UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Estimación robusta para un modelo de reducción de la dimensión

Tesis presentada para optar al título de Doctor de la Universidad de Buenos Aires en el área Ciencias Matemáticas

María Eugenia Szretter Noste

Director de tesis: Dr. Víctor J. Yohai
Consejero de estudios: Dr. Pablo A. Ferrari

Lugar de trabajo: Instituto de Cálculo, FCEyN, UBA

Buenos Aires, 1 de julio de 2017

Fecha de defensa: 6 de octubre de 2017

Estimación robusta para un modelo de reducción de la dimensión

Resumen

Las técnicas no paramétricas son herramientas flexibles para poder estudiar la relación entre una variable aleatoria continua y un vector de covariables explicativas, pero para su aplicación requieren un número de observaciones que crece exponencialmente con la dimensión p del vector de covariables.

Una manera de poder enfrentar la estimación no–paramétrica con una muestra de tamaño moderado y p grande, es obtener un número reducido de nuevas variables explicativas sin disminuir la información que ellas provean sobre la variable respuesta. Las diversas estrategias para encarar este problema se engloban en lo que se denomina *reducción de la dimensión.*

Cook [2007] introduce el concepto de *reducción suficiente,* y el modelo de *principal fitted components* (PFC). Cook y Forzani [2008] calcula el estimador de máxima verosimilitud (MV) para el modelo PFC suponiendo que los errores tienen distribución normal multivariada. Sin embargo, cuando hay contaminación o la distribución de los errores no es normal multivariada, los estimadores de MV se ven muy afectados y por lo tanto en estos casos, pueden ser muy poco informativos.

En esta tesis proponemos estimadores robustos de tipo τ para estimar el modelo PFC y por consiguiente para la correspondiente reducción suficiente. Estos estimadores están basados en una $\tau-$escala (ver Yohai y Zamar [1988]).

Definimos el $\tau-$funcional de estimación del cual se derivan los $\tau-$estimadores propuestos. Se demuestra que bajo condiciones generales estos estimadores son fuertemente consistentes.

A partir de las ecuaciones de estimación se obtiene una expresión para los $\tau-$estimadores similar a la de MV, excepto que las observaciones aparecen acompañadas por pesos que, a su vez, dependen de los parámetros. Esto sugiere un algoritmo iterativo natural para computar los τ. También se discute cómo obtener valores iniciales para este algoritmo.

Un estudio de Monte Carlo permite comparar los τ-estimadores y los estimadores de MV bajo el modelo PFC y bajo contaminación por outliers. Los resultados de esta simulación muestran claras ventajas para los τ-estimadores. También se presenta una propuesta de selección de la dimensión del espacio de reducción basada en validación cruzada.

Finalmente, ilustramos la aplicación del método con dos ejemplos de datos reales. Las demostraciones de los resultados se presentan en varios apéndices.

Palabras clave: reducción de la dimensión, *principal fitted components*, estimación robusta, $\tau-$estimador, reducción suficiente, regresión inversa, selección de la dimensión.

Robust estimation for a dimension reduction model

Abstract

Non-parametric estimating procedures are flexible tools to study the relationship between a continuous response and a vector of explanatory variables. However these methods require a number of observations that grows exponentially with the number of explanatory variables. One way to overcome this situation is to obtain a reduced number of new variables that contain the same information on the response than the original ones.

Several strategies have been proposed to achieve this *dimension reduction.* Cook [2007] introduces the concept of *sufficient reduction* and the *principal fitted components* (PFC) model. Cook and Forzani [2008] obtain the maximum likelihood (ML) for the PFC model assuming that the error distribution is multivariate normal. However these estimators may be much affected for outlier contamination or a non normal error distribution, and therefore, when this occurs, they may not be much informative.

In this thesis we propose robust estimators for the parameters of the PFC model based on a τ-scale (see Yohai and Zamar [1988]), and therefore we provide robust estimators for the sufficient reduction too.

We define the $\tau-$estimating functional that generate the $\tau-$estimators for the PFC model. We prove that under general assumptions this $\tau-$estimators are strongly consistent.

We obtain the estimating equations that characterize the critical points of the $\tau-$estimator. Using these equations we can express the $\tau-$estimators as a weighted ML estimator where the weight for each observation depends on the parameters. This expression suggests a natural iterative algorithm to compute the $\tau-$estimators. We also discuss how to obtain starting values for the parameters of the algorithm.

We perform a simulation study to compare the $\tau-$ and ML$-$estimators. The simulation results show that the $\tau-$estimators compare favorably with respect to the ML$-$estimator. We also propose a procedure based on cross validation to choose the dimension of the sufficient reduction.

Finally,we illustrate the advantages of the proposed estimation procedure using two real datasets. The proofs of the main results can be found in several appendices.

Key words: dimension reduction, principal fitted components, robust estimation, $\tau-$estimator, sufficient reduction, inverse regression, dimension selection.

Agradecimientos

A Víctor, el director de esta tesis. Especialmente por contagiarme su pasión para pensar. Y también por escuchar mis incansables preguntas con paciencia y responderlas con entusiasmo. Por impulsar este proyecto desde su inicio, con más certezas que las mías. Es un lujo aprender a su lado, con la libertad para debatir y proponer que él genera. Por su entereza como persona, también. Gracias.

A mi familia cotidiana, Andrés, Catalina, Iván, León y Nicolás. Porque la alegría que irradian es mi sostén indispensable, y sin ustedes no habría nada.

A mis papás, porque ellos me hicieron lo que soy, su amor incondicional y apoyo continuo me trajeron (feliz) hasta acá. Porque me enseñaron, con el ejemplo, a hacer las cosas lo mejor posible, siempre. Porque sembraron y estimularon la curiosidad como método, desde donde nacen mis recuerdos. Por haber priorizado a la familia en sus vidas. A mis abuelos y bisabuelos, porque su trabajo constante y sostenido fue abriendo el camino para que sus descendientes podamos disfrutar de una vida más cómoda y fácil. Porque ellos cosecharon en el campo, lavaron a mano, zurcieron, aprendieron el lenguaje, y trabajaron duro.

A mi hermana y mis sobrinas, por ser los tres soles que son. Por sus risas, sus programas, sus canciones y sus cuentos (no por sus chistes...) Por el amor que contagian.

A Andrea, porque nada supera a trabajar con una amiga. Porque pude entender muchas de las cosas que están acá escritas gracias a que las pensé con ella. Por el buen humor que imperó, a pesar de que las cuentas muchas veces nos fueron esquivas. Porque esta tesis no hubiera empezado sin su impulso. A Mariela y a Andrés, que escucharon mis dudas, leyeron mis cabos sueltos, pensaron conmigo muchas veces y me sugirieron muchas cosas. Admiro la gran soltura con la que ambos encaran los problemas, y la creatividad y energía sostenida que le inyectan a las cosas que encaran. A Dani Carando que una tarde resolvió los detalles analíticos que no me cerraban.

Al Departamento de Matemática y al Instituto de Cálculo de la FCEN. Al DM por haberme formado académicamente, agradezco la calidad y dedicación de sus profesores y auxiliares, que me enseñaron tanto. Además, le agradezco el cargo docente que me permitió mantenerme durante la concreción de este trabajo. Al IC por ser mi segunda casa este tiempo, al que le agradezco también haberse vuelto cada vez más interdisciplinario.

A mis compañeros de trabajo en el Instituto de Cálculo, por hacer que el día a día sea llevadero, Lucía, Marina, Daniela, Mariela, Inés, Agustín, Gonzalo, Manuel, Ezequiel, Pablo V, Leo, Pablo T, Stella, Julieta, Paula, Vero, Flor F, Flor S, y demás, que se me traspapelan. A Lucía, además, por compartir la oficina, las galletitas y las cotidianeidades. Por aguantarse tantas horas de Skype sin chistar.

Al grupo de estadística por ser un espacio permanente para conversar las dudas, y debatir los puntos de vista. Por la generosidad intelectual que siempre encuentro en sus miembros. También por ser un ámbito donde la igualdad de género no necesita imponerse porque es constitutiva.

A los probabilistas de la facultad, todos, con los que empecé este camino, me acompañaron y luego supieron entender el cambio de rumbo académico, con la sencillez que tienen para hacer las cosas. En especial a Pablo Ferrari.

A mis amigos de siempre. Por la alegría de que la vida nos vaya cambiando juntos. A Las difíciles, por ser inclaudicablemente tenaces y tremendamente diversas, pero siempre presentes: Chichi, Maru, Carla, Ely, Martuli, Gabi, Jose, Chole, Ceci, Caro. A Gaby G, por estar siempre cerquita y ser tan necesaria. A los Olivas, por la forma en que nuestros hijos crecen juntos. A los amigos que me dio la matemática: Eda, Manu, Vero, Dani C., Patu, Santi S., Mariela, Dani R, Pablo M. A sus parejas. A sus hijos. Algo ha de tener la matemática, para amontonar a personas tan valiosas, con las que trabajar ha sido siempre un placer, tanto como es disfrutar de los diversos programas que nos juntan por fuera de ella.

Al Jardín de Infantes "Mi Pequeña Ciudad", de la Facultad de Ciencias Exactas y Naturales, a todo su personal docente y no docente, que al cuidar, educar y querer a mis hijos permitió que yo pudiera dedicarme a trabajar en mi doctorado dos pisos más arriba. La tranquilidad de saberlos tan cerca, tan cuidados y mimados, es uno de los principales motivos que me posibilitaron concretar este plan. Ojalá todos los trabajadores y estudiantes de la FCEyN con hijos chiquitos tuvieran la oportunidad de disfrutarlo para sus hijos. El jardín aumenta aún más el orgullo de pertenecer a la Universidad pública que me formó y en la que trabajo. Agradezco también a la otra mirada que el jardín me dio de la facultad, de sus miembros. Y a la red de cariño y contención que tejimos con las distintas familias, los Souto, los Pastorino, los Kamienkowski, los Williams, los Pellegrino, por nombrar sólo los más cercanos.

A la escuela de mis hijos, que en esta última etapa se encarga de educarlos y estimularlos.

Y en especial a Andrés, porque cuando hay hijos chicos en la familia, los proyectos o son compartidos, o no son. Por haberme alentado permanentemente, y acompañado diariamente. "Porque estás en mis mañanas, y me río con vos".

Índice

1. Subespacios de Reducción

Sea $(\mathbf{x},y) \in \mathbb{R}^{p+1}$ un vector aleatorio, con $\mathbf{x} \in \mathbb{R}^p$. En muchas aplicaciones, el interés está en el modelado de la relación entre $\mathbf{x}$ e y. Para esto, un modelo frecuentemente usado es

$$y = g(\mathbf{x},\varepsilon)$$

donde g es una función desconocida y ε, el término del error, no es observable, pero se asume independiente de $\mathbf{x}$. La estimación de g se puede realizar a través de la regresión no paramétrica. Las técnicas de regresión no paramétrica permiten estimar a esta función g a partir de una muestra $\{(\mathbf{x}_i, y_i)\}$, con $1 \leq i \leq n$. Son herramientas más flexibles que los modelos más clásicos, como el modelo de regresión lineal, que imponen alguna estructura (en este caso, la linealidad) a la función g. Las técnicas no paramétricas (o también denominadas técnicas de suavizado no paramétrico) estiman la $E(y|\mathbf{x} = \mathbf{x}_0)$ a través de un promedio ponderado de los valores de la variable y observados en la muestra, donde los pesos o ponderadores de un valor y_i observado dependen, esencialmente, de la distancia entre $\mathbf{x}_0$ (el valor alrededor del cual uno quiere estimar) y $\mathbf{x}_i$ (el valor de las covariables observadas simultáneamente con y_i). Desafortunadamente, el tamaño de muestra n necesario para producir un suavizado no paramétrico apropiado es exponencial en p, el número de covariables. En las aplicaciones, no suele disponerse de esta cantidad de observaciones. En estadística, esto se conoce como la *maldición de la dimensionalidad.* Sin embargo, en muchos casos se tiene información redundante en las covariables, en el sentido de que existe una función $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^d$ con $d \leq p$ tal que la relación entre y y $\mathbf{x}$ puede ser modelada adecuadamente por

$$y = g_1(\mathbf{R}(\mathbf{x}), \varepsilon_1) \tag{1}$$

con ε_1 y $R(\mathbf{x})$ independientes. Si d es suficientemente pequeño podemos estimar el modelo (1) de manera adecuada con técnicas de regresión no paramétrica. El objetivo de este trabajo es estudiar y estimar a $\mathbf{R}$, dejando las cuestiones de cómo estimar a g_1 o describir sus características principales para un trabajo futuro.

Podemos formalizar esta idea de la siguiente manera.

Definición 1.1 *Una* ***reducción*** *es una función medible* $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^d$ *con* $d \leq p$.

Definición 1.2 *Los vectores aleatorios* $\mathbf{u}$ *y* $\mathbf{v}$ ***son condicionalmente independientes dado*** $\mathbf{z}$ *si para todas las funciones* f, g *tales que* $f(\mathbf{u}), g(\mathbf{v}) \in L^\infty(\Omega)$ *se cumple que*

$$E[f(\mathbf{u})\, g(\mathbf{v}) \mid \mathbf{z}] = E[f(\mathbf{u}) \mid \mathbf{z}]\, E[g(\mathbf{v}) \mid \mathbf{z}].$$

Lo notaremos de la siguiente forma, $\mathbf{u} \perp\!\!\!\perp \mathbf{v} \mid \mathbf{z}$.

En el Apéndice A.1 hacemos una introducción a la independencia condicional, y enunciamos las principales propiedades de ella.

Definición 1.3 *Una reducción* $\mathbf{R}$ *se dice* ***suficiente*** *si* $\mathbf{x} \perp\!\!\!\perp y \mid \mathbf{R}(\mathbf{x})$.

Observación 1.1 *Usando el Teorema A.5, podemos expresar el hecho de que* $\mathbf{R}$ *sea una reducción suficiente de las siguientes formas equivalentes*

i. $\mathbf{x} \perp\!\!\!\perp y \mid \mathbf{R}(\mathbf{x})$

ii. $y \mid \mathbf{x} \sim y \mid \mathbf{R}(\mathbf{x})$

iii. $\mathbf{x} \mid (y, \mathbf{R}(\mathbf{x})) \sim \mathbf{x} \mid \mathbf{R}(\mathbf{x})$

De acuerdo a *(ii)* podemos decir que una reducción suficiente $\mathbf{R}(\mathbf{x})$ contiene toda la información que $\mathbf{x}$ tiene acerca de y.

La siguiente definición corresponde a la reducción suficiente más simple que se puede obtener.

Definición 1.4 *Sea* $\mathbf{R}(\mathbf{x})$ *una reducción suficiente. Diremos que es* ***minimal*** *si para toda otra reducción suficiente* $\mathbf{T}(\mathbf{x})$ *resulta que* $\mathbf{R}(\mathbf{x})$ *es una función de* $\mathbf{T}(\mathbf{x})$.

Las técnicas de regresión implican el modelado (y la estimación) de lo que se suele llamar la reducción directa (*forward reduction*), expresada por la distribución de $y \mid \mathbf{x}$ (o alguna medida resumen de la misma que sea de interés). Por la Observación 1.1(ii), por supuesto, esto puede ser equivalentemente escrito en términos de modelar o resumir la distribución conjunta de $(\mathbf{x}, y)$ o de la reducción inversa, que es la distribución de $\mathbf{x} \mid y$. Esta libertad hace posible utilizar diferentes enfoques para encontrar reducciones suficientes. En la regresión directa, el modelo de regresión lineal estándar es el ejemplo de reducción más utilizado. En este modelo se propone que

$$y = \boldsymbol{\beta}^T \mathbf{x} + \varepsilon$$

con ε y $\mathbf{x}$ independientes, $E(\varepsilon) = 0$ y $\boldsymbol{\beta} \in \mathbb{R}^p$ un vector desconocido. Al modelar el vínculo entre $\mathbf{x}$ e y de esta forma, estamos imponiendo que $\mathbf{R}(\mathbf{x}) = \boldsymbol{\beta}^T \mathbf{x}$ $(d = 1)$ sea una reducción suficiente puesto que $y \mid \mathbf{x} \sim y \mid \mathbf{R}(\mathbf{x})$. La modelización de reducciones inversas, es decir, de la distribución de $\mathbf{x} \mid y$, se ha desarrollado mucho en los últimos años. Puede lidiar en forma sencilla con $d > 1$ y y resuelve los problemas vinculados con la maldición de la dimensionalidad cuando p es grande. Podemos mencionar dos enfoques generales para encontrar reducciones suficientes modelando la regresión inversa. El primero (que históricamente también surgió primero) es lo que se conoce como la *reducción inversa basada en momentos (moment-based approach)*. Este enfoque utiliza los momentos de las variables para estimar las reducciones suficientes. El segundo enfoque está basado en la *modelización de la regresión inversa (model-based inverse reduction)*, en la que se especifica un modelo paramétrico para la relación inversa de $\mathbf{x}$ en términos de y. En las Secciones 1.2 y 1.3 presentamos brevemente ambos enfoques.

1.0.1. Relación entre estadístico y reducción suficiente

En estadística, en un encuadre más general, a partir de una muestra $\mathbf{x}_1, \ldots, \mathbf{x}_n$ de vectores o variables aleatorios i.i.d. con distribución en la familia $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$ cualquier estadístico $T(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ define una forma de reducir la información contenida en la muestra.

Informalmente, un *estadístico* T es *suficiente* para θ si captura toda la información que la muestra pueda tener sobre el θ. Esta noción fue introducida por Fisher en 1922. La definición formal es la siguiente.

Definición 1.5 *Un estadístico $T(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ es suficiente para θ (o para la familia $\mathcal{F}$ de posibles distribuciones para $\mathbf{x}_i$) si la distribución condicional de la muestra dado $T(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, o sea $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \mid T(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, no depende de θ.*

¿Cómo se relaciona esta definición con la de reducción suficiente dada en la página 11? Aquí la familia de distribuciones que nos interesa es $\mathcal{F}$ que contiene a todas las distribuciones condicionales de $\mathbf{x} \mid y$. Dada una observación $(\mathbf{x}, y)$, en este caso, un estadístico $T(\mathbf{x}, y)$ será suficiente si la distribución condicional de $\mathbf{x} \mid (y, T(\mathbf{x}, y))$ no depende (matemáticamente) de $\mathcal{F}$. Si $\mathbf{R}(\mathbf{x})$ es una reducción suficiente, en el sentido de la Definición 1.3, entonces, por la equivalencia de la Observación 1.1, sabemos que $\mathbf{x} \mid (y, \mathbf{R}(\mathbf{x})) \sim \mathbf{x} \mid \mathbf{R}(\mathbf{x})$, luego $\mathbf{R}(\mathbf{x})$ será un estadístico suficiente para $\mathcal{F}$.

1.1. Subespacios de reducción de la dimensión y subespacio central

Por lo general, no hay una regla específica para guiar la elección de una reducción. Sin embargo, siempre es posible empezar restringiendo la atención a la clase de las reducciones lineales. Como $\mathbf{R}(\mathbf{x}) = \mathbf{x}$ es trivialmente suficiente, una reducción lineal siempre existe. Si $\mathbf{R}(\mathbf{x}) = \eta^T \mathbf{x}$ con $\eta \in \mathbb{R}^{p \times d}$ es una reducción $d-$dimensional, entonces también $\mathbf{R}_A(\mathbf{x}) = (\eta A)^T \mathbf{x}$ lo es para cada $A \in \mathbb{R}^{d \times d}$ de rango completo. En consecuencia, sólo el subespacio generado por las columnas de η, que vamos a denotar span(η), queda identificado. En este caso, el span(η) se llama un **subespacio de reducción de la dimensión**.

Si span(η) es un subespacio de reducción de la dimensión, entonces también lo es span(η, η_1) para cualquier $\eta_1 \in \mathbb{R}^{p \times d_1}$. Como consecuencia de esto, puede haber muchos subespacios lineales suficientes de reducción en una regresión particular; estaremos interesados en aquel que tenga la dimensión más pequeña. Entonces, podemos definir,

Definición 1.6 *Sea*

$$\mathcal{S}_{y|\mathbf{x}} = \bigcap_{\substack{\mathcal{S} \text{ es un subespacio} \\ \text{de reducción de la dimensión}}} \mathcal{S}.$$

$\mathcal{S}_{y|\mathbf{x}}$ siempre será un subespacio, aunque no necesariamente será un subespacio de reducción de la dimensión. Sin embargo, bajo condiciones bastante generales se prueba que $\mathcal{S}_{y|\mathbf{x}}$ resulta serlo, ver Cook [1994], Cook [1996] y Cook [1998]. Cuando esto sucede, a $\mathcal{S}_{y|\mathbf{x}}$ se lo denomina **subespacio central.**

Siguiendo a Cook, en este trabajo asumiremos que el subespacio central existe. En consecuencia, el objetivo de la inferencia en reducción suficiente será obtener $\mathcal{S}_{y|\mathbf{x}}$.

1.2. Reducción inversa basada en momentos

La primera propuesta de reducción inversa surge en el trabajo seminal de K.-C. Li [1991a], en el que se propone la *regresión inversa partida* (*Sliced Inverse Regression, SIR*) para la reducción de dimensión. En ese trabajo se utiliza la siguiente condición,

(Condición de Linealidad) Existe un $\eta_0 \in \mathbb{R}^{p\times d}$ tal que $E\left(\mathbf{x} \mid \eta_0^T\mathbf{x}\right)$ es una función lineal de $\mathbf{x}$.

K.-C. Li [1991a] demuestra que bajo dicha condición, $E(\mathbf{x} \mid y) - E(\mathbf{x})$ pertenece al subespacio de dimensión d, $\Sigma\text{span}(\eta_0)$, donde $\Sigma = \text{Var}(\mathbf{x})$ y $\text{span}(\eta_0)$ es el subespacio generado por las columnas de η_0. Si asumimos que $\mathcal{S}_{y|\mathbf{x}} = \text{span}(\eta_0)$, esta relación permite la estimación de $\mathcal{S}_{y|\mathbf{x}}$ a través de la estimación de la esperanza condicional de $\mathbf{x} \mid y$, la esperanza de $\mathbf{x}$ y la matriz de covarianza de $\mathbf{x}$.

La razón por la que el SIR es tan innovador es que propone invertir los roles de $\mathbf{x}$ e y: en vez de regresar a y en función de $\mathbf{x}$ (*forward regression*), Li es pionero en la propuesta de regresar a $\mathbf{x}$ en función de y (*inverse regression*). Con este cambio de roles se esquiva el problema de la dimensionalidad. Esto sucede pues la regresión inversa puede llevarse a cabo regresando cada coordenada de $\mathbf{x}$ en función de y. Es decir, enfrentamos p problemas de regresión univariada en vez del problema original de regresión múltiple con alta dimensión del espacio de las covariables.

Como no se impone un modelo, la estimación de $E(\mathbf{x} \mid y)$ se hace de forma no paramétrica de la manera más sencilla, a través de medias muestrales. Si y es discreta, de rango pequeño, la esperanza de $\mathbf{x}$ cuando $y = y_0$ se estima por la media calculada utilizando la distribución empírica, es decir, por el promedio de las observaciones $\mathbf{x}'s$ en la muestra cuyo valor y es igual a y_0. Si y es continua, su rango observado se divide en H intervalos disjuntos $(I_h)_{h=1,\ldots,H}$. Cada observación pertenece a uno de estos intervalos, la esperanza $E(\mathbf{x} \mid y \in I_h)$ se estima por la media muestral de las observaciones $\mathbf{x}'s$ en la muestra cuyo valor de y pertenece a I_h. $E(\mathbf{x})$ se estima por la media muestral de las $\mathbf{x}'s$. Σ se estima con la matriz de covarianza muestral $\widehat{\Sigma}$ (de nuevo, la covarianza calculada utilizando la distribución empírica) y el subespacio central se estima mediante la búsqueda de las d componentes principales de los valores estimados

$$\mathbf{m}_h = E_{P_n}(\mathbf{x} \mid y \in I_h) - E_{P_n}(\mathbf{x}) = E_P(\widehat{\mathbf{x} \mid y \in I_h}) - \widehat{E_P(\mathbf{x})} \tag{2}$$

donde P_n es la distribución empírica, $P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{(\mathbf{x}_i,y_i)}$ basada en la muestra $(\mathbf{x}_i, y_i)_{1\leq i\leq n}$. Así las H estimaciones calculadas en (2) yacen aproximadamente en el subespacio de interés, Li propone encontrar el subespacio de dimensión d más cercano (en la distancia habitual para espacios euclídeos) a los vectores $\{\mathbf{m}_1, \ldots, \mathbf{m}_H\}$. Este es el subespacio generado por las d primeras componentes principales de $\{\mathbf{m}_1, \ldots, \mathbf{m}_H\}$, donde cada $\mathbf{m}_h$ es ponderada por la proporción de observaciones de la muestra utilizadas para calcularla (w_h). Finalmente el estimador de $\mathcal{S}_{y|\mathbf{x}}$ es el espacio generado por los autovectores

asociados a los d mayores autovalores de

$$\widehat{\Sigma}^{-1}\left(\sum_{h=1}^{H}\mathbf{m}_h\mathbf{m}_h^T w_h\right) = \widehat{\Sigma}^{-1}\left[\mathbf{m}_1 \quad \cdots \quad \mathbf{m}_H\right] diag\left(w_1,\ldots,w_H\right)\begin{bmatrix}\mathbf{m}_1^T\\ \vdots \\ \mathbf{m}_H^T\end{bmatrix}.$$

En esta línea, otras propuestas de reducción de la dimensión surgen de modelar momentos condicionales de segundo orden. Entre ellas, la *Sliced Average Variance Estimator (SAVE)* es otra técnica de reducción de dimensión que fue propuesta por Cook y Weisberg [1991]. Bajo la condición de linealidad enunciada más arriba, sumada al supuesto de *condición de covarianza constante* definido como

$$\mathrm{Var}\left(\mathbf{x} \mid \eta_0^{\mathrm{T}}\mathbf{x}\right)$$

es constante para algun valor de $\eta_0 \in \mathbb{R}^{p\times d}$. Los autores observan que $\mathrm{span}\left(\Sigma - \mathrm{Var}\left(\mathbf{x} \mid y\right)\right) \in \Sigma\mathrm{span}\left(\eta_0\right)$ donde Σ es nuevamente la covarianza de $\mathbf{x}$, $\Sigma = \mathrm{Var}\left(\mathbf{x}\right)$ y span representa el subespacio generado por las columnas de la matriz. Luego, si asumimos que $\mathcal{S}_{y|\mathbf{x}} = \mathrm{span}\left(\eta_0\right)$, estimando apropiadamente a estas matrices se puede obtener un estimador para $\mathcal{S}_{y|\mathbf{x}}$. Este es el sustento poblacional para SAVE que se centra en la estimación no paramétrica de $\mathrm{Var}\left(\mathbf{x} \mid y\right)$ y Σ, y utiliza componentes principales. Algunos otros métodos han sido desarrollados en esta línea. Entre ellos podemos mencionar los métodos de *Principal Hessian Direction*, originalmente propuesto por K.-C. Li [1992], *inverse regression estimation*, por Cook y Ni [2005], *directional regression*, por B. Li y Wang [2007], *partial inverse regression with a categorical predictor* propuesta por Wen y Cook [2007].

Finalmente, vale la pena observar que tanto la condición de linealidad como la de covarianza constante imponen restricciones en la distribución conjunta de los predictores. Las distribuciones elípticamente simétricas satisfacen la condición de linealidad. Puede verse K.-C. Li [1991b], para una discusión sobre el tema.

Por otro lado, una limitación del SIR es la siguiente. Como observa K.-C. Li [1991a], bajo la condición de linealidad, podría pasar que

$$\mathrm{span}\left[E\left(\mathbf{x} \mid y\right) - E\left(\mathbf{x}\right)\right] \subsetneq \Sigma\mathrm{span}\left(\eta_0\right),$$

es decir, que la curva de regresión inversa esté incluida en un subespacio propio de $\Sigma\mathrm{span}\left(\eta_0\right)$, en ese caso no podremos recuperar a $\mathcal{S}_{y|\mathbf{x}}$ sólo estudiando la curva $E\left(\mathbf{x} \mid y\right)$. Es decir, hay casos en los que no basta con la esperanza condicional para caracterizar a la reducción suficiente. En ese marco, el subespacio generado por los autovectores calculados estará contenido en el subespacio de reducción, pero éstos podrían no generarlo, o sea, podría recuperarse una parte del subespacio de interés, no todo.

En el paquete de R `dr` Weisberg [2002], están programados el SIR, SAVE y otros estimadores de $\mathcal{S}_{y|\mathbf{x}}$ basados en momentos.

Tanto SIR como SAVE proporcionan estimadores consistentes de $\mathcal{S}_{y|\mathbf{x}}$ a tasa $\sqrt{n}$ bajo condiciones estándar. Varios artículos describen su desempeño en el caso de muestras finitas. En ellos, se describen las dificultades de SIR para encontrar las direcciones correctas en el caso en el que la $E\left(y \mid \mathbf{x}\right)$ depende de $\mathbf{x}$ de forma no lineal, por ejemplo

K.-C. Li [1991a], Cook y Weisberg [1991], H. Wang y Xia [2008] en el caso en el que la función g que vincula a y con $\mathbf{x}$ en el modelo (1) es simétrica respecto del origen. SAVE fue desarrollado para solucionar este problema, pero funciona peor que SIR en presencia de tendencias lineales. En H. Wang y Xia [2008] se exhibe un ejemplo de simulación en la que la varianza condicional $Var(y \mid \mathbf{x})$ no es constante, y muestra que ni el SIR ni el SAVE, ni *Principal Hessian Direction* (ni tampoco *Simple contour regression* que comentamos en la siguiente sección) encuentran las direcciones correctas. Por supuesto, todos ellos comparten la falta de resistencia a la presencia de valores atípicos, como se muestra por ejemplo en Gather, Hilker, y Becker [2002].

1.3. Reducción de la dimensión basada en modelos

El segundo enfoque para proponer estimadores de las reducciones o del subespacio central es lo que se llama reducción suficiente de la dimensión basada en modelos. La idea es proponer directamente un modelo para la distribución inversa, es decir, para la distribución de $\mathbf{x} \mid y$. El objetivo sigue siendo tratar de encontrar una reducción adecuada para que la estimación directa sea posible. En este escenario, la orientación hacia el método de estimación de $\mathcal{S}_{y|\mathbf{x}}$ adecuado proviene del propio modelo. Y también del modelo se deduce el comportamiento asintótico de los estimadores propuestos. La motivación para la introducción de una familia de distribuciones fue la posibilidad de derivar a través de ella nuevos estimadores que serían óptimos cuando las observaciones siguen la distribución supuesta, y luego estudiar su rendimiento cuando la verdadera distribución es otra. Otra motivación fue establecer condiciones precisas bajo las cuales existe el subespacio central y es único, y también para poder asegurar que la propuesta de estimación de la reducción es capaz de recuperar eficazmente todo el subespacio.

