1^T + \mathbf{R}_0\mathbf{R}_0^T = \mathbf{I}_r$, $\mathbf{R}_1^T\mathbf{R}_1 = \mathbf{I}_d$, $\mathbf{R}_0^T\mathbf{R}_0 = \mathbf{I}_{r-d}$ y $\mathbf{R}_0^T\mathbf{R}_1 = \mathbf{0}$.

Sea $\widehat{\mathbf{D}}$ el estimador de máxima verosimilitud de rango completo de $\mathbf{D}$ obtenido a través del algoritmo IRLS estandar y su SVD:

$$\widehat{\mathbf{D}} = \widehat{\mathbf{U}}^T \begin{pmatrix} \widehat{\mathbf{\Lambda}}_1 & 0 \\ 0 & \widehat{\mathbf{\Lambda}}_0 \end{pmatrix} \widehat{\mathbf{R}} \tag{6.10}$$

con $\widehat{\mathbf{U}}^T = (\widehat{\mathbf{U}}_1, \widehat{\mathbf{U}}_0)$ y $\widehat{\mathbf{R}}^T = (\widehat{\mathbf{R}}_1, \widehat{\mathbf{R}}_0)$ donde las particiones son las mismas que la SVD de $\mathbf{D}$. Las matrices $\widehat{\mathbf{\Lambda}}_1 = \text{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_d)$ y $\widehat{\mathbf{\Lambda}}_0$ de dimensiones $d \times d$ y $(k-d) \times (r-d)$ respectivamente contienen en su diagonal principal los valores singulares $\widehat{\lambda}_1, \ldots, \hat{\lambda}_{\min(k,r)}$ de $\widehat{\mathbf{D}}$ en orden decreciente. Observar que $\widehat{\mathbf{\Lambda}}_0$ tiene la forma:

$$\widehat{\mathbf{\Lambda}}_0 = \begin{pmatrix} \widehat{\lambda}_{d+1} & & \\ & \cdots & \\ & & \widehat{\lambda}_r \\ & \mathbf{0} & \end{pmatrix}, \text{ si } k > r. \quad \widehat{\mathbf{\Lambda}}_0 = \begin{pmatrix} \widehat{\lambda}_{d+1} & & \\ & \cdots & \mathbf{0} \\ & & \widehat{\lambda}_k \end{pmatrix}, \text{ si } k < r.$$

Además consideremos la SVD de $\widehat{\mathbf{D}}_{RR}$ definido como $\widehat{\mathbf{D}}_{RR} = \widehat{\mathbf{A}}\widehat{\mathbf{B}}$, donde $\widehat{\mathbf{A}}$ y $\widehat{\mathbf{B}}$ son los estimadores de máxima verosimilitud de $\mathbf{A}$ y $\mathbf{B}$ obtenidos vía el algoritmo en [**Yee and Hastie, 2003**]. Como $\widehat{\mathbf{D}}_{RR}$ fue construido tal que tenga rango d, entonces

$$\widehat{\mathbf{D}}_{RR} = \widehat{\mathbf{U}}_{RR}^T \begin{pmatrix} \widehat{\mathbf{\Lambda}}_{RR1} & 0 \\ 0 & 0 \end{pmatrix} \widehat{\mathbf{R}}_{RR} \tag{6.11}$$

con $\widehat{\mathbf{U}}_{RR}^T = (\widehat{\mathbf{U}}_{RR1}, \widehat{\mathbf{U}}_{RR0})$ y $\widehat{\mathbf{R}}_{RR}^T = (\widehat{\mathbf{R}}_{RR1}, \widehat{\mathbf{R}}_{RR0})$ donde las particiones son las mismas que la SVD de $\mathbf{D}$.

Estas descomposiciones en valores singulares de $\mathbf{D}$, $\widehat{\mathbf{D}}$ y $\widehat{\mathbf{D}}_{RR}$ fueron expresadas detalladamente con el objetivo de analizar las matrices de proyección sobre los subespacios generados por los estimadores propuestos. La siguiente proposición resume las distribuciones asintóticas de estas proyecciones.

Proposición 6.5. *Dado los estimadores de máxima verosimilitud* $\widehat{\mathbf{D}}$ *y* $\widehat{\mathbf{D}}_{RR} = \widehat{\mathbf{A}}\widehat{\mathbf{B}}$ *del modelo (6.1) con* $\boldsymbol{\eta}_Y$ *que satisface (6.2) y (6.3) respectivamente, las distribuciones asintóticas de las matrices de proyección* $\mathbf{P}_{\left(\widehat{\mathbf{D}}\right)_{(d)}}$ *y* $\mathbf{P}_{\widehat{\mathbf{A}}}$ *son las siguientes:*

$$\sqrt{n}(\mathbf{P}_{\left(\widehat{\mathbf{D}}\right)_{(d)}} - \mathbf{P}_{\mathbf{D}}) \to \mathcal{N}(0, \mathbf{W}_1)$$
$$\sqrt{n}(\mathbf{P}_{\widehat{\mathbf{A}}} - \mathbf{P}_{\mathbf{D}}) \to \mathcal{N}(0, \mathbf{W}_2)$$

con

$$\mathbf{W}_1 = (\mathbf{I}_{k^2} + \mathbf{K}_{kk})((\mathbf{AB})^\dagger \otimes \mathbf{Q}_{\mathbf{A}})^T \mathbf{V}_{\mathbf{D}}((\mathbf{AB})^\dagger \otimes \mathbf{Q}_{\mathbf{A}})(\mathbf{I}_{k^2} + \mathbf{K}_{kk}) \tag{6.12}$$
$$\mathbf{W}_2 = (\mathbf{I}_{k^2} + \mathbf{K}_{kk})((\mathbf{AB})^\dagger \otimes \mathbf{Q}_{\mathbf{A}})^T \mathbf{V}_{\mathbf{D}}^{red}((\mathbf{AB})^\dagger \otimes \mathbf{Q}_{\mathbf{A}})(\mathbf{I}_{k^2} + \mathbf{K}_{kk}), \tag{6.13}$$

donde $\mathbf{V}_{\mathbf{D}}$ *es la varianza asintótica del estimador de rango completo* $\widehat{\mathbf{D}}$ *y* $\mathbf{V}_{\mathbf{D}}^{red}$ *es la varianza asintótica del estimador de rango reducido reducido* $\widehat{\mathbf{D}}_{RR}$ *dada por* $\mathbf{V}_{\mathbf{D}}^{red} = \boldsymbol{\Delta}(\boldsymbol{\Delta}^T \mathbf{V}_{\mathbf{D}}^{-1} \boldsymbol{\Delta})^\dagger \boldsymbol{\Delta}^T$ *con* $\boldsymbol{\Delta} = \begin{pmatrix} \mathbf{B}^T \otimes \mathbf{I}_k & \mathbf{I}_r \otimes \mathbf{A} \end{pmatrix}$.

Los siguientes lemas previos serán usados en la demostración de la Proposición 6.5.

Lema 6.6. *Sean las matrices* $\Omega_n : k \times r$ *y* $\Omega : k \times r$ *y consideremos sus descomposición en valores singulares como en (6.10) y (6.9) respectivamente. Supongamos que* Ω_n *es asintóticamente normal:*

$$\sqrt{n}\mathrm{vec}(\Omega_n - \Omega) \to \mathcal{N}(0, \mathbf{V}).$$

Luego para $\widehat{\mathbf{H}} = \widehat{\mathbf{U}}_1 \widehat{\boldsymbol{\Lambda}}_1 \widehat{\mathbf{R}}_1^T \mathbf{R}_1 \boldsymbol{\Lambda}^{-1}$, *se tiene que*

$$\sqrt{n}\mathrm{vec}(\widehat{\mathbf{H}} - \mathbf{U}_1) \to \mathcal{N}(0, (\boldsymbol{\Lambda}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_k)\mathbf{V}(\mathbf{R}_1\boldsymbol{\Lambda}^{-1} \otimes \mathbf{I}_k)).$$

Demostración. Como $\Omega_n = \widehat{\mathbf{U}}_1 \widehat{\boldsymbol{\Lambda}}_1 \widehat{\mathbf{R}}_1^T + \widehat{\mathbf{U}}_0 \widehat{\boldsymbol{\Lambda}}_0 \widehat{\mathbf{R}}_0^T$,

$$\begin{aligned}\sqrt{n}(\boldsymbol{\Lambda}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_k)\mathrm{vec}(\Omega_n - \Omega) &= \sqrt{n}\mathrm{vec}(\widehat{\mathbf{U}}_1 \widehat{\boldsymbol{\Lambda}}_1 \widehat{\mathbf{R}}_1^T \mathbf{R}_1 \boldsymbol{\Lambda}^{-1} + \widehat{\mathbf{U}}_0 \widehat{\boldsymbol{\Lambda}}_0 \widehat{\mathbf{R}}_0^T \mathbf{R}_1 \boldsymbol{\Lambda}^{-1} - \mathbf{U}_1 \boldsymbol{\Lambda} \mathbf{R}_1^T \mathbf{R}_1 \boldsymbol{\Lambda}^{-1}) \\ &= \sqrt{n}\mathrm{vec}(\widehat{\mathbf{H}} - \mathbf{U}_1 + \widehat{\mathbf{U}}_0 \widehat{\boldsymbol{\Lambda}}_0 \widehat{\mathbf{R}}_0^T \mathbf{R}_1 \boldsymbol{\Lambda}^{-1}).\end{aligned}$$

Entonces

$$\sqrt{n}\mathrm{vec}(\widehat{\mathbf{H}} - \mathbf{U}_1) + \sqrt{n}\mathrm{vec}(\widehat{\mathbf{U}}_0 \widehat{\boldsymbol{\Lambda}}_0 \widehat{\mathbf{R}}_0^T \mathbf{R}_1 \boldsymbol{\Lambda}^{-1}) \to \mathcal{N}\left(0, \boldsymbol{\Sigma}_{\mathbf{U}}\right),$$

donde $\mathbf{\Sigma_U} = (\mathbf{\Lambda}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_k)\mathbf{V}(\mathbf{R}_1\mathbf{\Lambda}^{-1} \otimes \mathbf{I}_k)$. Pero como $\sqrt{n}\left(\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\right) = O_p(1)$ y $\mathbf{P}_{\mathbf{R}_1} = \left(\mathbf{P}_{\widehat{\mathbf{R}}_1} + O_p(n^{-1/2})\right)$, se tiene

$$\begin{aligned}\sqrt{n}(\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\mathbf{R}_1\mathbf{\Lambda}^{-1}) &= \sqrt{n}(\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T)\mathbf{P}_{\mathbf{R}_1}\mathbf{R}_1\mathbf{\Lambda}^{-1} \\ &= \sqrt{n}\left(\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\right)\left(\mathbf{P}_{\widehat{\mathbf{R}}_1} + O_p(n^{-1/2})\right)\mathbf{R}_1\mathbf{\Lambda}^{-1} \\ &= \sqrt{n}\left(\widehat{\mathbf{U}}_0\widehat{\mathbf{\Lambda}}_0\widehat{\mathbf{R}}_0^T\right)O_p(n^{-1/2})\mathbf{R}_1\mathbf{\Lambda}^{-1} = O_p(n^{-1/2})\end{aligned}$$

donde usamos que $\widehat{\mathbf{R}}_0^T\widehat{\mathbf{R}}_1 = \mathbf{0}$. Entonces $\sqrt{n}\text{vec}(\widehat{\mathbf{U}}_0\widehat{\mathbf{D}}_0\widehat{\mathbf{R}}_0^T\mathbf{R}_1\mathbf{\Lambda}^{-1}) \to \mathbf{0}$ en probabilidad y se obtiene el resultado. □

Lema 6.7. *Sean las matrices $\widehat{\mathbf{U}}$ y $\mathbf{U}$ de dimensiones: $k \times d$ con $d \leq k$ y $\mathbf{U}$ de rango completo d. Supongamos que $\widehat{\mathbf{U}}$ es asintóticamente normal:*

$$\sqrt{n}\text{vec}(\widehat{\mathbf{U}} - \mathbf{U}) \to \mathcal{N}(0, \mathbf{V}).$$

Luego, $\sqrt{n}\text{vec}(\mathbf{P}_{\widehat{\mathbf{U}}} - \mathbf{P}_{\mathbf{U}})$ converge a una distribución normal con media 0 y matriz de covarianza

$$(\mathbf{I}_{k^2} + \mathbf{K}_{kk})(\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1} \otimes \mathbf{Q}_{\mathbf{U}})\mathbf{V}((\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T \otimes \mathbf{Q}_{\mathbf{U}})(\mathbf{I}_{k^2} + \mathbf{K}_{kk}).$$

La demostración del Lema 6.7 se encuentra al final del capítulo.

Demostración de la Proposición 6.5: De acuerdo a la descomposición en valores singulares (6.10) y (6.11) de $\widehat{\mathbf{D}}$ y $\widehat{\mathbf{D}}_{RR}$ respectivamente, tenemos que $\left(\widehat{\mathbf{D}}\right)_{(d)} = \widehat{\mathbf{U}}_1$ pues fue definido como los primeros d autovectores de $\widehat{\mathbf{D}}\widehat{\mathbf{D}}^T$ y además $\text{span}(\mathbf{A}) = \text{span}(\widehat{\mathbf{U}}_{RR1})$. Por lo tanto, estudiar las distribuciones asintóticas de $\mathbf{P}_{\left(\widehat{\mathbf{D}}\right)_{(d)}}$ y de $\mathbf{P}_{\widehat{\mathbf{A}}}$ es lo mismo que estudiar las distribuciones asintóticas de $\mathbf{P}_{\widehat{\mathbf{U}}_1}$ y de $\mathbf{P}_{\widehat{\mathbf{U}}_{RR1}}$ respectivamente. Por otro lado, si definimos $\widehat{\mathbf{H}} = \widehat{\mathbf{U}}_1\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{R}}_1^T\mathbf{R}_1\mathbf{\Lambda}^{-1}$ y analogamente $\widehat{\mathbf{H}}_{RR} = \widehat{\mathbf{U}}_{RR1}\widehat{\mathbf{\Lambda}}_{RR1}\widehat{\mathbf{R}}_{RR1}^T\mathbf{R}_1\mathbf{\Lambda}^{-1}$ como en el Lema 6.6, las distribuciones asintóticas de $\widehat{\mathbf{H}}$ y de $\widehat{\mathbf{H}}_{RR}$ son

$$\begin{aligned}\sqrt{n}\text{vec}\left(\widehat{\mathbf{H}} - \mathbf{U}_1\right) &\to \mathcal{N}(0, \mathbf{\Sigma_U}) \text{ donde } \mathbf{\Sigma_U} = (\mathbf{\Lambda}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_k)\mathbf{V_D}(\mathbf{R}_1\mathbf{\Lambda}^{-1} \otimes \mathbf{I}_k) \\ \sqrt{n}\text{vec}\left(\widehat{\mathbf{H}}_{RR} - \mathbf{U}_1\right) &\to \mathcal{N}\left(0, \mathbf{\Sigma}_{\mathbf{U}}^{red}\right) \text{ donde } \mathbf{\Sigma}_{\mathbf{U}}^{red} = (\mathbf{\Lambda}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_k)\mathbf{V}_D^{red}(\mathbf{R}_1\mathbf{\Lambda}^{-1} \otimes \mathbf{I}_k).\end{aligned}$$

Además, observar que las matrices $\mathbf{\Lambda}$, $\widehat{\mathbf{\Lambda}}_1$, $\widehat{\mathbf{\Lambda}}_{RR1}$, $\widehat{\mathbf{R}}_1^T\mathbf{R}_1$ y $\widehat{\mathbf{R}}_{RR1}^T\mathbf{R}_1$ son matrices $d \times d$ invertibles. Luego tenemos que

$$\begin{aligned}\mathbf{P}_{\widehat{\mathbf{H}}} &= \mathbf{P}_{\widehat{\mathbf{U}}_1} \\ \mathbf{P}_{\widehat{\mathbf{H}}_{RR}} &= \mathbf{P}_{\widehat{\mathbf{U}}_{RR1}}.\end{aligned}$$

Por lo tanto, $\sqrt{n}(\mathbf{P}_{\widehat{\mathbf{H}}} - \mathbf{P}_{\mathbf{U}_1})$ tiene la misma distribución asintótica que $\sqrt{n}(\mathbf{P}_{\widehat{\mathbf{U}}_1} - \mathbf{P}_{\mathbf{U}_1})$ y $\sqrt{n}(\mathbf{P}_{\widehat{\mathbf{H}}_{RR}} - \mathbf{P}_{\mathbf{U}_1})$ tiene la misma distribución asintótica que $\sqrt{n}(\mathbf{P}_{\widehat{\mathbf{U}}_{RR1}} - \mathbf{P}_{\mathbf{U}_1})$. Aplicando el Lema 6.7 obtenemos que $\mathbf{P}_{\widehat{\mathbf{U}}_1}$ y $\mathbf{P}_{\widehat{\mathbf{U}}_{RR1}}$ son asintóticamente normales con las siguientes matrices de covarianzas

$$\begin{aligned}\mathbf{W}_1 &= (\mathbf{I}_{k^2} + \mathbf{K}_{kk})(\mathbf{U}_1\boldsymbol{\Lambda}^{-1}\mathbf{R}_1^T \otimes \mathbf{Q}_{\mathbf{U}_1})\mathbf{V}_{\mathbf{D}}(\mathbf{R}_1\boldsymbol{\Lambda}^{-1}\mathbf{U}_1^T \otimes \mathbf{Q}_{\mathbf{U}_1})(\mathbf{I}_{k^2} + \mathbf{K}_{kk}) \\ \mathbf{W}_2 &= (\mathbf{I}_{k^2} + \mathbf{K}_{kk})(\mathbf{U}_1\boldsymbol{\Lambda}^{-1}\mathbf{R}_1^T \otimes \mathbf{Q}_{\mathbf{U}_1})\mathbf{V}_{\mathbf{D}}^{red}(\mathbf{R}_1\boldsymbol{\Lambda}^{-1}\mathbf{U}_1^T \otimes \mathbf{Q}_{\mathbf{U}_1})(\mathbf{I}_{k^2} + \mathbf{K}_{kk}).\end{aligned}$$

Como $\mathbf{D}_2 = \mathbf{AB} = \mathbf{U}_1\boldsymbol{\Lambda}_1\mathbf{R}_1^T$ tenemos que $(\mathbf{AB})^\dagger = \mathbf{R}_1\boldsymbol{\Lambda}_1^{-1}\mathbf{U}_1^T$ y $\mathbf{Q}_{\mathbf{U}_1} = \mathbf{I}_r - \mathbf{P}_{\mathbf{U}_1} = \mathbf{I}_r - \mathbf{P}_{\mathbf{A}} = \mathbf{Q}_{\mathbf{A}}$. Remplazando las expresiones correspondientes llegamos a (6.12) y (6.13).