Una forma de introducir este enfoque es el siguiente. Como observara K.-C. Li [1991a], si la condición de linealidad es satisfecha, $E(\mathbf{x} \mid y) - E(\mathbf{x})$ pertenece a un subespacio. Además, por definición, esta cantidad debe depender de y. Uno puede modelar la distribución condicional (o inversa) de $\mathbf{x} \mid y$ tomando en cuenta estos hechos, y luego tratar de recuperar la reducción suficiente a partir de este subespacio. En Szretter y Yohai [2009] demostramos que bajo el supuesto de que la distribución de $\mathbf{x} \mid y$ es normal, con la media perteneciente a una variedad afín (desconocida), con matriz de covarianza constante (y desconocida), entonces podemos identificar y estimar el subespacio central mediante el método de máxima verosimilitud.

En Cook [2007] y en Cook y Forzani [2008] se propone un modelo más general para la distribución condicional: el modelo de componentes principales ajustadas (*principal fitted components,* PFC). Asumamos que $E(\mathbf{x} \mid y) - E(\mathbf{x})$ pertenece a un subespacio de dimensión d. Consideremos la matriz Γ (desconocida) de dimensión $p \times d$ que tiene por columnas a los generadores linealmente independientes de dicho subespacio. Entonces, para cada valor de y debería existir un vector de coordenadas $\boldsymbol{\nu}_y \in \mathbb{R}^d$ que satisfaga

$$E(\mathbf{x} \mid y) - E(\mathbf{x}) = \Gamma \boldsymbol{\nu}_y.$$

Los autores asumen que las observaciones satisfacen el siguiente modelo: condicional a y,

$$\mathbf{x} = \boldsymbol{\mu} + \Gamma\boldsymbol{\nu}_y + \sigma\boldsymbol{\epsilon},$$

donde $\boldsymbol{\mu} \in \mathbb{R}^p$, $\Gamma \in \mathbb{R}^{p\times d}$, $d < p$, $\Gamma^T\Gamma = I_d$, $\sigma > 0$, d es conocido, $\boldsymbol{\nu}_y \in \mathbb{R}^d$ es una función desconocida de y con matriz de covarianza definida positiva y media muestral igual a cero. El vector de errores $\boldsymbol{\epsilon} \in \mathbb{R}^p$ es independiente de y, con distribución normal de media $\mathbf{0}$ y matriz de covarianza identidad. $\boldsymbol{\nu}(y)$ son las coordenadas de $E(\mathbf{x} \mid y) - E(\mathbf{x})$ en la base dada por la columnas de Γ. De modo que $\boldsymbol{\nu} : \mathbb{R} \to \mathbb{R}^d$ es una función vectorial (o curva) desconocida que dependerá de la selección de la base de $\text{span}(\Gamma)$ realizada. La propuesta de Cook y Forzani [2008] es modelar esta función, sugiriendo guiarse por gráficos de los datos disponibles para hacerlo. La idea es encontrar una aproximación a $\boldsymbol{\nu}$ suficientemente buena usando, por ejemplo, un polinomio de grado r ($r \leq d$). Por ejemplo, para $r = 3$, se puede tomar

$$\boldsymbol{\nu}(y) = \beta \begin{pmatrix} y \\ y^2 \\ y^3 \end{pmatrix}$$

donde la matriz $\beta \in \mathbb{R}^{d\times r}$ es otro parámetro del modelo. Otra posibilidad inspirada en el SIR es dividir el rango de y en r fetas o intervalos disjuntos $(I_h)_{h=1,\ldots,r}$ y proponer

$$\boldsymbol{\nu}(y) = \beta \begin{pmatrix} I_{I_1}(y) \\ \vdots \\ I_{I_r}(y) \end{pmatrix},$$

donde I_A es la función indicadora del conjunto A. También se podrían utilizar funciones trigonométricas adaptadas de una serie de Fourier.

Otros ejemplos de reducciones suficientes basadas en modelos son *Simple contour regression*, y *Envelope models*, ambas propuestas por Cook, Li, y Chiaromonte [2010]. También *Model–based SIR*, propuesta por Scrucca [2011].

El trabajo de esta tesis se desarrolla de la siguiente forma. En el Capítulo 2 presentamos con detalle el modelo de PFC, incluyendo una descripción del espacio de parámetros. En el Capítulo 3 discutimos la relación entre estadístico suficiente y modelo de PFC. Cook [2007] exhibe una expresión para la reducción suficiente cuando los errores del modelo PFC tienen distribución normal multivariada. En este capítulo extendemos este resultado para el caso en el que la distribución del error pertenece a una familia de distribuciones más amplia que la normal multivariada.

En el Capítulo 4 presentamos el estimador de máxima verosimilitud (EMV) para el modelo PFC con error normal multivariado obtenido por Cook y Forzani [2008]. Estos estimadores son eficientes bajo normalidad, pero son sensibles a la presencia de *outliers*. En este capítulo presentamos una propuesta de estimación para superar esta limitación, basada en escalas de tipo τ. Las escalas de tipo τ, introducidas por Yohai y Zamar [1988] pueden ser altamente robustas y eficientes. A los estimadores propuestos los denominamos $\tau-$estimadores para el modelo PFC. Se basan en minimizar el $det(\Delta)$ sujeto a una condición sobre la $\tau-$escala de los residuos estandarizados de las observaciones.

En el Capítulo 5 construimos al $\tau-$funcional para el modelo PFC que da lugar a los $\tau-$estimadores, y probamos que resulta equivariante. También presentamos maneras alternativas de definirlo, siempre como aquella combinación de los parámetros en la que se minimiza una función objetivo.

En la Sección 6.1 exhibimos condiciones sobre la distribución de los errores que garantizan la existencia de dicho mínimo, tanto para el funcional como para el estimador. Estas condiciones incluyen a las distribuciones esféricas. En la Sección 6.2 refinamos dichas condiciones para garantizar, además, la unicidad del mínimo (y, por lo tanto, la buena definición del $\tau-$funcional). Y, además, probamos la Fisher consistencia del $\tau-$funcional para el modelo PFC.

En el Capítulo 7 probamos la consistencia fuerte de los $\tau-$estimadores. Para obtener este resultado, probamos la continuidad del funcional, por lo que también hemos verificado la robustez cualitativa asintótica del $\tau-$funcional.

En el Capítulo 8 hallamos las ecuaciones de estimación para obtener al $\tau-$funcional. A partir de ellas obtenemos una expresión para los $\tau-$estimadores análoga a la de máxima verosimilitud, pero con pesos dependiendo de los parámetros. Esto nos permite proponer un algoritmo iterativo para computar los $\tau-$estimadores en el que en cada paso se actualizan los estimadores y los ponderadores que dependen de las observaciones. Asimismo, presentamos un procedimiento para obtener valores iniciales para el algoritmo propuesto. Esto lo hacemos en el Capítulo 9, donde también mostramos los resultados de un estudio de Monte Carlo para comparar los $\tau-$estimadores con los EMV bajo el modelo PFC con errores normales y en distintos escenarios de contaminación. En la Sección 9.5 presentamos una propuesta basada en validación cruzada para seleccionar la dimensión del espacio de reducción, y evaluamos su performance en otra simulación. Ambos estudios de simulación muestran que los $\tau-$estimadores son altamente eficientes y robustos en muestras finitas. Finalmente, en la Sección 9.7, aplicamos el procedimiento propuesto (validación cruzada para seleccionar la dimensión y estimación robusta de tipo $\tau-$) a dos ejemplos de datos reales. Las demostraciones de los resultados se presentan en los apéndices correspondientes a los capítulos donde fueron enunciadas. También en los apéndices presentamos en forma resumida resultados que describen temas que este trabajo utiliza (independencia condicional, descomposición en valores singulares, normas de matrices, variedades de Grassman, DAGs y probabilidades, cálculo de ángulos principales).

2. Modelo PFC (Principal Fitted Components)

2.1. Modelo

Sea $(\mathbf{x}, y)$ un vector aleatorio con $\mathbf{x} \in \mathbb{R}^p$, $y \in \mathbb{R}$. Diremos que satisface el *modelo de componentes principales ajustadas* (Principal Fitted Components, PFC, por sus siglas en inglés) introducido por Cook [2007] si existe un vector $\boldsymbol{\mu}_0 \in \mathbb{R}^p$, una matriz $\Gamma_0 \in \mathbb{R}^{p\times d}$ con rango $(\Gamma_0) = d \leq p$, una matriz $\beta_0 \in \mathbb{R}^{d\times r}$ con $d \leq r$, una función $\mathbf{f} : \mathbb{R} \to \mathbb{R}^r$, y una matriz $\Delta_0 \in \mathbb{R}^{p\times p}$ definida positiva tales que

$$\mathbf{x} = \boldsymbol{\mu}_0 + \Gamma_0\beta_0\mathbf{f}(y) + \Delta_0^{1/2}\mathbf{u} \tag{3}$$

donde $\mathbf{u}$ es un vector aleatorio $p-$dimensional independiente de y. Los valores de los parámetros $\boldsymbol{\mu}_0, \Gamma_0, \beta_0$ y Δ_0 son desconocidos, pero la función $\mathbf{f}$ es conocida. El vector $\mathbf{u}$ no es observable. Al término $\Delta_0^{1/2}\mathbf{u}$ se lo denomina error del modelo.

Para completar el modelo, se suelen imponer condiciones sobre la distribución de los errores o sobre la distribución conjunta de $(\mathbf{x}, y)$. Cook [2007] propone que $\mathbf{u}$ tenga distribución conjunta normal multivariada con $E_F(\mathbf{u}) = 0$ y matriz de covarianza $E_F(\mathbf{u}\mathbf{u}^T) = I_p$. En ese artículo se prueba que para este modelo bajo distribución normal de los errores, $R(\mathbf{x}) = \Gamma_0^T\Delta_0^{-1}\mathbf{x}$ es una reducción suficiente. En Cook y Forzani [2008] se prueba que dicha reducción es minimal.

2.2. Matriz de regresión de rango reducido

Sea $\boldsymbol{\Theta}$ el espacio de posibles parámetros para el modelo PFC. Buscamos una descripción de este espacio que nos dé identificabilidad del modelo.

Observemos que Γ_0 y β_0 no están determinados, ya que si $A \in \mathbb{R}^{d\times d}$ es una matriz inversible, se puede cambiar a Γ_0 y β_0 por $\Gamma_0 A$ y $A^{-1}\beta_0$, respectivamente, y el modelo se mantiene igual. El subespacio de dimensión d en $\mathbb{R}^p$ generado por las columnas de Γ_0 es el que queda identificado por el modelo, ya que el subespacio generado por las columnas de Γ_0 y el generado por las columnas de $\Gamma_0 A$ coinciden. Lo denotaremos $\mathcal{S}_{\Gamma_0} = \text{span}(\Gamma_0)$.

Claramente $\boldsymbol{\mu} \in \mathbb{R}^p$ sin restricción alguna, y $\Delta \in PDS(p)$ donde $PDS(p)$ es la clase de todas las matrices $p \times p$ reales simétricas y definidas positivas. En cuanto a los coeficientes del modelo de regresión, esto es Γ y β, a través del siguiente lema podemos caracterizar a su producto, $B = \Gamma\beta \in \mathbb{R}^{p\times r}$. Las restricciones impuestas sobre las dimensiones de Γ_0 y β_0 restringen el rango de $\Gamma_0\beta_0$ a valer a lo sumo $d \leq \text{mín}(p, r)$.

Lema 2.1 *Sea $B \in \mathbb{R}^{p\times r}$ con* rango $(B) = d$. *Entonces existen matrices $U \in \mathbb{R}^{p\times d}, D \in \mathbb{R}^{d\times d}$ y $\Psi \in \mathbb{R}^{d\times r}$ tales que $U^TU = I_d, \Psi\Psi^T = I_d$, D es una matriz diagonal $D =$* diag $(\sigma_1, \ldots, \sigma_d) \in \mathbb{R}^{d\times d}$ *con $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$ y*

$$B = UD\Psi. \tag{4}$$

Este resultado es la Proposición 1.47 de Eaton [1983].

Luego, si llamamos $B_0 = \Gamma_0\beta_0$, resulta que $B_0 \in \mathbb{R}^{p\times r}$ con rango $(B) \leq d$, podemos escribir el modelo PFC de una forma alternativa

$$\mathbf{x} = \boldsymbol{\mu}_0 + B_0\mathbf{f}(y) + \Delta_0^{1/2}\mathbf{u}. \tag{5}$$

Asumiremos que $\mathbf{f}(y)$ no tiene ninguna fila constante (pues en tal caso no hay identificabilidad del parámetro $\boldsymbol{\mu}_0$ en el modelo). Alternativamente, podríamos agregar una fila con un uno a $\mathbf{f}(y)$ de modo de incluir a $\boldsymbol{\mu}_0$ en la matriz de coeficientes. De modo que si tomamos $\widetilde{\mathbf{f}} : \mathbb{R} \to \mathbb{R}^{r+1}$ definida por $\widetilde{\mathbf{f}}(y) = \begin{bmatrix} \mathbf{1} \\ \mathbf{f}(y) \end{bmatrix}$, y $D_0 = \begin{bmatrix} \boldsymbol{\mu}_0 & \Gamma_0\beta_0 \end{bmatrix} \in \mathbb{R}^{p\times(r+1)}$, podemos escribir

$$\mathbf{x} = D_0\mathbf{z} + \Delta_0^{1/2}\mathbf{u}, \tag{6}$$

donde $\mathbf{z} = \widetilde{\mathbf{f}}(y)$. Esta ecuación corresponde a un modelo de regresión lineal multivariado, con vector de respuestas $\mathbf{x}$, vector de covariables $\mathbf{z}$ y matriz de coeficientes de regresión D_0. Observemos que si $d < p$ y $r = d$, o bien $p = d$ tenemos directamente el modelo lineal multivariado. En los demás casos, la diferencia es que en el modelo PFC se impone reducir el rango de la matriz de coeficientes.

2.3. Espacio de parámetros del modelo

El espacio de posibles parámetros para el modelo PFC, $\boldsymbol{\Theta}$, puede escribirse como

$$\begin{aligned} \boldsymbol{\Theta} &= \left\{\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta) \in \mathbb{R}^p \times \mathbb{R}^{p\times r} \times PDS\,(p) : \text{rango}\,(B) \leq d\right\} && (7) \\ &= \Big\{\boldsymbol{\theta} = (\boldsymbol{\mu}, \Gamma\beta, \Delta) \in \mathbb{R}^p \times \mathbb{R}^{p\times r} \times PDS\,(p) : \Gamma \in \mathbb{R}^{p\times d}, && (8) \\ &\quad \beta \in \mathbb{R}^{d\times r}, \text{rango}\,(\Gamma) = d\Big\}, \end{aligned}$$

Como lo que queda identificado por el modelo es el subespacio span (Γ), y no la matriz Γ, siguiendo a Cook y Forzani [2008] utilizaremos la notación $\mathcal{S}_\Gamma = \text{span}\,(\Gamma)$. Definamos el espacio

$$\Omega = \left\{B \in \mathbb{R}^{p\times r} : \text{rango}\,(B) \leq d\right\},$$

de matrices de rango reducido. Dado $B \in \Omega$, por el Lema 2.1, podemos descomponer a B

$$B = UD\Psi. \tag{9}$$

Esta descomposción es única si asumimos que los valores singulares $\sigma_1, \ldots, \sigma_d$ de B son todos distintos y se adopta la convención de que la primera coordenada no nula de cada columna de U y Ψ sea positiva. La dimensión algebraica del conjunto de matrices

$$\left\{U \in \mathbb{R}^{p\times d} : U^TU = I_d\right\}$$

es $d\,(p-d) + \frac{d(d-1)}{2}$, ver Absil, Mahony, y Sepulchre [2009]. De ello resulta que la dimensión algebraica de Ω es $d\,(p+r-d)$. Si tomamos $\Gamma = U$ y $\beta = D\Psi$ se cumplen para

estas matrices las restricciones impuestas y se recupera a B. Luego (9) es una caracterización de Ω. A lo largo de este trabajo utilizaremos las dos notaciones alternativas para la matriz: $B = \Gamma\beta$.

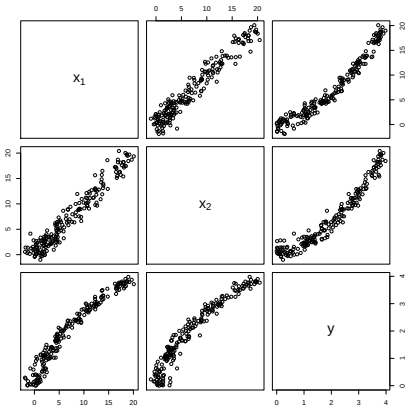
En el Apéndice B discutimos otras caracterizaciones del espacio de parámetros, con el objetivo de tener la misma notación que la utilizada en Cook y Forzani [2008].

2.4. Datos generados bajo el modelo PFC

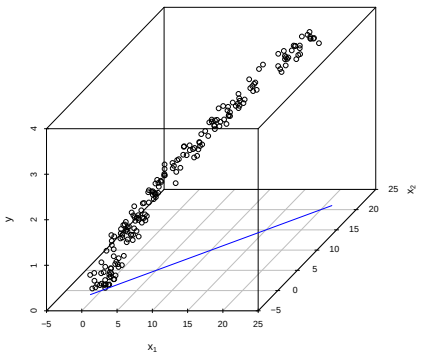
A modo ilustrativo, veamos cómo son los gráficos de dispersión de a pares, y en $\mathbb{R}^3$ para datos generados bajo el modelo PFC con $p = 2$, $r = 2$ y $d = 1$ ó 2. En la Figura 1a y 1b se ven, respectivamente, el gráfico de dispersión de a pares y el scatterplot en $\mathbb{R}^3$ para 200 observaciones simuladas bajo el modelo PFC con $d = 1$. En cambio, en las Figuras 1c y 1d apreciamos los mismos gráficos para datos simulados bajo el modelo PFC, pero en este caso para $d = 2$.

Figura 1: Las cuatro figuras muestran gráficos de doscientas observaciones generadas por el modelo PFC, con $p = 2$, $r = 2$, $\mathbf{f}(y) = \left(y, y^2\right)^T$, $\Delta = I$, $\boldsymbol{\mu} = (0,0)^T$. La variable y fue generada con distribución $U(0,4)$. En las figuras (a) y (b) tomamos $d = 1$ y $\Gamma = (1,1)^T$. En cambio, en (c) y (d) $d = 2$ y $\Gamma = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$.

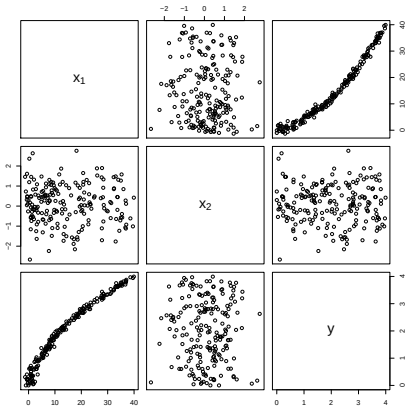
(a) Scatter plot de a pares, $d = 1$.



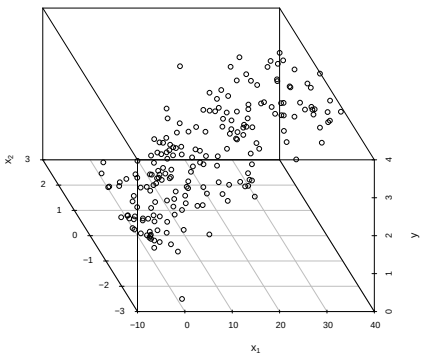
(b) Gráfico de dispersión en $\mathbb{R}^3, d = 1$.



(c) Scatter plot de a pares, $d = 2$.



(d) Gráfico de dispersión en $\mathbb{R}^3, d = 2$.



3. Reducción en el modelo PFC

Sea $\mathcal{V}$ un subespacio de dimensión d en $\mathbb{R}^p$, y Δ una matriz $p \times p$ simétrica de rango completo, escribimos $p_\Delta(\mathbf{x},\mathcal{V})$ para referirnos a la *proyección ortogonal de* $\mathbf{x}$ *en* $\mathcal{V}$*, con el producto escalar* $\langle\cdot,\cdot\rangle_\Delta$ *inducido por* Δ. Si $\{\mathbf{v}_1,\ldots,\mathbf{v}_d\}$ es una base ortonormal de $\mathcal{V}$ con el producto escalar $\langle\cdot,\cdot\rangle_\Delta$, podemos escribir

$$\begin{aligned} p_\Delta(\mathbf{x},\mathcal{V}) &= \sum_{i=1}^{d} \langle \mathbf{x},\mathbf{v}_i\rangle_\Delta \mathbf{v}_i = \sum_{i=1}^{d} \left(\mathbf{v}_i^T\Delta^{-1}\mathbf{x}\right)\mathbf{v}_i = \sum_{i=1}^{d} \mathbf{v}_i\left(\mathbf{v}_i^T\Delta^{-1}\mathbf{x}\right) \\ &= \left(\sum_{i=1}^{d}\mathbf{v}_i\mathbf{v}_i^T\right)\Delta^{-1}\mathbf{x} = [\mathbf{v}_1\cdots\mathbf{v}_d]\begin{bmatrix}\mathbf{v}_1^T\\ \vdots\\ \mathbf{v}_d^T\end{bmatrix}\Delta^{-1}\mathbf{x}. \end{aligned} \tag{10}$$

Recordemos que $\|\mathbf{x}\|_\Delta^2 = \|p_\Delta(\mathbf{x},\mathcal{V})\|_\Delta^2 + \|\mathbf{x}-p_\Delta(\mathbf{x},\mathcal{V})\|_\Delta^2$, donde $\|\mathbf{x}\|_\Delta^2 = \mathbf{x}^T\Delta^{-1}\mathbf{x}$, también $p_\Delta(\mathbf{x},\mathcal{V})+p_\Delta\left(\mathbf{x},\mathcal{V}^\perp\right) = \mathbf{x}$, donde $\mathcal{V}^\perp$ es el subespacio ortogonal a $\mathcal{V}$ con el producto escalar $\langle\cdot,\cdot\rangle_\Delta$. Si la base $\{\mathbf{v}_1,\ldots,\mathbf{v}_d\}$ no fuera ortonormal para el producto escalar $\langle\cdot,\cdot\rangle_\Delta$, y llamamos Γ a la matriz que tiene por columnas a los $\mathbf{v}_i$, es decir, si

$$\Gamma = \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_d\end{bmatrix},$$

entonces,

$$p_\Delta(\mathbf{x},\mathcal{V}) = p_\Delta(\mathbf{x},\operatorname{span}(\Gamma)) = \Gamma\left(\Gamma^T\Delta^{-1}\Gamma\right)^{-1}\Gamma^T\Delta^{-1}\mathbf{x}. \tag{11}$$

Desde aquí en adelante, consideramos un vector aleatorio $\mathbf{x}$ que satisface el modelo PFC, introducido en (3). Como observáramos, la matriz Γ_0 no es identificable, aunque sí lo es el subespacio generado por sus columnas, $\operatorname{span}(\Gamma_0)$. Por esta razón, asumiremos sin pérdida de generalidad que las columnas de Γ_0 son ortonormales, con el producto interno inducido por Δ_0, esto es $\Gamma_0^T\Delta_0^{-1}\Gamma_0 = I_{d\times d}$. Sea

$$\mp_1 = \left\{\Gamma_1 \in \mathbb{R}^{p\times(p-d)} : \Gamma_1^T\Delta_0^{-1}\Gamma_1 = I_{(p-d)\times(p-d)},\quad \Gamma_1^T\Delta_0^{-1}\Gamma_0 = 0_{(p-d)\times d}\right\},$$

es decir, $\mp_1$ es el conjunto de generadores ortonormales con el producto interno inducido por Δ_0 del subespacio $\operatorname{span}(\Gamma_0)^\perp$.

En el siguiente resultado, generalizamos el resultado probado por Cook [2007], Proposición 6, que afirma que $\mathbf{R}(\mathbf{x}) = \Gamma_0^T\Delta_0^{-1}\mathbf{x}$ es una reducción suficiente cuando se asume que la distribución de los errores es normal. En contraposición al modelo considerado en Cook [2007] y Cook y Forzani [2008], aquí no imponemos normalidad (de hecho, ni siquiera asumimos distribución elíptica de los errores, más aún, no pedimos siquiera que la distribución conjunta de las $\mathbf{x}$ sea continua).

Teorema 3.1 *Bajo el modelo PFC, si* $p_{\Delta_0}\left(\Delta_0^{1/2}\mathbf{u},\operatorname{span}(\Gamma_0)\right)$ *y* $p_{\Delta_0}\left(\Delta_0^{1/2}\mathbf{u},\operatorname{span}(\Gamma_0)^\perp\right)$ *son vectores independientes, entonces* $\mathbf{R}(\mathbf{x}) = \Gamma_0^T\Delta_0^{-1}\mathbf{x}$ *es una reducción suficiente.*

La demostración de este resultado está en el Apéndice C.2.

Observación 3.1 *Observemos que para todo* $\Gamma_1 \in \mp_1$,

$$\begin{aligned}
&\operatorname{cov}\left(p_{\Delta_0}\left(\Delta_0^{1/2}\mathbf{u}, \operatorname{span}(\Gamma_0)\right), p_{\Delta_0}\left(\Delta_0^{1/2}\mathbf{u}, \operatorname{span}(\Gamma_1)\right)\right) \\
&= \operatorname{cov}\left(\Gamma_0\Gamma_0^T\Delta_0^{-1/2}\mathbf{u}, \Gamma_1\Gamma_1^T\Delta_0^{-1/2}\mathbf{u}\right) \\
&= \Gamma_0\Gamma_0^T\Delta_0^{-1/2} cov(\mathbf{u}, \mathbf{u}) \left[\Gamma_1\Gamma_1^T\Delta_0^{-1/2}\right]^T \\
&= \Gamma_0\Gamma_0^T\Delta_0^{-1/2} I \Delta_0^{-1/2}\Gamma_1\Gamma_1^T \\
&= \Gamma_0\Gamma_0^T\Delta_0^{-1}\Gamma_1\Gamma_1^T = 0,
\end{aligned}$$

de modo que estos dos vectores aleatorios son ***siempre no correlacionados****.*

Observación 3.2 *Si* $\mathbf{u}$ *tiene distribución normal,* $\mathbf{u} \sim N_p(\mathbf{0}, I)$, *sea* $\Gamma_1 \in \mp_1$, *entonces, como* $\begin{bmatrix}\Gamma_0 & \Gamma_1\end{bmatrix}$ *tiene rango* p, *tenemos*

$$\begin{aligned}
\left(\Gamma_0^T\Delta_0^{-1/2}\mathbf{u}, \Gamma_1^T\Delta_0^{-1/2}\mathbf{u}\right) &= \begin{bmatrix}\Gamma_0 & \Gamma_1\end{bmatrix}^T \Delta_0^{-1/2}\mathbf{u} \\
&\sim N_p\left(\mathbf{0}, \begin{bmatrix}\Gamma_0 & \Gamma_1\end{bmatrix}^T \Delta_0^{-1/2}\left(\begin{bmatrix}\Gamma_0 & \Gamma_1\end{bmatrix}^T \Delta_0^{-1/2}\right)^T\right) \\
&= N_p\left(\mathbf{0}, \begin{bmatrix}\Gamma_0^T\Delta_0^{-1}\Gamma_0 & \Gamma_0^T\Delta_0^{-1}\Gamma_1 \\ \Gamma_1^T\Delta_0^{-1}\Gamma_0 & \Gamma_1^T\Delta_0^{-1}\Gamma_1\end{bmatrix}\right) = N_p(\mathbf{0}, I).
\end{aligned}$$

Esto significa que los vectores $\Gamma_0^T\Delta_0^{-1/2}\mathbf{u}$ *y* $\Gamma_1^T\Delta_0^{-1/2}\mathbf{u}$ *tienen distribución conjunta normal multivariada, y son independientes. De modo que si* $\mathbf{u}$ *tiene distribución normal,* $p_{\Delta_0}\left(\Delta_0^{1/2}\mathbf{u}, \operatorname{span}(\Gamma_0)\right) = \Gamma_0\Gamma_0^T\Delta_0^{-1/2}\mathbf{u}$ *y* $p_{\Delta_0}\left(\Delta_0^{1/2}\mathbf{u}, \operatorname{span}(\Gamma_1)\right) = \Gamma_1\Gamma_1^T\Delta_0^{-1/2}\mathbf{u}$ *también son independientes y la hipótesis del Teorema 3.1 se satisface, entonces en este caso* $\mathbf{R}(\mathbf{x})$ *es una reducción suficiente, recuperando lo que fuera probado por Cook [2007].*

Más aún, el Teorema 3.1 es más general que el probado por Cook. Para afirmarlo, debemos ser capaces de exhibir alguna situación para el modelo PFC en la que los errores no sean normales, y sin embargo estemos en las hipótesis de dicho lema para poder concluir que $\mathbf{R}(\mathbf{x})$ es una reducción suficiente. Las dos siguientes observaciones exhiben ejemplos.

Observación 3.3 *Sea* $\mathbf{u} = (u_1, \ldots, u_p)^T$ *un vector aleatorio con* $u_1, \ldots, u_p$ *variables aleatorias independientes e idénticamente distribuidas. Consideremos* $\Gamma_0 = \begin{bmatrix}\mathbf{e}_1 & \cdots & \mathbf{e}_d\end{bmatrix}$ *con* $\mathbf{e}_j$ *el* j*-ésimo vector canónico en* $\mathbb{R}^p$ *(con todas las coordenadas iguales a cero y un uno en la* j*-ésima) y* $\Delta_0 = I$. *En tal caso*

$$p_{\Delta_0}\left(\Delta_0^{1/2}\mathbf{u}, \operatorname{span}(\Gamma_0)\right) = (u_1, \ldots, u_d, 0, \ldots, 0)^T = \sum_{j=1}^{d} u_j\mathbf{e}_j$$

$$p_{\Delta_0}\left(\Delta_0^{1/2}\mathbf{u}, \operatorname{span}(\Gamma_0)^\perp\right) = (0, \ldots, 0, u_{d+1}, \ldots, u_p)^T = \sum_{j=d+1}^{p} u_j\mathbf{e}_j$$

que son claramente independientes. La distribución de cada u_j puede ser cualquiera (por ejemplo, t de Student con k grados de libertad). En este caso, el Teorema 3.1 garantiza que $\mathbf{R}(\mathbf{x})$ *será una reducción suficiente.*

Para el siguiente ejemplo, introducimos primero una familia de distribuciones (de variables y vectores aleatorios), para luego construir el ejemplo. La normal asimétrica (*skew-normal*) es una familia paramétrica de distribuciones univariadas que incluye a la normal estándar como caso particular.

Definición 3.1 *Sea* $\phi_p(\mathbf{z};\Omega)$ *la densidad de una variable aleatoria* $N_p(\mathbf{0},\Omega)$ *con* $\Omega \in \mathbb{R}^{p\times p}$ *definida positiva y sea* Φ *la función de distribución acumulada de una variable aleatoria normal estándar. Luego*

$$\Omega = C^T \omega^2 C$$

es la descomposición de Ω *en autovalores y autovectores, con* ω *una matriz con elementos positivos en la diagonal (los desvíos estándares asociados a la matriz de covarianza dada por* Ω*). Definimos la función* $f : \mathbb{R}^p \to \mathbb{R}$ *por*

$$f(\mathbf{y}) = 2\phi_p(\mathbf{y}-\boldsymbol{\xi};\Omega)\,\Phi\left(\boldsymbol{\alpha}^T\omega^{-1}(\mathbf{y}-\boldsymbol{\xi})\right),$$

donde $\boldsymbol{\alpha} \in \mathbb{R}^p$ *y* $\boldsymbol{\xi} \in \mathbb{R}^p$*. Si* $\mathbf{y}$ *es un vector aleatorio con densidad* f*, entonces decimos que* $\mathbf{y}$ *tiene* ***distribución normal asimétrica multivariada*** *(skew-normal) y escribimos* $\mathbf{y} \sim SN_p(\boldsymbol{\xi},\Omega,\boldsymbol{\alpha})$*. A los parámetros* $\boldsymbol{\alpha}$,$\boldsymbol{\xi}$ *y* Ω *se los denomina parámetro de inclinación (slant), posición y matriz de escala, respectivamente.*

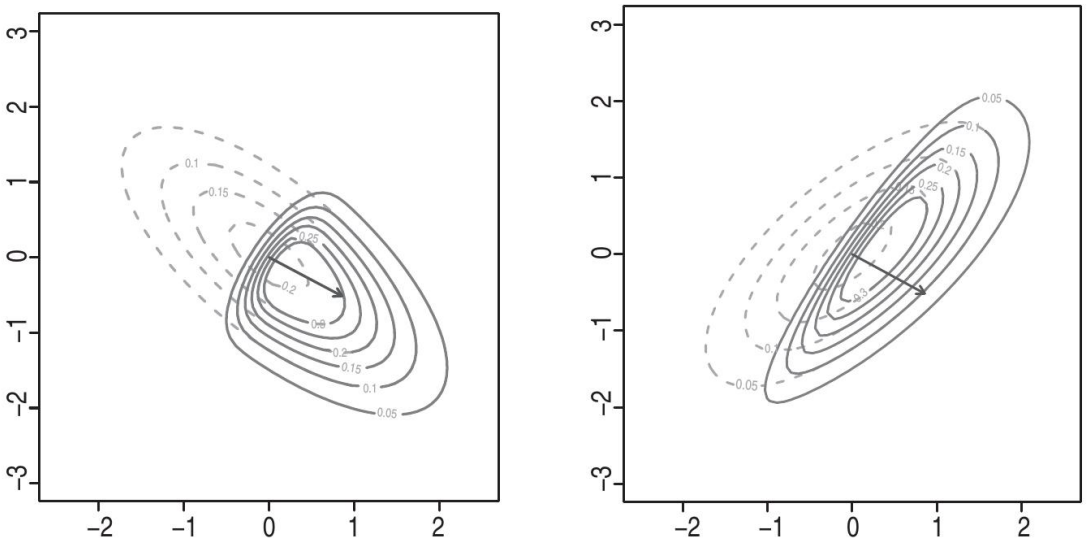
Observemos que ω puede escribirse de la siguiente forma

$$\omega = (\Omega \odot I_p)^{1/2}$$

donde $\odot$ denota el producto coordenada a coordenada (o producto de Hadamard). La matriz $\overline{\Omega} = \omega^{-1}\Omega\omega^{-1}$ es la matriz de correlación asociada a Ω. La forma de la densidad multivariada resultante depende de una combinación entre Ω y $\boldsymbol{\alpha}$. En la Figura 2 se pueden ver las curvas de nivel de la distribución normal asimétrica bivariada, en dos casos que comparten todos los valores de los parámetros excepto el Ω_{12}. En dicha figura se grafican las curvas de nivel de la normal bivariada de base, en línea punteada, y las de la normal asimétrica correspondiente en línea llena. Ambos gráficos comparten los valores de $\boldsymbol{\alpha}$,$\boldsymbol{\xi}$ y la diagonal de Ω. Vemos que es el efecto combinado entre $\boldsymbol{\alpha}$ y Ω el que afecta la forma de la densidad.

La familia de distribuciones normales multivariadas asimétricas es estudiada en varios artículos de Azzalini, en particular, Azzalini y Dalla Valle [1996]. La notación que presentamos aquí aparece en Azzalini y Capitanio [2003]. Un tratamiento sistemático de esta distribución, desarrollada en forma independiente de trabajos previos, en la que figuran los resultados importantes y varias generalizaciones puede encontrarse en el libro de Azzalini y Capitanio [2014]. Como explica Azzalini [2005], la razón por la cual se incluye el término aparentemente redundante ω^{-1} en el argumento de Φ es mantener el

Figura 2: Curvas de nivel de dos densidades normales asimétricas bivariadas de parámetros $\xi = \mathbf{0}, \boldsymbol{\alpha} = (5, -3)^T, \Omega_{11} = \Omega_{22} = 1$ en ambos casos, y $\Omega_{12} = -0{,}7$ para el gráfico de la izquierda y $\Omega_{12} = 0{,}7$ para el gráfico de la derecha. La flecha representa el vector $\boldsymbol{\alpha}$ dividido por su norma euclídea. Fuente Azzalini y Capitanio [2014].



parámetro de forma $\boldsymbol{\alpha}$ inalterado cuando se transforma linealmente al vector $\mathbf{y}$ definiendo $\mathbf{y}' = \mathbf{a} + B\mathbf{y}$ para una matriz *diagonal definida positiva* B : en ese caso cambia la posición y la escala pero no cambia $\boldsymbol{\alpha}$.

Se puede calcular la función generadora de momentos de un vector $\mathbf{y} \sim SN_p(\boldsymbol{\xi}, \Omega, \boldsymbol{\alpha})$ que está dada por

$$M(\mathbf{t}) = 2\exp\left(\boldsymbol{\xi}^T\mathbf{t} + \frac{1}{2}\mathbf{t}^T\Omega\mathbf{t}\right)\Phi\left(\boldsymbol{\delta}^T\omega\mathbf{t}\right), \quad \mathbf{t} \in \mathbb{R}^p,$$

donde

$$\boldsymbol{\delta} = \frac{\overline{\Omega}\boldsymbol{\alpha}}{\left(1 + \boldsymbol{\alpha}^T\overline{\Omega}\boldsymbol{\alpha}\right)^{1/2}},$$

y $\overline{\Omega}$ es la matriz de correlación asociada a Ω,

$$\overline{\Omega} = \omega^{-1}\Omega\omega^{-1}.$$

A partir de $M(\mathbf{t})$ se obtiene

$$E(\mathbf{y}) = \boldsymbol{\xi} + \omega\sqrt{\frac{2}{\pi}}\boldsymbol{\delta} \qquad \text{y} \qquad Var(\mathbf{y}) = \Omega - \frac{2}{\pi}\omega\boldsymbol{\delta}\boldsymbol{\delta}^T\omega. \tag{12}$$

Observemos que aunque el parámetro de posición $\boldsymbol{\xi}$ sea cero, esto no garantiza que el vector $\mathbf{y}$ tenga esperanza cero.

Hay varias construcciones explícitas que permiten obtener vectores aleatorios con distribución SN. Para los propósitos de la tesis será de interés el siguiente acoplamiento o construcción de un vector aleatorio con distribución normal asimétrica multivariada p-dimensional, a partir de un vector aleatorio $(p+1)-$dimensional normal multivariado.

Representación estocástica o acoplamiento. Sean $Y_0 \sim N(0,1)$ e

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix} \sim N_p\left(\mathbf{0}, \overline{\Psi}\right)$$

independientes, donde $\overline{\Psi}$ es una matriz de correlación de rango completo y de dimensión $p \times p$ (observar que la diagonal de $\overline{\Psi}$ tiene exclusivamente unos). Dado un vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)^T$ con todas sus coordenadas en el intervalo $(-1,1)$, se definen las variables aleatorias

$$Z_j = \delta_j |Y_0| + \sqrt{1-\delta_j^2} Y_j, \qquad j = 1, \ldots, p.$$

Podemos acomodar estas variables en el vector aleatorio $\mathbf{z} = (Z_1, \ldots, Z_p)^T$. Sea

$$D_{\boldsymbol{\delta}} = \left(I_p - diag\left(\boldsymbol{\delta}\right)^2\right)^{1/2},$$

donde $\mathrm{diag}(\boldsymbol{\delta})$ es una matriz diagonal, cuya diagonal es el vector $\boldsymbol{\delta}$. Con ella podemos escribir a $\mathbf{z}$ de una forma más compacta

$$\mathbf{z} = D_{\boldsymbol{\delta}} \mathbf{y} + \delta_j |Y_0|.$$

Entonces, $\mathbf{z} \sim SN_p\left(\mathbf{0}, \overline{\Omega}, \boldsymbol{\alpha}\right)$, donde $\overline{\Omega}$ y $\boldsymbol{\alpha}$ se definen a partir de $\overline{\Psi}$ y $\boldsymbol{\delta}$ por

$$\boldsymbol{\lambda} = D_{\boldsymbol{\delta}}^{-1} \boldsymbol{\delta} \tag{13}$$

$$\overline{\Omega} = D_{\boldsymbol{\delta}} \left(\overline{\Psi} + \boldsymbol{\lambda}\boldsymbol{\lambda}^T\right) D_{\boldsymbol{\delta}} = \Psi + D_{\boldsymbol{\delta}} \boldsymbol{\lambda}\boldsymbol{\lambda}^T D_{\boldsymbol{\delta}} \tag{14}$$

$$\boldsymbol{\alpha} = \frac{D_{\boldsymbol{\delta}}^{-1} \overline{\Psi}^{-1} \boldsymbol{\lambda}}{\left(1 + \boldsymbol{\lambda}^T \overline{\Psi}^{-1} \boldsymbol{\lambda}\right)^{1/2}}. \tag{15}$$

Observación 3.4 *Sea* $\mathbf{z}$ *un vector aleatorio con distribución normal asimétrica* $p-$*dimensional,* $\mathbf{z} \sim SN_p\left(\mathbf{0}, \overline{\Omega}, \boldsymbol{\alpha}\right)$, *con* $p = 2d$. *Tomemos* $c_1, c_2, \ldots, c_d \in (-1,1)$ *de modo que exista al menos un par* i, j *tales que* $c_i \neq c_j$, *sea*

$$\boldsymbol{\delta} = (c_1, c_1, c_2, c_2, \ldots, c_d, c_d) = \sum_{j=1}^{d} c_j \left(\mathbf{e}_{2j-1} + \mathbf{e}_{2j}\right)$$

y tomemos $\Psi = I$. *Sea*

$$\mathbf{u} = \mathbf{z} - E(\mathbf{z}) = \mathbf{z} - \sqrt{\frac{2}{\pi}} \boldsymbol{\delta},$$

por (12), y consideramos el modelo PFC con este error. Tomamos $\Delta_0 = I$, *y*

$$\Gamma_0 = \begin{pmatrix} \mathbf{e}_1 + \mathbf{e}_2 & \mathbf{e}_3 + \mathbf{e}_4 & \cdots & \mathbf{e}_{p-1} + \mathbf{e}_p \end{pmatrix} \frac{1}{\sqrt{2}},$$

donde $\mathbf{e}_i$ *es el iésimo vector canónico en* $\mathbb{R}^p$. *En este caso, podemos tomar*

$$\Gamma_1 = \begin{pmatrix} \mathbf{e}_1 - \mathbf{e}_2 & \mathbf{e}_3 - \mathbf{e}_4 & \cdots & \mathbf{e}_{p-1} - \mathbf{e}_p \end{pmatrix} \frac{1}{\sqrt{2}},$$

y, por el acoplamiento presentado, tenemos

$$\begin{aligned} p_{\Delta_0}\left(\Delta_0^{1/2}\mathbf{u}, \operatorname{span}(\Gamma_0)\right) &= \Gamma_0\Gamma_0^T\Delta_0^{-1/2}\mathbf{u} \\ &= \frac{1}{2}\sum_{j=1}^{d}\left(u_{2j-1} + u_{2j}\right)\left(\mathbf{e}_{2j-1} + \mathbf{e}_{2j}\right) \\ &= \frac{1}{2}\sum_{j=1}^{d}\left(Z_{2j-1} + Z_{2j} - E\left(Z_{2j-1} + Z_{2j}\right)\right)\left(\mathbf{e}_{2j-1} + \mathbf{e}_{2j}\right) \\ &= \frac{1}{2}\sum_{j=1}^{d}\left[2c_j\left|Y_0\right| + \sqrt{1-c_j^2}\left(Y_{2j-1} + Y_{2j}\right) - 2c_j\sqrt{\frac{2}{\pi}}\right]\left(\mathbf{e}_{2j-1} + \mathbf{e}_{2j}\right) \end{aligned}$$

y,

$$\begin{aligned} p_{\Delta_0}\left(\Delta_0^{1/2}\mathbf{u}, \operatorname{span}(\Gamma_0)^{\perp}\right) &= \Gamma_1\Gamma_1^T\Delta_0^{1/2}\mathbf{u} \\ &= \frac{1}{2}\sum_{j=1}^{d}\left(u_{2j-1} - u_{2j}\right)\left(\mathbf{e}_{2j-1} - \mathbf{e}_{2j}\right) \\ &= \frac{1}{2}\sum_{j=1}^{d}\left(Z_{2j-1} - Z_{2j} - E\left(Z_{2j-1} - Z_{2j}\right)\right)\left(\mathbf{e}_{2j-1} - \mathbf{e}_{2j}\right) \\ &= \frac{1}{2}\sum_{j=1}^{d}\sqrt{1-c_j^2}\left(Y_{2j-1} - Y_{2j}\right)\left(\mathbf{e}_{2j-1} - \mathbf{e}_{2j}\right) \end{aligned}$$

puesto que, por (12) se tiene

$$\begin{aligned} E\left(Z_{2j-1} + Z_{2j}\right) &= 2c_j\sqrt{\frac{2}{\pi}} \\ E\left(Z_{2j-1} - Z_{2j}\right) &= c_j\sqrt{\frac{2}{\pi}} - c_j\sqrt{\frac{2}{\pi}} = 0. \end{aligned}$$

Como $(Y_0, Y_1, \ldots, Y_p)^T \sim N_{p+1}(\mathbf{0}, I)$, *en particular sus coordenadas son independientes. Además, como el vector*

$$\left(Y_0, \frac{(Y_1+Y_2)}{\sqrt{2}}, \frac{(Y_1-Y_2)}{\sqrt{2}}, \frac{(Y_3+Y_4)}{\sqrt{2}}, \frac{(Y_3-Y_4)}{\sqrt{2}}, \ldots, \frac{(Y_{p-1}+Y_p)}{\sqrt{2}}, \frac{(Y_{p-1}-Y_p)}{\sqrt{2}}\right)^T$$

también tiene todas sus coordenadas independientes (por ser una transformación ortogonal de $(Y_0, Y_1, \ldots, Y_p)^T$*) y, como* $p_{\Delta_0}\left(\Delta_0^{1/2}\mathbf{u}, \operatorname{span}(\Gamma_0)\right)$ *es función de las coordenadas de índice par y de la primera, mientras que* $p_{\Delta_0}\left(\Delta_0^{1/2}\mathbf{u}, \operatorname{span}(\Gamma_0)^{\perp}\right)$ *es función de las coordenadas impares (mayores o iguales a tres), es decir, ambos son función de conjuntos disjuntos de coordenadas del vector aleatorio* $(Y_0, \mathbf{y})$*, resultan independientes. En este caso, la reducción es suficiente. Observemos que además, como* $c_i \neq c_j$ *para al menos un par de índices* i, j*, el vector* $\mathbf{z}$ *no tiene todas sus coordenadas idénticamente distribuidas, de modo que este ejemplo es distinto del presentado en la Observación 3.3.*

4. Estimador de Máxima Verosimilitud bajo normalidad

4.1. Estimador basado en una muestra

Supongamos que $(\mathbf{x}, y)$ satisface el modelo PFC y que $\mathbf{u}$ tiene distribución normal con media $\mathbf{0}$ y matriz de covarianza I. Sea $(\mathbf{x}_i, y_i), 1 \leq i \leq n$, una muestra i.i.d. de este modelo. En Cook y Forzani [2008], se derivan los estimadores de máxima verosimilitud (EMV) de los parámetros del modelo y consecuentemente también los del espacio de reducción suficiente.

Bajo normalidad, para todo $\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta) \in \Theta$ la función de verosimilitud está dada por

$$\begin{aligned} L(\boldsymbol{\theta}) &= L(\boldsymbol{\mu}, B, \Delta) \\ &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2}} |\Delta|^{-1/2} \exp\left\{ -\frac{1}{2} \left(\mathbf{x}_i - \boldsymbol{\mu} - B\mathbf{f}(y_i)\right)^T \Delta^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu} - B\mathbf{f}(y_i)\right) \right\}. \end{aligned} \tag{16}$$

El EMV $\widehat{\boldsymbol{\theta}}_{MV} = \left(\widehat{\boldsymbol{\mu}}_{MV}, \widehat{B}_{MV}, \widehat{\Delta}_{MV}\right)$ queda definido por ser aquella combinación de los parámetros $\boldsymbol{\theta} \in \Theta$ que maximiza a la función L. Es decir,

$$\widehat{\boldsymbol{\theta}}_{MV} = \arg \max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}).$$

Llamando

$$\boldsymbol{\mu}_i = \boldsymbol{\mu} + B\mathbf{f}(y_i), \tag{17}$$

el logaritmo de la función de verosimilitud está dado por

$$\ln\left(L(\boldsymbol{\mu}, B, \Delta)\right) = -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln|\Delta| - \frac{1}{2}\sum_{i=1}^{n} \left(\mathbf{x}_i - \boldsymbol{\mu}_i\right)^T \Delta^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}_i\right) \tag{18}$$

donde $|\Delta|$ indica el determinante de la matriz Δ.

4.2. Obteniendo el EMV

Se buscan valores de $\boldsymbol{\theta} \in \Theta$ que maximicen el $\ln\left(L(\boldsymbol{\theta})\right)$. Sean

$$\begin{aligned} \bar{\mathbf{x}} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i, \\ \mathbf{z}_i &= \mathbf{x}_i - \bar{\mathbf{x}}, \\ \bar{\mathbf{f}} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{f}(y_i), \end{aligned}$$

$\mathbb{X}$ la matriz de dimensión $n \times p$ que tiene a $\mathbf{z}_1^T, \ldots, \mathbf{z}_n^T$ por filas y $\mathbb{F}$ una matriz de dimensión $n \times r$ cuyas filas son $\left(\mathbf{f}\left(y_i\right) - \bar{\mathbf{f}}\right)^T$. Sean

$$\widehat{\Sigma}_{\text{fit}} = \frac{1}{n}\mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}\mathbb{F}^T\mathbb{X}$$

$$\widehat{\Sigma} = \frac{1}{n}\mathbb{X}^T\mathbb{X} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^T$$

y

$$\widehat{\Sigma}_{\text{res}} = \widehat{\Sigma} - \widehat{\Sigma}_{\text{fit}}.$$

Luego, $\widehat{\Sigma}$ resulta la matriz de covarianza muestral de las $\mathbf{x}'s$. Además $\widehat{A} = \mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}$ es la matriz de coeficientes ajustados para el modelo de regresión lineal multivariada de $\mathbf{x}_i$ en $\mathbf{f}\left(y_i\right)$. Obsérvese que $\widehat{A}$ no tiene restricciones de rango.

Sea C la matriz ortogonal $[\mathbf{c}_1 \cdots \mathbf{c}_p]$ donde $\mathbf{c}_i$ es un autovector de $\widehat{\Sigma}_{\text{res}}^{-1/2}\widehat{\Sigma}_{\text{fit}}\widehat{\Sigma}_{\text{res}}^{-1/2}$ correspondiente al autovalor π_i, y $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_p$. Denotemos por $\widehat{\Pi}$ a la matriz diagonal, $\widehat{\Pi} = \text{diag}\left(\pi_1, \ldots, \pi_p\right)$, entonces

$$\widehat{\Sigma}_{\text{res}}^{-1/2}\widehat{\Sigma}_{\text{fit}}\widehat{\Sigma}_{\text{res}}^{-1/2} = C\widehat{\Pi}C^T.$$

Sea C_h la matriz con las primeras h columnas de C.

Teorema 4.1 *Sean $(\mathbf{x}_i, y_i)_{1 \leq i \leq n}$ vectores de $\mathbb{R}^{p+1}$ independientes que cumplen el modelo PFC*

$$\mathbf{x}_i = \boldsymbol{\mu}_0 + \Gamma_0\beta_0\mathbf{f}(y_i) + \Delta_0^{1/2}\mathbf{u}$$

para $\mathbf{f}$ una función conocida, y los vectores $\mathbf{u}_i$ con distribución $\mathrm{N}_p\left(\mathbf{0}, I\right)$. Entonces,

(a) El EMV de Δ es

$$\widehat{\Delta} = \widehat{\Sigma}_{\text{res}} + \widehat{\Sigma}_{\text{fit}}^{1/2}C_{p-d}\ C_{p-d}^T\widehat{\Sigma}_{\text{fit}}^{1/2}.$$

(b) Sean $\widehat{\mathbf{t}}_i$, $1 \leq i \leq p$ los autovectores ortonormales de $\widehat{\Delta}^{-1/2}\widehat{\Sigma}_{\text{fit}}\widehat{\Delta}^{-1/2}$ correspondientes a los autovalores $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_p$. El EMV de $\mathcal{S}_\Gamma$, $\widehat{\mathcal{S}}_\Gamma$ es el subespacio generado por $\widehat{\Delta}^{1/2}\widehat{\mathbf{t}}_1, \ldots, \widehat{\Delta}^{1/2}\widehat{\mathbf{t}}_K$. El EMV de $\boldsymbol{\mu}$ es

$$\widehat{\boldsymbol{\mu}} = \bar{\mathbf{x}}.$$

Una vez fijadas las columnas de $\widehat{\Gamma}$ como cualquier base ortonormal de $\widehat{\mathcal{S}}_\Gamma$ resulta $\widehat{\beta} = \left(\widehat{\Gamma}^T\widehat{\Delta}^{-1}\widehat{\Gamma}\right)^{-1}\widehat{\Gamma}^T\widehat{\Delta}^{-1}\widehat{A}$ donde $\widehat{A} = \mathbb{X}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}$. Luego, en virtud de (11) resulta que

$$\widehat{B} = \widehat{\Gamma}\widehat{\beta} = \widehat{\Gamma}\left(\widehat{\Gamma}^T\widehat{\Delta}^{-1}\widehat{\Gamma}\right)^{-1}\widehat{\Gamma}^T\widehat{\Delta}^{-1}\widehat{A}$$

$$= p_{\widehat{\Delta}}\left(\widehat{A}, \text{span}\left(\widehat{\Gamma}\right)\right),$$

por lo que el EMV de B es la proyección ortogonal, en el producto interno inducido por $\widehat{\Delta}$, de $\widehat{A}$ en el subespacio $\widehat{\mathcal{S}}_\Gamma$ (observemos que $\widehat{B}$ así definida sí tiene rango reducido).

(c) $\widehat{\Delta}^{1/2}\widehat{\mathbf{t}}_i$ *es un autovector de* $\widehat{\Sigma}_{\text{fit}}\widehat{\Sigma}_{\text{res}}^{-1}$ *correspondiente al autovalor* $1/\pi_{p-i+1}$.

(d) $\widehat{\Delta}$ *también puede ser escrito de la siguiente forma*

$$\widehat{\Delta} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{B}\mathbf{f}\left(y_i\right)\right)\left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{B}\mathbf{f}\left(y_i\right)\right)^T \tag{19}$$

y satisface

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{B}\mathbf{f}\left(y_i\right)\right)^T\widehat{\Delta}^{-1}\left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{B}\mathbf{f}\left(y_i\right)\right) = p. \tag{20}$$

Los ítems (a) y (b) son parte del Teorema 3.1 de Cook y Forzani [2008], y de los resultados anteriores de la misma publicación. El ítem (c) es una combinación del Apéndice A.7 de Cook [2007] y el Corolario 3.2 de Cook y Forzani [2008]. La prueba de (d) está en el Apéndice D.1.

Maximizar la función de verosimilitud, con el requisito de que $\boldsymbol{\theta} \in \Theta$ es, por supuesto, equivalente a resolver el siguiente problema

$$\min_{\boldsymbol{\theta}\in\Theta}\left\{\ln|\Delta| + \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \boldsymbol{\mu}_i\right)^T\Delta^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu}_i\right)\right\}. \tag{21}$$

Si llamamos

$$g(\boldsymbol{\theta}) = \ln|\Delta|$$

y

$$h\left(\boldsymbol{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \boldsymbol{\mu}_i\right)^T\Delta^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu}_i\right),$$

el problema de hallar el estimador de máxima verosimilitud está dado por

$$\min_{\boldsymbol{\theta}\in\Theta} l(\boldsymbol{\theta}), \tag{22}$$

donde $l(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + h\left(\boldsymbol{\theta}\right)$.

En el Lema 4.1 mostramos una forma equivalente de escribir el problema de máxima verosimilitud para el modelo PFC.

Lema 4.1 *Son equivalentes*

i. El problema de máxima verosimilitud para PFC dado por (22).

ii. El problema

$$\min_{\boldsymbol{\theta}\in\Theta} g(\boldsymbol{\theta}) \tag{23}$$

sujeto a la condición

$$h\left(\boldsymbol{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \boldsymbol{\mu} - B\mathbf{f}(y_i)\right)^T \Delta^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu} - B\mathbf{f}(y_i)\right) = p. \tag{24}$$

La demostración de este resultado está en el Apéndice D.

4.3. Escalas

A continuación definimos una función de escala.

Definición 4.1 *Sea* $\mathbf{u} = (u_1, \ldots, u_n)$ *una muestra de tamaño* n*. Diremos que una función* $s : \mathbb{R}^n \to [0, +\infty)$ *es una* ***escala*** *si*

i. $s\left(\lambda\mathbf{u}\right) = \lambda s\left(\mathbf{u}\right)$*, para* $\lambda > 0$*.*

ii. $s\left(\mathbf{u}\right) = s\left(\left|u_1\right|, \ldots, \left|u_n\right|\right)$

iii. Si $\left|u_i\right| < \left|v_i\right|$*, para todo* $1 \leq i \leq n$*, entonces* $s\left(\mathbf{u}\right) \leq s\left(\mathbf{v}\right)$*.*

Las funciones de escala se utilizan para medir las magnitudes absolutas de un conjunto de números. Muchos estimadores (para distintos modelos) se definen minimizando una escala determinada de la diferencia entre el valor observado y el valor ajustado por el método propuesto (a esta diferencia suele denominársela residuos del modelo).

Veamos algunos ejemplos de escalas.

Ejemplo 4.2 ***Escala cuadrática.***

$$s_2\left(\mathbf{u}\right) = \frac{1}{n}\sum_{i=1}^{n} u_i^2. \tag{25}$$

Es la escala más utilizada. Los estimadores construidos en base a ella suelen denominarse estimadores de mínimos cuadrados. No es robusta.

Ejemplo 4.3 ***M-escalas.*** *Fueron introducidos por Huber [1981]. Se definen a partir de una función* $\rho : \mathbb{R} \to [0, +\infty)$ *que satisface*

A1. $\rho(0) = 0$*,* $\rho(u) = \rho(-u)$*,* ρ *es de clase* C^2*. Denotaremos por* ψ *a la derivada de* ρ*.*

A2. Existe una constante $c > 0$ *tal que* ρ *es estrictamente creciente en* $[0, c]$ *y constante en* $[c, +\infty)$*.*

Sea $a = \rho(c)$. *Sea* $\kappa \in (0, \sup \rho)$. *Una* ***M-escala*** $\mathbf{s}(u_1, \ldots, u_n)$ *se define (implícitamente) por*

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{u_i}{\mathbf{s}(u_1, \ldots, u_n)}\right) = \kappa. \tag{26}$$

La escala cuadrática puede ser vista como una M-escala, tomando $\rho(u) = u^2$ *y* $\kappa = 1$, *aunque en ese caso la* $\rho-$*función no cumple con ser acotada. Si la* $\rho-$*función cumple las condiciones A1 y A2, la M-escala resulta robusta.*

Ejemplo 4.4 *La mediana de los valores absolutos*

$$s(u_1, \ldots, u_n) = \text{mediana}\{|u_1|, \ldots, |u_n|\}$$

es una escala robusta.

Ejemplo 4.5 *La escala* L_1

$$s_1(u_1, \ldots, u_n) = \frac{1}{n}\sum_{i=1}^{n} |u_i|$$

es otra escala robusta.

Ejemplo 4.6 $\tau-$***escalas.*** *Fueron introducidas por Yohai y Zamar [1988]. Consideramos dos funciones* ρ_1 *y* ρ_2 *definidas en* $[0, +\infty)$ *que satisfacen las condiciones A1 y A2 (página 34). Además, imponemos la siguiente condición, sólo en la función* ρ_2,

A3. $2\rho_2(t) - \psi_2(t)\, t > 0$ *para* $t > 0$.