□

Observación 6.8. Notar que entre las matrices (6.12) y (6.13) existe la siguiente relación

$$\mathbf{W}_2 \leq \mathbf{W}_1.$$

En efecto, sabemos que $\mathbf{V}_{\mathbf{D}}^{red} \leq \mathbf{V}_{\mathbf{D}}$ y solo basta aplicar la Propiedad A.11.

6.3.3. Test asintóticos para la dimensión d

La dimensión de la reducción suficiente para la regresión de Y en $\mathbf{X}$ es $d = \dim(\text{span}\{\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}, Y \in \Omega_Y\})$, donde la regresión inversa de los predictores dada la respuesta $\mathbf{X}|Y$ es ajustada teniendo en cuenta que la distribución de $\mathbf{X}|Y$ pertenece a una familia exponencial dada por (6.1) y suponemos el modelo (6.2) para los parámetros naturales de la familia $\boldsymbol{\eta}_Y$.

Cuando utilizamos el segundo enfoque de estimación es necesario estimar el rango d de la matriz de coeficientes $\mathbf{D}$ o equivalentemente, la dimensión de la reducción suficiente para la regresión $Y|\mathbf{X}$. Para ello contamos con los test asintóticos que ya hemos presentado en el Capítulo 3. Estos test, propuestos en [**Bura and Yang, 2011**], requieren que el estimador con el que se cuenta sea asintóticamente normal y el conocimiento de su varianza asintótica.

En la Sección 1.3.3 hemos estudiado las propiedades asintóticas necesarias. Recordamos que en este caso tenemos el siguiente modelo más simple

$$\begin{aligned}\boldsymbol{\eta}_Y &= \overline{\boldsymbol{\eta}} + \mathbf{D}\mathbf{f}_Y = (\overline{\boldsymbol{\eta}}, \mathbf{D})(1, \mathbf{f}_Y^T)^T \\ &= (\widetilde{\mathbf{f}}_Y^T \otimes \mathbf{I}_k)\text{vec}\,((\overline{\boldsymbol{\eta}}, \mathbf{D})) = \mathbf{F}\boldsymbol{\Gamma}\end{aligned}$$

donde $\boldsymbol{\Gamma} = \text{vec}(\overline{\boldsymbol{\eta}}, \mathbf{D}) : k(r+1) \times 1$, $\widetilde{\mathbf{f}}_Y^T = (1, \mathbf{f}_Y^T)^T$ y $\mathbf{F} = (\widetilde{\mathbf{f}}_Y^T \otimes \mathbf{I}_k)$. Como consecuencia de la Proposición 1.2, $\sqrt{n}\,\text{vec}(\widehat{\mathbf{D}} - \mathbf{D}) \to \mathcal{N}(\mathbf{0}, \mathbf{V}_{\mathbf{D}})$ con $\mathbf{V}_{\mathbf{D}} = (\mathbf{M} \otimes \mathbf{I}_k)\mathbf{V}_{\boldsymbol{\Gamma}}(\mathbf{M}^T \otimes \mathbf{I}_k)$ y $\mathbf{M} = (0, \mathbf{I}_r)$. Por lo tanto, $\widehat{\mathbf{D}}$ es asintóticamente normal y podemos aplicar el test chi-cuadrado ponderado

asintótico o el test chi-cuadrado asintótico de Wald basados en los valores singulares más chicos de $\widehat{\mathbf{D}}$ desarrollados en **[Bura and Yang, 2011]** y explicados en detalle en la Sección 3.5 del Capítulo 3.

Cuando utilizamos el primer enfoque, en cambio, se propone estimar $\overline{\boldsymbol{\eta}}$ y $\mathbf{AB}$ para $d = 1, \ldots, \text{mín}(k, r)$ y comparar los modelos por ejemplo a través de AIC, BIC o LRT. Además, podemos directamente utilizar el valor de $\widehat{d}$ obtenido en el primer enfoque.

6.4. Conexión con otros métodos SDR para familias exponenciales

En esta sección ilustramos las conexiones, diferencias y ventajas de nuestro enfoque EF-DR con los métodos presentados en el Capítulo 5, usando el conjunto de datos Atletas de Autralia discutidos por **[Cook and Weisberg, 1994]** y **[Cook and Weisberg, 1999]** y analizados en **[Chiaromonte et al., 2002]**:

6.4.1. Datos: Atletas de Australia

Varios estudios se han llevado a cabo con el objetivo de investigar la relación entre el índice de masa corporal y varios predictores e indentificar los factores que están asociados al sobrepeso. El índice de masa corporal LBM es regresado en los logaritmos de las variables altura, peso, cantidad de glóbulos rojos, cantidad de glóbulos blancos y hemoglobina más una variable indicadora para el género. La muestra consta de 202 observaciones correspondientes a atletas de un instituto de deportes de Australia. Es decir, se tiene 6 variables predictoras de las cuales 5 son continuas, que representamos con $\mathbf{X}$ y una es categórica (el género) que se denotará como W con $W = 1$ para mujeres y $W = 0$ para los hombres. En **[Chiaromonte et al., 2002]** concluyen que una combinación lineal de los predictores continuos junto con la variable género W son suficientes para describir la regresión de Y en $(\mathbf{X}, W)$. La reducción en $\mathbf{X}$ fue estimada con la primera dirección obtenida con el método Sir Parcial, $\widehat{\boldsymbol{\zeta}}^T\mathbf{X}$. El gráfico de Y versus $\widehat{\boldsymbol{\zeta}}^T\mathbf{X}$ es dado en la Figura 1 con los dos géneros identificados por diferentes símbolos. En base a este gráfico, ellos infieren un modelo para la regresión directa y suponen que la relación entre Y y $\mathbf{X}$ es lineal en la primera dirección estimada por Sir Parcial pero estiman dos rectas con diferente perdiente y ordenada al origen, una para cada género. Para comparar con la metodología GPFC de **[Cook and Li, 2009]**, asumimos que $(\mathbf{X}, W)|Y$ tiene distribución perteneciente a una familia exponencial con todas las componentes de $\mathbf{X}$ y W independientes, dado Y. Esto es, $f_{(\mathbf{X},W)|Y} = f_{\mathbf{X}|Y} f_{W|Y} = f_{W|Y} \prod_{i=1}^{p} f_{X_i|Y}$. Consultando los gráficos marginales de X_j vs. Y, parece razonable

suponer un modelo de distribución normal para $X_j|Y$, $j = 1, \ldots, p$. Se asume que las varianzas no dependen de Y pero no son necesariamente iguales para todo j. Además, suponemos que $W|Y$ es una variable Bernoulli con $f_{W=w|Y=y} = p_y^w(1-p_y)^{1-w}$, $w = 0, 1$. En la Figura 2, graficamos Y versus las primeras dos reducciones suficientes obtenidas bajo este modelo. El gráfico de la primera dirección luce muy similar a la primera dirección obtenida por Sir Parcial pero notar que en este caso la reducción es de la forma $R(\mathbf{X}, W)$ y no $(R(\mathbf{X}), W)$ como en Sir Parcial. Los gráficos sugieren que necesitamos ambas direcciones de la reducción para modelar Y. Este resultado también es apoyado por ambos tests asintóticos de la dimensión, el test chi-cuadrado ponderado arrojó que la dimensión es 2, a nivel de 0.05 con un p-valor de 0.02 y el test chi-cuadrado estimó que la dimensión es mayor que 2, con un p-valor muy pequeño. Sin embargo, se sabe que este último test es muy conservador por lo que normalmente utilizamos el test chi-cuadrado ponderado para estimar d. Como se trata de un problema de dos dimensiones y es difícil de visualizar la relación entre la respuesta y las dos reducciones, simplemente ajustamos un modelo lineal en las dos reducciones y reportamos los resultados de este modelo en la siguiente sección, donde se compara la exactitud de la predicción.

Por último, asumimos que la distribución conjunta de $(\mathbf{X}, W)|Y$ pertenece a una familia exponencial más general. Observemos que la densidad condicional conjunta de los predictores $f_{(\mathbf{X},W)|Y}$ puede factorizarse de la siguiente forma

$$f_{(\mathbf{X},W|Y=y)}(\mathbf{x}, w|Y = y) = f_{(\mathbf{X}|W=w,Y=y)}(\mathbf{x}|W = w, Y = y) \cdot f_{(W|Y=y)}(w|Y = y).$$

Luego, modelamos $W|Y$ como Bernoulli(p_Y) y basándonos en los gráficos de $\mathbf{X}$ versus Y para los dos géneros, modelamos $\mathbf{X}|(Y, W)$ como una normal multivariada con varianza constante. Además, dichos gráficos indican que $\mathbf{E}(\mathbf{X}|(W, Y))$ depende de Y en forma lineal para cada sexo, i.e. $\mathbf{E}(\mathbf{X}|(W, Y)) = \boldsymbol{\mu}_\mathbf{X} + \mathbf{b}_1(f_Y - \bar{f}_Y) + \mathbf{b}_2(W - \mu_W)$, con $f_Y = Y$, $\bar{f}_Y = \mathbf{E}(f_Y)$ y $\mu_W = \mathbf{E}(W)$. Por lo tanto, modelamos la distribución condicional conjunta de $(\mathbf{X}, W)$ dado Y como

$$\begin{aligned} f(\mathbf{x}, w|y) = & \\ & \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_\mathbf{x} - \mathbf{b}_1(f_y - \bar{f}_y) - \mathbf{b}_2(w - \mu_w))^T \boldsymbol{\Delta}^{-1}(\mathbf{x} - \mu_\mathbf{x} - \mathbf{b}_1(f_y - \bar{f}_y) - \mathbf{b}_2(w - \mu_w)\right) \\ & \times \exp\left((w - \mu_w)\log\frac{p_y}{1-p_y} + \log(1-p_y) + \mu_w \log\frac{p_y}{1-p_y}\right)\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Delta}|^{1/2}} \\ = & \, e^{\boldsymbol{\eta}_y^T \mathbf{T}(\mathbf{x},w) - \psi(\boldsymbol{\eta}_y)} h(\mathbf{x}, w). \end{aligned} \tag{6.14}$$

La densidad (6.14) pertenece a una familia exponencial con

$$\begin{aligned}
\mathbf{T}(\mathbf{x}, w) &= (\mathbf{x} - \mu_{\mathbf{x}}, w - \mu_w) \\
\boldsymbol{\eta}_y &= (\boldsymbol{\eta}_1, \eta_2) = \left(\boldsymbol{\Delta}^{-1}\mathbf{b}_1(f - \bar{f}_y), \log \frac{p_y}{1-p_y} - \mathbf{b}_2^T \boldsymbol{\Delta}^{-1}\mathbf{b}_1(f - \bar{f}_y) \right) \\
\psi(\boldsymbol{\eta}_y) &= \frac{1}{2}\left(\mathbf{b}_1\left(f_y - \bar{f}_y\right)\right)^T \boldsymbol{\Delta}^{-1}\left(\mathbf{b}_1\left(f_y - \bar{f}_y\right)\right) - \log(1-p_y) - \mu_g \log \frac{p_y}{1-p_y} \\
h(\mathbf{x}, w) &= \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Delta}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})^T \boldsymbol{\Delta}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}}) - \frac{1}{2}(w-\mu_w)^2 \mathbf{b}_2^T \boldsymbol{\Delta}^{-1}\mathbf{b}_2 + (w-\mu_w)\mathbf{b}_2^T \boldsymbol{\Delta}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})}
\end{aligned}$$

donde $\boldsymbol{\eta}_1 = \boldsymbol{\Delta}^{-1}\mathbf{b}_1(f - \bar{f}_y)$ es $p \times 1$ y $\eta_2 = \log(p_y/(1-p_y)) - \mathbf{b}_2^T \boldsymbol{\Delta}^{-1}\mathbf{b}_1(f_y - \bar{f}_y)$ es un escalar. Si $\boldsymbol{\alpha} = \text{span}\{\boldsymbol{\eta}_Y - \overline{\boldsymbol{\eta}} : Y \in \Omega_Y\}^T$, $\boldsymbol{\alpha}^T\mathbf{T}(\mathbf{X}, W)$ es la reducción suficiente minimal de Y en $\mathbf{X}$ y W. El algoritmo para estimar la reducción suficiente está en la Sección 6.6 del Apéndice. Mediante los test dados en la Sección 6.3.3, la dimensión estimada con ambos test fue 1. El test chi-cuadrado ponderado arrojó un p-valor 0 para el primer test $d = 0$ versus $d \geq 1$ y 0.83 para el segundo test $d = 1$ versus $d = 2$. El test chi-cuadrado arrojó un p-valor 0 para el primer test $d = 0$ versus $d = 1$ y 0.053 para el segundo test $d = 1$ versus $d \geq 2$. En la Figura 3, la respuesta LBM es graficada versus la reducción suficiente estimada por EF-DR. La figura indica que una función cuadrática en $R(\mathbf{X}, W)$ puede ser ajustada para predecir LBM a partir de los predictores, tanto cuantitativos como cualitativos sin discriminar por sexo. Por lo tanto, el modelo de regresión para predecir LBM es

$$E(Y|\mathbf{X}, W) = E(Y|R(\mathbf{X}, W)) = \gamma_1 + \gamma_2 R(\mathbf{X}, W) + \gamma_3 R^2(\mathbf{X}, W).$$

Este resultado contrasta el análisis de los métodos Sir Parcial y el de familias exponenciales independientes de Cook y Li, que estiman que la dimensión de la reducción es 2. Teniendo en cuenta la estructura de dependencia, EF-DR no sólo proporciona una caracterización más exacta de la relación entre la respuesta y los predictores, sino que también reduce la complejidad de la regresión.

Además aplicamos KSIR, KDR y COIR al conjunto de datos de los atletas autralianos. No se han propuesto test de dimensión para estos métodos, sin embargo **Yeh et al., 2009** ordena la importancia de los KSIR predictores de acuerdo a sus autovalores asociados y sugiere usar la primera o la dos primeras direcciones de acuerdo a una inspección empírica de los autovalores. En **Fukumizu et al., 2009** no considera el problema de inferir acerca de la dimensión de la reducción KDR obtenida pues este método frecuentamente se usa para una exploración gráfica de los datos, donde el analista tal vez desee explorar vistas de diferentes dimensionalidades. En las Figuras 4, 5 y 6 graficamos la respuesta versus las primeras KSIR, KDR y COIR direcciones

aplicado a los predictores continuos. Para estos métodos kernel fue usado el núcleo gaussiano, para el cual se eligió el ancho de banda a través de validación cruzada. Basándonos en estas figuras, modelamos Y como en el caso de Sir Parcial

$$E(Y|\mathbf{X}, W) = E(Y|R(\mathbf{X}), W) = (\alpha_1 + \alpha_2 I(W = 1)) + (\beta_1 + \beta_2 I(W = 1))R(\mathbf{X}),$$

donde I es la función indicadora, y $R(\mathbf{X})$ es la reducción para cada método.

En el Cuadro 1 reportamos el vector $\widehat{\boldsymbol{\alpha}}$ de la reducción para los métodos aplicados: para Sir Parcial y KDR, la reducción es $\widehat{\boldsymbol{\alpha}}^T\mathbf{X}$ y para EF-DR y GPFC, es $\widehat{\boldsymbol{\alpha}}^T(\mathbf{X}, W)$. En el Cuadro 2 reportamos el error de predicción $\sum_{i=1}^{n}(y_i - \widehat{y}_{(-i)})^2/n$, donde $\widehat{y}_{(-i)}$ indica la predicción de la i-ésima respuesta usando todos los datos excepto la i-ésima observación. EF-DR tiene el mejor desempeño con respecto al error de predicción, seguido por Sir Parcial, KDR, KSIR y COIR. Esta comparación confirma lo expuesto en [**van der Maaten and van den Herik, 2009**], que afirma que en general los métodos kernel tienen un funcionamiento inferior que los métodos de reducción de dimensiones globales, basado en momentos o cuando se cuenta con un modelo.

Para examinar cómo la estructura de dependencia afecta al rendimiento de EF-DR, llevamos a cabo simulaciones adicionales utilizando la misma estructura de los datos de los atletas australianos variando la correlación de $\mathbf{X}|(Y, W)$. Es decir, hemos generado una respuesta normal, predictores continuos y un predictor binario que satisfacen (6.14). Los resultados concuerdan con nuestras observaciones en el ejemplo de atletas de Australia. EF-DR mostró en general un rendimiento superior para las diferentes estructuras de dependencia, tanto en estimación como en predicción. Un comentario aparte es que el costo computacional para la aplicación de los métodos kernel, el cual fue superior en comparación con EF-DR: el tiempo necesario para KSIR o KDR varió de 150 a 1.000 veces el de EF-DR. No se incluyó COIR en los cálculos, ya que en general, dió peores resultados y alto costo computacional.