Definimos una $\tau-$ ***escala*** *para la muestra* $\mathbf{u} = (u_1, \ldots, u_n)$ *por*

$$\tau^2(\mathbf{u}) = s^2(\mathbf{u}) \left[\frac{1}{n}\sum_{i=1}^{n} \rho_2\left(\frac{u_i}{s(\mathbf{u})}\right)\right], \tag{27}$$

donde $s(\mathbf{u})$ *es una M-escala basada en* ρ_1 *y* κ_1 *(en lugar de* ρ *y* κ*) definida en (26).*

Una $\tau-$*escala puede verse como un múltiplo adaptativo de una M-escala. Observemos que si se elige* $\rho_2(u) = u^2$, *resulta que* $\tau^2(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} u_i^2$, *por lo que las* $\tau-$*escalas pueden verse como una generalización de la escala cuadrática. Nuevamente, esta elección de* ρ_2 *resulta no acotada. Sin embargo, tomando a* ρ_2 *como una función acotada, que cerca del origen se parezca a la cuadrática, puede probarse que la* $\tau-$*escala construida con esta función resulta eficiente. Asimismo, si la función* ρ_1 *satisface las condiciones requeridas, y* κ_1 *es elegido de manera apropiada, la* $\tau-$*escala resulta robusta.*

Tomando en cuenta las definiciones de función de escala realizada, podemos reescribir la definición del EMV para el modelo PFC.

Sean $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ vectores aleatorios independientes idénticamente distribuidos que satisfacen el modelo PFC. Para cada observación $(\mathbf{x}_i, y_i)$ de la muestra, y cada $\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta) \in \Theta$ se puede definir la distancia

$$d_i(\boldsymbol{\theta}) = d(\mathbf{x}_i, y_i, \boldsymbol{\theta}) = \left\{ (\mathbf{x}_i - \boldsymbol{\mu} - B\mathbf{f}(y_i))^T \Delta^{-1} (\mathbf{x}_i - \boldsymbol{\mu} - B\mathbf{f}(y_i)) \right\}^{1/2}. \quad (28)$$

A la distancia d_i dada en (28) la denominaremos *distancia de Mahalanobis*, de la i-ésima observación, haciendo abuso de notación, ya que en realidad es la efectiva distancia de Mahalanobis cuando el parámetro $\boldsymbol{\theta}$ toma el valor verdadero, es decir, cuando $\boldsymbol{\theta} = \boldsymbol{\theta}_0 = (\boldsymbol{\mu}_0, B_0, \Delta_0)$. Observemos que

$$d\left(\mathbf{x}_i, y_i, \widehat{\boldsymbol{\theta}}_{MV}\right) = \left\| \widehat{\Delta}_{MV}^{-1/2} \left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{MV} - \widehat{B}_{MV}\mathbf{f}(y_i) \right) \right\|, \quad (29)$$

es decir, la distancia de Mahalanobis de la i-ésima observación, tomada con los parámetros estimados es igual a la norma de los residuos estandarizados del modelo PFC ajustado.

La equivalencia dada en el Lema 4.1 permite escribir al EMV del siguiente modo:

$$\widehat{\boldsymbol{\theta}}_{MV} = \arg \min_{\boldsymbol{\theta} \in \Theta} |\Delta| \quad (30)$$

sujeto a la condición

$$h(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} d^2(\mathbf{x}_i, y_i, \boldsymbol{\theta}) = p. \quad (31)$$

Una manera alternativa de escribir la condición (31) es

$$s_2(d_1(\boldsymbol{\theta}), \ldots, d_n(\boldsymbol{\theta})) = p \quad (32)$$

donde la escala s_2 es la escala cuadrática definida en el Ejemplo 4.2.

Escrito de este modo, el problema de hallar estimadores para el modelo PFC puede ser visto como un problema de encontrar los valores de los parámetros que minimizan el determinante de la matriz de covarianza de los errores (es decir, una medida del tamaño de los errores) sujeto a la condición (32) que impone que la escala (es decir, una medida del tamaño) de las distancias de Mahalanobis de las observaciones, sea fija. Observemos que esta condición (32) evita que el mínimo se alcance en un valor degenerado de los parámetros del modelo. Otra forma de verlo es pensar que el EMV elige el valor del estimador como el que minimiza una función objetivo, entre muchos competidores que comparten el valor de la escala cuadrática.

4.4. Propuesta de $\tau-$estimador para el modelo PFC

La escala cuadrática considerada para definir al EMV es sensible a la presencia de observaciones atípicas. Nosotros proponemos un estimador robusto para el modelo PFC cambiando esta escala por una escala más robusta, es decir, una $\tau-$escala que tiene alto punto de ruptura y a la vez, alta eficiencia bajo normalidad. Ahora, podemos formalizar la definición de $\tau-$estimadores para el modelo PFC.

Fijemos $\kappa_2 \in [0, \sup \rho_2]$.

Definición 4.2 *Para las observaciones* $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ *independientes idénticamente distribuidas que satisfacen el modelo PFC, definimos los* τ ***estimadores para el modelo PFC*** *por los vectores y matrices* $\left(\widehat{\boldsymbol{\mu}}, \widehat{B}, \widehat{\Delta}\right)$ *que resuelven el siguiente problema*

$$\min_{\boldsymbol{\theta} \in \Theta} |\Delta| \tag{33}$$

sujeto a la condición

$$\tau^2 \left(\mathbf{d}\left(\boldsymbol{\theta}\right)\right) = \kappa_2, \tag{34}$$

donde $\mathbf{d}\left(\boldsymbol{\theta}\right) = \left(d_1\left(\boldsymbol{\theta}\right), \ldots, d_n\left(\boldsymbol{\theta}\right)\right)$ *y cada coordenada* $d_i\left(\boldsymbol{\theta}\right)$ *fue definida en (28).*

Observemos que en virtud del Lema 4.1, esta definición coincide con la de máxima verosimilitud bajo normalidad bajo el modelo PFC, cuando tomamos como τ^2 a la definida en (25).

En la siguiente sección hacemos una introducción a los estimadores definiéndolos a partir de funcionales de estimación. Esta aproximación permite escribir las demostraciones de algunos resultados de forma más sencilla.

5. Definición del $\tau-$funcional y propiedades

Supongamos que se observa un vector $\mathbf{z}$ en $\mathbb{R}^q$ y se quiere estimar un parámetro $\boldsymbol{\theta}$ que varía en Θ usando una muestra aleatoria $\mathbf{z}_1, \ldots, \mathbf{z}_n$. Una forma de definir un estimador de $\boldsymbol{\theta}$ es utilizando un funcional $\mathbf{T} : \mathcal{P}(\mathbb{R}^q) \to \Theta$, donde si E es un espacio medible, entonces $\mathcal{P}(E)$ es el espacio de distribuciones sobre E. Luego un estimador de $\boldsymbol{\theta}$ puede ser definido como $\widehat{\boldsymbol{\theta}} = \mathbf{T}(F_n)$, donde F_n es la distribución empírica de la muestra. En la sección 5.1 damos la definición de un funcional que genera los τ-estimadores para el modelo de PFC. En las secciones 5.2 y 5.3 estudiamos las propiedades de este funcional.

5.1. Definición del $\tau-$funcional de estimación

Consideramos dos funciones ρ_1 y ρ_2, que serán dos $\rho-$funciones definidas en $[0, +\infty)$ que satisfacen las condiciones A1, A2 y A3 establecidas en las páginas 34 y 35.

Sea P_0 una probabilidad definida en $\mathbb{R}_{\geq 0}$, sean

$$\kappa_i := E_{P_0}\left(\rho_i\left(d\right)\right), \; i = 1, 2 \tag{35}$$

Antes de definir el $\tau-$funcional de estimación para el modelo PFC, definimos los M$-$funcionales de escala y los $\tau-$funcionales de escala.

Definición 5.1 *Un* ***M$-$funcional de escala*** $S_M : \mathcal{P}\left([0, +\infty)\right) \to \mathbb{R}_{>0}$ *está dado por el valor* $S_M\left(H\right)$ *que satisface*

$$E_H\left(\rho_1\left(\frac{v}{S_M\left(H\right)}\right)\right) = \kappa_1, \tag{36}$$

cuando v *tiene distribución* H*. De esta forma,* $S_M\left(P_0\right) = 1$*.*

Definición 5.2 *Un* ***$\tau-$funcional de escala*** $S_\tau : \mathcal{P}\left([0, +\infty)\right) \to \mathbb{R}_{>0}$ *está dado por*

$$S_\tau^2\left(H\right) = S_M^2\left(H\right) E_H\left(\rho_2\left(\frac{v}{S_M\left(H\right)}\right)\right) \tag{37}$$

donde $S_M\left(H\right)$ *es un M$-$funcional de escala definido previamente, basado en* ρ_1*.*

Recordemos el espacio de posibles parámetros para el modelo PFC, dado por

$$\boldsymbol{\Theta} = \left\{\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta) \in \mathbb{R}^p \times \mathbb{R}^{p \times r} \times PDS\left(p\right) : \text{rango}\left(B\right) \leq d\right\},$$

donde $PDS\left(p\right)$ es la clase de todas las matrices $p \times p$ reales simétricas y definidas positivas. Como ya dijimos, la matriz B también puede ser representada por el producto $\Gamma\beta$. Sea $(\mathbf{x}, y)$ un vector aleatorio que satisface el modelo PFC, y $\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta) \in \boldsymbol{\Theta}$. Luego definimos

$$d\left(\mathbf{x}, y, \boldsymbol{\theta}\right) = \left(\left(\mathbf{x} - \boldsymbol{\mu} - B\mathbf{f}(y)\right)^T \Delta^{-1} \left(\mathbf{x} - \boldsymbol{\mu} - B\mathbf{f}(y)\right)\right)^{1/2}. \tag{38}$$

Para cada $P \in \mathcal{P}\left(\mathbb{R}^{p+1}\right)$ y $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, definimos $H_{P,\boldsymbol{\theta}}$ como la distribución de $d\left(\mathbf{x}, y, \boldsymbol{\theta}\right)$ cuando $(\mathbf{x}, y)$ tiene distribución P.

Definición 5.3 *Un* $\tau-$***funcional de estimación para el modelo PFC*** *que notaremos por* $\boldsymbol{\theta}_\tau(P)$, $\boldsymbol{\theta}_\tau : \mathcal{P}\left(\mathbb{R}^{p+1}\right) \to \boldsymbol{\Theta}$, *está dado por el valor de* $\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta) \in \boldsymbol{\Theta}$ *que minimiza* $|\Delta|$ *sujeto a*

$$S_\tau^2\left(H_{P,\boldsymbol{\theta}}\right) = \kappa_2, \tag{39}$$

donde $|\Delta|$ *indica determinante de* Δ.

5.2. Formulaciones equivalentes del $\tau-$funcional

En esta sección mostraremos algunas definiciones equivalentes del τ-funcional. Para hacerlo, veamos primero algunas propiedades de los funcionales de tipo M y τ recién definidos.

Lema 5.1 *Sea* $\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta) \in \boldsymbol{\Theta}$ *y* $\lambda > 0$. *Llamemos* $\boldsymbol{\theta}_\lambda := (\boldsymbol{\mu}, B, \lambda\Delta)$.

i. La condición

$$E_{H_{P,\boldsymbol{\theta}}}\left(\rho_1(v)\right) = \kappa_1. \tag{40}$$

es equivalente a la condición $S_M\left(H_{P,\boldsymbol{\theta}}\right) = 1$. *Además, para todo* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, *si elegimos* $\lambda = S_M^2\left(H_{P,\boldsymbol{\theta}}\right)$ *entonces* $\boldsymbol{\theta}_\lambda$ *satisface la condición (40), o sea* $E_{H_{P,\boldsymbol{\theta}_\lambda}}\left(\rho_1(v)\right) = \kappa_1$.

ii. Para todo $\lambda > 0$, $S_M^2\left(H_{P,\boldsymbol{\theta}_\lambda}\right) = \frac{1}{\lambda} S_M^2\left(H_{P,\boldsymbol{\theta}}\right)$.

iii. Para todo $\lambda > 0$, $S_\tau^2\left(H_{P,\boldsymbol{\theta}_\lambda}\right) = \frac{1}{\lambda} S_\tau^2\left(H_{P,\boldsymbol{\theta}}\right)$.

La demostración de este resultado está en el Apéndice E.

En lo que sigue, mantendremos la convención de notar $\boldsymbol{\theta}_\lambda := (\boldsymbol{\mu}, B, \lambda\Delta)$ para todo $\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta) \in \boldsymbol{\Theta}$ y $\lambda > 0$. Consideremos los siguientes cuatro funcionales $\boldsymbol{\theta}_A(P)$,$\boldsymbol{\theta}_B(P)$, $\boldsymbol{\theta}_C(P)$,$\boldsymbol{\theta}_D(P)$,

A $\boldsymbol{\theta}_A(P)$ está definido como en la Definición 5.3. Es decir, $\boldsymbol{\theta}_A(P) = \boldsymbol{\theta}_\tau(P)$.

B $\boldsymbol{\theta}_B(P)$ se define como el valor $\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta) \in \boldsymbol{\Theta}$ tal que minimiza

$$\Phi_{B,P}(\boldsymbol{\theta}) = |\Delta| \left[S_\tau^2\left(H_{P,\boldsymbol{\theta}}\right)\right]^p \tag{41}$$

y además satisface (39). Obsérvese que si $\boldsymbol{\theta}_1 = (\boldsymbol{\mu}, B, \Delta)$ y $\boldsymbol{\theta}_2 = (\boldsymbol{\mu}, B, \lambda\Delta)$, donde $\lambda > 0$ entonces $\Phi_{B,P}(\boldsymbol{\theta}_1) = \Phi_{B,P}(\boldsymbol{\theta}_2)$. Luego, por el Lema 5.1(iii), entre los valores que minimizan $\Phi_{B,P}(\boldsymbol{\theta})$, siempre se puede encontrar uno que satisface (39).

C Para definir a $\boldsymbol{\theta}_C(P)$, definimos antes un funcional auxiliar. $\widetilde{\boldsymbol{\theta}}_C(P)$ se define como el valor $\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta) \in \boldsymbol{\Theta}$ que minimiza

$$\Phi_{C,P}(\boldsymbol{\theta}) = |\Delta| \left[E_{H_{P,\boldsymbol{\theta}}}\left(\rho_2(v)\right)\right]^p \tag{42}$$

sujeto a

$$E_{H_{P,\boldsymbol{\theta}}}\left(\rho_1(v)\right) = \kappa_1. \tag{43}$$

Sea $q = \frac{S_\tau^2\left(H_{P,\widetilde{\boldsymbol{\theta}}_C(P)}\right)}{\kappa_2}$. Entonces, definimos a $\boldsymbol{\theta}_C(P)$ de la siguiente forma

$$\boldsymbol{\theta}_C(P) := \left(\widetilde{\boldsymbol{\theta}}_C(P)\right)_q = (\boldsymbol{\mu}, B, q\Delta)\,.$$

D Primero se define el funcional auxiliar $\widetilde{\boldsymbol{\theta}}_D(P)$ como el valor $\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta) \in \boldsymbol{\Theta}$ que minimiza $S_\tau^2\left(H_{P,\boldsymbol{\theta}}\right)$ con la restricción

$$|\Delta| = 1. \tag{44}$$

Sea $u = \frac{S_\tau^2\left(H_{P,\widetilde{\boldsymbol{\theta}}_D(P)}\right)}{\kappa_2}$. Entonces,

$$\begin{aligned}\boldsymbol{\theta}_D(P) &:= \left(\widetilde{\boldsymbol{\theta}}_D(P)\right)_u = \left(\widetilde{\boldsymbol{\mu}}_D(P), \widetilde{B}_D(P), \widetilde{\Delta}_D(P)\right)_u \\ &= \left(\widetilde{\boldsymbol{\mu}}_D(P), \widetilde{B}_D(P), u\widetilde{\Delta}_D(P)\right).\end{aligned}$$

Observación 5.1 *Observemos que en virtud del Lema 5.1(iii), tanto $\boldsymbol{\theta}_C(P)$ como $\boldsymbol{\theta}_D(P)$ satisfacen*

$$S_\tau^2\left(H_{P,\boldsymbol{\theta}_C(P)}\right) = S_\tau^2\left(H_{P,\boldsymbol{\theta}_D(P)}\right) = \kappa_2.$$

Observación 5.2 *Para todo $\lambda > 0$ se tiene que $\Phi_{B,P}(\boldsymbol{\theta}_\lambda) = \Phi_{B,P}(\boldsymbol{\theta})$.*

Observación 5.3 *Para el problema definido en B, observemos que el valor del mínimo de $\Phi_{B,P}$ es el mismo, ya sea que se lo busque con o sin la restricción (39), por la Observación 5.2. Dicha restricción sólo se impone para poder identificar una solución al problema de minimizar, ya que una vez obtenido dicho valor de $\boldsymbol{\theta}$ el mínimo se alcanza también en $\boldsymbol{\theta}_\lambda$ para todo $\lambda > 0$.*

Los $\tau-$estimadores fueron originalmente propuestos por Yohai y Zamar [1988] para estimar los parámetros del modelo lineal univariado con errores homoscedásticos. En dicho artículo, se presenta a los $\tau-$estimadores como los valores que minimizan una escala τ de los residuos, como en el funcional D, pero sin ninguna restricción, ya que buscan los estimadores de los coeficientes del modelo lineal, y en ese contexto la solución al problema se alcanza en un valor no degenerado de los parámetros. En García Ben, Martínez, y Yohai [2006] proponen $\tau-$estimadores para los coeficientes de regresión y la matriz de covarianza de los errores para el modelo de regresión de respuesta multivariada. Estos estimadores se presentan en forma similar al caso A. La propuesta de Lopuhaä [1991] de estimadores robustos de tipo τ para el modelo de posición y escala multivariado los escribe en los mismos términos de nuestro funcional C. El Lema 5.2 que sigue establece la equivalencia de las cuatro formulaciones bajo el modelo PFC.

Lema 5.2 *Sea $P \in \mathcal{P}\left(\mathbb{R}^{p+1}\right)$, entonces $\boldsymbol{\theta}_A(P) = \boldsymbol{\theta}_B(P) = \boldsymbol{\theta}_C(P) = \boldsymbol{\theta}_D(P)$.*

La demostración de este resultado está en el Apéndice E.

5.3. Equivarianza del $\tau-$funcional de estimación

En esta sección, probaremos que el funcional $\boldsymbol{\theta}_B$ definido en la Sección 5.2 es afín equivariante. Por las equivalencias probadas en la Sección 5.2 esto se traducirá en la equivarianza del $\tau-$ funcional de estimación para el modelo PFC.

Obsérvese que si $(\mathbf{x},y)$ sigue un modelo de PFC con $\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta)$, entonces si definimos a $\mathbf{x}^*$ por $\mathbf{x}^* = A\mathbf{x} + \mathbf{b}$, donde A es una matriz $p \times p$ no singular y $\mathbf{b} \in \mathbb{R}^p$ entonces $(\mathbf{x}^*, y)$ sigue un modelo de PFC con $\boldsymbol{\theta}^* = \left(A\boldsymbol{\mu} + \mathbf{b}, AB, A\Delta A^T\right)$. El siguiente teorema prueba la equivarianza de $\boldsymbol{\theta}_B$.

Teorema 5.1 *Sea P la distribución de $(\mathbf{x},y)$ y P^* la distribución de $(A\mathbf{x}+\mathbf{b},y)$ con $A \in \mathbb{R}^{p\times p}$ no singular y $\mathbf{b} \in \mathbb{R}^p$. Entonces $\boldsymbol{\theta}_B(P) = (\boldsymbol{\mu}(P), B(P), \Delta(P))$ implica que $\boldsymbol{\theta}_B(P^*) = \left(A\boldsymbol{\mu}(P) + \mathbf{b}, AB(P), A\Delta(P)A^T\right)$.*

La demostración de este resultado está en el Apéndice E.

5.4. Distribución empírica y $\tau-$estimador

Como decimos al principio de este capítulo, una manera alternativa de definir a los $\tau-$estimadores es aplicar el $\tau-$funcional de estimación definido en la Sección 5.1 a la distribución empírica que se deduce de una muestra aleatoria obtenida bajo el modelo. Recordemos que la medida empírica P_n o probabilidad inducida por una muestra se define para subconjuntos A medibles de $\mathbb{R}^{p+1}$ por

$$P_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_A(\mathbf{x}_i,y_i) = \frac{1}{n}\sum_{i=1}^{n} \delta_{(\mathbf{x}_i,y_i)}(A)$$

donde I_A es la función indicadora y $\delta_{(\mathbf{x}_i,y_i)}$ es la medida de Dirac, dada por

$$\delta_{(\mathbf{x}_i,y_i)}(A) = \begin{cases} 1 & \text{si } (\mathbf{x}_i,y_i) \in A \\ \\ 0 & \text{si } (\mathbf{x}_i,y_i) \notin A. \end{cases}$$

A partir de la medida empírica, la función de distribución empírica puede escribirse como

$$F_n(\mathbf{x},y) = P_n\left(\{(\mathbf{s},t) : s_j \leq x_j \text{ para } j = 1,\ldots,p \text{ y } t \leq y\}\right).$$

si $\mathbf{x} = (x_1,\ldots,x_p)$.

Luego, los $\tau-$estimadores para el modelo PFC definidos en 4.2 cumplen $\left(\widehat{\boldsymbol{\mu}}, \widehat{B}, \widehat{\Delta}\right) = \boldsymbol{\theta}_\tau(P_n)$.

En virtud de las equivalencias probadas en la Sección 5.2 podemos interpretar al $\tau-$estimador para el modelo PFC mediante el funcional $\boldsymbol{\theta}_D$. A través de este funcional resulta que $\boldsymbol{\theta}_\tau$ puede verse como el valor de los parámetros que minimiza la $\tau-$escala de las distancias de Mahalanobis de las observaciones, sujeto a una restricción que impide

que el mínimo se alcance en un valor degenerado de los parámetros. Como observáramos en (29) en el contexto de observaciones del modelo PFC, la distancia de Mahalanobis de una observación a un candidato $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $d(\mathbf{x}_i, y_i, \boldsymbol{\theta})$ puede interpretarse como la norma del residuo estandarizado cuando consideramos a $\boldsymbol{\theta}$ como el verdadero parámetro del modelo PFC. Luego, tanto el EMV como el estimador robusto propuesto pueden verse como aquellos valores que minimizan una escala (cuadrática en el caso del EMV, robusta de tipo τ para el $\tau-$estimador) de la norma de los residuos estandarizados, sujeto a una restricción sobre el determinante de la matriz de covarianza de los errores que excluye a las matrices de rango menor a p como posibles.

6. Propiedades del $\tau-$funcional

6.1. Existencia

Los resultados probados en esta sección son similares a los de la la Sección 2.2 de Lopuhaä [1991], Llamemos P a la distribución de $(\mathbf{x}, y)$ y supongamos que $\mathbf{f} = (f_1, ..., f_r) : \mathbb{R} \to \mathbb{R}^r$ ha sido fijada. Sea $D_{\mathbf{f}}$ el conjunto de discontinuidades de $\mathbf{f}$,

$$\begin{aligned} D_{\mathbf{f}} &= \left\{(\mathbf{x}, y) \in \mathbb{R}^{p+1} : \ \mathbf{f} \text{ es discontinua en } y\right\} \\ &= \left\{(\mathbf{x}, y) \in \mathbb{R}^{p+1} : h(\mathbf{x},y) = (\mathbf{x}, \mathbf{f}(y)) \text{ es discontinua en } (\mathbf{x}, y)\right\} \\ &= \left\{\mathbf{z} \in \mathbb{R}^{p+1} : h(\mathbf{z}) \text{ es discontinua en } \mathbf{z}\right\} \\ &= \bigcup_{\varepsilon} \bigcap_{\delta} A_{\varepsilon,\delta} \end{aligned} \tag{45}$$

donde $A_{\varepsilon,\delta}$ es el conjunto de $\mathbf{z}$ en $\mathbb{R}^{p+1}$ para los cuales existen puntos $\mathbf{z}_1$ y $\mathbf{z}_2$ en $\mathbb{R}^{p+1}$ tales que $\|\mathbf{z} - \mathbf{z}_1\| < \delta$, $\|\mathbf{z} - \mathbf{z}_2\| < \delta$, y $\|h(\mathbf{z}_1) - h(\mathbf{z}_2)\| \geq \varepsilon$. Entonces $A_{\varepsilon,\delta}$ es un conjunto abierto, y podemos tomar las operaciones de unión e intersección en (45) restringidas a los racionales, entonces $D_{\mathbf{f}}$ es un conjunto de Borel (aun en el caso en el que h resulte ser una función de $\mathbb{R}^{p+1}$ a $\mathbb{R}^s$ no medible). Pedimos que

$$P(D_{\mathbf{f}}) = 0. \tag{46}$$

Observemos que la condición (46) se verifica trivialmente en el caso en el que $\mathbf{f}$ es continua, pero también en el caso en el que la $\mathbf{f}$ es apropiada para usar con el método SIR, es decir, cuando $\mathbf{f}(y) = \sum_{h=1}^{H} I_{I_h}(y)\, \mathbf{e}_h$ donde $r = H$ y los $\mathbf{e}_h$ son vectores canónicos en $\mathbb{R}^r$ y P es absolutamente continua, pues en tal caso

$$D_{\mathbf{f}} = \left\{(\mathbf{x}, y) \in \mathbb{R}^{p+1} : y = a_h, h = 1, ..., H\right\} \text{ donde tomamos } I_h = [a_h, a_{h+1}),$$

que es un un conjunto de probabilidad P cero. Esta condición también se verifica en el caso en el que y es una variable aleatoria discreta, de rango finito, ya que en tal caso podemos tomar a $\mathbf{f}$ como $\mathbf{f}(y) = \sum_{h=1}^{H} I_{I_h}(y)\, \mathbf{e}_h$, donde elegimos a los I_h de modo que contengan exactamente un valor del rango de y en su interior, de modo que $D_{\mathbf{f}}$ sea el mismo de antes, que tenga probabilidad cero pero por un motivo diferente.

Existirán soluciones para nuestro problema si P, la probabilidad de $(\mathbf{x}, y)$, y la función $\mathbf{f} : \mathbb{R} \to \mathbb{R}^r$ se combinan de forma tal de no ubicar demasiada masa concentrada en alguna región muy angosta. Formalicemos este requisito. Para eso, definimos para $\mathbf{a} \in \mathbb{R}^p, \mathbf{b} \in \mathbb{R}^r, c \in \mathbb{R}$, con $\mathbf{a}$ ó $\mathbf{b}$ no nulos

$$H_{\mathbf{a},\mathbf{b},c} = \left\{(\mathbf{x}, y) : \mathbf{a}^T\mathbf{x} + \mathbf{b}^T\mathbf{f}(y) = c\right\}. \tag{47}$$

Al evento $H_{(\mathbf{a},\mathbf{b},c)}$ lo denominaremos "*hiperplano*". Si pensamos en la porción de $\mathbb{R}^{p+r}$ representada por $\left\{(\mathbf{x}, \mathbf{z}) \in \mathbb{R}^{p+r} : \mathbf{a}^T\mathbf{x} + \mathbf{b}^T\mathbf{z} = c\right\}$, esto será, de hecho, un hiperplano, y la región

$$\left\{(\mathbf{x}, \mathbf{z}) \in \mathbb{R}^{p+r} : \mathbf{a}^T\mathbf{x} + \mathbf{b}^T\mathbf{z} = c, \quad \mathbf{z} = \mathbf{f}(y), \quad y \in \mathbb{R}\right\}$$

representará una superficie contenida en él. En este sentido, mantenemos el nombre hiperplano (sin las comillas) para referirnos a él.

Para ε entre 0 y 1 suficientemente pequeño, definimos la siguiente propiedad:

(S_ε) La probabilidad P tiene la propiedad (S_ε) si para todo $H_{(\mathbf{a},\mathbf{b},c)}$ se verifica que $P\left(H_{(\mathbf{a},\mathbf{b},c)}\right) < \varepsilon$.

Obsérvese que la propiedad (S_ε) se cumple si ε es una cota superior para la probabilidad de todo $H_{(\mathbf{a},\mathbf{b},c)}$.

Definición 6.1 *Sea* $\boldsymbol{\alpha} = \left(\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T\right)$, *donde* $\boldsymbol{\alpha}_1 \in \mathbb{R}^p$, $\boldsymbol{\alpha}_2 \in \mathbb{R}^r, l \in \mathbb{R}$, $\delta \in \mathbb{R}_{>0}$. *Luego definimos* ***faja de tamaño*** δ ***alrededor del hiperplano*** $H_{(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,l)}$ *a*

$$H\left(\boldsymbol{\alpha}, l, \delta\right) = \left\{(\mathbf{x}, y) : l \leq {\boldsymbol{\alpha}_1}^T \mathbf{x} + {\boldsymbol{\alpha}_2}^T \mathbf{f}\left(y\right) \leq l + \delta, \quad \boldsymbol{\alpha} = \left(\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T\right)\right\}. \tag{48}$$

Para simplificar la notación, usaremos además de

$$\boldsymbol{\Theta} = \left\{\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta) \in \mathbb{R}^p \times \mathbb{R}^{p\times r} \times PDS\left(p\right) : \text{rango}\left(B\right) \leq d\right\},$$

la siguiente descripción del espacio de parámetros $\boldsymbol{\Theta}$,

$$\widetilde{\boldsymbol{\Theta}} = \left\{\boldsymbol{\theta} = (D, \Delta) \in \mathbb{R}^{p\times(r+1)} \times PDS\left(p\right) : \text{rango}\left(D\right) \leq d+1\right\}$$

donde D es el resultado de concatenar a $\boldsymbol{\mu}$ y B en una sola matriz, $D = \begin{bmatrix}\boldsymbol{\mu} & B\end{bmatrix} \in \mathbb{R}^{p\times(r+1)}$, seguiremos denominando $\boldsymbol{\theta}$ a cada elemento de ambos conjuntos. En (6) proponíamos agregar una fila con un uno a $\mathbf{f}(y)$ para poder incluir al término independiente del modelo PFC en un producto de matrices. Si tomamos $\widetilde{\mathbf{f}} : \mathbb{R} \to \mathbb{R}^{r+1}$ definida por $\widetilde{\mathbf{f}}(y) = \begin{bmatrix}\mathbf{1} \\ \mathbf{f}(y)\end{bmatrix}$, y $D_0 = \begin{bmatrix}\boldsymbol{\mu}_0 & B_0\end{bmatrix} \in \mathbb{R}^{p\times(r+1)}$, escribimos

$$\mathbf{x} = D_0\mathbf{z} + \Delta_0^{1/2}\mathbf{u},$$

donde $\mathbf{z} = \widetilde{\mathbf{f}}(y)$.

Definición 6.2 *Sea* $\boldsymbol{\theta} = (D, \Delta) \in \widetilde{\boldsymbol{\Theta}}$, $c > 0$. *Definimos al conjunto* $E\left(D, \Delta, c\right)$ *por*

$$E\left(D, C, c\right) = \left\{(\mathbf{x}, y) : \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right] \leq c^2\right\}. \tag{49}$$

Lo denominaremos un ***elipsoide generalizado, centrado en*** $D\widetilde{\mathbf{f}}(y)$ ***y de radio*** c.

Ahora tenemos las herramientas para establecer el resultado de existencia de solución para el problema que define al $\tau-$funcional.

Teorema 6.1 *Sea* $(\mathbf{x}, y)$ *que cumple el modelo PFC y sea* P *su distribución. Supongamos que* P *satisface* (S_ε) *para algún* $0 < \varepsilon \leq 1 - \frac{\kappa_1}{a_1}$, *siendo* $a_1 = max\rho_1$, *y que* $\mathbf{f}$ *es continua o acotada. Entonces, existe al menos un valor de* $\boldsymbol{\theta}$ *que cumple con las condiciones impuestas por el funcional* $\boldsymbol{\theta}_C(P)$.

La demostración de este resultado está en el Apéndice F.

Observación 6.1 *Cuando el vector* $\mathbf{u}$ *es absolutamente continuo e independiente de* y, *se satisface* (S_ε) *para todo* $0 < \varepsilon \leq 1$, *puesto que en tal caso tenemos*

$$\begin{aligned} P\left(H_{(\mathbf{a},\mathbf{b},c)}\right) &= P\left(\left\{(\mathbf{x}, y) : \mathbf{a}^T\mathbf{x} + \mathbf{b}^T\mathbf{f}(y) = c\right\}\right) \\ &= P\left(\left\{(\mathbf{u}, y) : \mathbf{a}^T\left(\boldsymbol{\mu}_0 + B_0\mathbf{f}(y) + \Delta_0^{1/2}\mathbf{u}\right) + \mathbf{b}^T\mathbf{f}(y) = c\right\}\right) \\ &= P\left(\left\{(\mathbf{u}, y) : \mathbf{a}^T\Delta_0^{1/2}\mathbf{u} = c - \mathbf{a}^T\boldsymbol{\mu}_0 - \left(B_0 + \mathbf{b}^T\right)\mathbf{f}(y)\right\}\right) \\ &= E\left[P\left(\left\{(\mathbf{u}, y) : \mathbf{a}^T\Delta_0^{1/2}\mathbf{u} = c - \mathbf{a}^T\boldsymbol{\mu}_0 - \left(B_0 + \mathbf{b}^T\right)\mathbf{f}(y)\right\} \mid \mathbf{f}(y)\right)\right]. \end{aligned}$$

La probabilidad condicional vale cero, en el caso en el que el vector $\mathbf{u}$ *es absolutamente continuo, por lo que* $P\left(H_{(\mathbf{a},\mathbf{b},c)}\right) = 0$, *y* P *cumple* S_ε *para todo* $0 < \varepsilon \leq 1$.

Observación 6.2 *¿Qué condiciones debe cumplir la muestra para poder asegurar que el* $\tau-$*estimador exista? Sea* k_n *el máximo número de observaciones* $(\mathbf{x}_i, y_i)$ *tales que* $(\mathbf{x}_i, \mathbf{f}(y_i))$ *caen en un hiperplano. En general, por definición se tiene* $k_n \geq p + r$. *Se dice que un conjunto de al menos* $d + 1$ *puntos en* $\mathbb{R}^d$ *está en* ***posición general*** *si ningún hiperplano contiene más de* d *puntos. Si los puntos* $(\mathbf{x}_i, \mathbf{f}(y_i)) \in \mathbb{R}^{p+r}$ *están en posición general, resulta que* $k_n = p + r$. *Los puntos deben arreglarse especialmente para no estar en posición general, por ejemplo teniendo tres en una línea en* $\mathbb{R}^2$ *o cuatro en un plano en* $\mathbb{R}^3$. *Sea* P_n *la probabilidad empírica inducida por la muestra, para cualquier* $\mathbf{a} \in \mathbb{R}^p, \mathbf{b} \in \mathbb{R}^r, c \in \mathbb{R}$ *se tiene*

$$P_n\left(H_{(\mathbf{a},\mathbf{b},c)}\right) \leq \frac{k_n}{n}.$$

Si $k_n < n\varepsilon < n\left(1 - \frac{\kappa_1}{a_1}\right)$, P_n *está en las hipótesis del Teorema 6.1 y hay al menos una solución del problema que define a* $\boldsymbol{\theta}_C$ *para* P_n. *Por lo tanto, una condición suficiente para que el estimador* $\boldsymbol{\theta}_C(P_n)$ *de los parámetros del modelo PFC exista, es que las observaciones de la muestra* $\{(\mathbf{x}_i, \mathbf{f}(y_i))\}_{1 \leq i \leq n}$ *estén en posición general.*

6.2. Unicidad y Fisher consistencia del $\tau-$funcional

Necesitaremos imponer algunas condiciones sobre la distribución de los errores del modelo PFC para obtener resultados de unicidad en la definición del τ-funcional. Por supuesto, queremos resultados generales, que incluyan a la distribución normal multivariada de los errores como caso particular. Comenzamos recordando la definición de simetría esférica.

Definición 6.3 *Diremos que una función* $f : \mathbb{R}^p \to \mathbb{R}$ *es* ***esférica o esféricamente simétrica*** *si*

$$f(\mathbf{u}) = f^*(\mathbf{u}^T\mathbf{u}) \ ,$$

para alguna función $f^* : [0, +\infty) \to [0, +\infty)$ *que integra uno.*

Definición 6.4 *Diremos que una función* $f : \mathbb{R}^p \to \mathbb{R}$ *es* ***elíptica o elípticamente simétrica*** *si*

$$f(\mathbf{u}) = cf^*((\mathbf{u}-\boldsymbol{\mu})^T \Sigma (\mathbf{u}-\boldsymbol{\mu})),$$

para alguna $f^* : [0, +\infty) \to [0, +\infty)$ *con integral finita,* Σ *una matriz* $p \times p$ *definida positiva,* $\boldsymbol{\mu} \in \mathbb{R}^p$ *y* c *una constante de normalización que se incluye para garantizar que* f *sea una función de densidad.*

Definición 6.5 *Diremos que una función* $f : \mathbb{R}^p \to \mathbb{R}$ *es* ***simétrica y unimodal*** *si:*

(a) $f(\mathbf{x}) = f(-\mathbf{x})$ *para todo* $\mathbf{x} \in \mathbb{R}^p$.

(b) $H_a = \{\mathbf{x} : f(\mathbf{x}) \geq a\}$ *es convexa para todo* $0 \leq a$.

Si una función f es simétrica y unimodal entonces es radialmente decreciente. Esto es, para todo $\mathbf{x} \in \mathbb{R}^p$ si $\alpha_1 > \alpha_2 \geq 0$ resulta que $f(\alpha_1\mathbf{x}) \leq f(\alpha_2\mathbf{x})$. Si la segunda desigualdad es siempre estricta, entonces decimos que f es radialmente decreciente en sentido estricto. También nos referiremos a estas funciones diciendo que son estrictamente unimodales. Las distribuciones elípticas y unimodales satisfacen las condiciones de la Definición 6.5. Podremos probar la unicidad del $\tau-$funcional para familias más amplias que las que acabamos de definir.

Sea $\mathcal{O}(p)$ el grupo de matrices $p \times p$ ortogonales, y sea W_p el subgrupo de $\mathcal{O}(p)$ generado por las matrices de permutación (es decir, matrices que tienen una entrada de la matriz igual a 1 en cada fila y cada columna, y ceros en los restantes lugares) y por las matrices de reflexión (esto es, matrices diagonales que tienen en la diagonal o bien un $+1$ o un -1). La siguiente definición es de Anderson [1955].

Definición 6.6 *Diremos que una función* $f : \mathbb{R}^p \to \mathbb{R}$ *es* W_p-***invariante*** *si*

$$f(\mathbf{x}) = f(A\mathbf{x})$$

para todo A *en* W_p *y* $\mathbf{x}$ *en* $\mathbb{R}^p$.

Entre los ejemplos de funciones W_p-invariantes se incluyen la densidad de un vector aleatorio esféricamente simétrico, la densidad de un vector aleatorio con coordenadas simétricas independientes e idénticamente distribuidas, o en general, la densidad de un vector aleatorio de coordenadas simétricas intercambiables. Observemos que las funciones W_p-invariantes quedan determinadas por su definición en $\mathbb{R}^p_+$.

Las siguientes definiciones de mayorización, M-mayorización y M-concavidad aparecen en Marshall y Olkin [1979].

Definición 6.7 *Dado el vector* $\mathbf{x} = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$, *denotemos por* $x_{[1]} \geq \cdots \geq x_{[p]}$ *a las coordenadas de* $\mathbf{x}$ *en orden decreciente. Para* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$,

$$\mathbf{x} \prec \mathbf{y} \text{ si } \begin{cases} \sum_{j=1}^{k} x_{[j]} \leq \sum_{j=1}^{k} y_{[j]} & \text{para } 1 \leq k \leq p-1, \text{ y} \\ \\ \sum_{j=1}^{p} x_{[j]} = \sum_{j=1}^{p} y_{[j]}. \end{cases}$$

Decimos que $\mathbf{x}$ *está* ***mayorado*** *por* $\mathbf{y}$.

Definición 6.8 *Para* $\mathbf{x}, \mathbf{y} \in (0, +\infty)^p$,

$$\mathbf{x} \prec_M \mathbf{y} \text{ si } \begin{cases} \prod_{j=1}^{k} x_{[j]} \leq \prod_{j=1}^{k} y_{[j]} & \text{para } 1 \leq k \leq p-1, \text{ y} \\ \\ \prod_{j=1}^{p} x_{[j]} = \prod_{j=1}^{p} y_{[j]}. \end{cases}$$

Decimos que $\mathbf{x}$ *está* ***M-mayorado*** *por* $\mathbf{y}$.

Una definición alternativa de M-mayorización es la siguiente: para $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p_+$,

$$\mathbf{x} \prec_M \mathbf{y} \text{ si y sólo si } \log(\mathbf{x}) \prec \log(\mathbf{y}),$$

donde $\log(\mathbf{x}) = (\log(x_1), \ldots, \log(x_p))$ y el símbolo $\prec$ se refiere a la mayorización definida previamente. Observemos que existe una correspondencia entre los puntos $\mathbf{x} \in (0, +\infty)^p$ y las elipses con semi ejes $x_1, \ldots, x_p$. Intuitivamente, $\mathbf{x} \prec_M \mathbf{y}$ significa que la elipse correspondiente a $\mathbf{x}$ es "menos dispersa" que la elipse (con el mismo volumen) correspondiente a $\mathbf{y}$. La "elipse menos dispersa" entre todas las elipses con el mismo volumen es la esfera. Siguiendo a Tatsuoka y Tyler [2000], definimos la siguiente familia de funciones.

Definición 6.9 *Diremos que una función* $f : \mathbb{R}^p_+ \to \mathbb{R}$ *o que una función* W_p-*invariante* $f : \mathbb{R}^p \to \mathbb{R}$ *es* ***M-cóncava*** *si para* $\mathbf{x} \prec_M \mathbf{y}$, *resulta* $f(\mathbf{x}) \geq f(\mathbf{y})$. *Diremos que la función* f *es* ***estrictamente M-cóncava*** *si para* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p_+$ *y* $\mathbf{x} \prec_M \mathbf{y}$ *implica que vale la desigualdad estricta salvo que exista una permutation* A *en* W_p *tal que* $A\mathbf{x} = \mathbf{y}$.

A las funciones M-cóncavas se las denomina M-decrecientes. Definimos la clase de medidas de probabilidad $\mathcal{P}_p(W_p, M) \subset \mathcal{P}(\mathbb{R}^p)$ como la clase de distribuciones definidas en $\mathbb{R}^p$ que tienen densidades que son W_p-invariantes y M-cóncavas. La clase $\mathcal{P}_p(W_p, M)$ incluye a las distribuciones unimodales y esféricamente simétricas, como así también una gran clase de distribuciones cuyas coordenadas son independientes e idénticamente distribuidas, con distribución univariada simétrica, como por ejemplo la distribución *t de Student*. Es un resultado conocido que la única distribución dentro de la clase de distribuciones esféricas que tiene coordenadas independientes e idénticamente distribuidas es la

distribución normal multivariada esférica. Es por eso que la clase $\mathcal{P}_p(W_p, M)$ puede ser vista como una generalización de las normales esféricas, que incluye a las distribuciones unimodales esféricamente simétricas.

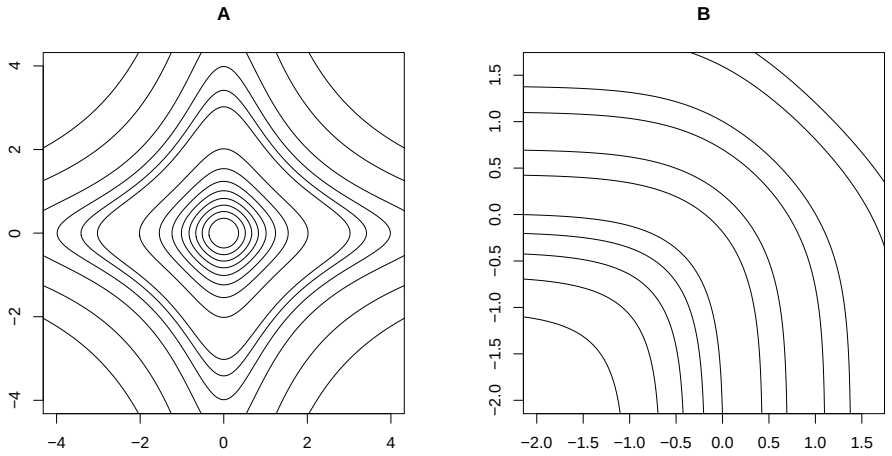
Un ejemplo de una función M-cóncava que no es esféricamente simétrica, ni simétrica y unimodal es la densidad conjunta de v.a.i.i.d. con distribución t de Student con $k > 0$ grados de libertad, es decir

$$f(x_1, \ldots, x_p) = c \prod_{i=1}^{p} \left(1 + \frac{x_i^2}{k}\right)^{-(k+1)/2}.$$

Esto está probado en Tatsuoka y Tyler [2000], en el Apéndice, como consecuencia de los Lemas A1 y A2 (página 1241 de dicho artículo).

Gráficamente, una densidad W_p-invariante es M-cóncava si y sólo si las curvas de nivel en $(0, +\infty)^p$ son convexas cuando son graficadas en escala logarítmica. En la Figura 3 podemos ver las curvas de nivel de la densidad conjunta de un vector aleatorio de dos variables aleatorias independientes idénticamente distribuídas, con distribución Cauchy (es decir, distribución t de Student con un grado de libertad).

Figura 3: Curvas de nivel de la densidad de dos variables aleatorias independientes con distribución Cauchy (es decir, t de Student con un grado de libertad). El gráfico A está en escala usual, en el gráfico B la escala es logarítmica.



Probaremos que el $\tau-$funcional para el modelo PFC tiene única solución bajo el supuesto de que la distribución de los errores pertenezca a la clase $\mathcal{P}_p(W_p, M)$. Como estableciéramos en la Observación 3.3, cuando el error del modelo PFC tiene coor-

denadas independientes idénticamente distribuidas, hemos probado que la reducción $R(\mathbf{x}) = \Gamma_0^T \Delta_0^{-1} \mathbf{x}$ es una reducción suficiente.

Si asumimos que $(\mathbf{x}, y)$ satisface el modelo PFC y que la distribución del error $\mathbf{u}$ está en la clase $\mathcal{P}_p(W_p, M)$ entonces, como

$$\mathbf{x} = \boldsymbol{\mu}_0 + B_0 \mathbf{f}(y) + \Delta_0^{1/2} \mathbf{u}$$

donde $\mathbf{u}$ e y son independientes, entonces P, la probabilidad inducida por $(\mathbf{x}, y)$ puede obtenerse a partir de las distribuciones de $F_{\mathbf{u}}$ y F_y si conociéramos a ambas y si también conociéramos los valores de los parámetros verdaderos $(\boldsymbol{\mu}_0, B_0, \Delta_0)$. Recordemos que para cualquier $\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta)$ que está en el espacio de parámetros del modelo PFC, notamos por $H_{P,\boldsymbol{\theta}}$ a la distribución de

$$d(\mathbf{x}, y, \boldsymbol{\theta}) = \left\{ (\mathbf{x} - \boldsymbol{\mu} - B\mathbf{f}(y))^T \Delta^{-1} (\mathbf{x} - \boldsymbol{\mu} - B\mathbf{f}(y)) \right\}^{1/2}.$$

Teorema 6.2 *Supongamos que ρ_1 y ρ_2 satisfacen A1-A2, y además ρ_2 satisface A3. Sea $(\mathbf{x}, y)$ que satisface el modelo PFC y asumamos que la distribución del error $\mathbf{u}$ está en la clase $\mathcal{P}_p(W_p, M)$ y que es independiente de y. Llamemos P_0 a la distribución de probabilidad de $(\mathbf{x}, y)$. Entonces, el problema de minimizar el* $\det(\Delta)$ *sujeto a que $\boldsymbol{\theta} \in \Theta$ y*

$$S_\tau^2(H_{P,\boldsymbol{\theta}}) = \kappa_2.$$

tiene a $(\boldsymbol{\theta}_0)_c = (\boldsymbol{\mu}_0, B_0, c\Delta_0)$ con $c = \frac{1}{\kappa_2} S_\tau^2(H_{P_0,\boldsymbol{\theta}_0})$ como única solución.

La demostración de este resultado está en el Apéndice F.

Observemos que, en otras palabras, el Teorema 6.2 nos dice que bajo estas hipótesis, el $\tau-$funcional de estimación para el modelo PFC está bien definido, en el sentido de que devuelve un único valor en el espacio de parámetros. Más aún, el mínimo se alcanza en el valor verdadero de los parámetros con la (posible) excepción de una constante multiplicativa para la matriz de covarianza. Esta constante vale uno si se toman a κ_1 y κ_2 como en (35); pues, en tal caso, $S_M(H_{P_0,\boldsymbol{\theta}_0}) = 1$ y $S_\tau^2(H_{P_0,\boldsymbol{\theta}_0}) = \kappa_2$.

Recordemos la siguiente definición.

Definición 6.10 *Sea $\mathcal{A} \subset \mathcal{P}(\mathbb{R}^d)$ un subconjunto de probabilidades definidas en $\mathbb{R}^d$ parametrizadas por el conjunto de parámetros Θ. Decimos que el funcional $T : \mathcal{A} \to \Theta$ es **Fisher consistente para** Θ si $T(P_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$.*

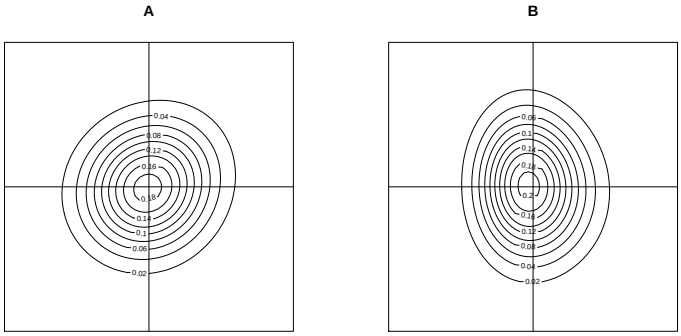
Observemos que el Teorema 6.2 prueba la Fisher consistencia del $\tau-$funcional de estimación para el modelo PFC, $\boldsymbol{\theta}_\tau$, en este caso, ya que $\boldsymbol{\theta}_\tau(P_0) = \boldsymbol{\theta}_0$.

Observación 6.3 (error t de Student) *En particular, por los comentarios posteriores a la Definición 6.9, el Teorema 6.2 garantiza la unicidad del $\tau-$funcional en el caso en el que el error tiene distribución normal multivariada y también cuando sus coordenadas son i.i.d. con distribución t de Student con k grados de libertad.*

Observación 6.4 (P_n empírica) *En el caso del estimador basado en una muestra de n observaciones que no están en posición general (i.e. cuando las observaciones no satisfacen más relaciones lineales de las que deben) hemos probado la existencia de al menos una solución al problema que define a $\boldsymbol{\theta}_C$ para P_n en la Observación 6.2, no hemos probado unicidad de la misma (y posiblemente no lo sea).*

Observación 6.5 (error Skew normal) *Cabe preguntarse si la densidad skew-normal descripta en el Ejemplo 3.4, para la cual ya hemos visto que hay identificabilidad del subespacio de reducción suficiente de la dimensión, pertenece a $\mathcal{P}_p(W_p, M)$. En la Figura 4 pueden verse las curvas de nivel de dos densidades skew-normal para dos conjuntos de parámetros (correspondientes a los casos $\boldsymbol{\delta}$ constante o no). Ninguna de las densidades resultan W_2-invariantes.*

Figura 4: Curvas de nivel de la densidad skew normal bivariada, descripta en la Definición 3.1, con la parametrízación dada por (13), (14) y (15) del Ejemplo 3.4. En ambos $\overline{\Psi} = I$.La densidad del gráfico A, $\boldsymbol{\delta}^{\mathrm{T}} = (0{,}5, 0{,}5)$por lo que resulta $\boldsymbol{\lambda} = (1,1)\frac{1}{\sqrt{3}}$,$\boldsymbol{\alpha} = (0{,}5164, 0{,}5164)$y $\overline{\Omega} = \begin{bmatrix} 1 & 0{,}25 \\ 0{,}25 & 1 \end{bmatrix}$. Para el gráfico B, $\boldsymbol{\delta}^{\mathrm{T}} = (0{,}8, -0{,}1)$por lo que resulta $\boldsymbol{\lambda} = \left(\frac{4}{3}, -0{,}1005\right)$,$\boldsymbol{\alpha} = (1{,}33092, -0{,}0605)$y $\overline{\Omega} = \begin{bmatrix} 1 & -0{,}08 \\ -0{,}08 & 1 \end{bmatrix}$.Ninguna de las dos densidades resulta W_2-invariante. En ambos casos, $\xi = 0$.



Hemos probado la unicidad del $\tau-$funcional de estimación para el modelo PFC bajo una amplia familia de distribuciones unimodales y simétricas. Esto no implica que las

ecuaciones de estimación asociadas tengan una única solución, sólo dicen que el valor que optimiza el criterio dado por el $\tau-$funcional es único. Para los estimadores, como discutimos en la Observación 6.4, no tenemos certeza de unicidad.

7. Consistencia

7.1. Notación y preliminares

Recordemos que $\mathbf{f} = (f_1, ..., f_r) : \mathbb{R} \to \mathbb{R}^r$ está fija y sea P la medida de probabilidad definida en $\mathbb{R}^{p+1}$. Asumimos que $P(D_{\mathbf{f}}) = 0$, donde $D_{\mathbf{f}}$ es el conjunto de discontinuidades de $\mathbf{f}$. Estamos interesados en las siguientes dos clases de subconjuntos de $\mathbb{R}^{p+1}$: la clase $\mathcal{E}$ de "*elipsoides generalizados*" definida en (49) para una función $\mathbf{f}$ fija, esto es

$$\mathcal{E} = \{E(D, \Delta, c) : (D, \Delta) \in \mathbf{\Theta}, c > 0\} \tag{50}$$

donde

$$E(D, \Delta, c) = \left\{(\mathbf{x}, y) : \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right] \leq c^2\right\}. \tag{51}$$

y la clase $\mathcal{H}$ de "*hiperplanos generalizados*" definida en la condición S_ε, (página 46) para una función $\mathbf{f}$ fija, esto es

$$\mathcal{H} = \left\{H_{(\mathbf{a},\mathbf{b},c)} : \mathbf{a} \in \mathbb{R}^p, \mathbf{b} \in \mathbb{R}^r, c \in \mathbb{R}\right\} \tag{52}$$

donde

$$H_{(\mathbf{a},\mathbf{b},c)} = \left\{(\mathbf{x}, y) : \mathbf{a}^T\mathbf{x} + \mathbf{b}^T\mathbf{f}(y) = c\right\}.$$

Podemos escribir estas dos clases de conjuntos de una manera apropiada para probar los resultados que nos interesan. Sea $\mathcal{G}$ el subespacio lineal de dimensión finita generado por las siguientes funciones reales definidas en $\mathbb{R}^{p+1}$,

$$\{1, x_h x_l,\ x_h,\ f_s(y),\ f_s(y) f_t(y),\ x_h f_s(y) :\quad 1 \leq h, l \leq p,\quad 1 \leq s, t \leq r\}. \tag{53}$$

Sea $\mathcal{D}_1$ la clase de subconjuntos definido de la siguiente manera: $C \in \mathcal{D}_1$ si existe $g \in \mathcal{G}$ tal que $C = \{\mathbf{z} = (y, x_1, ..., x_p) : g(\mathbf{z}) \geq 0\}$ y sea $\mathcal{D}_2$ definida por

$$\mathcal{D}_2 = \{C \cap D : C \in \mathcal{D}_1,\ D \in \mathcal{D}_1\}. \tag{54}$$

Claramente, $\mathcal{D}_1$ y $\mathcal{D}_2$ están contenidos en los borelianos de $\mathbb{R}^{p+1}$. Entonces, $\mathcal{E} \subset \mathcal{D}_1$, y $\mathcal{H} \subset \mathcal{D}_2$ porque

$$\begin{gathered} H_{(\mathbf{a},\mathbf{b},c)} = \{g_1 \geq 0\} \cap \{-g_1 \geq 0\} \\ \text{donde } g_1(\mathbf{x}, y) = \mathbf{a}^T\mathbf{x} + \mathbf{b}^T\mathbf{f}(y) - c, \quad g_1 \in \mathcal{G}. \end{gathered} \tag{55}$$

Utilizaremos los dos siguientes resultados, del libro de Billingsley [1968] relacionados con la convergencia débil de probabilidades en espacios métricos. Los utilizaremos en la demostración del Teorema 7.4. También agregamos el Lema 7.1 que particulariza el segundo resultado al contexto en el que trabajamos, y será utilizado también en la prueba del Teorema 7.4.

Teorema 7.1 (Teorema 5.1, Billingsley [1968]) *Sean S, S' espacios métricos, $\mathcal{S}, \mathcal{S}'$ las $\sigma-$álgebras de Borel respectivas (i.e. las $\sigma-$álgebras generadas por los abiertos de cada espacio métrico) y sea $(P_k)_k$ una sucesión de medidas de probabilidad definidas en S, que converge débilmente a P. Sea $h : S \to S'$ una función medible tal que $P(D_h) = 0$, donde D_h es el conjunto de discontinuidades de h, entonces la sucesión de medidas de probabilidades inducidas $P_k h^{-1}$ converge débilmente a Ph^{-1}. (En particular, $\int h dP_k \to \int h dP$).*

Recordemos que Ph^{-1} (o $P \circ h^{-1}$) es la probabilidad inducida en $(S', \mathcal{S}')$, definida por $Ph^{-1}(A) = P\left(h^{-1}(A)\right)$ para todo $A \in \mathcal{S}'$. Ver el Apéndice, Sección E.2.

Teorema 7.2 (Teorema 5.5 Billingsley [1968]) *Sean S, S' espacios métricos, S' separable, $\mathcal{S}$, $\mathcal{S}'$ las $\sigma-$álgebras de Borel respectivas y sean h_k y h funciones medibles, $(h_k)_{k\in\mathbb{N}}$, $h : S \to S'$. Sea $(P_k)_k$ una sucesión de medidas de probabilidad definidas en S, que converge débilmente a P. Definimos*

$$E = \{\mathbf{x} \in S : \exists (\mathbf{x}_k)_k \text{ tales que } \mathbf{x}_k \to \mathbf{x} \text{ y } h_k(\mathbf{x}_k) \nrightarrow h(\mathbf{x})\},$$

supongamos que $P(E) = 0$, entonces $P_k h_k^{-1}$ converge débilmente a Ph^{-1}.

Lema 7.1 *Sea $(P_k)_k$ una sucesión de medidas de probabilidad definidas en $\mathbb{R}^{p+1}$ que converge débilmente a una medida de probabilidad P cuando $k \to \infty$. Sea $(\boldsymbol{\theta}_k)_k$ una sucesión dada por $\boldsymbol{\theta}_k = (D_k, \Delta_k) \in \boldsymbol{\Theta}$ tal que $\boldsymbol{\theta}_k \to \boldsymbol{\theta}_L$. Sea*

$$g(\mathbf{x}, y, \boldsymbol{\theta}) = \rho_1\left(\left[\left(\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right)^T \Delta^{-1}\left(\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right)\right]^{1/2}\right) \tag{56}$$

para $\boldsymbol{\theta} = (D, \Delta)$, $\widetilde{\mathbf{f}}$ y P que satisfacen condición (46). Entonces,

$$\lim_{k\to\infty} \int g(\mathbf{x}, y, \boldsymbol{\theta}_k)\, dP_k(\mathbf{x}, y) = \int g(\mathbf{x}, y, \boldsymbol{\theta}_L)\, dP(\mathbf{x}, y).$$

La demostración de este resultado está en el Apéndice G.

7.2. Convergencia débil de probabilidades y convergencia uniforme

Para poder probar la consistencia, necesitamos relacionar la convergencia débil de una sucesión de medidas de probabilidad con la convergencia uniforme de las integrales sobre ciertos conjuntos (o, lo que es lo mismo, la convergencia uniforme de las probabilidades de ciertos subconjuntos). Para ello, damos una breve introducción de convergencias de funciones en espacios euclídeos para luego enunciar el resultado de Rao [1962] que utilizaremos en la prueba de consistencia del $\tau-$funcional definido.