También estudiamos que sucede cuando aumenta el tamaño de la muestra y el número de predictores, utilizando nuevamente la estructura (6.14). Como era de esperar, el aumento de tamaño de las muestras dan como resultado un mejor rendimiento. Además, cuanto mayor es la correlación entre los predictores, EF-DR realiza un mejor desempeño a través de los diferentes tamaños de muestra y número de predictores. Estos resultados se encuentran al final del capítulo.

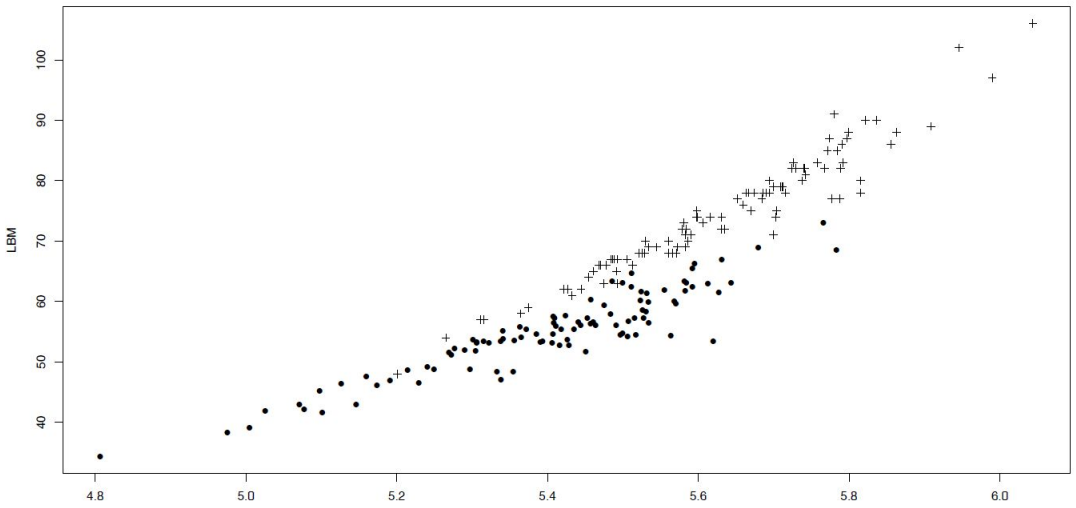


FIGURA 1. LBM versus la reducción suficiente estimada por SIR Parcial (Puntos para sexo femenino, cruces para sexo masculino).

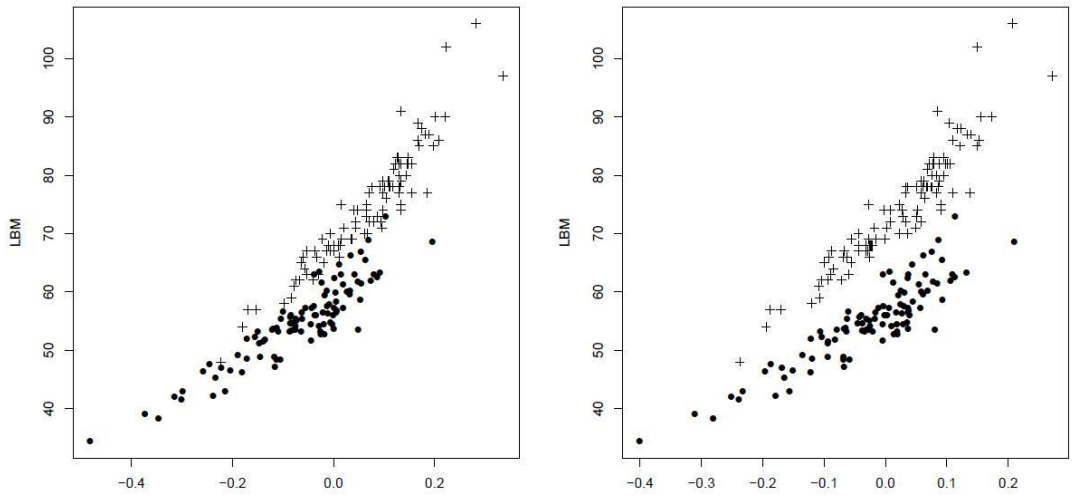


FIGURA 2. LBM versus las dos primeras reducciónes suficientes obtenidas por GPFC (Puntos para sexo femenino, cruces para sexo masculino).

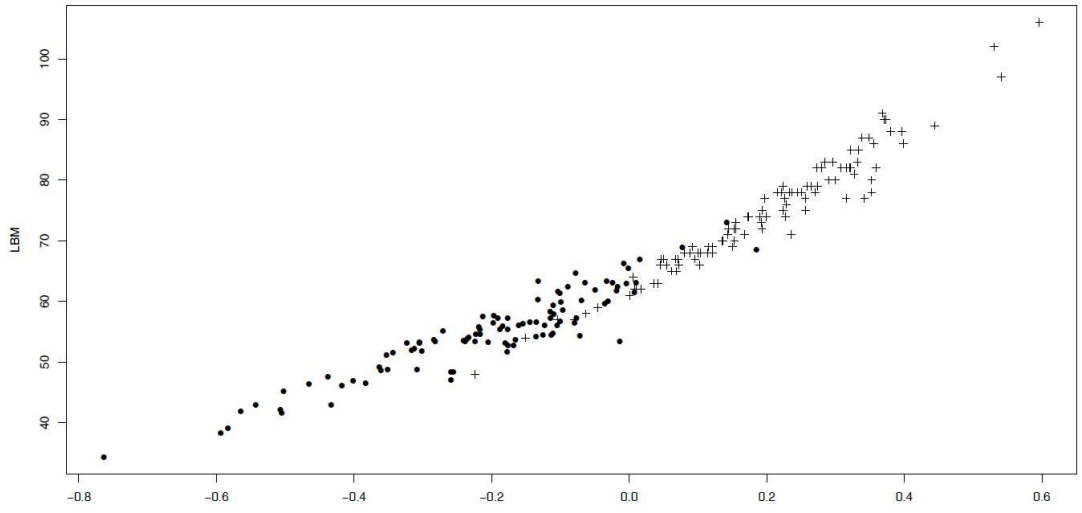


FIGURA 3. LBM versus la reducción suficiente estimada por el método EF-DR (Puntos para sexo femenino, cruces para sexo masculino).

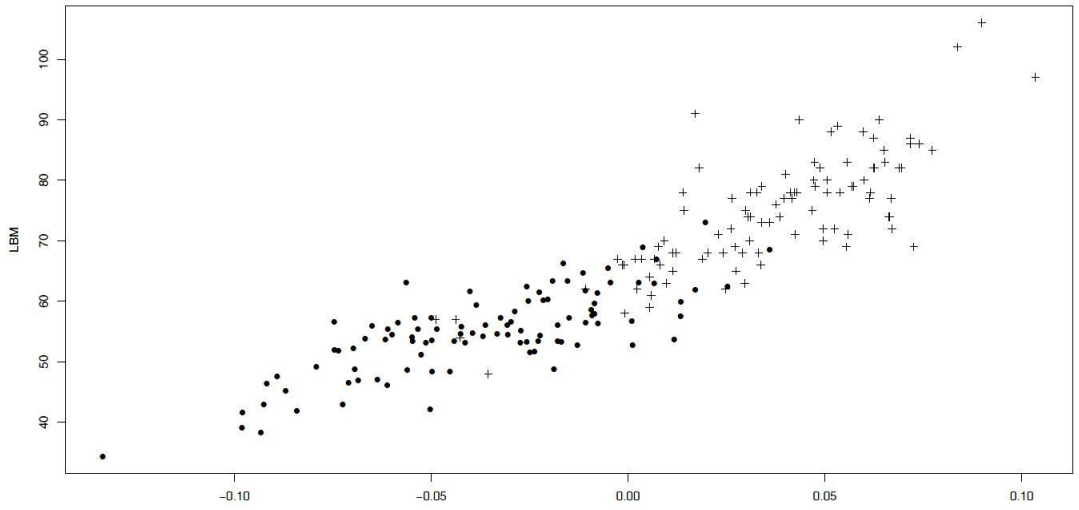


FIGURA 4. LBM versus el primer predictor KSIR (Puntos para sexo femenino, cruces para sexo masculino).

6.5. Distribución Bernoulli multivariada

En muchos campos de estudios y aplicaciones, las variables predictoras que intentan explicar y predecir una variable respuesta son categóricas o binarias. Por ejemplo, el estudio genético y asociaciones (**Peng et al., 2009**; **Wang et al., 2011**), el precesamiento de imágenes (**Hassner and Sklansky, 1980**; **Woods, 1978**),

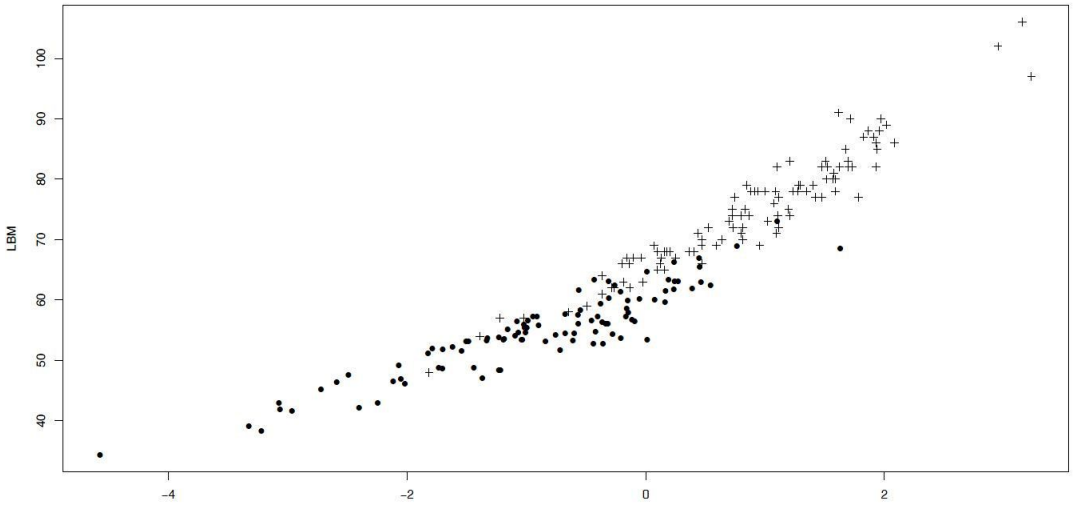


FIGURA 5. LBM versus el primer predictor KDR (Puntos para sexo femenino, cruces para sexo masculino).

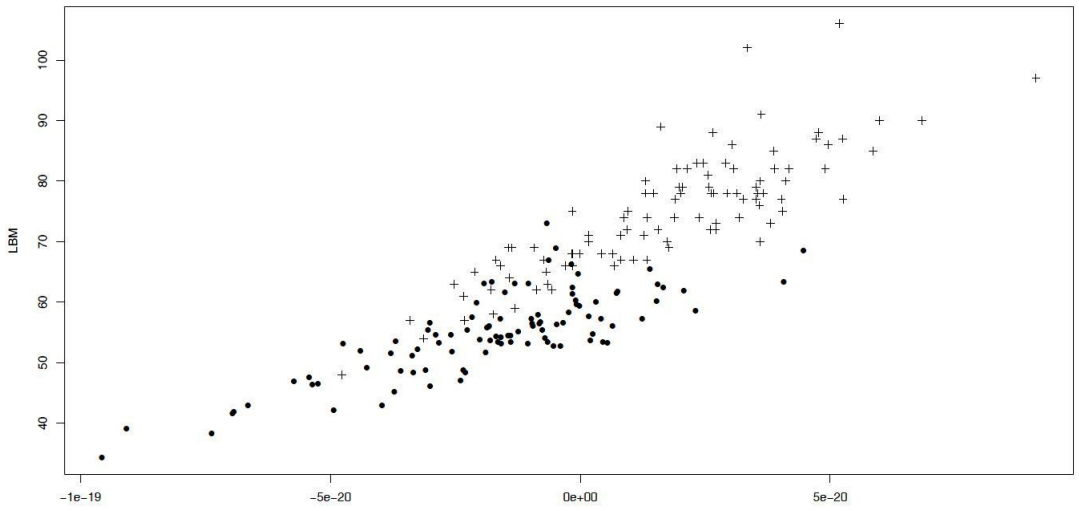


FIGURA 6. LBM versus el primer predictor COIR(Puntos para sexo femenino, cruces para sexo masculino).

el procesamiento del lenguaje natural ([**Manning and Schutze, 1999**]), redes sociales ([**Wasserman and Pattison, 1996**]; [**Handcock, 2003**]), estadística espacial ([**Besag, 1974**]), en economía ([**Pierce and Cleveland, 1984**]), o en modelos basados en interaciones ([**Brock and Durlauf, 2001**]) se analizan empíricamente un gran número de interaciones. La distribución Bernoulli multivariada permite modelar variables binarias correlacionadas y es miembro de la familia exponencial como vimos en la Sección 1.2.2. En esta

	$\widehat{\boldsymbol{\alpha}}$	W
Sir Parcial		
	$(0.18, 0.98, 0.01, 0.02, 0.12)$	
GPFC		
	$(0.88, 0.45, 0.10, 0.02, 0.11)$	0.04
	$(0.92, 0.37, 0.07, 0.02, 0.08)$	0.10
EF-DR		
	$(0.27, 0.92, -0.03, -0.04, 0.21)$	-0.17
KDR		
	$(-0.29, -0.91, -0.03, 0.08, -0.29)$	

CUADRO 1. Coeficientes de la reducción suficiente.

SIR Parcial	GPFC	EF-DR	KDR	KSIR	COIR
2.660	3.563	2.573	3.160	5.580	8.804

CUADRO 2. Errores de predicción.

sección, dirigimos nuestra atención a regresiones inversas $\mathbf{X}|Y$ tal que los predictores dada la respuesta, se distribuyen conjuntamente como Bernoulli multivariado.

6.5.1. El modelo Ising

La función de probabilidad puntual de la distribución Bernoulli multivariada involucra términos que representan momentos de orden 3 o más grandes aún. Los modelos gráficos han sido usados para representar la distribución conjunta de variables categóricas. Mientras estos gráficos pueden fácilmente capturar y representar correlaciones de grado dos (es decir entre dos componentes del vector), aquellas interaciones de mayor grado (entre más de dos componentes) son extremadamente complejas de representar. Además, la estimación es computacionalmente inviable para tamaños de redes realistas.

El modelo Ising **Ising, 1925** es un modelo gráfico no dirigido que permite representar los efectos de las interacciones hasta grado dos y ha sido usado para modelar datos binarios multivariados en gran medida. A pesar de que el modelo de Ising es un caso especial de la distribución Bernoulli multivariada, **Wainwright and Jordan, 2008** mostraron que estructuras de dependencia más generales, que incluyan por ejemplo las interacciones de orden superior, se

pueden expresar como interacciones de grado 2 a través de la introducción de variables adicionales (véase también **Ravikumar et al., 2010**).

En el análisis de datos de la Sección 6.5.3 modelamos la regresión inversa de los predictores binarios usando el modelo Ising. Supongamos que contamos con p predictores binarios $X_1|Y, \ldots, X_p|Y$ con $X_j|Y \in \{0,1\}$, $1 \leq j \leq p$, cuya función de probabilidad puntual conjunta, que fue presentada en la Sección 1.2.2, es

$$p(x_1, \ldots, x_p|Y = y) = \exp\left(\sum_{j=1}^{p} \eta_y^{jj} x_j + \sum_{1 \leq j \leq j' \leq p} \eta_y^{jj'} x_j x_{j'} - \psi(\boldsymbol{\eta}_y)\right), \tag{6.15}$$

donde ψ es la función dada en (1.9) y $\boldsymbol{\eta}_y = \text{vech}\left((\eta_y^{jj})_{p\times p}\right)$ es una vector que especifica la estructura de correlaciones. Los parámetros η_y^{jj}, $1 \leq j \leq p$, corresponde a los efectos principales de la variable $X_j|Y$ y $\eta_y^{jj'}$, $1 \leq j < j' \leq p$, corresponde a los efectos de la interacción entre $X_j|Y$ y $X_{j'}|Y$.

Los parámetros $\eta_y^{jj'}$ están asociados directamente a la estructura subyacente del modelo gráfico y en **Cheng et al., 2012** se muestra que estos están conectados a probabilidades condicionales de la siguiente forma:

$$\log \frac{P(X_j = 1|\mathbf{X}_{-j}, y)}{1 - P(X_j = 1|\mathbf{X}_{-j}, y)} = \eta_y^{jj} + \sum_{j' \neq j} \eta_y^{jj'} X_{j'}$$

donde $\mathbf{X}_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$. Además condicionando en $\mathbf{X}_{-j,-j'} = \mathbf{0}$, **Cheng et al., 2012** obtienen que

$$\eta_y^{jj'} = \log \frac{P(X_j = 1, X_{j'} = 1|\mathbf{X}_{-j,-j'}, Y) P(X_j = 0, X_{j'} = 0|\mathbf{X}_{-j,-j'}, Y)}{P(X_j = 1, X_{j'} = 0|\mathbf{X}_{-j,-j'}, Y) P(X_j = 0, X_{j'} = 1|\mathbf{X}_{-j,-j'}, Y)},$$

lo que implica que X_j y $X_{j'}$ son condicionalmente independientes dado Y y las restantes componentes de $\mathbf{X}$ si y solo si $\eta_y^{jj'} = 0$ y por lo tanto sus correspondientes nodos no están conectados.

Estimación de la reducción suficiente

La función de probabilidad (6.15) pertenece a una familia exponencial con parámetro natural $\eta_y = (\eta_y^{11}, \eta_y^{22}, \ldots, \eta_y^{pp}, \eta_y^{12}, \ldots, \eta_y^{(p-1)p})$ y estadístico suficiente:

$$\mathbf{T}(\mathbf{X}) = (X_1, \ldots, X_p, X_1X_2, \ldots, X_1X_p, \ldots, X_{p-1}X_p)^T,$$

ambos con $p + p(p-1)/2$ elementos. Aplicando el Teorema 6.1, la reducción suficiente para la regresión de Y en $\mathbf{X}$ es $\boldsymbol{\alpha}^T(\mathbf{T}(\mathbf{X}) - \mathbf{E}(\mathbf{T}(\mathbf{X}))$, donde $\text{span}(\boldsymbol{\alpha}) = \text{span}\{\boldsymbol{\eta}_Y - \overline{\boldsymbol{\eta}} : y \in \Omega_Y\}$. Para estimar los parámetros naturales $\boldsymbol{\eta}_Y$ aplicamos regresión logística multivariada. Supongamos que contamos con n muestras de las variables $\mathbf{X}$ e Y, y denotamos estas $\mathbf{y} = (y_1, \ldots, y_n)$ y

$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$. La función objetivo de máxima verosimilitud para el modelo lineal generalizado es

$$l(\mathbf{x}, y) = \sum_{i=1}^{n} \left(\sum_{j=1}^{p} x_{ij}\eta_{jj} + \sum_{j<j'} \eta_{jj'} x_{ij} x_{ij'} - \psi(\boldsymbol{\eta}_y) \right). \tag{6.16}$$

La maximización de (6.16) con respecto a los coeficientes de la componente $\mathbf{f}_y$ en el modelo lineal generalizado para los parámetros naturales $\boldsymbol{\eta}_y$ se puede hacer como en la Sección 6.3, y el algoritmo IRLS se puede utilizar para estimar $\boldsymbol{\alpha}$ por medio del ajuste de un modelo de regresión logística multivariado. La dimensión de $\boldsymbol{\alpha}$ puede estimarse ya sea con los test asintóticos en la Sección 6.3.3, o con un criterio basado en la información, tales como AIC o BIC. La reducción suficiente de la regresión de Y en los $\mathbf{X}$ es entonces $(\widehat{\boldsymbol{\alpha}}_1, \ldots, \widehat{\boldsymbol{\alpha}}_{\widehat{d}})^T \left(\mathbf{T}(\mathbf{X}) - \overline{\mathbf{T}}(\mathbf{X})\right)$, donde $\overline{\mathbf{T}}$ es la media muestral de $\mathbf{T}(\mathbf{X})$ y $\widehat{d}$ es la dimensión estimada. En las siguientes dos secciones, 6.5.2 y 6.5.3, ilustramos cómo aplicar nuestra metodología con predictores inversamente Bernoulli.

6.5.2. Simulaciones

Supongamos que Y es normal con media cero y desviación estandar 0.5. Dado Y, sea $\mathbf{X} = (X_1, \ldots, X_p)^T$ una vector Bernoulli multivariado de dimensión p, donde la estructura de correlaciones de grado 2 entre sus componentes se da solamemente entre los pares contiguos. Es decir, están correlacionas X_1 con X_2, X_3 con X_4, ..., X_{p-1} con X_p y todas las demás interacciones de grado 2 o mayor no están presentes. Por lo tanto, $\boldsymbol{\eta}_Y = (\eta^{11}, \eta^{22}, \ldots, \eta^{pp}, \eta^{12}, \eta^{34}, \ldots, \eta^{(p-1)p})$. El estadístico suficiente es $\mathbf{T}(\mathbf{X}) = (X_1, X_2, \ldots, X_p, X_1X_2, X_3X_4, \ldots, X_{p-1}X_p)$. El parámetro natural $\boldsymbol{\eta}_Y$ es generado como

$$\boldsymbol{\eta}_Y = \mathbf{AB}(\mathbf{f}_Y - \mathbf{E}\mathbf{f}_Y)$$

donde $\mathbf{B} = 1$ y $\mathbf{f}_Y = Y$. Para $p = 4$, $\mathbf{A} = (1, 1, 1, 1, 10, 10)^T/\sqrt{204}$ y para $p > 4$, completamos $\mathbf{A}$ con ceros de formal tal que para todo p, la reducción suficiente minimal es $[(X_1 - \mathbf{E}(X_1)) + (X_2 - \mathbf{E}(X_2)) + (X_3 - \mathbf{E}(X_3)) + (X_4 - \mathbf{E}(X_4)) + 10(X_1X_2 - \mathbf{E}(X_1X_2)) + 10(X_3X_4 - \mathbf{E}(X_3X_4))]/\sqrt{204}$. Calculamos la precisión de la estimación de la reducción suficiente en tres casos:

- **(a)** asumiendo la verdadera estructura de correlaciones,
- **(b)** asumiendo que las componentes de $\mathbf{X}$ son independientes dado Y, que es el enfoque de **Cook and Li, 2009**,
- **(c)** asumiendo que todas las correlaciones de grado 2 están presentes, que es el modelo Ising completo.

Esta precisión es medida como el promedio de los ángulos entre el verdadero subespacio generado por $\boldsymbol{\eta}_Y - \overline{\boldsymbol{\eta}}$, $Y \in \Omega_Y$ y el estimado en N repeticiones. En estas simulaciones utilizamos $n = 100, 200, 300, 500$ y $p = 4, 6, 10, 12$. La estimación bajo (b) y (c) es llevada a cabo usando el paquete MVB desarrollado por **[Dai, 2013]**. **[Dai et al., 2013]** estudia la distribución Bernoulli multivariada y la estimación en GLMM que incluye interacciones entre los predictores. En el caso (a), aplicamos nuestro algoritmo. La versión actual del paquete MVB no permite especificar como cero algunas interacciones de grado 2 o mayores, es por esto que nuestro algoritmo IRLS es la herramienta computacional disponible para estimar en el caso (a).

En el Cuadro 4 informamos el tiempo requerido para calcular las reducciones bajo (a), (b) y (c) para los diferentes tamaños de muestra y números de predictores binarios en una replica. Como era de esperar, el modelo completo Ising es la más exigente computacionalmente, seguido por nuestro método bajo la estructura de correlación verdadera. El modelo de independencia condicional (b) es el más rápido, ya que se calculan estimadores en una estructura menor. Los

	$n = 100$	$n = 200$	$n = 300$	$n = 500$	N
$p = 4$					
(a)	35.12 (16.65)	26.29 (12.14)	21.94 (8.50)	17.49 (8.09)	100
(b)	43.69 (16.30)	39.29 (12.52)	37.64 (10.49)	37.19 (7.33)	100
(c)	45.81 (15.47)	41.12 (13.78)	30.95 (13.62)	29.56 (12.24)	100
$p = 6$					
(a)	40.42 (12.10)	34.82 (13.15)	28.76 (16.96)	23.31 (13.24)	50
(b)	64.24 (15.84)	63.16 (23.91)	58.08 (25.59)	55.42 (30.12)	50
(c)	51.06 (14.56)	42.51 (13.25)	38.21 (13.78)	36.80 (12.89)	50
$p = 10$					
(a)	50.38 (10.78)	40.43 (10.27)	33.97 (10.47)	28.84 (9.16)	50
(b)	75.23 (10.86)	72.16 (11.93)	73.14 (12.11)	74.45 (9.90)	50
(c)	57.12 (13.41)	55.63 (12.71)	50.22 (12.03)	47.01 (10.82)	50
$p = 12$					
(a)	57.96 (13.10)	42.56 (7.78)	35.36 (6.27)	30.29 (7.05)	50
(b)	76.65 (9.09)	78.80 (7.91)	78.23 (9.97)	74.61 (12.72)	50
(c)	59.14 (13.21)	58.09 (4.87)	51.70 (7.80)	46.34 (10.11)	50

CUADRO 3. Promedios y desviaciones estandar, entre paréntesis, de los ángulos entre el verdadero subespacio y el estimado.

promedios de los ángulos y sus desviaciones estándar, entre paréntesis, entre las reducciones reales y estimadas en (a), (b) y (c) se reportan en el Cuadro 3. La columna encabezada por N informa del número de repeticiones para cada combinación de tamaño de muestra y número de variables predictoras. En el Cuadro 3 vemos que cuando suponemos la verdadera estructura de correlación de los predictores, la exactitud de la estimación de la reducción suficiente aumenta rápidamente a medida que aumenta el tamaño muestral y es uniformemente mejor que (b) y (c). También es de destacar que el modelo completo Ising (sobre-parametrizado en este caso) que contiene todas las interacciones de orden 2, da mejores resultados en comparación con el modelo sencillo (b) de independencia condicional.

	$n = 100$	$n = 200$	$n = 300$	$n = 500$
$p = 4$				
(a)	20.99	42.09	63.94	105.9
(b)	1.98	4.15	5.8	11.14
(c)	30.12	52.23	89.21	200.98
$p = 6$				
(a)	87.36	183.76	292.87	492.14
(b)	2.35	5.32	8.38	12.85
(c)	102.23	207.69	387.21	605.41
$p = 10$				
(a)	2264.08	3888.49	5327.56	11962.96
(b)	2.73	3.34	8.25	13.1
(c)	3675.45	4671.90	6713.13	13945.12
$p = 12$				
(a)	6158.65	13925.21	18243.74	26580.66
(b)	3.26	5.74	8.89	14.50
(c)	8712.21	15092.01	20193.12	28143.81

CUADRO 4. Tiempos computacionales (en segundos) para una replica de la estimación de la reducción bajo (a), (b) y (c) para distintos valores de n y p.

Cuando la estructura de correlación condicional de los predictores se incorpora en el proceso de estimación, la precisión mejora significativamente y es mucho mejor en ambos casos. Surge entonces la pregunta de cómo se puede deducir la estructura de correlación en datos reales y

problemas de análisis de predictores binarios. Un enfoque, también apoyado por los resultados de la simulación del Cuadro 3, es proponer a un modelo Ising completo, que tiene $p+p(p-1)/2$ parámetros a estimar, y testear la significancia de los efectos de los coeficientes. Pero para valores de p grandes, esto sería prácticamente imposible. Por ejemplo, aunque sea un número relativamente pequeño de predictores como $p = 10$, hay 55 parámetros para estimar y 210 para $p = 20$.

Alternativamente, se puede detectar la red de dependencias condicionales de los predictores. Como en el caso de la distribución normal multivariada, si los parámetros que corresponden a la interacción en una distribución Bernoulli multivariada son cero, las variables correspondientes son condicionalmente independientes ([**Dai et al., 2013**]). Utilizando este hecho, se puede controlar la estructura de dependencia entre los componentes de un vector de Bernoulli mediante el establecimiento de los correspondientes parámetros del vector natural como cero. Este enfoque puede ser muy útil cuando el vector Bernoulli tiene dimensión grande y el número de parámetros a estimar hace que el problema no sea factible para tamaños de muestra realistas. En el trabajo [**Cheng et al., 2012**] propusieron una penalización L_1 para las interacciones y la selección de variables en modelos Ising. En el análisis de los datos Zoo que siguen, aplicamos este método con el fin de identificar qué efectos principales y de segundo orden se incluirán en el modelo lineal generalizado y luego usamos nuestro algoritmo IRLS de estimación.

6.5.3. Datos: Zoo

Los datos Zoo consisten en 101 animales clasificados en 7 categorías: anfibios, aves, peces, insectos, invertebrados, mamíferos y reptiles. Los animales son distribuidos entre estas categorías siendo las cantidades por cada clase respectivamente: 4, 20, 13, 8, 10, 41 y 5. Se cuenta con 16 predictores categóricos los cuales fueron registrados para cada animal, estos son: pelo, plumas, huevos, leche, aéreo, acuático, predador, dientes, vertebrados, respiración, venenosos, aletas, cola, doméstico, cat-size y patas. Los primeros 15 predictores son dicotómicos y el último predictor es policotómico. Este conjunto de datos es analizado en [**Cook and Li, 2009**]. Los 15 predictores binarios, $\mathbf{X}$, tienen como objetivo clasificar los animales en alguna de las 7 categorías, Y. El último predictor es omitido en este análisis con el fin de focalizarnos en la precisión de la clasificación bajo la hipótesis de predictores Bernoulli condicionalemente independientes o correlacionados. Bajo la independencia condicional de los predictores $\mathbf{X}$ dado la respuesta Y, el modelo logístico multivariado se ajusta de manera sencilla por medio del método iterativo IRLS. Como la respuesta es categórica, definimos (de acuerdo a lo visto en la Sección 5.3) $\mathbf{f}_y$

con la k-ésima coordenada dada por:

$$f_{yk} = I(y = k) - \frac{n_k}{n} \text{ para } k = 1, \ldots, r,$$

donde $r = 7 - 1 = 6$ (7 es el número de valores distintos que toma la variable respuesta), $I(\cdot)$ es la función indicadora y n_k es el número de observaciones distintas en la categoría k. Los parámetros naturales $\boldsymbol{\eta}_y$ en (6.15) son expresados como función lineal de $\mathbf{f}_y$

$$\eta_y^{jj} = \gamma_{0j} + \gamma_{1j} f_{y1} + \ldots + \gamma_{kj} f_{yr}, \tag{6.17}$$

donde $\boldsymbol{\gamma}_j = (\gamma_{0j}, \ldots, \gamma_{rj})^T$ es el vector de coeficientes que debe ser estimado para $j = 1, \ldots, p$. Es decir, que para el caso de predictores condicionalmente independientes hay $7 \times 15 = 105$ coeficientes para estimar. Si contamos con n muestras de $\mathbf{X}$ e Y, expresamos el modelo (6.17) como

$$\boldsymbol{\eta}_n = \mathbf{F}_n \boldsymbol{\Gamma}, \tag{6.18}$$

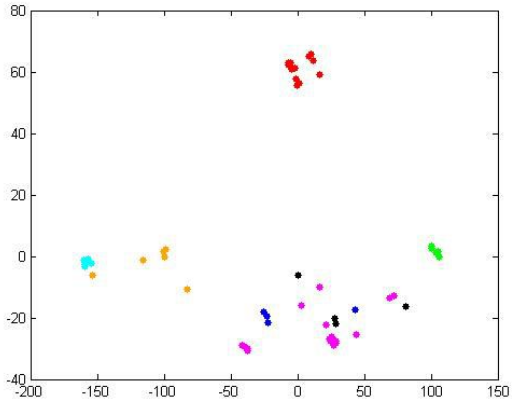
donde $\boldsymbol{\eta}_n = (\eta_i^{jj})$ es una matriz de parámetros de dimensiones $n \times p$ con $\eta_i^{jj} = \eta_{y_i}^{jj}$, $\mathbf{F}_n = (f_{il}) = (f_{y_i,l})$ una matriz fija $n \times (r+1)$ y $\boldsymbol{\Gamma} = (\gamma_{lj})$ una matriz de coeficientes $(r+1) \times p$. Observar que $\text{rank}(\mathbf{F}_n \boldsymbol{\Gamma}) = \text{rank}(\boldsymbol{\Gamma}^T \mathbf{F}_n \mathbf{F}_n^T \boldsymbol{\Gamma}) = \text{rank}(\boldsymbol{\Gamma})$, pues $\mathbf{F}_n^T \mathbf{F}_n$ es definida positiva (ver Sección A4.4 de [**Seber, 1977**]). Luego, de (6.18), el espacio de los parámetros naturales es generado por las filas de $\mathbf{F}_n \boldsymbol{\Gamma}$, i.e. $\text{span}(\boldsymbol{\eta}_y) = \text{span}(\boldsymbol{\Gamma}^T \mathbf{F}_n^T) = \text{span}(\boldsymbol{\Gamma}^T)$. En consecuencia, la inferencia acerca del la dimensión d del subespacio $\text{span}(\boldsymbol{\eta}_y)$ se basa en $\boldsymbol{\Gamma}$ en el sentido que un estimador del rango de $\boldsymbol{\Gamma}$ constituye un estimador de la dimensión de $\text{span}(\boldsymbol{\eta}_y)$. Los d vectores sigulares a izquierda del estimador IRLS de $\boldsymbol{\Gamma}^T$ que corresponde a los d valores singulares más grandes proporcionan una estimación de vectores bases de $\text{span}(\boldsymbol{\eta}_y)$ y las d combinaciones lineales de $\mathbf{T}(\mathbf{X})$ centrado constituyen los predictores suficientes. Por medio del paquete MVB del software estadístico [R] se obtiene una estimación para $\boldsymbol{\Gamma}$, que posiblemente será un tanto imprecisa ya que $n = 101$. Bajo condicional independencia, el paquete MVB proporciona los mismos estimadores que [**Cook and Li, 2009**] por lo que los dos métodos pueden aplicanse. En la Figura 7, las reducciones 1, 3, 5 y 6 son graficadas. Los colores sirven para identificar las diferentes categorías: azul para anfibios ($Y = 1$), rojo para aves ($Y = 2$), verde para peces ($Y = 3$), celeste para insectos ($Y = 4$), rosado para moluscos ($Y = 5$), dorado para mamíferos ($Y = 6$) y negro para reptiles ($Y = 7$). En estos gráficos vemos que bajo la independencia condicional de los predictores Bernoulli, en el panel (a) el primer predictor suficiente versus el tercero puede separar rojo (de aves) y verde (de peces), pero a pesar de que los puntos celestes (de insectos) y los puntos dorados (de mamíferos) parecen estar separados, un punto dorado

está muy cercano a los celestes. En el panel (b) puede separar verde (de peces) y azul (de anfibios) fácilmente. Además, hay separación entre el rojo (de aves) y negro (de reptiles), se puede trazar una curva cerrada alrededor de todos los puntos negros sin intersección con otro grupo. Esto se puede hacer usando algoritmos de aprendizaje automático para la clasificación de patrones que se basan en técnicas de segmentación de imagen y de optimización utilizados en conjuntos de nivel (véase por ejemplo, **Varshney and Willsky, 2010**). El panel (c) puede separar rojo (de aves), magenta (de moluscos), celeste (de insectos) y verde (de peces) con facilidad, y además se puede dibujar curvas cerradas para los puntos dorados (de mamíferos) pero los puntos azules (de anfibios) y negros (de reptiles), están demasiado cerca para diferenciarlos. A pesar de que reportamos los resultados de algunas reducciones, los valores singulares de $\widehat{\mathbf{\Gamma}}^T$ son 120.87, 100.81, 67.03, 48.25, 35.52 y 22.75, todos sustancialmente lejos de cero. Esto indica que la dimensión del problema es seis. Las reducciones suficientes que se exponen en la Figura 7 fueron seleccionadas ya que proporcionan una separación visual más clara. Señalamos que las reducciones no están clasificadas según su potencial de separación (es decir, las primeras direcciones separan más que las últimas) y varias fueron necesarias.