Sea $\mathcal{X}$ un espacio métrico, y sea $C_s(\mathcal{X})$ el espacio de todas las funciones continuas de $\mathcal{X}$ en $\mathbb{R}^s$, $f : \mathcal{X} \to \mathbb{R}^s$. Para cada subconjunto compacto K de $\mathcal{X}$, podemos definir la seminorma $\| \quad \|_{C_s(K)}$ en $C_s(\mathcal{X})$ por $\|f\|_{C_s(K)} := \sup_{x\in K} |f(x)|$. La topología generada por todas estas seminormas (cuando K varía por todos los subconjuntos compactos de $\mathcal{X}$) se denomina la *topología de la convergencia uniforme sobre compactos*; es más fuerte que la topología de la convergencia puntual, pero más débil que la topología de la convergencia uniforme. De hecho, una sucesión $(f_k)_k \in C_s(\mathcal{X})$ converge a $f \in C_s(\mathcal{X})$ en esta topología si y sólo si f_k converge uniformemente a f en cada conjunto compacto, es decir, si

$$\lim_{k\to\infty} \sup_{x\in K} |f_k(x) - f(x)| = 0,$$

para cada subconjunto compacto K de $\mathcal{X}$. A esta convergencia se la denomina usualmente *convergencia uniforme sobre compactos* (u.c.c. por sus siglas en inglés, *uniform convergence on compact sets*).

La caracterización de subconjuntos compactos de $C_s(\mathcal{X})$ es bien conocida y está dada por el Teorema de Arzelá-Ascoli en el caso de los conjuntos localmente compactos y separables, es decir cuando el espacio métrico $\mathcal{X}$ puede escribirse de la siguiente forma, $\mathcal{X} = \bigcup_{i=1}^{\infty} K_i$ donde K_i es compacto y $K_i \subset interior\,(K_{i+1})$ para todo $i \in \mathbb{N}$. En particular, los espacios euclídeos $\mathbb{R}^s$ y los $\mathcal{X}$ compactos son localmente compactos y separables. El Teorema de Arzelá-Ascoli da la siguiente equivalencia en espacios localmente compactos y separables: una familica $\mathcal{A} \subset C(\mathcal{X})$ es condicionalmente compacta (i.e., su clausura es compacta), si y sólo si las dos condiciones siguientes se satisfacen:

(i) para cada x, $\sup_{f\in\mathcal{A}} |f(x)| < \infty$.

(ii) $\mathcal{A}$ es equicontinua en cada $x \in \mathcal{X}$, i.e., para cada $\varepsilon > 0$ existe una vecindad N de x tal que $|f(x) - f(y)| < \varepsilon$, para todo $y \in N$ y toda $f \in \mathcal{A}$.

Utilizaremos el Teorema 3.4 de Rao [1962], que enunciamos a continuación.

Teorema 7.3 (Teorema 3.4 Rao [1962]) *Sea μ una medida (es decir, una función definida en una $\sigma-$álgebre de conjuntos, finita, no negativa y $\sigma-$aditiva) en un espacio métrico separable $\mathcal{X}$ y sea $\mathcal{A}$ una familia de funciones continuas de $\mathcal{X}$ a $\mathbb{R}^s$, que satisfacen las dos condiciones siguientes:*

(i) $\mathcal{A}$ es (u.c.c.) compacta y,

(ii) μg^{-1} tiene distribuciones marginales continuas para cada $g \in \mathcal{A}$.

Si μ_k *converge débilmente a* μ, *entonces*

$$\lim_{k\to\infty} \sup_{A} |\mu_k(A) - \mu(A)| = 0, \tag{57}$$

donde el supremo *en (57) se toma sobre todos los conjuntos A de la forma*

$$A = \{x : g_j(x) \le a_j,\ j = 1, 2, \ldots, s\}$$

donde $g(x) = (g_1(x), \ldots, g_s(x)) \in \mathcal{A}$ *y* $(a_1, \ldots, a_s)$ *es un vector arbitrario de* $\mathbb{R}^s$.

Este resultado establece la continuidad del funcional $\widetilde{\boldsymbol{\theta}}_C$, o, equivalentemente del $\tau-$funcional, y es esencial para probar la consistencia de los $\tau-$estimadores.

Teorema 7.4 *Sea* $(P_k)_{k\geq 1}$ *una sucesión de medidas de probabilidad, y sea* P *otra probabilidad absolutamente continua, todas definidas en* $\mathbb{R}^{p+1}$ *tales que* P_k *converge débilmente a* P. *Supongamos que* P *satisface* (S_ε) *para algún* $0 < \varepsilon < 1 - \frac{\kappa_1}{c_1}$ *y que* $\widetilde{\boldsymbol{\theta}}_C(P) = \left(\widetilde{D}_C(P), \widetilde{\Delta}_C(P)\right)$ *se define de forma única. Entonces, para todo* k *suficientemente grande, existe al menos un valor* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *que cumple las condiciones que definen al funcional* $\widetilde{\boldsymbol{\theta}}_C(P_k)$, *y para cualquier sucesión de soluciones* $\widetilde{\boldsymbol{\theta}}_C(P_k)$, $k \geq 1$, *resulta que* $\lim_{k\to\infty} \widetilde{\boldsymbol{\theta}}_C(P_k) = \widetilde{\boldsymbol{\theta}}_C(P)$.

La demostración de este resultado está en el Apéndice G.

Por el estrecho vínculo entre los funcionales $\widetilde{\boldsymbol{\theta}}_C$ y $\boldsymbol{\theta}_C$, el Teorema 7.4 podría enunciarse en términos de $\boldsymbol{\theta}_C$ en vez de $\widetilde{\boldsymbol{\theta}}_C$.

7.3. Consistencia del $\tau-$estimador

El objetivo es utilizar el Teorema 7.4 para $(P_k)_k$ la secuencia de medidas empíricas obtenidas a partir de muestras (observaciones independientes e idénticamente distribuidas, *iid*) que verifican el modelo PFC, para probar la consistencia del $\tau-$estimador. Para ello transcribimos algunos resultados de Pollard [1984].

Definición 7.1 (Pollard [1984], p.17) *Sea* $\mathcal{D}$ *una clase de subconjuntos de un espacio* S. *Decimos que la clase tiene* ***discriminación polinomial*** *(de grado* ν*) si existe un polinomio* $\rho(\cdot)$ *(de grado* ν*) tal que, para todo* n *y todo subconjunto* S_0 *de* S *que consiste en* n *puntos , la clase de conjuntos*

$$\{S_0 \cap D : D \in \mathcal{D}\} \tag{58}$$

(contenida en $\mathcal{P}(S_0)$, *el conjunto de partes de* S_0*) tiene cardinalidad a lo sumo* $\rho(n)$. ρ *se denomina polinomio discriminante para* $\mathcal{D}$.

Una consecuencia de esta definición es que si $\mathcal{D}_1 \subset \mathcal{D}$ y $\mathcal{D}$ tiene discriminación polinomial, también la tiene $\mathcal{D}_1$ (porque el mismo polinomio discriminante establece una cota para la cardinalidad de (58) cuando sólo consideramos $D \in \mathcal{D}_1$).

Recordamos la siguiente generalización de Glivenko–Cantelli de la misma fuente.

Teorema 7.5 (Teorema 14, Pollard [1984]) *Sea* P *una medida de probabilidad definida en un espacio de medida* S *y sea* P_k *la medida empírica obtenida a partir de una muestra iid de tamaño* k *de* P. *Para cada clase* $\mathcal{D}$ *de subconjuntos de* S *con discriminación polinomial, tal que* $\sup_{D\in\mathcal{D}} |P_k(D) - P(D)|$ *es medible, tenemos*

$$\sup_{D\in\mathcal{D}} |P_k(D) - P(D)| \underset{k\to\infty}{\to} 0 \quad \textit{casi seguramente.} \tag{59}$$

Observación 7.1 *Pollard [1984] da un tratamiento más general a la clase $\mathcal{D}$ de subconjuntos de S para los cuales $\sup_{D\in\mathcal{D}}|P_k(D)-P(D)|$ es medible, definiendo una* ***clase permisible*** *de subconjuntos que incluye la condición de medibilidad establecida más arriba como caso particular.*

El siguiente resultado establece la consistencia de los $\tau-$estimadores.

Teorema 7.6 *Sea P una medida de probabilidad que satisface (S_ε) para algún $0<\varepsilon<1-\frac{\kappa_1}{c_1}$, y supongamos que $\widetilde{\boldsymbol{\theta}}_C(P)=\left(\widetilde{D}_C(P),\widetilde{\Delta}_C(P)\right)$ está univocamente definido. Sea P_k la medida empírica obtenida a partir de una muestra iid de tamaño k de observaciones $(\mathbf{x}_i,y_i)_{i\geq 1}$ de P, entonces $\lim_{k\to\infty}\widetilde{\boldsymbol{\theta}}_C(P_k)=\widetilde{\boldsymbol{\theta}}_C(P)$ casi seguramente.*

La demostración de este resultado está en el Apéndice G.

Para garantizar la consistencia del $\tau-$estimador de Δ, que difiere de $\widetilde{\Delta}_C(P)$ en una constante multiplicativa, todavía resta demostrar que dicha sucesión de constantes por la cual se corrige a $\widetilde{\boldsymbol{\theta}}_C(P_k)$ para obtene r el $\tau-$estimador de Δ para P_k, converge a la adecuada.

Corolario 7.1 *Bajo las condiciones del Corolario 7.6, se tiene que $\lim_{k\to\infty}\boldsymbol{\theta}_\tau(P_k)=\boldsymbol{\theta}_\tau(P)$ casi seguramente.*

La demostración de este resultado está en el Apéndice G.

Observación 7.2 *Supongamos que ρ_1 y ρ_2 satisfacen A1–A2, y además ρ_2 satisface A3. Sea $(\mathbf{x},y)$ que satisface el modelo PFC y asumamos que la distribución del vector $\mathbf{u}$ está en la clase $\mathcal{P}_p(W_p,M)$ y que es independiente de y. Como ya observamos en la prueba del Teorema 6.2, en este caso P, la distribución de probabilidad de $(\mathbf{x},y)$, estará en las condiciones del Corolario 7.1 y la consistencia de los $\tau-$estimadores del modelo PFC (al único valor posible del $\tau-$funcional) queda garantizada.*

Observación 7.3 *Enfaticemos que en el camino para probar la consistencia, hemos probado la continuidad del $\tau-$funcional para el modelo PFC. En virtud de esto, logramos ver la robustez cualitativa asintótica del funcional.*

8. Ecuaciones de estimación para el $\tau-$ funcional

8.1. Planteo y notación

Por el resultado de equivalencias probado en el Teorema 5.2 para obtener las ecuaciones del $\tau-$funcional podemos basarnos en los distintos funcionales definidos en la Sección 5.2. Una posibilidad es hallar el mínimo de la función $\Phi_{C,P}$ sujeto a la restricción dada por (43) utilizando multiplicadores de Lagrange. Este enfoque es el que sigue Lopuhaä [1991], por ejemplo, para el caso de $\tau-$estimadores de posición y covarianza multivariados. Otra posibilidad es obtener las ecuaciones que resultan de minimizar la función $\Phi_{B,P}$ y agregarle la ecuación de la restricción,

$$S_\tau^2\left(H_{P,\boldsymbol{\theta}}\right) = \kappa_2.$$

Este enfoque es el que se sigue en García Ben et al. [2006] para el cálculo del estimador del modelo lineal con respuesta multivariada. En esta sección, resolvemos el problema de caracterizar a $\boldsymbol{\theta}_B\left(P\right)$ para una probabilidad P arbitraria que cumple el modelo PFC.

En esta Sección, trabajaremos con la parametrización del modelo PFC dada por $(\boldsymbol{\mu},\Gamma,\beta,\Delta) \in \boldsymbol{\Theta}$, con $\boldsymbol{\mu} \in \mathbb{R}^p, \Gamma \in \mathbb{R}^{p\times d}, \beta \in \mathbb{R}^{d\times r}, \Delta \in PDS\left(p\right)$ y pedimos que el rango de Γ sea d. Lo seguiremos notando por la letra griega $\boldsymbol{\theta}$. En las demostraciones se agregarán las demás restricciones que necesitamos para tener identificabilidad del modelo, como discutimos en el Apéndice B.4. Sea $\boldsymbol{\theta} = (\boldsymbol{\mu},\Gamma,\beta,\Delta) \in \boldsymbol{\Theta}$ y recordemos que notábamos

$$d\left(\mathbf{x},y,\boldsymbol{\theta}\right) = \left\{\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}(y)\right)^T\Delta^{-1}\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}(y)\right)\right\}^{1/2}, \tag{60}$$

a la distancia de Mahalanobis que se obtiene utilizando esos parámetros. Luego,

$$\Phi_{B,P}\left(\boldsymbol{\theta}\right) = |\Delta|\left(S_\tau^2\left(H_{P,\boldsymbol{\theta}}\right)\right)^p = |\Delta|\left(S_M^2\left(H_{P,\boldsymbol{\theta}}\right)\right)^p\left(\int\rho_2\left(\frac{d\left(\mathbf{x},y,\boldsymbol{\theta}\right)}{S_M\left(H_{P,\boldsymbol{\theta}}\right)}\right)dP\left(\mathbf{x},y\right)\right)^p,$$

donde $S_M\left(H_{P,\boldsymbol{\theta}}\right)$ esta (implícitamente) definido por

$$E_{H_{P,\boldsymbol{\theta}}}\left(\rho_1\left(\frac{v}{S_M\left(H_{P,\boldsymbol{\theta}}\right)}\right)\right) = \kappa_1. \tag{61}$$

Definimos a $\boldsymbol{\theta}_B\left(P\right) \in \boldsymbol{\Theta}$ como aquel valor de los parámetros que minimiza a $\Phi_{B,P}\left(\boldsymbol{\theta}\right)$ entre todos los $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ que cumplen

$$S_\tau^2\left(H_{P,\boldsymbol{\theta}}\right) = \kappa_2. \tag{62}$$

Como observamos en el Capítulo 5, la restricción (62) sólo se impone para identificar una solución al problema planteado, ya que $\Phi_{B,P}\left(\boldsymbol{\theta}_\lambda\right) = \Phi_{B,P}\left(\boldsymbol{\theta}\right)$ para todo $\lambda > 0$. De modo que $\boldsymbol{\theta}_B\left(P\right)$ también satisface el problema de minimización sin restricciones. Luego, buscaremos el mínimo de $\Phi_{B,P}\left(\boldsymbol{\theta}\right)$ para $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, o, equivalentemente, queremos minimizar el $\ln\left(\Phi_{B,P}\left(\boldsymbol{\theta}\right)\right)$,

$$\ln\left(\Phi_{B,P}\left(\boldsymbol{\theta}\right)\right) = \ln|\Delta| + 2p\ln\left(S_M\left(H_{P,\boldsymbol{\theta}}\right)\right) + p\ln\left(\int\rho_2\left(\frac{d\left(\mathbf{x},y,\boldsymbol{\theta}\right)}{S_M\left(H_{P,\boldsymbol{\theta}}\right)}\right)dP\left(\mathbf{x},y\right)\right).$$

Establecemos la siguiente notación para exhibir el resultado de un modo compacto. Sean

$$\begin{aligned} a(d) &= 2\rho_2(d) - \psi_2(d)\, d \\ b(d) &= \psi_1(d)\, d \\ u(\boldsymbol{\theta}) &= E_P\left(\rho_2\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_M(H_{P,\boldsymbol{\theta}})}\right)\right) \end{aligned}$$

Definamos

$$\begin{aligned} w^{(1)}(\boldsymbol{\theta}) &= \frac{pE_P\left[a\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_M(H_{P,\boldsymbol{\theta}})}\right)\right]}{2u(\boldsymbol{\theta})\, S_M^2(H_{P,\boldsymbol{\theta}})\, E_P\left[b\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_M(H_{P,\boldsymbol{\theta}})}\right)\right]} \\ w^{(2)}(\boldsymbol{\theta}) &= \frac{p}{2u(\boldsymbol{\theta})\, S_M^2(H_{P,\boldsymbol{\theta}})}, \end{aligned} \tag{63}$$

Finalmente definimos los pesos

$$w(d,\boldsymbol{\theta}) = \left\{ w^{(1)}(\boldsymbol{\theta})\frac{\psi_1(d)}{d} + w^{(2)}(\boldsymbol{\theta})\frac{\psi_2(d)}{d} \right\}. \tag{64}$$

Obsérvese que

$$w\left(\tfrac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_M(H_{P,\boldsymbol{\theta}})},\boldsymbol{\theta}\right) = w^{(1)}(\boldsymbol{\theta})\,\psi_1\left(\tfrac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_M(H_{P,\boldsymbol{\theta}})}\right)\tfrac{S_M(H_{P,\boldsymbol{\theta}})}{d(\mathbf{x},y,\boldsymbol{\theta})} + w^{(1)}(\boldsymbol{\theta})\,\psi_2\left(\tfrac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_M(H_{P,\boldsymbol{\theta}})}\right)\tfrac{S_M(H_{P,\boldsymbol{\theta}})}{d(\mathbf{x},y,\boldsymbol{\theta})}.$$

Sea

$$W(\boldsymbol{\theta}) = E_P\left[w\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_M(H_{P,\boldsymbol{\theta}})},\boldsymbol{\theta}\right)\right]. \tag{65}$$

Como probaremos en la página 190, los pesos $w(d,\boldsymbol{\theta})$ son positivos y acotados, por lo que W está bien definida. Luego podemos definir una nueva distribución que depende del $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ dada por,

$$P_{\boldsymbol{\theta}}^*(A) := \int_A \frac{1}{W(\boldsymbol{\theta})} w\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_M(H_{P,\boldsymbol{\theta}})},\boldsymbol{\theta}\right) dP(\mathbf{x},y), \tag{66}$$

para todo A boreliano de $\mathbb{R}^{p+1}$. De esta forma, $P_{\boldsymbol{\theta}}^*$ es una distribución para $(\mathbf{x},y)$, y para toda función g medible tenemos

$$E_P\left[\frac{1}{W(\boldsymbol{\theta})} w\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_M(H_{P,\boldsymbol{\theta}})},\boldsymbol{\theta}\right) g(\mathbf{x},y)\right] = E_{P_{\boldsymbol{\theta}}^*}[g(\mathbf{x},y)].$$

Luego si $\mathrm{cov}_P(\mathbf{v},\mathbf{z})$ indica la matriz de covarianza entre los vectores $\mathbf{v}$ y $\mathbf{z}$ cuando la probabilidad inducida por ellos es P y $\mathrm{var}_p(\mathbf{v}) = \mathrm{cov}_p(\mathbf{v},\mathbf{v})$, llamemos

$$\Pi(\boldsymbol{\theta}) = \mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x},\mathbf{f}(y)]\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}\left(\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x},\mathbf{f}(y)]\right)^T. \tag{67}$$

Claramente $\Pi(\boldsymbol{\theta})$ resulta una matriz simétrica y definida positiva.

8.2. Caracterizando los puntos críticos y el mínimo

Consideremos la siguiente condición sobre las $\rho-$funciones.

A4 Las funciones $\psi_k(t)/t$ son no crecientes $(k = 1, 2)$.

Observemos que los pesos w que se utilizan en el cálculo del $\tau-$funcional dependen de las distancias a través de las funciones $\psi_k(t)/t$, $(k = 1, 2)$, luego, para que el $\tau-$funcional resulte robusto es razonable pedir que las observaciones con mayores distancias de Mahalanobis tengan menores pesos en el cómputo de los estimadores. Con la notación introducida al final de la Sección precedente, en (64), (65), (66) y (67), el siguiente resultado permite caracterizar tanto a los puntos críticos de la función objetivo Φ_B, que da una manera alternativa de definir al $\tau-$funcional para el modelo PFC, como a dicho $\tau-$funcional.

Teorema 8.1 *Supongamos que ρ_1 y ρ_2 satisfacen A1 y A2, y además ρ_2 satisface A3. Sea $(\mathbf{x}, y)$ que satisface el modelo PFC y asumamos que la distribución del vector $\mathbf{u}$ está en la clase $\mathcal{P}_p(W_p, M)$ y que es independiente de y. Sea P la distribución de probabilidad de $(\mathbf{x}, y)$.*

i. Los valores de $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Gamma, \beta, \Delta) \in \Theta$ críticos de $\ln \Phi_{B,P}(\boldsymbol{\theta})$ satisfacen las siguientes ecuaciones de estimación,

$$\int w(\boldsymbol{\theta})(\mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y))\, dP(\mathbf{x},y) = 0 \tag{68}$$

$$\int w(\boldsymbol{\theta}) \left[-(\mathbf{x} - \boldsymbol{\mu})\mathbf{f}(y)^T + \Gamma\beta\mathbf{f}(y)\mathbf{f}(y)^T \right] dP(\mathbf{x},y) = 0$$

$$\Delta - \int w(\boldsymbol{\theta})(\mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y))(\mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y))^T dP(\mathbf{x},y) = 0$$

$$\int w(\boldsymbol{\theta})(\mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y))\mathbf{f}^T(y)\beta^T dP(\mathbf{x},y) = 0.$$

ii. Luego, los valores de $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Gamma, \beta, \Delta) \in \Theta$ críticos de $\ln(\Phi_{B,P}(\boldsymbol{\theta}))$ satisfacen las siguientes ecuaciones

$$\begin{aligned} \boldsymbol{\mu} &= \frac{1}{W(\boldsymbol{\theta})} E_P \left[w\left(\frac{dE_{P^*_{\boldsymbol{\theta}}}[g(\mathbf{x},y)]}{S_M(H_{P,\boldsymbol{\theta}})}, \boldsymbol{\theta} \right)(\mathbf{x} - \Gamma\beta\mathbf{f}(y)) \right] \\ &= E_{P^*_{\boldsymbol{\theta}}}(\mathbf{x} - \Gamma\beta\mathbf{f}(y)). \end{aligned} \tag{69}$$

Las columnas de $\Delta^{-1/2}\Gamma$ son d autovectores ortogonales de la matriz simétrica

$$\Delta^{-1/2}\Pi(\boldsymbol{\theta})\Delta^{-1/2},$$

donde $\Pi(\boldsymbol{\theta})$ fue definida en (67). Sin pérdida de generalidad, los podemos elegir ortonormales, es decir satisfaciendo

$$\Delta^{-1/2}\Pi(\boldsymbol{\theta})\Delta^{-1/2}\left(\Delta^{-1/2}\Gamma\right) = \Delta^{-1/2}\Gamma\Omega(\boldsymbol{\theta}), \tag{70}$$

donde $\Omega(\boldsymbol{\theta})$ *es diagonal, y*

$$\left(\Delta^{-1/2}\Gamma\right)^T\left(\Delta^{-1/2}\Gamma\right) = I_d,$$

o equivalentemente

$$\Gamma^T\Delta^{-1}\Gamma = I_d \tag{71}$$

Además

$$\begin{aligned}\Delta &= E_P\left[w\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_M(H_{P,\boldsymbol{\theta}})},\boldsymbol{\theta}\right)(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}(y))(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}(y))^T\right] \\ &= W(\boldsymbol{\theta})\,\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}-\Gamma\beta\mathbf{f}(y)]\end{aligned} \tag{72}$$

y

$$\beta = \Gamma^T\Delta^{-1}\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}. \tag{73}$$

iii. Supongamos que ρ_1 *y* ρ_2 *también satisfacen A4. Entonces, para que el valor de* $\boldsymbol{\theta} = (\boldsymbol{\mu},\Gamma,\beta,\Delta)$ *sea mínimo local de* $\ln(\Phi_{B,P}(\boldsymbol{\theta}))$ *además de cumplir las condiciones establecidas en (ii), en las columnas de* $\Delta^{-1/2}\Gamma$ *se deben poner los autovectores de la matriz simétrica* $\Delta^{-1/2}\Pi(\boldsymbol{\theta})\Delta^{-1/2}$ *asociados a los mayores autovalores de dicha matriz, donde* $\Pi(\boldsymbol{\theta})$ *fue definida en (67).*

iv. El $\tau-$*funcional de estimación* $\boldsymbol{\theta}_\tau(P)$ *se obtiene corrigiendo el valor* $\boldsymbol{\theta} = (\boldsymbol{\mu},\Gamma,\beta,\Delta)$ *obtenido en (iii) del siguiente modo*

$$\boldsymbol{\theta}_\tau(P) = (\boldsymbol{\mu},\Gamma,\beta,m\Delta)$$

con

$$m = \frac{1}{\kappa_2}S^2_\tau(H_{P,\boldsymbol{\theta}}).$$

La demostración de este resultado está en el Apéndice H.

Observación 8.1 *Es inmediato que si* $\boldsymbol{\theta} = (\boldsymbol{\mu},\Gamma,\beta,\Delta) \in \boldsymbol{\Theta}$ *cumple (72), entonces* $\boldsymbol{\theta}_t = (\boldsymbol{\mu},\Gamma,\beta,t\Delta)$ *también lo cumple, para todo* $t > 0$*. Por lo tanto si* $\boldsymbol{\theta}$ *es un punto crítico de* $\ln(\Phi_{B,P}(\boldsymbol{\theta}))$*, entonces* $\boldsymbol{\theta}_t$ *también lo es, para todo* $t > 0$*.*

Observación 8.2 *El Teorema 8.1 (i) permite expresar al* $\tau-$*funcional de estimación para el modelo PFC como un M-funcional, excepto por el hecho de que la función* $\frac{w(d,\boldsymbol{\theta})}{W(\boldsymbol{\theta})}$ *que da los pesos, resulta ser un promedio pesado de dos funciones* ψ*, y los pesos, en este caso, dependen de las variables* $(\mathbf{x},y)$*.*

Observación 8.3 *Para permitir la identificación de* Γ *y de* β*, a lo largo del cálculo surgieron condiciones de normalización comentadas en la página 140, al caracterizar el espacio de parámetros como* $\boldsymbol{\Theta}^*$*. Estas restricciones resultaron: (1) las columnas de* $\Delta^{-1/2}\Gamma$ *son ortogonales, ya que son autovectores de la matriz simétrica* $\Delta^{-1/2}\Pi(\boldsymbol{\theta})\Delta^{-1/2}$*, o*

sea $\Gamma^T\Delta^{-1}\Gamma = I_d$ *y (2)* $\beta var_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\,\beta^T$ *es diagonal, con elementos decrecientes en la diagonal. Esto es consecuencia de que, por (73),*

$$\beta = \Gamma^T\Delta^{-1}\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1},$$

luego

$$\begin{aligned}
\beta\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\,\beta^T &= \Gamma^T\Delta^{-1}\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)] \\
&\quad\cdot\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y),\mathbf{x}]\,\Delta^{-1}\Gamma \\
&= \Gamma^T\Delta^{-1}\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y),\mathbf{x}]\,\Delta^{-1}\Gamma \\
&= \Gamma^T\Delta^{-1}\Pi(\boldsymbol{\theta})\,\Delta^{-1}\Gamma \\
&= \left(\Gamma^T\Delta^{-1/2}\right)\left(\Delta^{-1/2}\Pi(\boldsymbol{\theta})\,\Delta^{-1/2}\right)\Delta^{-1/2}\Gamma \\
&= \left(\Gamma^T\Delta^{-1/2}\right)\Delta^{-1/2}\Gamma\Omega(\boldsymbol{\theta}) = \Gamma^T\Delta^{-1}\Gamma\Omega(\boldsymbol{\theta}) = \Omega(\boldsymbol{\theta}).
\end{aligned}$$

En lo que sigue daremos una interpretación de los $\tau-$funcionales obtenidos.

Observación 8.4 (Relación con el modelo lineal multivariado sin restricciones) *Si miramos el problema con la probabilidad* $P^*_{\boldsymbol{\theta}}$ *que se obtiene a través de* P *con los pesos (64) calculados en el punto crítico, resulta que el valor propuesto para la matriz* $B = \Gamma\beta$ *de coeficientes es*

$$\Gamma\beta = \Gamma\Gamma^T\Delta^{-1}\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}.$$

Para el modelo lineal multivariado sin restricciones de rango en la matriz de coeficientes, el $\tau-$*funcional de* $\Gamma\beta$ *resulta ser*

$$\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1},$$

ver García Ben et al. [2006]. Entonces, para el modelo PFC, el $\tau-$*funcional correspondiente resulta ser la proyección del* $\tau-$*funcional sin restricciones de rango al subespacio generado por las columnas de* Γ*, en la norma inducida por* Δ*, ver (10).*

Observación 8.5 *Para interpretar las condiciones que satisface el* $\mathrm{span}(\Gamma)$*, primero recordemos el problema de correlaciones canónicas, que describimos a continuación. Sean* $\mathbf{x}_1 \in \mathbb{R}^p$ *y* $\mathbf{x}_2 \in \mathbb{R}^r$ *vectores aleatorios con esperanza* $\mathbf{0}$ *y sean* $\Sigma_{11} = var[\mathbf{x}_1]$*,* $\Sigma_{22} = var[\mathbf{x}_2], \Sigma_{12} = cov[\mathbf{x}_1,\mathbf{x}_2]$ *y* $\Sigma_{21} = \Sigma_{12}^T$*. Se tiene el siguiente resultado, que es el Teorema 5.9 de Seber [1984].*

Teorema 5.9 de Seber [1984]. *Sean* $1 > \rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_m^2 > 0$, *donde* $m = \text{rango}\,(\Sigma_{12})$, *los* d *autovalores no negativos de* $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ *(y también de* $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$*). Sean* $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ *y* $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m$ *los correspondientes autovectores de* $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ *y de* $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, *respectivamente. Sean* $\boldsymbol{\alpha}$ *y* $\boldsymbol{\beta}$ *vectores arbitrarios tales que para* $s \leq m-1$, $\boldsymbol{\alpha}^T\mathbf{x}_1$ *no está correlacionada con cada* $\mathbf{a}_j^T\mathbf{x}_1$ $(j = 1, 2, ..., s)$ *y* $\boldsymbol{\beta}^T\mathbf{x}_2$ *no está correlacionada con cada* $\mathbf{b}_j^T\mathbf{x}_2$ $(j = 1, 2, ..., s)$. *Entonces:*

(i) La máxima correlación al cuadrado entre $\boldsymbol{\alpha}^T\mathbf{x}_1$ *y* $\boldsymbol{\beta}^T\mathbf{x}_2$ *está dada por* ρ_{s+1}^2 *y ocurre cuando* $\boldsymbol{\alpha} = \mathbf{a}_{s+1}$ *and* $\boldsymbol{\beta} = \mathbf{b}_{s+1}$.

(ii) $cov[\mathbf{a}_j^T\mathbf{x}_1, \mathbf{a}_k^T\mathbf{x}_1] = 0$, $j \neq k$, *y* $cov[\mathbf{b}_j^T\mathbf{x}_2, \mathbf{b}_k^T\mathbf{x}_2] = 0$, $j \neq k$.