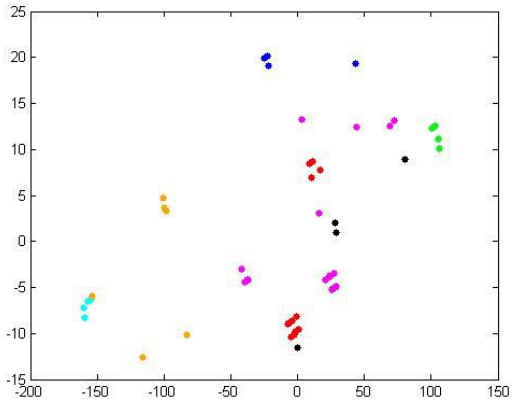
Como $n = 101$ no es posible ajustar el modelo Ising completo, pues el número de parámetros naturales de este modelo es de $p + \binom{p}{2} = 120$ y ajustar el modelo correspondiente a (6.17) requeriría la estimación de $7 \times 120 = 840$ parámetros con 101 observaciones. Por esta razón, se supone que el modelo de Ising es sparse; es decir, que algunos parámetros naturales η_y^{jk} son cero, para algunos $j \leq k$.

La penalización L_1 propuesta en **Cheng et al., 2012** inducen sparsidad tanto en el vector de parámetros naturales $\boldsymbol{\eta}_y = (\eta_y^{11}, \ldots, \eta_y^{pp}, \eta_y^{12}, \ldots, \eta_y^{p-1,p})^T$ como en la matriz de coeficientes $\mathbf{\Gamma}$, que ahora es de dimensión $(r+1) \times \left(p + \binom{p}{2}\right) = 7 \times 120$ en el modelo ampliado de (6.17). Se aplicó el *algoritmo de estimación conjunta* de **Cheng et al., 2012** a los datos Zoo. El código Matlab fue proporcionado por el Dr. Cheng. Los valores singulares del $\mathbf{\Gamma}$ estimado son 12.07, 10.09, 9.91, 9.62, 7.91 y 7.46. El tercer y cuarto, y el quinto y sexto valor singular son muy similares, lo que indica que los correspondientes vectores singulares a izquierda definen el mismo subespacio característico.

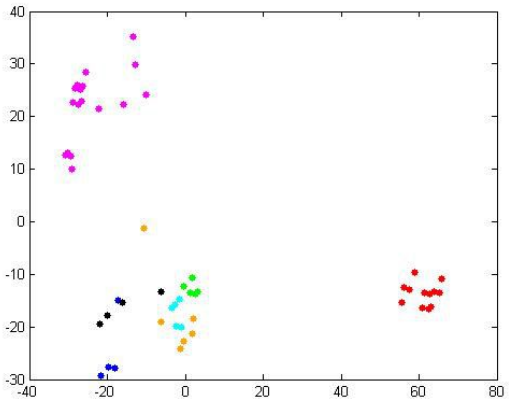
Al principio, es decir, antes de aplicar la selección de variables, $\mathbf{T}(\mathbf{X}) = (X_1, \ldots, X_{15}, X_1X_2, \ldots, X_{14}X_{15})$ con 120 términos. Después de seleccionar variables, 66 de sus términos fueron retenidos en el cálculo de la reducción suficiente. Como todavía es cuantioso el número de términos, no reportamos de manera explícita $\widehat{\boldsymbol{\alpha}}$ o la reducción suficiente, pero resaltamos que los efectos de las variables X_1, X_5, X_7, X_{11} y X_{14} están activos en la reducción



(a) SR_1 versus SR_3 bajo independencia



(b) SR_1 versus SR_6 bajo independencia



(c) SR_3 versus SR_5 bajo independencia

FIGURA 7. Gráficos de las reducciones suficientes 1, 3, 5, 6 bajo independencia condicional de las componentes de $\mathbf{X}$ dado Y.

suficiente. Estas variables indican si los animales tienen pelo, están en el aire, son depredadores, venenosos y domésticos, respectivamente. Todos los otros 61 efectos son interacciones, lo que sugiere que los predictores son dependientes.

En la figura 8 graficamos las dos primeras reducciones suficientes. Podemos ver que todos los colores están separados por curvas cerradas simples, incluidos los puntos negros y azules. Es decir que con dos reducciones suficientes podemos separar las siete clases y por lo tanto, se estima que la dimensión de la regresión puede ser 2, en contraste con las 6 reducciones bajo el supuesto de independencia. Para los datos Zoo, EF-DR ofrece perfecta clasificación dentro de la muestra y reduce significativamente la complejidad.

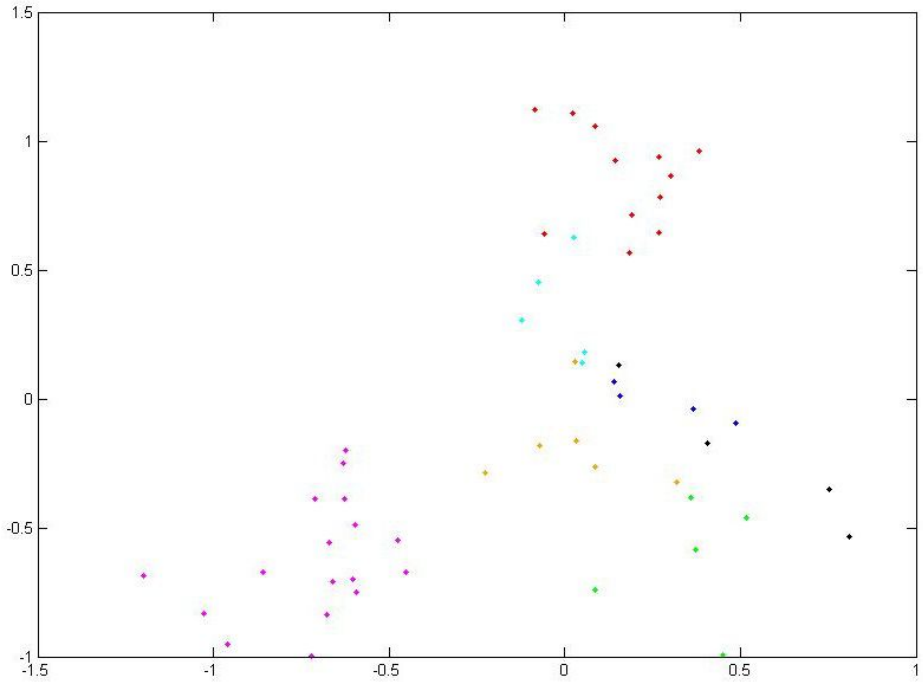


FIGURA 8. EF-DR: SR_1 versus SR_2 bajo dependencia

	EF-DR	KDR
$d = 2$		
LDA	0.139	0.297
DQDA	0.158	0.257
$d = 3$		
LDA	0.119	0.158
DQDA	0.109	0.139

CUADRO 5. Error de clasificación en LDA y dQDA

También aplicamos KDR con kernel gaussiano a los datos Zoo. Con las dos primeras direcciones, las siete clases no se pueden separar con curvas cerradas, como en EF-DR. Para comparar numéricamente la precisión de la clasificación de EF-DR y KDR, se utilizó el análisis discriminante lineal (LDA) y la versión simple del análisis discriminante cuadrático (DQDA), que utiliza estimaciones de la matriz de covarianza diagonal, ya que algunas matrices de covarianza de las clases eran singulares. Esta forma de clasificador QDA opera bajo el supuesto de independencia condicional de las variables dentro de cada clase, las cuales, a pesar de que no es cierto en general, se ha encontrado que funciona bien en la práctica de muchos conjuntos de

datos. KDR se aplicó directamente a los datos sin imponer que ciertos coeficientes sean cero, como en la aplicación de EF-DR. Se puede observar en el Cuadro 5, donde se reportan los promedios de clasificación errónea para LDA y DQDA con los dos métodos y los valores de la dimensión de la reducción ($d = 2, 3$), que EF-DR es más preciso con ambos clasificadores.

6.6. Demostraciones y resultados de las simulaciones del Capítulo 6

Demostración del Lema 6.7: El resultado de este lema se obtiene aplicando el Lema A.7 del Apéndice y el Delta método multivariado. En efecto, sea la función $g(\mathbf{U})$ definida en el subespacio de las matrices $k \times d$ de rango completo d tal que $g(\mathbf{U}) = \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T = \mathbf{P_U}$, luego de acuerdo al Lema A.7 tenemos que el gradiente de g es no nulo e igual a

$$\nabla g(\mathbf{U}) = \frac{\partial \mathbf{P_U}}{\partial \text{vec}^T(\mathbf{U})} = (\mathbf{I}_{k^2} + \mathbf{K}_{kk})(\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1} \otimes \mathbf{Q_U}).$$

Luego, como $\sqrt{n}\text{vec}(\widehat{\mathbf{U}} - \mathbf{U}) \to \mathcal{N}(0, \mathbf{V})$, aplicando el Delta método tenemos que

$$\sqrt{n}\left(g(\widehat{\mathbf{U}}) - g(\mathbf{U})\right) \to \mathcal{N}\left(0, \nabla g(\mathbf{U})\mathbf{V}\nabla^T g(\mathbf{U})\right).$$

Es decir que,

$$\left(\mathbf{P}_{\widehat{\mathbf{U}}} - \mathbf{P_U}\right) \to \mathcal{N}(0, (\mathbf{I}_{k^2} + \mathbf{K}_{kk})(\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1} \otimes \mathbf{Q_U})\mathbf{V}((\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}_1^T \otimes \mathbf{Q_U})(\mathbf{I}_{k^2} + \mathbf{K}_{kk})).$$

□

Valores iniciales y algoritmo EF-DR para los datos Atletas de Australia

Suponemos para cada $Y = y$, $\boldsymbol{\eta}_y = \overline{\boldsymbol{\eta}} + \mathbf{D}(\mathbf{f}_y - \overline{\mathbf{f}}_y) \in \mathbb{R}^{p+1}$. Bajo este modelo, sea $\boldsymbol{\Gamma}^T = (\overline{\boldsymbol{\eta}}, \mathbf{D})$ la matriz de coeficientes desconocidos de dimenesión $(r+1)\times(p+1)$ y $\boldsymbol{\gamma} = \text{vec}(\boldsymbol{\Gamma}) = \text{vec}\begin{pmatrix}\overline{\boldsymbol{\eta}}^T \\ \mathbf{D}^T\end{pmatrix} \in \mathbb{R}^{(p+1)(r+1)\times 1}$, luego

$$\boldsymbol{\eta}_y = \begin{pmatrix}\boldsymbol{\eta}_1 \\ \eta_2\end{pmatrix} = \left(\mathbf{I}_{p+1} \otimes (1, (\mathbf{f}_y - \overline{\mathbf{f}}_y)^T)\right)\boldsymbol{\gamma} = \mathbf{F}_y\boldsymbol{\gamma}$$

con $\boldsymbol{\eta}_1 \in \mathbb{R}^{p\times 1}$, $\eta_2 \in \mathbb{R}$ y $\mathbf{F}_y = \mathbf{I}_{p+1} \otimes (1, (\mathbf{f}_y - \overline{\mathbf{f}}_y)^T) : (p+1) \times (p+1)(r+1)$. Notar que $\overline{\boldsymbol{\eta}}$ es $(p+1) \times 1$ y $\mathbf{D}$ es de orden $(p+1) \times r$, donde p en el número de predictores y la dimensión adicional es debido a la variable binaria W.

Los párametos de interés son $\mathbf{D}$ y $\overline{\boldsymbol{\eta}}$, mientras que $\boldsymbol{\Delta}$ y $\mathbf{b}_2$ son parámetros extras y son fijados en los valores iniciales. $\mathbf{D}$ y $\overline{\boldsymbol{\eta}}$ son actualizados en cada iteración del algoritmo.

Valores iniciales y algoritmo:

1. *a)* Realizar la regresión $\mathbf{X}_{ctd} = \mathbf{X} - \overline{\mathbf{X}} : p \times 1$ en los valores centrados de $\mathbf{f}_y : r \times 1$ y W,

$$\mathbf{X}_{ctd} = \mathbf{b}_1(\mathbf{f}_y - \overline{\mathbf{f}}_y) + \mathbf{b}_2(W - \overline{W}) + \boldsymbol{\epsilon} \tag{6.19}$$

para obtener los estimadores $\widehat{\mathbf{b}}_2 : p \times 1$ y $\widehat{\mathbf{b}}_1 : p \times r$. Sea $\widehat{\boldsymbol{\Delta}}$ la matriz de covarianza de los residuos del ajuste de (6.19).

b) Realizar la regresión logística de W en $\mathbf{f}_y - \overline{\mathbf{f}}_y$,

$$\log \frac{p_y}{1-p_y} = h_1 + \mathbf{h}_2^T(\mathbf{f}_y - \overline{\mathbf{f}}_y)$$

y obtener los estimadores $\widehat{h}_1$ y $\widehat{\mathbf{h}}_2$.

2. Los valores iniciales para $\boldsymbol{\gamma}$ son

$$\boldsymbol{\gamma}_0 = \text{vec}\begin{pmatrix} \overline{\boldsymbol{\eta}}_0^T \\ \mathbf{D}_0^T \end{pmatrix}$$

donde

a) $\overline{\boldsymbol{\eta}}_0 = (\mathbf{A}_{10}^T, A_{20})^T$, $\mathbf{A}_{10} : p \times 1$, $A_{20} \in \mathbb{R}$ y $\widehat{\mathbf{D}}_0 = (\mathbf{D}_{10}, \mathbf{D}_{20})^T$, con $\mathbf{D}_{10} : r \times p$, $\mathbf{D}_{20} : r \times 1$.

b) $\mathbf{A}_{10} = \mathbf{0}$ y $\widehat{\mathbf{D}}_{10}^T = \widehat{\boldsymbol{\Delta}}^{-1}\widehat{\mathbf{b}}_1$.

c) $A_{20} = \widehat{h}_1$ y $\mathbf{D}_{20} = \widehat{\mathbf{h}}_2 - \widehat{\mathbf{b}}_2^T\widehat{\mathbf{D}}_{10}^T$.

3. Actualizar $\boldsymbol{\gamma}$

$$\boldsymbol{\gamma}^{(t+1)} = \left(\sum_{i=1}^n \mathbf{F}_{y_i}^T \mathbf{W}_{y_i}^{(t)} \mathbf{F}_{y_i}\right)^{-1} \sum_{i=1}^n \mathbf{F}_{y_i}^T \mathbf{W}_{y_i}^{(t)} \left(\mathbf{F}_{y_i}\boldsymbol{\gamma}^{(t)} + (\mathbf{W}_{y_i}^{(t)})^{-1}(\mathbf{T}(\mathbf{x}, w) - d_{y_i}^{(t)})\right)$$

donde para cada $i = 1, \ldots, n$,

a)

$$d_{y_i}^{(t)} = \begin{pmatrix} \widehat{\boldsymbol{\Delta}}\boldsymbol{\eta}_{1i}^{(t)} + \widehat{\mathbf{b}}_2(p_{y_i}^{(t)} - \overline{w}) \\ p_{y_i}^{(t)} - \overline{w} \end{pmatrix}$$

usando que $\mathbf{E}(\mathbf{T}(\mathbf{X}, W)|Y) = \nabla\psi(\boldsymbol{\eta}_y)$ y $\partial l/\partial\boldsymbol{\eta}_y = \mathbf{T}(\mathbf{X}, W) - \nabla\psi(\boldsymbol{\eta}_y)$,

b)

$$\mathbf{W}_{y_i}^{(t)} = \begin{pmatrix} \widehat{\boldsymbol{\Delta}} + \widehat{\mathbf{b}}_2\widehat{\mathbf{b}}_2^T p_{y_i}^{(t)}(1-p_{y_i}^{(t)}) & p_{y_i}^{(t)}(1-p_{y_i}^{(t)})\widehat{\mathbf{b}}_2 \\ \widehat{\mathbf{b}}_2^T p_{y_i}^{(t)}(1-p_{y_i}^{(t)}) & p_{y_i}^{(t)}(1-p_{y_i}^{(t)}) \end{pmatrix}$$

usando $\text{var}(\mathbf{T}(\mathbf{X}, W)|Y) = \partial^2\psi(\boldsymbol{\eta}_y)/\partial\boldsymbol{\eta}_y^T\partial\boldsymbol{\eta}_y$, donde

- $\boldsymbol{\eta}_{y_i}^{(t)} = \begin{pmatrix} \boldsymbol{\eta}_{1i}^{(t)} \\ \eta_{2i}^{(t)} \end{pmatrix} = \mathbf{F}_{y_i}\boldsymbol{\gamma}^{(t)}$, $\boldsymbol{\eta}_{1i}^{(t)} \in \mathbb{R}^{p+1}$ and $\eta_{2i}^{(t)} \in \mathbb{R}$
- $p_{y_i}^{(t)} = \dfrac{e^{\eta_{2i}^{(t)} + \widehat{\mathbf{b}}_2^T\boldsymbol{\eta}_{1i}^{(t)}}}{1 + e^{\eta_{2i}^{(t)} + \widehat{\mathbf{b}}_2^T\boldsymbol{\eta}_{1i}^{(t)}}}$.

4. Repetir el paso 3 hasta obtener convergencia.

Tablas de resultados de las simulaciones

Como dijimos en la Sección 6.4 utilizamos la estructura de los datos de los atletas australianos para diseñar un experimento de simulación con el fin de comparar nuestro método, EF-DR, con Sir Parcial, GPFC, KDR y KSIR. Suponemos que la respuesta es continua y que tenemos un predictor binario, W y p predictores continuos, $\mathbf{X} \in \mathbb{R}^p$.

Sea $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. La función de distribución condicional $(\mathbf{X}, W)|Y$ es

$$f_{(\mathbf{X},W|Y=y)}(\mathbf{x}, w|Y = y) = f_{(\mathbf{X}|W=w,Y=y)}(\mathbf{x}|W = w, Y = y) \cdot f_{(W|Y=y)}(w|Y = y).$$

Sea $W|Y$ una variable aleatoria Bernoulli con,

$$p_y = P(W = 1|Y = y) = \frac{e^{a(y-\mu_y)}}{1 + e^{a(y-\mu_y)}}.$$

Además, sea $\mathbf{X}|(Y, W) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}|(W,Y)} = \boldsymbol{\mu}_{\mathbf{X}} + \mathbf{b}_1(Y - \mu_Y) + \mathbf{b}_2(W - \mu_W), \boldsymbol{\Delta}_{\mathbf{X}|(W,Y)})$ con $\mu_W = \mathbf{E}(W)$ y $\boldsymbol{\mu}_{\mathbf{X}} = \mathbf{E}(\mathbf{X})$. Bajo este modelo, la reducción suficiente minimal esta dada por

$$\boldsymbol{\alpha} = \left(\boldsymbol{\Delta}_{\mathbf{X}|(W,Y)}^{-1}\mathbf{b}_1, \quad \mathbf{b}_2^T\boldsymbol{\Delta}_{\mathbf{X}|(W,Y)}^{-1}\mathbf{b}_1 + a\right)^T.$$

Simulación 1: Comparación de los métodos

En los experimentos de simulación realizados analizamos (a) la estimación de la reducción y (b) la predicción de la reducción. Cada réplica del experimento se llevó a cabo de la siguiente manera: n muestras se han extraído de $\mathbf{X}$, W y Y que satisfacen la estructura de la distribución. Se obtienen las reducciones bajo nuestro método EF-DR, Sir Parcial, GPFC y KDR y, posteriormente, se calculan los ángulos (en grados) entre $\widehat{\boldsymbol{\alpha}}$ y el verdadero $\boldsymbol{\alpha}$. Para KSIR, $\widehat{\boldsymbol{\alpha}}$ no puede ser calculado. A continuación, una muestra de validación $(Y', \mathbf{X}', W')$ de tamaño n se extrae, y se calcula el ángulo de la predicción $\widehat{\boldsymbol{\alpha}}^T(\mathbf{X}', W)$ y la verdadera predicción $\boldsymbol{\alpha}^T(\mathbf{X}', W)$. KSIR proporciona la predicción de la reducción KSIR, por lo que este número sí se informa en los cuadros. Las dos etapas se repiten 50 veces y los ángulos promedio junto con las desviaciones estándar correspondientes se calculan y se presentan en los cuadros que están a continuación. También se calcula el tiempo de ejecución (en segundos) para una repetición de los dos pasos para cada método. Los resultados son reportados para $n = 200$, $p = 5$, $\mu_Y = 64.87$, $\sigma_Y = 13.07$, $\boldsymbol{\mu}_{\mathbf{X}} = (5.19, 4.3, 1.55, 1.93, 2.67)$, $\mathbf{b}_1 = (0.003, 0.02, 0.0007, 0.002, 0.001)$, $\mathbf{b}_2 = (0.012, 0.14, -0.12, 0.02, -0.12)$, $a = -0.01$ y $\mu_W = 0.5$. Estos valores fueron seleccionados para que coincida con las estimaciones de los

parámetros correspondientes en el ejemplo de los atletas de Australia. Para tener en cuenta diferentes estructuras de dependencia, consideramos lo siguientes casos para $\boldsymbol{\Delta} = \boldsymbol{\Delta}_{\mathbf{X}|(W,Y)}$,

- Caso 1: $\boldsymbol{\Delta} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}|LBM,Sex}$, la matriz de covarianza muestral de los residuos de la regresión $\mathbf{X}|(LBM, Sex)$ en el ejemplo de los atletas de Australia.
- Caso 2: $\boldsymbol{\Delta} = \mathbf{I}_p$
- Caso 3: $\boldsymbol{\Delta} = 0.03 * \mathbf{I}_p$
- Caso 4: $\boldsymbol{\Delta} = 0.8 * \mathbf{I}_p$
- Caso 5: $\boldsymbol{\Delta} = \text{diag}(0.01, 0.1, 0.5, 0.7, 1)$

y los siguientes sub-casos en los que se añade correlación ρ fuera de la diagonal en los casos 2-5

- Sub-caso 1: $\rho = 0.1$
- Sub-caso 2: $\rho = 0.3$
- Sub-caso 3: $\rho = 0.9$

Notar que incluso cuando $\text{var}(\mathbf{X}|(Y, W))$ es diagonal, como en los casos 2-5, los predictores $\mathbf{X}|Y$ no son condicionalmente independientes ya que

$$\begin{aligned} \text{var}(\mathbf{X}|Y) &= \mathbf{E}_W(\boldsymbol{\Delta}_{\mathbf{X}|(Y,W)}) + \text{var}_W(\mathbf{E}(\mathbf{X}|(Y,W))) \\ &= \boldsymbol{\Delta}_{\mathbf{X}|(Y,W)} + \text{var}_W(\boldsymbol{\mu}_{\mathbf{X}} + \mathbf{b}_1(Y - \mu_Y) + \mathbf{b}_2(W - \mu_W)) \\ &= \boldsymbol{\Delta}_{\mathbf{X}|(Y,W)} + p_y(1 - p_y)\mathbf{b}_2\mathbf{b}_2^T. \end{aligned}$$

En el Cuadro 6 se informa el promedio de los ángulos y la desviación estándar, entre paréntesis, entre $\boldsymbol{\alpha}$ y la reducción estimada $\widehat{\boldsymbol{\alpha}}$ para EF-DR y GPFC, y entre $\boldsymbol{\alpha}_1$ y $\widehat{\boldsymbol{\alpha}}_1$ para Sir Parcial y KDR. El promedio de los ángulos y las correspondiente desviaciones estándar entre la verdadera reducción y la predicción de la reducción $\widehat{\mathbf{R}}(\mathbf{X}', W')$, donde $(\mathbf{X}', W')$ es una nueva muestra, para todos los métodos están reportadas en el Cuadro 7. Por último, el Cuadro 8 expone los tiempos computacionales, en segundos, requeridos para la estimación de la reducción en una réplica.

	Sir Parcial	GPFC	EF-DR	KDR
Caso 1	14.22 (7.17)	43.46 (2.95)	10.32 (4.98)	8.95 (4.24)
Caso 2	49.26 (9.29)	37.37 (6.41)	37.85 (6.44)	42.15 (17.08)
Caso 2.1	48.24 (20.94)	39.92 (12.19)	40.57 (13.90)	49.35 (22.87)
Caso 2.2	40.54 (18.50)	40.01 (10.57)	33.58 (11.74)	37.99 (15.47)
Caso 2.3	16.43 (11.83)	50.68 (15.81)	13.41 (11.44)	16.61 (22.56)
Caso 3	10.92 (4.53)	10.31 (2.09)	9.84 (3.25)	9.22 (2.48)
Caso 3.1	11.15 (3.93)	17.87 (3.37)	10.12 (3.39)	11.57 (3.53)
Caso 3.2	10.37 (3.70)	28.95 (4.47)	9.22 (3.39)	14.37 (5.01)
Caso 3.3	9.18 (3.55)	42.16 (4.59)	7.12 (2.88)	32.74 (5.71)
Caso 4	46.31 (21.64)	33.43 (10.17)	32.44 (11.11)	39.10 (21.04)
Caso 4.1	44.65 (17.81)	36.98 (12.36)	38.31 (12.69)	36.51 (16.34)
Caso 4.2	42.28 (19.91)	41.94 (13.08)	33.55 (11.99)	38.15 (20.29)
Caso 4.3	14.81 (6.70)	49.49 (16.05)	12.22 (5.12)	16.40 (15.49)
Caso 5	6.68 (5.74)	6.41 (2.62)	5.39 (3.21)	36.65 (6.86)
Caso 5.1	6.98 (4.96)	6.72 (2.21)	6.23 (4.12)	34.21 (6.23)
Caso 5.2	7.29 (5.63)	9.92 (2.34)	7.07 (4.68)	34.89 (6.21)
Caso 5.3	7.91 (6.05)	15.06 (1.77)	5.92 (3.48)	34.39 (3.20)

CUADRO 6. Promedio de los ángulos y desviaciones estándar, entre paréntesis, obtenidos entre la redución verdadera y la estimada por los métodos EF-DR, GPFC, Sir Parcial y KDR.

Simulación 2: EF-DR para diferentes n y p

Nuevamente usamos la estructura de los datos de la Sección 6.6, donde los parámetros son adaptados de acuerdo a los diferentes valores de p, para estudiar el comportamiento de EF-DR para varios valores de p y n. Se consideran cuatro tamaños de muestra $n = 200, 400, 600$ y 800 y tres valores de p (5, 10 y 20). Las medias y las desviaciones estándar de los ángulos entre la reducción verdadera y la estimada se informa en el Cuadro 9, mientras que en el Cuadro 10 se reporta entre la predicción verdadera y la predicción estimada. Por último los tiempos computacionales para una sola réplica se informan en el Cuadro 11.

	Sir Parcial	GPFC	EF-DR	KDR	KSIR
Caso 1	3.32 (1.52)	15.34 (1.45)	2.70 (1.13)	5.27 (1.57)	14.82 (2.74)
Caso 2	49.26 (8.89)	37.37 (14.24)	37.85 (7.19)	41.18 (16.58)	65.00 (25.96)
Caso 2.1	47.75 (21.55)	36.29 (11.30)	35.66 (11.70)	49.56 (24.48)	61.94 (24.64)
Caso 2.2	39.98 (19.02)	39.09 (10.79)	30.29 (10.46)	41.83 (19.00)	77.07 (12.08)
Caso 2.3	15.41 (6.63)	72.15 (14.35)	12.55 (4.53)	16.80 (23.72)	78.86 (11.40)
Caso 3	6.70 (2.82)	13.07 (2.98)	6.42 (2.34)	8.74 (2.57)	15.69 (3.42)
Caso 3.1	6.70 (2.40)	19.94 (3.63)	6.44 (2.22)	11.50 (3.59)	19.62 (6.39)
Caso 3.2	6.33 (2.18)	30.31 (5.33)	5.73 (2.48)	16.22 (5.13)	46.69 (5.96)
Caso 3.3	3.16 (1.16)	53.97 (4.74)	2.12 (0.74)	35.74 (9.40)	55.42 (6.56)
Caso 4	45.28 (21.33)	31.14 (11.24)	32.45 (11.72)	39.14 (21.04)	53.87 (26.64)
Caso 4.1	44.18 (17.91)	33.79 (11.98)	34.12 (11.59)	39.86(16.34)	51.78 (24.32)
Caso 4.2	41.91 (21.60)	40.49 (12.39)	30.55 (11.98)	37.25(17.60)	70.63 (16.77)
Caso 4.3	14.05 (5.23)	71.42 (13.89)	11.56 (4.22)	17.99(18.49)	72.80 (13.57)
Caso 5	10.52 (4.40)	12.57 (3.81)	10.15 (3.95)	26.58 (6.41)	23.22 (4.14)
Caso 5.1	10.84 (4.99)	14.91 (4.29)	9.84 (3.87)	25.93 (5.13)	21.23 (3.50)
Caso 5.2	9.31 (3.57)	28.17 (5.14)	9.81 (3.27)	27.43 (5.12)	27.40 (9.42)
Caso 5.3	4.28 (1.51)	58.28 (5.44)	3.66 (1.15)	45.88(4.34)	45.07 (6.32)

CUADRO 7. Promedio de los ángulos y desviaciones estándar, entre paréntesis, obtenidos en predicción para los métodos EF-DR, GPFC, Sir Parcial y KDR.

	Sir Parcial	GPFC	EF-DR	KDR	KSIR
Caso 1	0.81	1.18	1.38	1.9725+e004	352.18
Caso 2	0.71	1.20	1.34	4.5679e+003	265.55
Caso 2.1	0.75	1.16	1.45	1.0643e+004	259.45
Caso 2.2	0.15	1.18	1.36	1.7975e+004	239.99
Caso 2.3	0.14	1.24	1.46	1.8017e+004	237.81
Caso 3	0.17	1.23	1.40	9.7337e+003	238.78
Caso 3.1	0.14	1.39	1.40	8.8374e+003	232.24
Caso 3.2	0.16	1.10	1.43	1.1857e+004	263.83
Caso 3.3	0.15	1.27	1.36	1.8649e+004	265.25
Caso 4	0.17	1.19	1.43	7.4192e+003	287.61
Caso 4.1	0.17	1.16	1.47	7.6829e+003	303.81
Caso 4.2	0.11	1.18	1.41	7.7203e+003	271.18
Caso 4.3	0.14	1.22	1.34	8.4938e+003	234.71
Caso 5	0.16	1.19	1.40	1.3910e+004	241.61
Caso 5.1	0.14	1.20	1.39	1.1967e+004	248.23
Caso 5.2	0.16	1.21	1.42	1.1832e+004	242.19
Caso 5.3	0.13	1.26	1.50	4.6339e+003	221.46

CUADRO 8. Tiempos computacionales (en segundos) para una replicacion para EF-DR, GPFC, Sir Parcial, KDR, y KSIR.

		Caso 3	Caso 3.1	Caso 3.2	Caso 3.3	Caso 5	Caso 5.1	Caso 5.2	Caso 5.3
$p = 5$	$n = 200$	9.37	10.18	8.38	5.98	5.99	6.08	6.04	5.51
		(3.53)	(3.22)	(3.61)	(2.98)	(4.32)	(2.50)	(3.43)	(2.83)
	$n = 400$	6.69	7.11	6.52	3.89	4.15	4.04	3.85	3.67
		(2.71)	(2.85)	(2.17)	(1.81)	(2.34)	(1.70)	(3.52)	(3.03)
	$n = 600$	5.60	5.93	5.45	3.22	3.38	3.52	3.32	3.05
		(2.32)	(2.08)	(2.03)	(1.47)	(1.95)	(1.66)	(2.59)	(2.43)
	$n = 800$	4.87	4.97	4.81	2.53	2.86	2.95	2.84	2.41
		(1.90)	(1.83)	(1.74)	(1.13)	(1.32)	(1.48)	(1.95)	(1.72)
$p = 10$	$n = 200$	13.47	14.46	13.40	11.18	4.99	4.67	4.46	4.45
		(3.02)	(2.42)	(3.07)	(3.21)	(2.43)	(2.02)	(2.81)	(2.29)
	$n = 400$	9.04	10.53	9.69	8.18	3.23	3.13	3.01	2.90
		(2.42)	(1.29)	(2.40)	(1.98)	(1.97)	(1.70)	(1.53)	(1.43)
	$n = 600$	7.31	8.50	8.00	6.46	2.64	2.61	2.69	2.56
		(1.55)	(1.14)	(1.87)	(1.57)	(1.61)	(1.45)	(1.17)	(1.41)
	$n = 800$	6.50	7.45	6.88	5.42	2.26	2.21	2.37	2.22
		(1.28)	(1.02)	(1.55)	(1.33)	(1.36)	(1.38)	(1.27)	(1.20)
$p = 20$	$n = 200$	25.18	27.18	26.04	19.13	9.73	8.97	8.35	7.52
		(4.53)	(4.01)	(3.98)	(2.86)	(2.92)	(3.22)	(2.29)	(2.49)
	$n = 400$	17.70	19.17	17.69	12.83	6.29	6.14	6.12	4.96
		(2.89)	(3.48)	(2.45)	(2.43)	(1.71)	(2.08)	(1.76)	(1.62)
	$n = 600$	15.14	15.17	14.51	10.92	4.91	4.92	5.09	4.25
		(2.50)	(2.78)	(1.65)	(1.87)	(1.61)	(1.61)	(1.68)	(1.97)
	$n = 800$	13.01	13.37	13.01	9.35	4.22	4.22	4.33	3.80
		(1.43)	(2.30)	(1.43)	(1.38)	(1.45)	(1.40)	(1.37)	(1.94)

CUADRO 9. Promedio de los ángulos y desviaciones estándar, entre paréntesis, obtenidos entre la reducción estimada y la verdadera usando EF-DR para diferentes n y p.