El valor $\sqrt{\rho_j^2} = |\rho_j|$ *para que valor absoluto si* $\rho_j \geq 0$. *se denomina la* ***j-ésima correlación canónica****, y* $u_j = \mathbf{a}_j^T\mathbf{x}_1$ *y* $v_j = \mathbf{b}_j^T\mathbf{x}_2$ *son las* ***j-ésimas variables canónicas****. Llamaremos a* $\mathbf{a}_j$ *y* $\mathbf{b}_j$ *las* ***j-ésimas direcciones canónicas****, asociadas a* $\mathbf{x}_1$ *y* $\mathbf{x}_2$ *respectivamente. Más allá de los signos, éstas serán únicas si las correlaciones canónicas son distintas.*

Volviendo a los τ*−funcionales propuestos, observemos que* $\Delta^{-1/2}\Pi(\boldsymbol{\theta})\Delta^{-1/2}$ *es simétrica y semidefinida positiva. Luego, sean*

$$\lambda_1\left(\Delta^{-1/2}\Pi(\boldsymbol{\theta})\Delta^{-1/2}\right) \geq \cdots \geq \lambda_p\left(\Delta^{-1/2}\Pi(\boldsymbol{\theta})\Delta^{-1/2}\right)$$

sus autovalores, sea $\widetilde{\Omega}(\boldsymbol{\theta}) \in \mathbb{R}^{p\times p}$ *la matriz que los tienen en su diagonal ordenados en forma decreciente y sea* $U \in \mathbb{R}^{p\times p}$ *ortonormal tales que*

$$\begin{aligned}
\Delta^{-1/2}\Pi(\boldsymbol{\theta})\Delta^{-1/2} &= U\widetilde{\Omega}(\boldsymbol{\theta})U^T \\
&= \begin{bmatrix}\Delta^{-1/2}\Gamma & \Delta^{-1/2}\Gamma_1\end{bmatrix}\begin{bmatrix}\Omega(\boldsymbol{\theta}) & 0\\ 0 & \Omega_1(\boldsymbol{\theta})\end{bmatrix}\begin{bmatrix}\Gamma^T\Delta^{-1/2}\\ \Gamma_1^T\Delta^{-1/2}\end{bmatrix} \\
&= \Delta^{-1/2}\begin{bmatrix}\Gamma & \Gamma_1\end{bmatrix}\begin{bmatrix}\Omega(\boldsymbol{\theta}) & 0\\ 0 & \Omega_1(\boldsymbol{\theta})\end{bmatrix}\begin{bmatrix}\Gamma^T\\ \Gamma_1^T\end{bmatrix}\Delta^{-1/2} \\
&= \Delta^{-1/2}\Gamma\Omega(\boldsymbol{\theta})\Gamma^T\Delta^{-1/2} + \Delta^{-1/2}\Gamma_1\Omega_1(\boldsymbol{\theta})\Gamma_1^T\Delta^{-1/2}.
\end{aligned}$$

Sea $\boldsymbol{\theta}$ *el punto crítico de* $\ln\Phi_{B,P}$ *donde se alcanza el mínimo de la función. Por el Teorema 8.1 (iii) sabemos que en el valor de* $\boldsymbol{\theta}$ *que alcanza el mínimo de* $\Phi_{B,P}$, *los autovectores de* $\Delta^{-1/2}\Pi(\boldsymbol{\theta})\Delta^{-1/2}$ *que se deben utilizar como columnas de* $\Delta^{-1/2}\Gamma$ *son aquellos que corresponden a los* d *mayores autovalores. Luego, por las igualdades (70) y (71) sabemos que*

$$\widetilde{\Omega}(\boldsymbol{\theta}) = \begin{bmatrix}\Omega(\boldsymbol{\theta}) & 0\\ 0 & \Omega_1(\boldsymbol{\theta})\end{bmatrix},$$

y que las primeras d *columnas de* U *(las que están asociadas a los mayores autovalores) están compuestas por* $\Delta^{-1/2}\Gamma$. *Sea* $\Gamma_1 \in \mathbb{R}^{p\times(p-d)}$ *tal que*

$$U = \Delta^{-1/2}\begin{bmatrix}\Gamma & \Gamma_1\end{bmatrix},$$

es decir, llamamos Γ_1 *a las* $p-d$ *últimas columnas de* $\Delta^{-1/2}U$. *Resulta,*

$$\begin{aligned}\Gamma_1^T\Delta^{-1}\Gamma_1 &= I_{p-d},\\ \Gamma_1^T\Delta^{-1}\Gamma &= 0_{p-d,d}, \quad \Gamma^T\Delta^{-1}\Gamma = I_d.\end{aligned}$$

Ahora daremos simplemente un argumento heurístico para respaldar este hecho. Calculemos las correlaciones canónicas entre $\mathbf{x}_1 = \mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y)$ *y* $\mathbf{x}_2 = \mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]$, *con dicho* $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Gamma, \beta, \Delta)$. *En ese caso,* $\mathbf{x}_1$ *representará el residuo del modelo y* $\mathbf{x}_2$ *será el vector de las covariables explicativas apropiadamente centradas. El ajuste será satisfactorio si la mayor correlación canónica entre ellas es muy pequeña. Calculemos dicha correlación. El cuadrado de la mayor correlación canónica entre* $\mathbf{x}_1$ *y* $\mathbf{x}_2$ *será igual al mayor autovalor de la matriz* $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ *donde*

$$\begin{aligned}\Sigma_{11} &= \mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y)]\\ \Sigma_{22} &= \mathrm{var}_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right] = \mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\\ \Sigma_{12} &= \mathrm{cov}_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{x} - \Gamma\beta\mathbf{f}(y) - \boldsymbol{\mu}, \mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right].\end{aligned}$$

Como $\boldsymbol{\theta}$ *es crítico, satisface (69) y (72), luego* $\boldsymbol{\mu} = E_{P_{\boldsymbol{\theta}}^*}[\mathbf{x} - \Gamma\beta\mathbf{f}(y)]$ *y tanto* $\mathbf{x}_1$ *como* $\mathbf{x}_2$ *tienen esperanza cero. Además, resulta*

$$\begin{aligned}\Sigma_{11} &= \frac{1}{W(\boldsymbol{\theta})}\Delta\\ \Sigma_{12} &= \mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x}, \mathbf{f}(y)] - \Gamma\Gamma^T\Delta^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x}, \mathbf{f}(y)]\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\\ &= \left(I - \Gamma\Gamma^T\Delta^{-1}\right)\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x}, \mathbf{f}(y)]\end{aligned}$$

Luego,

$$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Delta^{-1}\left(I - \Gamma\Gamma^T\Delta^{-1}\right)\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x}, \mathbf{f}(y)]\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1} \tag{74}$$

$$\cdot\,\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y), \mathbf{x}]\left(I - \Gamma\Gamma^T\Delta^{-1}\right)^T \tag{75}$$

$$= \Delta^{-1}\left(I - \Gamma\Gamma^T\Delta^{-1}\right)\Pi(\boldsymbol{\theta})\left(I - \Delta^{-1}\Gamma\Gamma^T\right).$$

Observemos que

$$\Delta^{-1}\left(I - \Gamma\Gamma^T\Delta^{-1}\right)\Pi(\boldsymbol{\theta})\Delta^{-1}\Gamma = 0$$

por lo que de (74) tenemos

$$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Delta^{-1}\left(\Gamma_1\Omega_1(\boldsymbol{\theta})\Gamma_1^T\right).$$

Esta matriz tiene por autovectores a las columnas de $\Delta^{-1}\Gamma_1$ *con autovalores correspondientes a la diagonal de* $\Omega_1(\boldsymbol{\theta})$, *pues*

$$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\left(\Delta^{-1}\Gamma_1\right) = \Delta^{-1}\left(\Gamma_1\Omega_1(\boldsymbol{\theta})\Gamma_1^T\right)\left(\Delta^{-1}\Gamma_1\right) = \left(\Delta^{-1}\Gamma_1\right)\Omega_1(\boldsymbol{\theta}).$$

Luego, las correlaciones canónicas al cuadrado entre $\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}(y)$ *y* $\mathbf{f}(y)-E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]$ *son los autovalores de* $\Delta^{-1/2}\Pi(\boldsymbol{\theta})\Delta^{-1/2}$ *que no se utilizaron para definir a los autovectores de* $\Delta^{-1/2}\Pi(\boldsymbol{\theta})\Delta^{-1/2}$ *que conforman a* Γ*. En ese sentido, si para elegir las columnas de* Γ *se tomaron los* d *mayores autovalores de* $\Delta^{-1/2}\Pi(\boldsymbol{\theta})\Delta^{-1/2}$ *entonces la mayor correlación canónica entre* $\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}(y)$ *y* $\mathbf{f}(y)-E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]$ *será lo menor posible, y valdrá* $\lambda_{d+1}(\Pi(\boldsymbol{\theta}))$*. Observemos que al identificar el subespacio tal que al proyectar los residuos, éstos quedan lo menos correlacionados posibles con la respectiva proyección* $d-$*dimensional de las* $\mathbf{f}(y)$*, por ortogonalidad, también estamos identificando al subespacio generado por las columnas de* $\Delta^{-1}\Gamma$ *que es el que reduce la dimensión. Dicha matriz resulta tener por columnas al subespacio ortogonal a las direcciones canónicas entre los residuos entre* $\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}(y)$ *y las explicativas,* $\mathbf{f}(y)-E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]$*, asociadas a las elección que hace que las correlaciones canónicas sean lo más chicas posibles.*

9. Estudio de los $\tau-$estimadores para el modelo PFC para muestras finitas

9.1. Ecuaciones de estimación para el $\tau-$estimador basado en una muestra

Calculemos el estimador basado en una muestra aleatoria $\{(\mathbf{x}_i,y_i), 1 \leq i \leq n\}$, es decir, el valor del $\tau-$funcional de estimación evaluado en la medida empírica P_n correspondiente. Recordemos que la medida empírica P_n se define para subconjuntos A medibles de $\mathbb{R}^{p+1}$ por

$$P_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_A(\mathbf{x}_i,y_i) = \frac{1}{n}\sum_{i=1}^{n} \delta_{(\mathbf{x}_i,y_i)}(A)$$

donde I_A es la función indicadora y $\delta_{(\mathbf{x}_i,y_i)}$ es la medida de Dirac. A partir de la medida empírica, la función de distribución empírica puede escribirse como

$$F_n(\mathbf{x},y) = P_n(\{(\mathbf{s},t) : s_j \leq x_j \text{ para } j = 1,\ldots,p \text{ y } t \leq y\}).$$

Para describir el algoritmo, adaptaremos la notación introducida en la Sección 8.1. Para toda función s medible definida en el espacio $p+1$ dimensional, tenemos

$$E_{P_n}[s(\mathbf{x},y)] = \frac{1}{n}\sum_{i=1}^{n} s(\mathbf{x}_i,y_i).$$

Recordemos que para cada $\boldsymbol{\theta} \in \Theta$, denotamos por $H_{P_n,\boldsymbol{\theta}}$ a la distribución de $d(\mathbf{x},y,\boldsymbol{\theta})$ cuando $(\mathbf{x},y)$ tiene distribución P_n. En particular $S_M(H_{P_n,\boldsymbol{\theta}})$, que por brevedad notaremos $S_n(\boldsymbol{\theta})$, se define por

$$E_{P_n}\left(\rho_1\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_M(H_{P_n,\boldsymbol{\theta}})}\right)\right) = E_{H_{P_n,\boldsymbol{\theta}}}\left(\rho_1\left(\frac{v}{S_M(H_{P_n,\boldsymbol{\theta}})}\right)\right) = \kappa_1$$

Es decir $S_n(\boldsymbol{\theta})$ satisface

$$\sum_{i=1}^{n}\frac{1}{n}\rho_1\left(\frac{d(\mathbf{x}_i,y_i,\boldsymbol{\theta})}{S_n(\boldsymbol{\theta})}\right) = \kappa_1. \tag{76}$$

Luego,

$$u_n(\boldsymbol{\theta}) = E_{P_n}\left(\rho_2\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_n(\boldsymbol{\theta})}\right)\right) = \frac{1}{n}\sum_{i=1}^{n}\rho_2\left(\frac{d(\mathbf{x}_i,y_i,\boldsymbol{\theta})}{S_n(\boldsymbol{\theta})}\right). \tag{77}$$

Hemos agregado el subíndice n para enfatizar la dependencia de la muestra. De igual modo se computan

$$E_{P_n}\left[a\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_n(\boldsymbol{\theta})}\right)\right],\ E_{P_n}\left[b\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_n(\boldsymbol{\theta})}\right)\right],$$

donde las funciones a y b están definidas en (79) y (80). Notaremos los pesos por

$$\begin{aligned} w_i(\boldsymbol{\theta}) &= w\left(\frac{d(\mathbf{x}_i,y_i,\boldsymbol{\theta})}{S_n(\boldsymbol{\theta})},\boldsymbol{\theta}\right) \\ &= w^{(1)}(\boldsymbol{\theta})\,\psi_1\left(\frac{d(\mathbf{x}_i,y_i,\boldsymbol{\theta})}{S_n(\boldsymbol{\theta})}\right)\frac{S_n(\boldsymbol{\theta})}{d(\mathbf{x}_i,y_i,\boldsymbol{\theta})} \\ &+ w^{(2)}(\boldsymbol{\theta})\,\psi_2\left(\frac{d(\mathbf{x}_i,y_i,\boldsymbol{\theta})}{S_n(\boldsymbol{\theta})}\right)\frac{S_n(\boldsymbol{\theta})}{d(\mathbf{x}_i,y_i,\boldsymbol{\theta})}. \end{aligned} \tag{78}$$

donde

$$\begin{aligned} w^{(1)}(\boldsymbol{\theta}) &= \frac{pE_{P_n}\left[a\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_n(\boldsymbol{\theta})}\right)\right]}{2u_n(\boldsymbol{\theta})\,S_n^2(\boldsymbol{\theta})\,E_{P_n}\left[b\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_n(\boldsymbol{\theta})}\right)\right]} \\ w^{(2)}(\boldsymbol{\theta}) &= \frac{p}{2u_n(\boldsymbol{\theta})\,S_n^2(\boldsymbol{\theta})}, \end{aligned}$$

y

$$a(d) = 2\rho_2(d) - \psi_2(d)\,d \tag{79}$$

$$b(d) = \psi_1(d)\,d. \tag{80}$$

Sea

$$\begin{aligned} W_n(\boldsymbol{\theta}) &= E_{P_n}\left[w\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_n(\boldsymbol{\theta})},\boldsymbol{\theta}\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} w\left(\frac{d(\mathbf{x}_i,y_i,\boldsymbol{\theta})}{S_n(\boldsymbol{\theta})},\boldsymbol{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} w_i(\boldsymbol{\theta}). \end{aligned}$$

Obsérvese que si $\boldsymbol{\theta} = (\boldsymbol{\mu},\Gamma,\beta,\Delta)$ y $\boldsymbol{\theta}^* = (\boldsymbol{\mu},\Gamma,\beta,\lambda\Delta)$, entonces

$$w_i(\boldsymbol{\theta}^*) = \lambda w_i(\boldsymbol{\theta}). \tag{81}$$

Para todo $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, a partir de la medida empírica podemos definir la medida inducida $P^*_{n,\boldsymbol{\theta}}$ (que es una medida aleatoria discreta) dada por,

$$\begin{aligned} P^*_{n,\boldsymbol{\theta}}(A) &:= \frac{1}{W_n(\boldsymbol{\theta})\,n}\sum_{i=1}^{n} w\left(\frac{d(\mathbf{x}_i,y_i,\boldsymbol{\theta})}{S_n(\boldsymbol{\theta})},\boldsymbol{\theta}\right)\delta_{(\mathbf{x}_i,y_i)}(A) \\ &= \frac{1}{W_n(\boldsymbol{\theta})\,n}\sum_{i=1}^{n} w_i(\boldsymbol{\theta})\,\delta_{(\mathbf{x}_i,y_i)}(A), \end{aligned}$$

para todo A conjunto medible en $\mathbb{R}^{p+1}$. Luego para cada función $g : \mathbb{R}^{p+1} \to \mathbb{R}$, podemos escribir

$$E_{P^*_{n,\boldsymbol{\theta}}}[g(\mathbf{x},y)] = E_{P_n}\left[\frac{1}{W_n(\boldsymbol{\theta})}w\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_n(\boldsymbol{\theta})},\boldsymbol{\theta}\right)g(\mathbf{x},y)\right] = \frac{1}{nW_n(\boldsymbol{\theta})}\sum_{i=1}^{n} w_i(\boldsymbol{\theta})\,g(\mathbf{x}_i,y_i).$$

De este modo, sean

$$\overline{\mathbf{x}}_w(\boldsymbol{\theta}) = E_{P^*_{n,\boldsymbol{\theta}}}[\mathbf{x}] = \frac{1}{nW_n(\boldsymbol{\theta})}\sum_{i=1}^{n} w_i(\boldsymbol{\theta})\,\mathbf{x}_i$$

$$\overline{\mathbf{f}}_w(\boldsymbol{\theta}) = E_{P^*_{n,\boldsymbol{\theta}}}[\mathbf{f}(y)] = \frac{1}{nW_n(\boldsymbol{\theta})}\sum_{i=1}^{n} w_i(\boldsymbol{\theta})\,\mathbf{f}(y_i)$$

$$\mathbb{X}_{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{x}_1^T - \overline{\mathbf{x}}_w^T(\boldsymbol{\theta}) \\ \vdots \\ \mathbf{x}_n^T - \overline{\mathbf{x}}_w^T(\boldsymbol{\theta}) \end{bmatrix} \in \mathbb{R}^{n\times p},$$

$$\mathbb{F}_{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{f}(y_1)^T - \overline{\mathbf{f}}_w^T(\boldsymbol{\theta}) \\ \vdots \\ \mathbf{f}(y_n)^T - \overline{\mathbf{f}}_w^T(\boldsymbol{\theta}) \end{bmatrix} \in \mathbb{R}^{n\times r},$$

$$\mathbb{W}_{\boldsymbol{\theta}} = \frac{1}{W_n(\boldsymbol{\theta})}\mathrm{diag}\left(w_1(\boldsymbol{\theta}),\ldots,w_n(\boldsymbol{\theta})\right) \in \mathbb{R}^{n\times n}.$$

Recordemos que $\mathrm{cov}_P(\mathbf{v},\mathbf{z})$ indica la matriz de covarianza entre los vectores $\mathbf{v}$ y $\mathbf{z}$ cuando la probabilidad inducida por ellos es P y $\mathrm{var}_p(\mathbf{v}) = \mathrm{cov}_p(\mathbf{v},\mathbf{v})$. Tenemos que

$$\begin{aligned}\mathrm{cov}_{P^*_{n,\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)] &= \frac{1}{nW_n(\boldsymbol{\theta})}\sum_{i=1}^{n} w_i(\boldsymbol{\theta})\left(\mathbf{x}_i - \overline{\mathbf{x}}_w(\boldsymbol{\theta})\right)\left(\mathbf{f}(y_i) - \overline{\mathbf{f}}_w(\boldsymbol{\theta})\right)^T \\ &= \frac{1}{n}\mathbb{X}_{\boldsymbol{\theta}}^T\mathbb{W}_{\boldsymbol{\theta}}\mathbb{F}_{\boldsymbol{\theta}} \\ \mathrm{var}_{P^*_{n,\boldsymbol{\theta}}}[\mathbf{f}(y)] &= \frac{1}{nW_n(\boldsymbol{\theta})}\sum_{i=1}^{n} w_i(\boldsymbol{\theta})\left(\mathbf{f}(y_i) - \overline{\mathbf{f}}_w(\boldsymbol{\theta})\right)\left(\mathbf{f}(y_i) - \overline{\mathbf{f}}_w(\boldsymbol{\theta})\right)^T \\ &= \frac{1}{n}\mathbb{F}_{\boldsymbol{\theta}}^T\mathbb{W}_{\boldsymbol{\theta}}\mathbb{F}_{\boldsymbol{\theta}}\end{aligned}$$

$$\begin{aligned}\Pi_n(\boldsymbol{\theta}) &= \mathrm{cov}_{P^*_{n,\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\mathrm{var}_{P^*_{n,\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\left(\mathrm{cov}_{P^*_{n,\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\right)^T \\ &= \frac{1}{n}\mathbb{X}_{\boldsymbol{\theta}}^T\mathbb{W}_{\boldsymbol{\theta}}\mathbb{F}_{\boldsymbol{\theta}}\left(\mathbb{F}_{\boldsymbol{\theta}}^T\mathbb{W}_{\boldsymbol{\theta}}\mathbb{F}_{\boldsymbol{\theta}}\right)^{-1}\left(\mathbb{X}_{\boldsymbol{\theta}}^T\mathbb{W}_{\boldsymbol{\theta}}\mathbb{F}_{\boldsymbol{\theta}}\right)^T.\end{aligned}$$

Luego, la versión empírica del Teorema 8.1 es la siguiente. La función objetivo basada en la muestra observada es

$$\begin{aligned}\ln\left(\Phi_{B,P_n}(\boldsymbol{\theta})\right) &= \ln\left|\Delta\left(S_n^2(\boldsymbol{\theta})\right)\right| \\ &\quad + p\ln\left(E_{P_n}\left(\rho_2\left(\left\{\frac{(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}(y))^T\Delta^{-1}(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}(y))}{S_n^2(\boldsymbol{\theta})}\right\}^{1/2}\right)\right)\right) \\ &= \ln|\Delta| + p\ln\left(S_n^2(\boldsymbol{\theta})\right) + p\ln\left(\frac{1}{n}\sum_{i=1}^{n}\rho_2\left(\left\{\frac{d(\mathbf{x}_i,y_i,\boldsymbol{\theta})}{S_n^2(\boldsymbol{\theta})}\right\}^{1/2}\right)\right). \qquad (82)\end{aligned}$$

Los valores de $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Gamma, \beta, \Delta)$ que son puntos críticos de la función $\ln\left(\Phi_{B,P_n}(\boldsymbol{\theta})\right)$ satisfacen las ecuaciones (83), (84), (85) y (86) que figuran más abajo

$$\begin{aligned}\boldsymbol{\mu} &= E_{P^*_{n,\boldsymbol{\theta}}}\left[\mathbf{x} - \Gamma\beta\mathbf{f}(y)\right] \\ &= \frac{1}{nW_n(\boldsymbol{\theta})}\sum_{i=1}^{n} w_i(\boldsymbol{\theta})\left[\mathbf{x}_i - \Gamma\beta\mathbf{f}(y_i)\right].\end{aligned} \tag{83}$$

Las columnas de $\Delta^{-1/2}\Gamma$ son los autovectores asociados a los mayores d autovalores de la matriz simétrica $\Delta^{-1/2}\Pi_n(\boldsymbol{\theta})\Delta^{-1/2}$, y serán elegidos ortonormales. Luego satisfacen

$$\begin{gathered}\Delta^{-1/2}\Pi_n(\boldsymbol{\theta})\Delta^{-1/2}\left(\Delta^{-1/2}\Gamma\right) = \Delta^{-1/2}\Gamma\Omega_n(\boldsymbol{\theta}), \\ \Gamma\Delta^{-1}\Gamma = I_{d\times d}\end{gathered} \tag{84}$$

donde $\Omega_n(\boldsymbol{\theta}) \in \mathbb{R}^{d\times d}$ es la matriz diagonal, conteniendo los d mayores autovalores de la matriz $\Delta^{-1/2}\Pi_n(\boldsymbol{\theta})\Delta^{-1/2}$ ordenados en orden decreciente.

$$\begin{aligned}\Delta &= W_n(\boldsymbol{\theta})\, E_{P^*_{n,\boldsymbol{\theta}}}\left[\left(\mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y)\right)\left(\mathbf{x} - \boldsymbol{\mu} - \Gamma\boldsymbol{\beta}\mathbf{f}(y)\right)^T\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} w_i(\boldsymbol{\theta})\left[\mathbf{x}_i - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y_i)\right]\left[\mathbf{x}_i - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y_i)\right]^T\end{aligned} \tag{85}$$

$$\begin{aligned}\beta &= \Gamma^T\Delta^{-1}\mathrm{cov}_{P^*_{n,\boldsymbol{\theta}}}\left[\mathbf{x}, \mathbf{f}(y)\right]\left(\mathrm{var}_{P^*_{n,\boldsymbol{\theta}}}\left[\mathbf{f}(y)\right]\right)^{-1} \\ &= \Gamma^T\Delta^{-1}\mathbb{X}^T_{\boldsymbol{\theta}}\mathbb{W}_{\boldsymbol{\theta}}\mathbb{F}_{\boldsymbol{\theta}}\left(\mathbb{F}^T_{\boldsymbol{\theta}}\mathbb{W}_{\boldsymbol{\theta}}\mathbb{F}_{\boldsymbol{\theta}}\right)^{-1}.\end{aligned} \tag{86}$$

Por (81), si $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Gamma, \beta, \Delta)$ satisface (83), (84), (85) y (86), también las satisface $\boldsymbol{\theta}^* = (\boldsymbol{\mu}, \Gamma, \beta, \lambda\Delta)$ para todo $\lambda > 0$. Luego, a partir de $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Gamma, \beta, \Delta)$ un punto crítico de $\Phi_{B,P_n}(\boldsymbol{\theta})$, construimos a $\boldsymbol{\theta}^* = (\boldsymbol{\mu}, \Gamma, \beta, \lambda\Delta)$ tomando a

$$\lambda = \frac{S^2_\tau\left(H_{P_n,\boldsymbol{\theta}}\right)}{\kappa_2} = \frac{\tau^2(\mathbf{d}(\boldsymbol{\theta}))}{\kappa_2}.$$

De esta manera nos aseguramos que $\boldsymbol{\theta}^*$ cumple la restricción (39) que debe satisfacer el $\tau-$ estimador. Finalmente el $\tau-$estimador resulta ser

$$\boldsymbol{\theta}^* = (\boldsymbol{\mu}, \Gamma, \beta, \lambda\Delta) = \left(\boldsymbol{\mu}, \Gamma, \beta, \frac{\tau^2(\mathbf{d}(\boldsymbol{\theta}))}{\kappa_2}\Delta\right). \tag{87}$$

9.2. Algoritmo para computar el $\tau-$estimador

Sea una muestra aleatoria de n observaciones $\{(\mathbf{x}_i, y_i), 1 \le i \le n\}$ con $(\mathbf{x}_i, y_i) \in \mathbb{R}^{p+1}$ que suponemos satisface el modelo PFC.

Describiremos un algoritmo iterativo para calcular el $\tau-$estimador. Más detalles sobre la definición de los estimadores, y el cálculo de las constantes involucradas puede verse en las secciones que siguen.

Lo importante, como en la mayoría de los procedimientos iterativos que conducen al cálculo de un estimador robusto, será elegir un estimador inicial cercano al mínimo del $\tau-$funcional. Presentamos nuestra elección del estimador inicial en la sección siguiente.

Supongamos que tenemos un estimador inicial $\boldsymbol{\theta}^{(0)} = \left(\boldsymbol{\mu}^{(0)}, \Gamma^{(0)}, \beta^{(0)}, \Delta^{(0)}\right)$. El estimador calculado en el paso k será denotado por $\boldsymbol{\theta}^{(k)} = \left(\boldsymbol{\mu}^{(k)}, \Gamma^{(k)}, \beta^{(k)}, \Delta^{(k)}\right)$. Para describir el algoritmo bastará describir como obtener $\boldsymbol{\theta}^{(k+1)}$ a partir de $\boldsymbol{\theta}^{(k)}$. Para esto se seguirán los siguientes pasos.