		Caso 3	Caso 3.1	Caso 3.2	Caso 3.3	Caso 5	Caso 5.1	Caso 5.2	Caso 5.3
$p = 5$	$n = 200$	6.22	6.77	5.24	3.57	10.52	9.58	8.89	3.77
		(2.61)	(2.31)	(2.37)	(0.95)	(3.21)	(3.17)	(3.37)	(1.16)
	$n = 400$	4.32	4.60	3.89	2.67	7.36	6.71	6.27	2.61
		(1.99)	(1.89)	(1.32)	(0.69)	(2.48)	(1.89)	(2.97)	(0.91)
	$n = 600$	3.77	3.70	3.21	2.10	6.05	5.57	5.25	2.13
		(1.53)	(1.35)	(1.37)	(0.57)	(2.02)	(1.61)	(2.05)	(0.79)
	$n = 800$	3.21	3.15	2.81	1.80	5.18	5.03	4.55	1.85
		(1.21)	(1.16)	(1.11)	(0.43)	(1.52)	(1.64)	(1.91)	(0.66)
$p = 10$	$n = 200$	6.03	6.54	6.31	3.99	5.06	4.83	4.26	1.94
		(2.18)	(1.77)	(2.10)	(1.50)	(1.21)	(1.38)	(1.01)	(0.74)
	$n = 400$	3.96	4.63	4.52	2.57	3.39	3.36	3.06	1.32
		(1.01)	(1.32)	(1.08)	(1.14)	(0.75)	(0.83)	(0.77)	(0.42)
	$n = 600$	3.18	3.75	3.78	2.15	2.77	2.80	2.59	1.06
		(0.74)	(1.08)	(0.89)	(1.11)	(0.52)	(0.58)	(0.69)	(0.31)
	$n = 800$	2.80	3.28	3.31	1.75	2.40	2.43	2.32	0.90
		(0.58)	(0.75)	(0.69)	(0.98)	(0.49)	(0.57)	(0.56)	(0.26)
$p = 20$	$n = 200$	17.54	19.17	18.66	8.02	7.01	6.74	6.68	2.97
		(3.61)	(3.32)	(1.28)	(1.60)	(1.28)	(1.71)	(1.04)	(1.05)
	$n = 400$	12.25	13.46	12.53	5.58	4.71	4.76	4.59	1.92
		(2.26)	(2.66)	(0.93)	(1.23)	(0.93)	(0.97)	(0.75)	(0.56)
	$n = 600$	10.51	10.63	10.49	4.69	3.87	3.97	3.79	1.77
		(1.92)	(2.14)	(0.82)	(1.02)	(0.82)	(0.60)	(0.79)	(1.52)
	$n = 800$	9.06	9.46	9.41	4.11	3.48	3.51	3.28	1.41
		(1.43)	(1.72)	(0.70)	(0.77)	(0.70)	(0.46)	(0.65)	(0.38)

CUADRO 10. Promedio de los ángulos y desviaciones estandár, entre paréntesis, obtenidos en predicción con el método EF-DR para diferentes n y p.

		Caso 3	Caso 3.1	Caso 3.2	Caso 3.3	Caso 5	Caso 5.1	Caso 5.2	Caso 5.3
$p = 5$	$n = 200$	1.63	1.17	0.98	1.21	0.83	1.06	1.02	0.94
	$n = 400$	2.50	1.78	1.92	1.78	2.53	1.76	1.74	1.92
	$n = 600$	3.51	2.60	2.12	2.51	2.89	2.07	2.51	2.51
	$n = 800$	1.62	0.87	1.53	1.53	1.32	1.50	1.51	1.47
$p = 10$	$n = 200$	1.44	1.17	1.05	1.48	1.45	1.62	1.04	1.28
	$n = 400$	1.78	1.95	1.91	3.57	1.87	2.28	2.11	2.76
	$n = 600$	2.54	2.86	2.30	3.76	2.15	2.75	2.62	2.79
	$n = 800$	1.20	0.97	0.92	0.89	1.53	0.90	1.51	0.92
$p = 20$	$n = 200$	2.14	1.06	1.26	0.92	1.60	1.25	1.30	1.53
	$n = 400$	4.10	2.17	2.88	1.85	2.47	1.86	2.19	2.99
	$n = 600$	5.19	2.65	3.62	3.90	3.34	3.35	4.21	4.32
	$n = 800$	2.34	1.10	1.13	1.21	1.44	1.11	1.06	1.07

CUADRO 11. Tiempo computacional (en segundos) para una replica de EF-DR para diferentes n y p.

CONCLUSIONES GENERALES

Para los modelos lineales generalizados multivariados de rango reducido hemos obtenido la distribución asintótica de los estimadores de máxima verosimilitud basándonos en un teorema general de **Shapiro, 1986** e imitando el procedimiento presentado en **Cook et al., 2015** para los modelos de regresión lineal de rango reducido. De esta forma, hemos complementado el trabajo de **Yee and Hastie, 2003** obteniendo intervalos de confianza asintóticos de los parámetros de este modelo.

Además, teniendo en cuenta que para obtener los estimadores de máxima verosimilitud de los GLMM es necesario realizar un procedimiento iterativo de dos pasos, hemos propuestos dos estimadores de rango reducido para este modelo que son sencillos de obtener. En particular, para el estimador de **minimimización cuadrática**, hemos demostrado que es tan eficiente como el estimador de máxima verosimilitud y hemos comprobado su desempeño por medio de simulaciones, el cual resultó satisfactorio pues se obtuvieron errores de estimación similares a los que arrojaron los estimadores de máxima verosimilitud.

Las metodologías SDR basadas en modelos surgieron de la conexión que Cook **Cook, 2007** señala entre los estadísticos suficientes y reducciones suficientes. Basados en esta conexión, hemos calculado las reducciones suficientes en regresiones donde la distribución de los predictores dada la respuesta pertenece a una familia exponencial. Las características más atractivas de las reducciones suficientes que hemos derivado a través del método EF-DR son que (1) son exhaustiva, (2) son funciones lineales de las estadísticos suficientes, que pueden ser tanto funciones lineales o no lineales de los predictores, y (3) se cuenta con una forma funcional explícita. Además, hemos propuesto estimaciones de máxima verosimilitud de las reducciones suficientes y por lo tanto asintóticamente eficientes.

Un aspecto potencialmente difícil de nuestra metodología deriva del hecho de que es necesario realizar una regresión multivariada de un vector de predictores, que puede ser cualquier mezcla de variables continuas y categóricas, en funciones de la variable respuesta y requiere el ajuste del algoritmo IRLS a la distribución conjunta específica de $\mathbf{X}|Y$. Estas posibles dificultades

computacionales se han reflejado en el análisis del conjunto de datos Atletas de Australia en la Sección 6.4.

BIBLIOGRAFÍA

[Adragni et al., 2012] Adragni, K., Cook, R. D., and Wu., S. (2012). Grassmann optim: An r package for grassmann manifold optimization. *Journal of Statistical Software*, 50:1–18.

[Adragni and Cook, 2009] Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 367:4385–4405.

[Akaho, 2001] Akaho, S. (2001). *A kernel method for canonical correlation analysis*. Springer, Tokio.

[Anderson, 1984] Anderson, J. A. (1984). Regression and ordered categorical variables. *J. Roy. Statist. Soc. Ser. B*, 46:1–30.

[Anderson, 1951] Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statistics*, pages 327–351.

[Anderson, 1999] Anderson, T. W. (1999). Asymptotic distribution of the reduced rank regression estimator under general conditions. *Ann. Statist.*, 27:1141–1154.

[Anderson, 2003] Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition.

[Bach and Jordan, 2002] Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48.

[Berger, 1982] Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24:295–300.

[Besag, 1974] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B*, pages 192–236. With discussion by D. R. Cox, A. G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J. M. Hammersley, and M. S. Bartlett and with a reply by the author.

[Brock and Durlauf, 2001] Brock, W. and Durlauf, S. (2001). *Interactions Based Models*. Handbook of Econometrics. J. Heckman and E. Learner. Amsterdam: North-Holland.

[Bura and Cook, 2001] Bura, E. and Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 63:393–410.

[Bura and Forzani, 2015] Bura, E. and Forzani, L. (2015). Sufficient Reductions in Regressions With Elliptically Contoured Inverse Predictors. *J. Amer. Statist. Assoc.*, 110:420–434.

[Bura and Yang, 2011] Bura, E. and Yang, J. (2011). Dimension estimation in sufficient dimension reduction: a unifying approach. *J. Multivariate Anal.*, 102:130–142.

[Casella and Berger, 1990] Casella, G. and Berger, R. L. (1990). *Statistical inference*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.

[Cheng et al., 2012] Cheng, J., Levina, E., Wang, P., and Zhu, J. (2012). A sparse ising model with covariates. *Biometrics*, 70:943–953.

[Chiaromonte et al., 2002] Chiaromonte, F., Cook, R. D., and Li, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *Ann. Statist.*, 30:475–497.

[Chikuse, 2002] Chikuse, Y., S. (2002). *Statistics on Spacial Manifolds*. Springer, New York.

[Cook, 1998] Cook, R. D. (1998). *Regression Graphics: Ideas for studying regressions through graphics*. New York: Wiley.

[Cook, 2007] Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.*, 22:1–26.

[Cook and Forzani, 2008] Cook, R. D. and Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statist. Sci.*, 23:485–501.

[Cook and Forzani, 2009] Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *J. Amer. Statist. Assoc.*, 104:197–208.

[Cook et al., 2015] Cook, R. D., Forzani, L., and Zhang, X. (2015). Envelopes and reduced rank-regression. *Biometrika*, 102:439–456.

[Cook and Li, 2002] Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.*, 30:455–474.

[Cook and Li, 2009] Cook, R. D. and Li, L. (2009). Dimension reduction in regressions with exponential family predictors. *J. Comput. Graph. Statist.*, 18:774–791.

[Cook and Ni, 2005] Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *J. Amer. Statist. Assoc.*, 100:410–428.

[Cook and Weisberg, 1991] Cook, R. D. and Weisberg, S. (1991). Discussion of sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, 86:328–332.

[Cook and Weisberg, 1994] Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*. New York: Wiley.

[Cook and Weisberg, 1999] Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. New York: Wiley.

[Dai, 2013] Dai, B. (2013). Mvb: Multivariate bernoulli log-linear model. R package version 1.1.

[Dai et al., 2013] Dai, B., Ding, S., and Wahba, G. (2013). Multivariate Bernoulli distribution. *Bernoulli*, 19:1465–1483.

[Dmitrienko et al., 2010] Dmitrienko, A., Tamhane, A. C., and Bretz, F. (2010). Multiple testing problems in pharmaceutical statistics.

[Eaton, 2007] Eaton, M. L. (2007). *Multivariate statistics*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 53. Institute of Mathematical Statistics, Beachwood, OH. A vector space approach, Reprint of the 1983 original [MR0716321].

[Edelman et al., 1998] Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20.

[Fahrmeir and Kaufmann, 1985] Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.*, 13:342–368.

[Fahrmeir and Kaufmann, 1986] Fahrmeir, L. and Kaufmann, H. (1986). Asymptotic inference in discrete response models. *Statist. Hefte (N.F.)*, 27:179–205.

[Fahrmeir and Tutz, 2001] Fahrmeir, L. and Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models*. Springer Series in Statistics. Springer-Verlag, New York, second edition. With contributions by Wolfgang Hennevogl.

[Forzani, 2007] Forzani, L. (2007). Sufficient dimension reduction based on normal and wishart models. Ph.D thesis, School of Statistics, University of Minnesota.

[Fukumizu et al., 2007] Fukumizu, K., Bach, F. R., and Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *J. Mach. Learn. Res.*, 8:361–383.

[Fukumizu et al., 0304] Fukumizu, K., Bach, F. R., and Jordan, M. I. (2003/04). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.*, 5:73–99.

[Fukumizu et al., 2009] Fukumizu, K., Bach, F. R., and Jordan, M. I. (2009). Kernel dimension reduction in regression. *Ann. Statist.*, 37:1871–1905.

[Green, 1984] Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J. Roy. Statist. Soc. Ser. B*, 46:149–192. With discussion.

[Haberman, 1977] Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.*, 5:815–841.

[Handcock, 2003] Handcock, M. (2003). Statistical models for social networks: inference and degeneracy. *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers. National Academies Press*, pages 191–203.

[Hassner and Sklansky, 1980] Hassner, M. and Sklansky, J. (1980). The use of markov random fields as models of texture. *Computer Graphics Image Processing*, 12:357–370.

[Hsing and Ren, 2009] Hsing, T. and Ren, H. (2009). An RKHS formulation of the inverse regression dimension-reduction problem. *Ann. Statist.*, 37:726–755.

[Ising, 1925] Ising, E. (1925). Beitrag zur theorie des ferromagnetismus [contribution to the theory of ferromagnetism]. *Zeitschrift für Physik*, 31:253–258.

[Izenman, 1975] Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *J. Multivariate Anal.*, pages 248–264.

[Kim and Pavlovic, 2013] Kim, M. and Pavlovic, V. (2013). Central subspace dimensionality reduction using covariance operators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:657–670.

[Lehmann and Casella, 1998] Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.

[Li and Wang, 2007] Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.*, 102:997–1008.

[Li et al., 2005] Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *Ann. Statist.*, 33:1580–1616.

[Li, 1991] Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, 86:316–342. With discussion and a rejoinder by the author.

[Li, 1992] Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Amer. Statist. Assoc.*, 87:1025–1039.

[Lindsey, 1997] Lindsey, J. K. (1997). *Applying Generalized Linear Models*. New York: Springer.

[Magnus, 1988] Magnus, J. R. (1988). *Linear structures*. Charles Griffin & Co., Ltd., London; The Clarendon Press, Oxford University Press, New York.

[Magnus and Neudecker, 1999] Magnus, J. R. and Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and econometrics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester. Revised reprint of the 1988 original.

[Manning and Schutze, 1999] Manning, C. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT press.

[McCulloch and Searle, 2001] McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.

[McCulloch and Nelder, 2001] McCulloch, P. and Nelder, J. A. (2001). Generalized linear models. statistics in the 21st century. *CRC Press LLC and American Statistical Association, Boca Raton, Alexandria*, pages 387–396.

[Muirhead, 1982] Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. John Wiley & Sons, Inc., New York. Wiley Series in Probability and Mathematical Statistics.

[Nelder and Wedderburn, 1972] Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 135:370–384.

[Peng et al., 2009] Peng, J., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.*, 104:735–746.

[Pierce and Cleveland, 1984] Pierce, D. A., G. M. R. and Cleveland, W. P. (1984). Seasonal adjustment of weekly monetary aggregates: A model-based approach. *Journal of Business and Economics Statistics*, 2:260–270.

[Ravikumar et al., 2010] Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.*, 38:1287–1319.

[Reinsel and Velu, 1998] Reinsel, G. C. and Velu, R. P. (1998). *Multivariate reduced-rank regression*. Springer-Verlag, New York.

[Seber, 1977] Seber, G. A. F. (1977). *Linear regression analysis*. John Wiley & Sons, New York-London-Sydney. Wiley Series in Probability and Mathematical Statistics.

[Shapiro, 1985] Shapiro, A. (1985). Asymptotic equivalence of minimum discrepancy function estimators to GLS estimators. *South African Statist. J.*, 19:73–81.

[Shapiro, 1986] Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *J. Amer. Statist. Assoc.*, 81:142–149.

[Spokoiny and Dickhaus, 2015] Spokoiny, V. and Dickhaus, T. (2015). *Basics of modern mathematical statistics*. Springer Texts in Statistics. Springer, Heidelberg.

[Srivastava and Khatri, 1979] Srivastava, M. S. and Khatri, C. G. (1979). *An introduction to multivariate statistics*. North-Holland, New York-Oxford.

[Stoica and Viberg, 1996] Stoica, P. and Viberg, M. (1996). Maximum likelihood parameter and rank estimation in reduced-rank multivariate linear regression. *IEEE Trans. Signal Process.*, 44:3069–3079.

[van der Maaten and van den Herik, 2009] van der Maaten, L., P. E. and van den Herik, H. (2009). Dimensionality reduction: A comparative review. Technical Report TiCC-TR 2009-005, Tilburg University.

[van der Vaart, 1998] van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

[Varshney and Willsky, 2010] Varshney, K. R. and Willsky, A. S. (2010). Classification using geometric level sets. *J. Mach. Learn. Res.*, 11.

[Wainwright and Jordan, 2008] Wainwright, M. and Jordan, M. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305.

[Wang et al., 2011] Wang, P., Chao, D. L., and Hsu, L. (2011). Learning oncogenic pathways from binary genomic instability data. *Biometrics*, 67:164–173.

[Wasserman and Pattison, 1996] Wasserman, S. and Pattison, P. (1996). Logit models and logistic regressions for social networks. I. An introduction to Markov graphs and *p*. *Psychometrika*, 61:401–425.

[Wedderburn, 1976] Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63:27–32.

[Westfall and Krishen, 2001] Westfall, P. H. and Krishen, A. (2001). Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *J. Statist. Plann. Inference*, 99:25–40.

[Woods, 1978] Woods, J. (1978). Markov image modeling. *IEEE Transactions on Automatic Control*, 23:846–850.

[Wu, 2008] Wu, Q. (2008). Kernel sliced inverse regression with applications to classification. *J. Comput. Graph. Statist.*, 17:590–610.

[Wu et al., 2008] Wu, Q., Liang, F., and Mukherjee, S. (2008). Regularized sliced inverse regression for kernel models. Technical report, Duke Univ., Durham, NC.

[Xia et al., 2002] Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64:363–410.

[Yee and Hastie, 2003] Yee, T. W. and Hastie, T. J. (2003). Reduced-rank vector generalized linear models. *Stat. Model.*, 3:15–41.

[Yee and Wild, 1996] Yee, T. W. and Wild, C. J. (1996). Vector generalized additive models. *J. Roy. Statist. Soc. Ser. B*, 58:481–493.

[Yeh et al., 2009] Yeh, Y. R., Huang, S. Y., and Lee, Y. Y. (2009). Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Transactions on Knowledge and Data Engineering*, 21:1590–1031.

[Zhu and Li, 2011] Zhu, H. and Li, L. (2011). Biological pathway selection through nonlinear dimension reduction. *Biostatistics*, 12:429–444.

[Zhu and Zeng, 2006] Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Amer. Statist. Assoc.*, 101:1638–1651.

APÉNDICE A

RESULTADOS Y HERRAMIENTAS ÚTILES

En este capítulo presentaremos propiedades y resultados que serán utilizados a lo largo de toda la tesis.

Propiedad A.1. $\text{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\text{vec}(\mathbf{B})$ *para matrices* $\mathbf{A}$, $\mathbf{B}$ *y* $\mathbf{C}$ *tales que se puedan multiplicar.*

Propiedad A.2.

1. *Si* $\mathbf{M} : m \times n$ *es de rango completo* m*, luego* $\mathbf{M}^\dagger = \mathbf{M}^T(\mathbf{M}\mathbf{M}^T)^{-1}$.
2. *Si* $\mathbf{M} : m \times n$ *es de rango completo* n*, luego* $\mathbf{M}^\dagger = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$.
3. *Si* $\mathbf{A}$ *o* $\mathbf{B}$ *es de rango completo,* $(\mathbf{AB})^\dagger = \mathbf{B}^\dagger\mathbf{A}^\dagger$.

Propiedad A.3.

1. $\mathbf{K}_{pd}^T = \mathbf{K}_{dp}$.
2. $\mathbf{K}_{pd}^T\mathbf{K}_{pd} = \mathbf{K}_{pd}\mathbf{K}_{pd}^T = \mathbf{I}_{pd}$.
3. *Si* $\mathbf{A} : r_1 \times r_2$*, y* $\mathbf{B} : r_3 \times r_4$*, luego*

$$r_3 r_1(\mathbf{A} \otimes \mathbf{B})\mathbf{K}_{r_2 r_4} = \mathbf{B} \otimes \mathbf{A},$$

   y por el item 1. y 2. también es cierto que

$$r_3 r_1(\mathbf{A} \otimes \mathbf{B}) = (\mathbf{B} \otimes \mathbf{A})\mathbf{K}_{r_4 r_2}.$$

4. $\frac{1}{2}(\mathbf{I}_{r^2} + \mathbf{K}_{rr})(\mathbf{M} \otimes \mathbf{M})\frac{1}{2}(\mathbf{I}_{r^2} + \mathbf{K}_{rr}) = \frac{1}{2}(\mathbf{I}_{r^2} + \mathbf{K}_{rr})(\mathbf{M} \otimes \mathbf{M}) = (\mathbf{M} \otimes \mathbf{M})\frac{1}{2}(\mathbf{I}_{r^2} + \mathbf{K}_{rr})$ *para cualquier matriz* $\mathbf{M} : r \times r$.

Propiedad A.4.

1. $\mathbf{D}_r\mathbf{E}_r = \frac{1}{2}(\mathbf{I}_{r^2} + \mathbf{K}_{rr})$.

2. $\mathbf{K}_{rr}\mathbf{D}_r = \mathbf{D}_r$.
3. $(\mathbf{D}_r^T(\mathbf{M}\otimes\mathbf{M})\mathbf{D}_r)^{-1} = \mathbf{E}_r(\mathbf{M}^{-1}\otimes\mathbf{M}^{-1})\mathbf{E}_r^T$ *para cualquier matriz* $\mathbf{M}$ *invertible.*
4. $(\mathbf{E}_r(\mathbf{M}\otimes\mathbf{M})\mathbf{D}_r)^{-1} = \mathbf{E}_r(\mathbf{M}^{-1}\otimes\mathbf{M}^{-1})\mathbf{D}_r$ *para cualquier matriz* $\mathbf{M}$ *invertible.*

Propiedad A.5. ***Identidades de Woodbury.***

$$\begin{aligned}(\mathbf{A}+\mathbf{C}^T\mathbf{B}\mathbf{C})^{-1} &= \mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{C}(\mathbf{B}^{-1}+\mathbf{C}^T\mathbf{A}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{A}^{-1}\\ (\mathbf{A}+\mathbf{U}\mathbf{B}\mathbf{V})^{-1} &= \mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{U}(\mathbf{B}^{-1}+\mathbf{V}\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}\mathbf{A}^{-1}.\end{aligned}$$

Propiedad A.6.

1. *Sea* $\mathbf{X}$ *una matriz de dimensiones arbitrarias,* $\mathbf{F}(\mathbf{X}) : m\times p$ *y* $\mathbf{G}(\mathbf{X}) : p\times q$ *funciones diferenciables de* $\mathbf{X}$. *Luego*

$$\frac{\partial \text{vec}[\mathbf{F}(\mathbf{X})\mathbf{G}(\mathbf{X})]}{\partial \text{vec}^T(\mathbf{X})} = (\mathbf{G}^T\otimes\mathbf{I}_m)\frac{\partial \text{vec}[\mathbf{F}(\mathbf{X})]}{\partial \text{vec}^T(\mathbf{X})} + (\mathbf{I}_q\otimes\mathbf{F})\frac{\partial \text{vec}[\mathbf{G}(\mathbf{X})]}{\partial \text{vec}^T(\mathbf{X})}. \tag{A.1}$$

2. *Supongamos que* $\mathbf{F}(\mathbf{X}) = \mathbf{X}^T$ *y* $\mathbf{G}(\mathbf{X}) = \mathbf{X}$ *con* $\mathbf{X} : p\times q$, *aplicando la propiedad anterior*

$$\frac{\partial \text{vec}(\mathbf{X}^T\mathbf{X})}{\partial^T \text{vec}(\mathbf{X})} = (\mathbf{I}_{q^2}+\mathbf{K}_{qq})(\mathbf{I}_q\otimes\mathbf{X}^T) \tag{A.2}$$

$$\frac{\partial \text{vec}(\mathbf{X}^T\mathbf{X})^{-1}}{\partial^T \text{vec}(\mathbf{X})} = -((\mathbf{X}^T\mathbf{X})^{-1}\otimes(\mathbf{X}^T\mathbf{X})^{-1}))\frac{\partial \text{vec}(\mathbf{X}^T\mathbf{X})}{\partial^T \text{vec}(\mathbf{X})}. \tag{A.3}$$

Lema A.7. *Sea* $\mathbf{\Gamma}$ *una matriz* $p\times d$ *de rango* d *y* $\mathbf{P}_{\mathbf{\Gamma}} = \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T$ *la matriz de poyección sobre el espacio generado por las columnas de* $\mathbf{\Gamma}$. *Luego,*

$$\frac{\partial \mathbf{P}_{\mathbf{\Gamma}}}{\partial \text{vec}^T(\mathbf{\Gamma})} = (\mathbf{I}_{p^2}+\mathbf{K}_{pp})(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\otimes\mathbf{Q}_{\mathbf{\Gamma}}) \tag{A.4}$$

donde $\mathbf{K}_{pp}$ *es la matriz de permutación de dimensión* $p^2\times p^2$.

Demostración. Aplicando la igualdad (A.1) tenemos

$$\begin{aligned}\frac{\partial \text{vec}\mathbf{P}_{\mathbf{\Gamma}}}{\partial \text{vec}^T(\mathbf{\Gamma})} &= \frac{\partial \text{vec}(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T)}{\partial \text{vec}^T(\mathbf{\Gamma})}\\ &= (\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\otimes\mathbf{I}_p)\frac{\partial \text{vec}(\mathbf{\Gamma})}{\partial \text{vec}^T(\mathbf{\Gamma})} + (\mathbf{I}_p\otimes\mathbf{\Gamma})\frac{\partial \text{vec}((\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T)}{\partial \text{vec}^T(\mathbf{\Gamma})}\\ &= (\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\otimes\mathbf{I}_p) + (\mathbf{I}_p\otimes\mathbf{\Gamma})\frac{\partial \text{vec}((\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T)}{\partial \text{vec}^T(\mathbf{\Gamma})}.\end{aligned}$$

Sea $\mathbf{H} = \dfrac{\partial \text{vec}((\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T)}{\partial \text{vec}^T(\mathbf{\Gamma})}$, luego aplicando (A.1), (A.2) y (A.3)

$$\begin{aligned}\mathbf{H} &= (\mathbf{\Gamma}\otimes\mathbf{I}_d)\frac{\partial \text{vec}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}}{\partial \text{vec}^T(\mathbf{\Gamma})} + (\mathbf{I}_p\otimes(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1})\mathbf{K}_{pd}\\ &= -(\mathbf{\Gamma}\otimes\mathbf{I}_d)((\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\otimes(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}))(\mathbf{I}_{d^2}+\mathbf{K}_{dd})(\mathbf{I}_d\otimes\Gamma^T) + (\mathbf{I}_p\otimes(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1})\mathbf{K}_{pd}\end{aligned}$$

Luego,

$$\begin{aligned}
\frac{\partial \text{vec}\mathbf{P}_{\mathbf{\Gamma}}}{\partial \text{vec}^T(\mathbf{\Gamma})} &= (\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes \mathbf{I}_p) + (\mathbf{I}_p \otimes \mathbf{\Gamma})[-(\mathbf{\Gamma} \otimes \mathbf{I}_d)((\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes (\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1})(\mathbf{I}_{d^2} + \mathbf{K}_{dd})(\mathbf{I}_d \otimes \mathbf{\Gamma}^T) \\
&\quad +(\mathbf{I}_p \otimes (\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1})\mathbf{K}_{pd}] \\
&= (\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes \mathbf{I}_p) + (\mathbf{I}_p \otimes \mathbf{\Gamma}(\mathbf{\Gamma}\mathbf{\Gamma})^{-1})\mathbf{K}_{pd} - (\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T) \\
&\quad -(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1})\mathbf{K}_{dd}(\mathbf{I}_d \otimes \mathbf{\Gamma}^T) \\
&= (\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes \mathbf{I}_p) - (\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T) \\
&\quad -(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1})(\mathbf{\Gamma}^T \otimes \mathbf{I}_d)\mathbf{K}_{pd} \\
&= (\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes \mathbf{I}_p) - (\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes \mathbf{P}_{\mathbf{\Gamma}}) \\
&= (\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes \mathbf{I}_p - \mathbf{P}_{\mathbf{\Gamma}}) \\
&= (\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1} \otimes \mathbf{Q}_{\mathbf{\Gamma}}).
\end{aligned}$$

□

Lema A.8. *Sea* $\mathbf{\Sigma}$ *la matriz de covarianza del vector* $\mathbf{x}^T = (\mathbf{x}_1^T, \mathbf{x}_2^T)$ *y consideremos la partición de* $\mathbf{\Sigma}$ *conforme a la partición de* $\mathbf{x}^T$*:*

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}.$$

Luego,

$$\mathbf{\Sigma}^{\ddagger} = \begin{pmatrix} \mathbf{\Sigma}_{11.2}^{\ddagger} & -\mathbf{\Sigma}_{11.2}^{\ddagger}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{\ddagger} \\ -\mathbf{\Sigma}_{22}^{\ddagger}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11.2}^{\ddagger} & \mathbf{\Sigma}_{22}^{\ddagger} + \mathbf{\Sigma}_{22}^{\ddagger}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11.2}^{\ddagger}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{\ddagger} \end{pmatrix}$$

es una inversa generalizada de la matriz $\mathbf{\Sigma}$*, donde* $\mathbf{\Sigma}_{11.2} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{\ddagger}\mathbf{\Sigma}_{21}$*.*

Demostración. De acuerdo a la definición de inversa generalizada, tenemos que ver que se cumpla $\mathbf{\Sigma}\mathbf{\Sigma}^{\dagger}\mathbf{\Sigma} = \mathbf{\Sigma}$.

Para ello vamos a considerar la siguiente propiedad (Proposición 2.16 de [**Eaton, 2007**]):

$$\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{\dagger}\mathbf{\Sigma}_{22} = \mathbf{\Sigma}_{12} \tag{A.5}$$

Luego, aplicando (A.5) tenemos que

$$\begin{aligned}
\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\dagger} &= \begin{pmatrix} \boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{11.2}^{\dagger} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{\dagger}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11.2}^{\dagger} & -\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{11.2}^{\dagger}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{\dagger} + \boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22}^{\dagger} + \boldsymbol{\Sigma}_{22}^{\dagger}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11.2}^{\dagger}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{\dagger}) \\ \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11.2}^{\dagger} - \boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{\dagger}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11.2}^{\dagger} & -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11.2}^{\dagger}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{\dagger} + \boldsymbol{\Sigma}_{22}(\boldsymbol{\Sigma}_{22}^{\dagger} + \boldsymbol{\Sigma}_{22}^{\dagger}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11.2}^{\dagger}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{\dagger}) \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{\Sigma}_{11.2}\boldsymbol{\Sigma}_{11.2}^{\dagger} & -\boldsymbol{\Sigma}_{11.2}\boldsymbol{\Sigma}_{11.2}^{\dagger}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{\dagger} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{\dagger} \\ 0 & \boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{\dagger} \end{pmatrix}
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\dagger}\boldsymbol{\Sigma} &= \begin{pmatrix} \boldsymbol{\Sigma}_{11.2}\boldsymbol{\Sigma}_{11.2}^{\dagger}\boldsymbol{\Sigma}_{11} - (\boldsymbol{\Sigma}_{11.2}\boldsymbol{\Sigma}_{11.2}^{\dagger}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{\dagger} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{\dagger})\boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{\dagger}\boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{\dagger}\boldsymbol{\Sigma}_{22} \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{\Sigma}_{11.2}\boldsymbol{\Sigma}_{11.2}^{\dagger}\boldsymbol{\Sigma}_{11.2} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{\dagger}\boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} = \boldsymbol{\Sigma}
\end{aligned}$$

donde para el bloque superior izquierdo utilizamos que

$$\begin{aligned}
(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\dagger}\boldsymbol{\Sigma})_{12} &= \boldsymbol{\Sigma}_{11.2}\boldsymbol{\Sigma}_{11.2}^{\dagger}\boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_{11.2}\boldsymbol{\Sigma}_{11.2}^{\dagger}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{\dagger}\boldsymbol{\Sigma}_{22} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{\dagger}\boldsymbol{\Sigma}_{22} \\
&= \boldsymbol{\Sigma}_{11.2}\boldsymbol{\Sigma}_{11.2}^{\dagger}\boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_{11.2}\boldsymbol{\Sigma}_{11.2}^{\dagger}\boldsymbol{\Sigma}_{12} + \boldsymbol{\Sigma}_{12} \\
&= \boldsymbol{\Sigma}_{12}.
\end{aligned}$$

□

Lema A.9. *Para matrices arbitrarias* $\mathbf{P}$ *y* $\mathbf{Q}$ *tal que* $\mathbf{P}\mathbf{X}^T\mathbf{X} = \mathbf{Q}\mathbf{X}^T\mathbf{X}$*, se tiene que* $\mathbf{P}\mathbf{X}^T = \mathbf{Q}\mathbf{X}^T$*.*

Demostración. Observemos que

$$\begin{aligned}
(\mathbf{P}\mathbf{X}^T - \mathbf{Q}\mathbf{X}^T)(\mathbf{P}\mathbf{X}^T - \mathbf{Q}\mathbf{X}^T)^T &= \mathbf{P}\mathbf{X}^T\mathbf{X}\mathbf{P}^T - \mathbf{P}\mathbf{X}^T\mathbf{X}\mathbf{Q}^T - \mathbf{Q}\mathbf{X}^T\mathbf{X}\mathbf{P}^T + \mathbf{Q}\mathbf{X}^T\mathbf{X}\mathbf{Q}^T \\
&= (\mathbf{P}\mathbf{X}^T\mathbf{X} - \mathbf{Q}\mathbf{X}^T\mathbf{X})\mathbf{P}^T - (\mathbf{P}\mathbf{X}^T\mathbf{X} - \mathbf{Q}\mathbf{X}^T\mathbf{X})\mathbf{Q}^T \\
&= (\mathbf{P}\mathbf{X}^T\mathbf{X} - \mathbf{Q}\mathbf{X}^T\mathbf{X})(\mathbf{P} - \mathbf{Q})^T = 0.
\end{aligned}$$

Cada elemento de la diagonal de $(\mathbf{P}\mathbf{X}^T - \mathbf{Q}\mathbf{X}^T)(\mathbf{P}\mathbf{X}^T - \mathbf{Q}\mathbf{X}^T)^T$ es la suma de los cuadrados de los elementos de cada columna de la matriz $\mathbf{P}\mathbf{X}^T - \mathbf{Q}\mathbf{X}^T$ respectivamente. Es decir que $\mathbf{P}\mathbf{X}^T - \mathbf{Q}\mathbf{X}^T = 0$. □

Lema A.10. *Sea* $\mathbf{G}$ *una inversa generalizada de la matriz* $\mathbf{X}^T\mathbf{X}$*, entonces se tiene:*

(i) $\mathbf{G}^T$ *también es una inversa generalizada de* $\mathbf{X}^T\mathbf{X}$.

(ii) $\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{X} = \mathbf{X}$, *es decir* $\mathbf{G}\mathbf{X}^T$ *es una inversa generalizada de* $\mathbf{X}$.

(iii) $\mathbf{X}\mathbf{G}\mathbf{X}^T$ *es invariante respecto a* $\mathbf{G}$.

Demostración. De acuerdo a la definición de inversa generalizada, $\mathbf{X}^T\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{X} = \mathbf{X}^T\mathbf{X}$ y solamente basta trasponer para obtener $\mathbf{X}^T\mathbf{X}\mathbf{G}^T\mathbf{X}^T\mathbf{X} = \mathbf{X}^T\mathbf{X}$ por lo que es cierto **(i)**. Luego, si aplicamos a esta última igualdad el Lema A.9 con $\mathbf{P} = \mathbf{X}^T\mathbf{X}\mathbf{G}^T$ y $\mathbf{Q} = \mathbf{I}$ obtenemos **(ii)**.

Para ver que vale **(iii)**, sea F otra inversa generalizada de $\mathbf{X}^T\mathbf{X}$. Luego por **(ii)** tenemos que $\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{X} = \mathbf{X} = \mathbf{X}\mathbf{F}\mathbf{X}^T\mathbf{X}$, lo que implica $\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{X} = \mathbf{X}\mathbf{F}\mathbf{X}^T\mathbf{X}$. Aplicando nuevamente el Lema A.9 con $\mathbf{P} = \mathbf{X}\mathbf{G}$ y $\mathbf{Q} = \mathbf{X}\mathbf{F}$ obtenemos que $\mathbf{X}\mathbf{G}\mathbf{X}^T = \mathbf{X}\mathbf{F}\mathbf{X}^T$. □

Propiedad A.11. *Sean* $\mathbf{A}$ *y* $\mathbf{B}$ *dos matrices* $m \times m$ *tal que* $\mathbf{A} \leq \mathbf{B}$, *luego para cualquier matriz* $\mathbf{M} : m \times n$ *se tiene que* $\mathbf{M}^T\mathbf{A}\mathbf{M} \leq \mathbf{M}^T\mathbf{B}\mathbf{M}$.