1. Calculemos las distancias de Mahalanobis usando $\boldsymbol{\theta}^{(k)}$

$$\begin{aligned} d_i^{(k)} &= d\left(\mathbf{x}_i, y_i, \boldsymbol{\theta}^{(k)}\right) \\ &= \sqrt{\left(\mathbf{x}_i - \boldsymbol{\mu}^{(k)} - \Gamma^{(k)}\beta^{(k)}\mathbf{f}\left(y_i\right)\right)^T \left(\Delta^{(k)}\right) \left(\mathbf{x}_i - \boldsymbol{\mu}^{(k)} - \Gamma^{(k)}\beta^{(k)}\mathbf{f}\left(y_i\right)\right)}, \end{aligned}$$

calculemos el valor del $S_M^{(k)} = S_M\left(H_{P_n,\boldsymbol{\theta}^{(k)}}\right)$ que cumple

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\frac{d\left(\mathbf{x}_i, y_i, \boldsymbol{\theta}^{(k)}\right)}{S_M\left(H_{P_n,\boldsymbol{\theta}^{(k)}}\right)}\right) = \kappa_1$$

y los pesos iniciales, sea $w_i^{(k)}$ el peso calculado en base a la $i-$ésima observación, $1 \leq i \leq n$. Esto es

$$\begin{aligned} w_i^{(k)} &= w\left(\frac{d\left(\mathbf{x}_i, y_i, \boldsymbol{\theta}^{(k)}\right)}{S_M\left(H_{P_n,\boldsymbol{\theta}^{(k)}}\right)}, \boldsymbol{\theta}^{(k)}\right) \\ &= w_1\left(\boldsymbol{\theta}^{(k)}\right) \psi_1\left(\frac{d\left(\mathbf{x}_i, y_i, \boldsymbol{\theta}^{(k)}\right)}{S_M\left(H_{P_n,\boldsymbol{\theta}^{(k)}}\right)}\right) \frac{S_M\left(H_{P_n,\boldsymbol{\theta}^{(k)}}\right)}{d\left(\mathbf{x}_i, y_i, \boldsymbol{\theta}^{(k)}\right)} \\ &+ w_2\left(\boldsymbol{\theta}^{(k)}\right) \psi_2\left(\frac{d\left(\mathbf{x}_i, y_i, \boldsymbol{\theta}^{(k)}\right)}{S_M\left(H_{P_n,\boldsymbol{\theta}^{(k)}}\right)}\right) \frac{S_M\left(H_{P_n,\boldsymbol{\theta}^{(k)}}\right)}{d\left(\mathbf{x}_i, y_i, \boldsymbol{\theta}^{(k)}\right)}. \end{aligned} \tag{88}$$

donde

$$\begin{aligned} w_1\left(\boldsymbol{\theta}^{(k)}\right) &= \frac{pE_{P_n}\left[a\left(\frac{d\left(\mathbf{x},y,\boldsymbol{\theta}^{(k)}\right)}{S_M\left(H_{P_n,\boldsymbol{\theta}^{(k)}}\right)}\right)\right]}{2u_n\left(\boldsymbol{\theta}^{(k)}\right) S_M^2\left(H_{P_n,\boldsymbol{\theta}^{(k)}}\right) E_{P_n}\left[b\left(\frac{d\left(\mathbf{x},y,\boldsymbol{\theta}^{(k)}\right)}{S_M\left(H_{P_n,\boldsymbol{\theta}^{(k)}}\right)}\right)\right]} \\ w_2\left(\boldsymbol{\theta}^{(k)}\right) &= \frac{p}{2u_n\left(\boldsymbol{\theta}^{(k)}\right) S_M^2\left(H_{P_n,\boldsymbol{\theta}^{(k)}}\right)}, \end{aligned}$$

y u_n fue definida en (77). Calculamos también

$$W_n\left(\boldsymbol{\theta}^{(k)}\right) = \frac{1}{n}\sum_{i=1}^{n} w_i^{(k)}.$$

2. Con ellos, calculamos

$$\overline{\mathbf{x}}_w^{(k)} = E_{P^*_{n,\boldsymbol{\theta}^{(k)}}}\left[\mathbf{x}\right] = \frac{1}{nW_n\left(\boldsymbol{\theta}^{(k)}\right)}\sum_{i=1}^{n} w_i^{(k)}\mathbf{x}_i$$
$$\overline{\mathbf{f}}_w^{(k)} = E_{P^*_{n,\boldsymbol{\theta}^{(k)}}}\left[\mathbf{f}\left(y\right)\right] = \frac{1}{nW_n\left(\boldsymbol{\theta}^{(k)}\right)}\sum_{i=1}^{n} w_i^{(k)}\mathbf{f}\left(y_i\right)$$

$$\mathbb{X}^{(k)} = \begin{bmatrix} \mathbf{x}_1^T - \overline{\mathbf{x}}_w^{(k)T} \\ \vdots \\ \mathbf{x}_n^T - \overline{\mathbf{x}}_w^{(k)T} \end{bmatrix} \in \mathbb{R}^{n\times p},$$
$$\mathbb{F}^{(k)} = \begin{bmatrix} \mathbf{f}\left(y_1\right)^T - \overline{\mathbf{f}}_w^{(k)T} \\ \vdots \\ \mathbf{f}\left(y_n\right)^T - \overline{\mathbf{f}}_w^{(k)T} \end{bmatrix} \in \mathbb{R}^{n\times r},$$
$$\mathbb{W}^{(k)} = \frac{1}{W_n\left(\boldsymbol{\theta}^{(k)}\right)}\mathrm{diag}\left(w_1^{(k)},\ldots,w_n^{(k)}\right) \in \mathbb{R}^{n\times n}$$

$$\mathrm{cov}^{(k)}_{P^*_{n,\boldsymbol{\theta}^{(k)}}}\left[\mathbf{x},\mathbf{f}\left(y\right)\right] = \frac{1}{n}\mathbb{X}^{(k)T}\mathbb{W}^{(k)}\mathbb{F}^{(k)}$$
$$\mathrm{var}^{(k)}_{P^*_{n,\boldsymbol{\theta}^{(k)}}}\left[\mathbf{f}\left(y\right)\right] = \frac{1}{n}\mathbb{F}^{(k)T}\mathbb{W}^{(k)}\mathbb{F}^{(k)}$$

$$\begin{aligned}\widehat{\Pi}^{(k)} &= \mathrm{cov}^{(k)}_{P^*_{n,\boldsymbol{\theta}^{(k)}}}\left[\mathbf{x},\mathbf{f}\left(y\right)\right]\left(\mathrm{var}^{(k)}_{P^*_{n,\boldsymbol{\theta}^{(k)}}}\left[\mathbf{f}\left(y\right)\right]\right)^{-1}\left(\mathrm{cov}^{(k)}_{P^*_{n,\boldsymbol{\theta}^{(k)}}}\left[\mathbf{x},\mathbf{f}\left(y\right)\right]\right)^{T} \\ &= \frac{1}{n}\mathbb{X}^{(k)T}\mathbb{W}^{(k)}\mathbb{F}^{(k)}\left(\mathbb{F}^{(k)T}\mathbb{W}^{(k)}\mathbb{F}^{(k)}\right)^{-1}\left(\mathbb{X}^{(k)T}\mathbb{W}^{(k)}\mathbb{F}^{(k)}\right)^{T}\end{aligned}$$

3. Calculamos

$$\boldsymbol{\mu}^{(k+1)} = \frac{1}{nW_n\left(\boldsymbol{\theta}^{(k)}\right)}\sum_{i=1}^{n} w_i(\boldsymbol{\theta}^{(k)})\left[\mathbf{x}_i - \Gamma^{(k)}\beta^{(k)}\mathbf{f}\left(y_i\right)\right].$$

4. Luego, hacemos la descomposición espectral de $\left(\Delta^{(k)}\right)^{-1/2}\widehat{\Pi}^{(k)}\left(\Delta^{(k)}\right)^{-1/2}$. Sean $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_d$ los autovectores correspondientes a sus máximos d autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ y construimos la matriz $C^{(k+1)}$

$$C^{(k+1)} = \begin{pmatrix} \mathbf{c}_1 & \cdots & \mathbf{c}_d \end{pmatrix}$$

$\in \mathbb{R}^{p\times d}$, es decir, la matriz cuyas columnas son los autovectores asociados a los d mayores autovalores. Luego

$$\Gamma^{(k+1)} = \left(\Delta^{(k)}\right)^{1/2} C^{(k+1)}.$$

Ahora estamos en condiciones de actualizar β y Δ. Sea

$$\beta^{(k+1)} = \left(\Gamma^{(k+1)}\right)^T \left(\Delta^{(k)}\right)^{-1} \mathbb{X}^{(k)T}\mathbb{W}^{(k)}\mathbb{F}^{(k)} \left(\mathbb{F}^{(k)T}\mathbb{W}^{(k)}\mathbb{F}^{(k)}\right)^{-1}.$$

Antes de actualizar Δ, recalculamos los pesos mediante

$$d_i^{(k+1)} = d\left(\mathbf{x}_i, y_i, \boldsymbol{\theta}^{\circledast(k+1)}\right)$$
$$S_M^{(k+1)} = S_M\left(H_{P_n, \boldsymbol{\theta}^{\circledast(k+1)}}\right)$$

$$w_i^{(k+1)} = w\left(\frac{d_i^{(k+1)}}{S_M^{(k+1)}}, \boldsymbol{\theta}_1^{\circledast(k+1)}\right),$$

donde

$$\boldsymbol{\theta}^{\circledast(k+1)} = \left(\boldsymbol{\mu}^{(k+1)}, \Gamma^{(k+1)}, \widehat{\beta}^{(k+1)}, \widehat{\Delta}^{(k)}\right)$$

Luego, primero definimos el nuevo valor de Δ sin corregir por

$$\Delta^{\circledast(k+1)} = \frac{1}{n}\sum_{i=1}^{n} w_i^{(k+1)} \left[\mathbf{x}_i - \boldsymbol{\mu}^{(k+1)} - \Gamma^{(k+1)}\beta^{(k+1)}\mathbf{f}\left(y_i\right)\right] \cdot \left[\mathbf{x}_i - \boldsymbol{\mu}^{(k+1)} - \Gamma^{(k+1)}\beta^{(k+1)}\mathbf{f}\left(y_i\right)\right]^T.$$

Entonces definimos el nuevo valor de $\boldsymbol{\theta}$ con Δ sin corregir por

$$\boldsymbol{\theta}^{\circledast\circledast(k+1)} = \left(\boldsymbol{\mu}^{(k+1)}, \Gamma^{(k+1)}, \beta^{(k+1)}, \Delta^{\circledast(k+1)}\right).$$

5. Con estos estimadores de paso uno, actualizamos las distancias de Mahalanobis para cada observación

$$\begin{aligned} d_i^{(k+1)} &= d\left(\mathbf{x}_i, y_i, \boldsymbol{\theta}^{\circledast\circledast(k+1)}\right) \\ &= \sqrt{\left(\mathbf{x}_i - \boldsymbol{\mu}^{(k+1)} - \Gamma^{(k+1)}\beta^{(k+1)}\mathbf{f}\left(y_i\right)\right)^T \left(\Delta^{\circledast(k+1)}\right)^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}^{(k+1)} - \Gamma^{(k+1)}\beta^{(k+1)}\mathbf{f}\left(y_i\right)\right)} \end{aligned}$$

o, escritos de manera vectorial, $\mathbf{d}^{(k+1)} = \left(d_1^{(k+1)}, \ldots, d_n^{(k+1)}\right)^T$. Actualizamos el valor del S-estimador de escala $S_M^{(k+1)} = S_M\left(H_{P_n,\boldsymbol{\theta}^{\circledast\circledast(k+1)}}\right)$ como aquella cantidad que cumple con la siguiente ecuación

$$\sum_{i=1}^{n} \frac{1}{n} \rho_1 \left(\frac{d\left(\mathbf{x}_i, y_i, \boldsymbol{\theta}^{\circledast\circledast(k+1)}\right)}{S_M\left(H_{P_n,\boldsymbol{\theta}^{\circledast\circledast(k+1)}}\right)} \right) = \kappa_1$$

Entonces $\Delta^{(k+1)} = \frac{1}{\kappa_2} S_\tau^2\left(H_{P_n,\boldsymbol{\theta}^{\circledast(k+1)}}\right) \Delta^{\circledast(k+1)}$, con

$$S_\tau^2\left(H_{P_n,\boldsymbol{\theta}^{\circledast\circledast(k+1)}}\right) = S_M^2\left(H_{P_n,\boldsymbol{\theta}^{\circledast\circledast(k+1)}}\right) E_{H_{P_n,\boldsymbol{\theta}^{\circledast\circledast(k+1)}}}\left(\rho_2\left(\frac{v}{S_M\left(H_{P_n,\boldsymbol{\theta}^{\circledast\circledast(k+1)}}\right)}\right)\right)$$

$$= \left(S_M^{(k+1)}\right)^2 \frac{1}{n} \sum_{i=1}^{n} \rho_2 \left(\frac{d\left(\mathbf{x}_i, y_i, \boldsymbol{\theta}^{\circledast\circledast(k+1)}\right)}{S_M^{(k+1)}} \right)$$

y definimos $\boldsymbol{\theta}^{(k+1)} = \left(\boldsymbol{\mu}^{(k+1)}, \Gamma^{(k+1)}, \beta^{(k+1)}, \Delta^{(k+1)}\right)$.

El algoritmo se detiene en la iteración $(k+1)-$ésima cuando el cambio en la función objetivo es menor que una cota prefijada $\delta > 0$, es decir

$$\left| \ln \Phi_{B,P_n}\left(\boldsymbol{\theta}^{(k)}\right) - \ln \Phi_{B,P_n}\left(\boldsymbol{\theta}^{(k+1)}\right) \right| < \delta.$$

donde $\ln \Phi_{B,P_n}$ está escrita en la ecuación (82).

9.2.1. Estimadores iniciales

Comenzamos con un estimador robusto inicial para los parámetros. Para ello, utilizamos el modelo de regresión univariado para cada coordenada del vector $\mathbf{x}$,

$$x_j = \mathbf{b}_j^T \begin{bmatrix} \mathbf{f}(y) \\ 1 \end{bmatrix} + \varepsilon_j \tag{89}$$

con $x_j, \varepsilon_j \in \mathbb{R}$, $\mathbf{b}_j \in \mathbb{R}^{r+1}$, $j = 1, \ldots, p$. Estimamos cada vector de coeficientes con un método robusto para la regresión lineal, un MM-estimador, ver Yohai [1987]. Este estimador se puede calcular, por ejemplo, utilizando la rutina `lmRob` del paquete `robust` para R desarrollada por J. Wang et al. [2014] del paquete `R`, (R Core Team [2015]). Observemos que el modelo (89) incluye un término de intercept, ya que la última fila de la covariable es igual a uno. Luego construimos una matriz de coeficientes estimados, $\boldsymbol{\Theta} \in \mathbb{R}^{p\times(r+1)}$ concantenando en una matriz los vectores estimados, por fila, de la siguiente forma,

$$\boldsymbol{\Theta} = \begin{bmatrix} \mathbf{b}_1^T \\ \vdots \\ \mathbf{b}_p^T \end{bmatrix} \in \mathbb{R}^{p\times(r+1)}.$$

Como estimador inicial de $\Gamma\beta$ tomaremos la matriz formada por las primeras r columnas de $\boldsymbol{\Theta}$, y como estimador inicial $\boldsymbol{\mu}^{(0)}$ para $\boldsymbol{\mu}$ la última columna de dicha matriz. Con ellos, computamos los residuos iniciales

$$\mathbf{r}_i^{(0)} = \mathbf{x}_i - \boldsymbol{\mu}^{(0)} - (\Gamma\beta)^{(0)}\,\mathbf{f}\,(y_i)\,.$$

Como estimador inicial de Δ sin corregir tomamos un S-estimador de covarianza de los residuos iniciales $\mathbf{r}_i^{(0)}$, computados con la función bicuadrada de Tukey usando un algoritmo rápido para su cómputo, que se encuentra programado en el software R Core Team [2015], bajo la función `CovSest`, de la librería `rrcov`, desarrollada por Todorov y Filzmoser [2009], es decir

$$\Delta^{*(0)} = \text{CovSest}\left(\mathbf{r}_1^{(0)}, \ldots, \mathbf{r}_n^{(0)}\right).$$

El algoritmo rápido es similar al propuesto en Salibian-Barrera y Yohai [2006]. Cuando corregimos el valor de la matriz de covarianza de la forma establecida en (87) obtenemos el estimador inicial de $\boldsymbol{\theta}$

$$\widehat{\boldsymbol{\theta}}^{(0)} = \left(\boldsymbol{\mu}^{(0)}, \left(\Gamma^{(0)}\beta^{(0)}\right), \frac{S_\tau^2\left(H_{P_n,\left(\widehat{\boldsymbol{\mu}}^{(0)},\Gamma^{(0}\beta^{(0)},\Delta^{*(0)}\right)}\right)}{\kappa_2}\Delta^{*(0)}\right).$$

Obsérvese que en el valor inicial no se estiman separadamente Γ y β sino su producto. Más detalles sobre el MM-estimador de regresión y el S-estimador de covarianza en la Sección 9.2.2.

9.2.2. Algunos comentarios sobre los estimadores iniciales

Para obtener un estimador inicial de la matriz de coeficientes de regresión, realizamos p regresiones lineales univariadas, utilizando un estimador robusto con alto punto de ruptura y alta eficiencia. Utilizamos la función `lmRob` de la librería `robust` de J. Wang et al. [2014] del paquete `R` (R Core Team [2015]). El estimador se calcula a partir de un S- estimador inicial con punto de ruptura igual a 0.5. El estimador final es un MM-estimador, propuesto por Yohai, Stahel, y Zamar [1991], que utiliza una M-escala de los residuos del estimador inicial (ver la definición de M-escalas en la página 34). Para calcular este M-estimador se utiliza un algoritmo de mínimos cuadrados ponderados iterados, tomando al estimador inicial como punto de partida.

Para la elección de los parámetros correspondientes al MM-estimador, tenemos en cuenta las recomendaciones de Maronna, Martin, y Yohai [2006], Secciones 5.5 y 5.9. Ellos muestran que hay un trade-off para los estimadores entre eficiencia bajo normalidad y sesgo bajo contaminación: a mayor eficiencia asintótica, mayor sesgo. Por eso recomiendan elegir los parámetros de modo que el estimador tenga eficiencia de 0,85 (en vez de tomar valores mayores) de modo de controlar el sesgo manteniendo una eficiencia suficientemente alta.

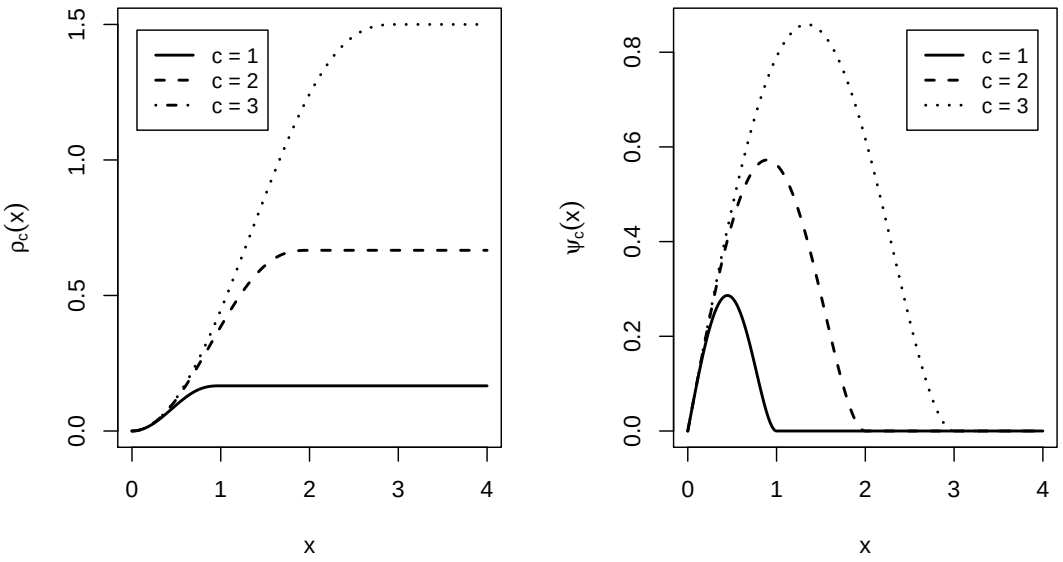
Los S-estimadores de posición multivariada y matriz de dispersión fueron introducidos por Davies [1987] y luego estudiados por Lopuhaa [1989]. Para una descripción de estos estimadores, puede verse también Maronna et al. [2006].

La librería `rrcov` implementa cuatro algoritmos distintos para calcular el S-estimador, en la función `CovSest`. Nosotros utilizamos la opción llamada "bisquare" basada en una $\rho-$función en la familia de la bicuadrada de Tukey, es decir, una función ρ dada por

$$\rho_c(u) = \begin{cases} \dfrac{u^2}{2}\left(1 - \dfrac{u^2}{c^2} + \dfrac{u^4}{3c^4}\right) & \text{si } |u| \leq c \\ \dfrac{c^2}{6} & \text{si } |u| > c, \end{cases} \tag{90}$$

donde la constante c es positiva. Observemos que, para cada u fijo, cuando c aumenta, la $\rho-$función tiende a $\frac{u^2}{2}$. Es fácil verificar que estas $\rho-$funciones satisfacen las hipótesis $A1 - A3$. En la Figura 5 puede verse un gráfico de la $\rho-$función y de su derivada ψ para algunos valores de la constante c. Utilizamos la función `CovSest` de la librería `rrcov` del R, con la opción `method="bisquare"`, que calcula los S-estimadores de posición y escala de la forma que describimos a continuación, como proponen Maronna et al. [2006], Secciones 6.7.3 y 6.7.5.

Figura 5: Función ρ_c bicuadrada definida en (90) y su derivada ψ_c, para distintos valores de la constante c.



9.2.3. Detalles sobre la implementación del algoritmo

Cuando la función ρ_2 se elige como $\rho_2(u) = u^2$, el $\tau-$estimador para el modelo PFC resulta ser el estimador de máxima verosimilitud para el caso normal. Luego, para combinar robustez con eficiencia, podemos elegir la función ρ_2 de modo que coincida con el cuadrado en una amplia zona alrededor del cero. Para ello se utiliza una familia de funciones ρ_c propuesta en Muler y Yohai [2002] y que es una aproximación polinomial simple a la función ρ óptima derivada en Yohai y Zamar [1997]. Esta familia tiene la forma

$$\rho_c(u) = \begin{cases} 1{,}38\left(\frac{u}{c}\right)^2 & \text{si } \left|\frac{u}{c}\right| \leq \frac{2}{3} \\ 4{,}04\left(\frac{u}{c}\right)^8 - 11{,}66\left(\frac{u}{c}\right)^6 + 10{,}76\left(\frac{u}{c}\right)^4 - 2{,}69\left(\frac{u}{c}\right)^2 + 0{,}55 & \text{si } \frac{2}{3} < \left|\frac{u}{c}\right| \leq 1 \\ 1 & \text{si } \left|\frac{u}{c}\right| > 1 \end{cases} \tag{91}$$

y la denominamos $\rho-$casi-óptima. En la implementación del algoritmo, las constantes se eligen para tener simultáneamente, robustez (para eso se elige la función ρ_1) y alta eficiencia bajo normalidad (para eso se elige la función ρ_2), como detallamos a continuación. En la Figura 7 además, vemos que tanto la familia de $\rho-$funciones bicuadrada como la familia de $\rho-$funciones que denominamos casi-óptima satisfacen la hipótesis $A4$, ya que en ambas familias la función $\psi(u)/u$ resulta decreciente. En dicha figura podemos ver que para la $\rho-$casi-óptima la función $\psi(u)/u$ que define los pesos es una versión suavizada de la función escalón (indicadora o heaviside), es decir que básicamente, dependiendo de si la distancia de Mahalanobis de una observación a los valores calculados de los parámetros es pequeña o no, las observaciones o bien entran al cálculo de los estimadores (básicamente todas con el mismo peso) o bien no entran en dichos cálculos.

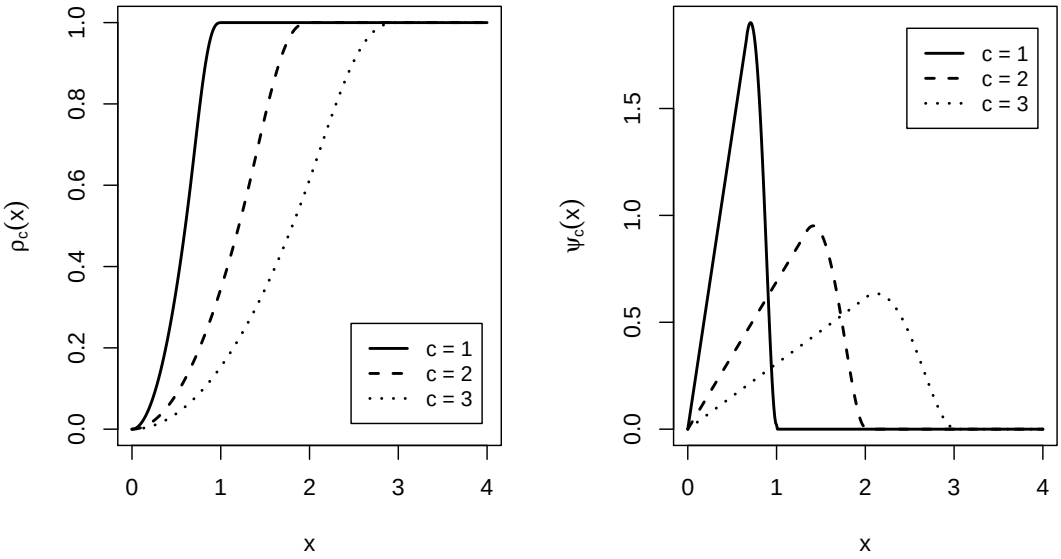
García Ben et al. [2006], proponen $\tau-$estimadores para el modelo lineal multivariado (tanto para la matriz de coeficientes del modelo, como para la matriz de covarianza de los errores), y en el Teorema 2 prueban que el máximo punto de ruptura que pueden alcanzar dichos estimadores es cercano a 0,5 cuando el tamaño de muestra n es grande (es en realidad $0{,}5-(p+q-1)/(2n)$, donde p es es la dimensión del vector de observaciones independientes y q es la dimensión del vector de las variables dependientes en el modelo lineal multivariado (MLM)) y se alcanza cuando

$$\frac{\kappa_1}{\text{máx}_u\, \rho_1(u)} = \frac{1}{2}. \tag{92}$$

En el mismo artículo, se prueba en el Corolario al Teorema 5, que, bajo condiciones de regularidad de las $\rho-$funciones, de regularidad de la distribución del vector de covariables y elipticidad de la distribución de los errores, la eficiencia relativa asintótica (ARE, por sus siglas en inglés) del $\tau-$estimador de la matriz de los coeficientes del MLM respecto del estimador de máxima verosimilitud de dicha matriz de coeficientes, está dada por

$$ARE(\psi_1, \psi_2, H_0) = \frac{E_{H_0}(v^2)}{c_0^2 k_0^2 E_{H_0}\left(\psi^{*2}\left(\frac{v}{k_0}\right)\right)} \tag{93}$$

Figura 6: Función ρ_c polinomial (casi óptima) definida en (91) y su derivada ψ_c, para distintos valores de la constante c.



donde $\psi_i = \rho_i'$, H_0 es la verdadera distribución de las distancias de Mahalanobis del MLM. Bajo normalidad de los errores del MLM en $\mathbb{R}^p$, H_0 resulta ser la distribución de la raíz cuadrada de una variable $\chi_p^2 = \Gamma\left(\frac{p}{2}, \frac{1}{2}\right)$ (chi cuadrado con p grados de libertad), es decir, tiene densidad

$$f(v) = \frac{1}{2^{\frac{p}{2}-1}\Gamma\left(\frac{p}{2}\right)} v^{p-1} e^{-v^2/2} I_{(0,+\infty)}(v).$$

La función ψ^* se define por

$$\psi^*(v) = C\psi_1(v) + D\psi_2(v)$$

con

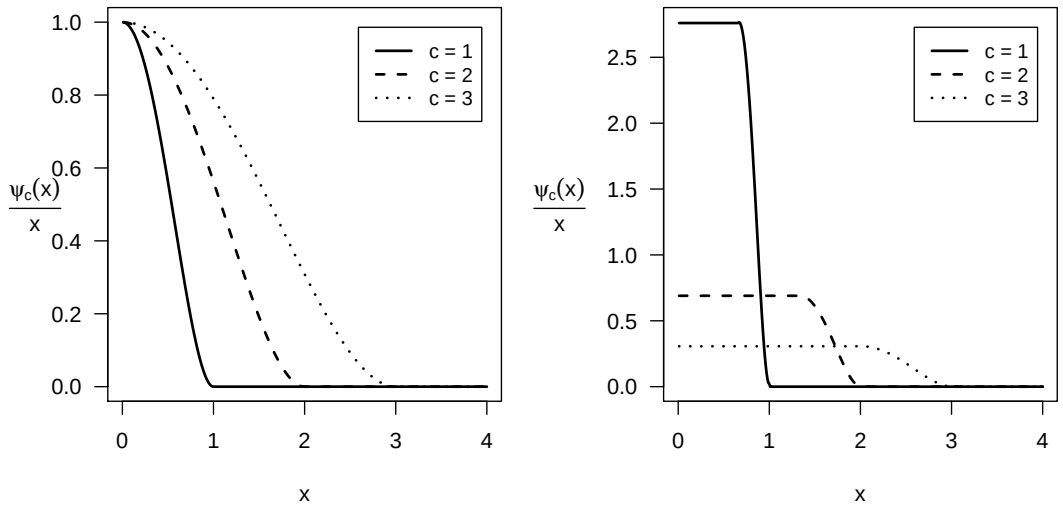
$$C = E_{H_0}\left(2\rho_2\left(\frac{v}{k_0}\right) - \psi_2\left(\frac{v}{k_0}\right)\frac{v}{k_0}\right) \tag{94}$$

$$D = E_{H_0}\left(\psi_1\left(\frac{v}{k_0}\right)\frac{v}{k_0}\right) \tag{95}$$

y

$$k_0 = S(H_0)$$

Figura 7: Gráfico de $\psi_c(x)/x$, para la familia bicuadrada definida en (90) a la izquierda y para la polinomial (casi óptima) definida en (91) a la derecha, para distintos valores de la constante c.



definido por

$$E_{H_0}\left(\rho_1\left(\frac{v}{S(H_0)}\right)\right) = \kappa_1 \tag{96}$$

y

$$c_0 = \frac{p}{E_{H_0}\left((p-1)\,\frac{\psi^*\left(\frac{v}{k_0}\right)}{\frac{v}{k_0}} + \psi^{*\prime}\left(\frac{v}{k_0}\right)\right)}.$$

Asumiendo que la distribución de los errores del MLM es gaussiana, podemos calcular las constantes κ_1 y c_1 a partir de (92) y (97), a continuación

$$\kappa_1 = E_{H_0}\left(\rho_1(v)\right), \tag{97}$$

ya que $\text{máx}_u\, \rho_1(u) = 1$. Con esa elección de constantes, por (96) resulta $k_0 = S(H_0) = 1$. A partir de ρ_1 calculamos D según la ecuación (95). Para obtener un $\tau-$estimador que sea eficiente, podemos elegir el valor de c_2 para que

$$ARE(\psi_1, \psi_2, H_0) = 0{,}90,$$

bajo normalidad de los errores del MLM. Luego, calculamos κ_2 como aquel valor que cumple

$$\kappa_2 = E_{H_0}\left(\rho_2(v)\right).$$

A continuación figuran algunos de los valores de las constantes que consideramos.

p	2		5		10	
eficiencia	0.85	0.90	0.85	0.90	0.85	0.90
c_1	2.0803	2.0803	3.6076	3.6076	5.2108	5.2108
κ_1	0.50	0.50	0.50	0.50	0.50	0.50
c_2	3.0193	3.2332	3.7252	3.9759	2.5762	2.1767
κ_2	0.2898	0.2567	0.4744	0.4242	0.9739	0.9920

9.3. Cómo evaluar la estimación de un subespacio

Nos interesa resumir la bondad de la estimación obtenida con distintos métodos, para poder comparar el ajuste que se obtiene usando los $\tau-$estimadores del subespacio de reducción y usando los EMV en distintos escenarios simulados, con muestras finitas, ya que, como hemos visto, lo que queda identificado es el subespacio y no sus generadores. En esta sección definiremos ángulos entre subespacios y distancias entre los mismos.

Distintas propuestas estadísticas para reducir la dimensión han utilizado distintos métodos para evaluar la efectividad obtenida. En el caso de las simulaciones, cuando se conocen los parámetros con los que fueron obtenidos los datos, se tiene por un lado Δ_0 y Γ_0 verdaderos y por el otro $\widehat{\Delta}$ y $\widehat{\Gamma}$ estimados. Citemos algunos ejemplos. Cook [2007] analiza el caso $d = 1$ y propone utilizar el ángulo entre Γ_0 y $\widehat{\Gamma}$. Cook y Forzani [2008] también simulan el caso $d = 1$ pero utilizan, en cambio, el ángulo entre $\Delta_0^{-1}\Gamma_0$ y $\widehat{\Delta}^{-1}\widehat{\Gamma}$, es decir, entre los generadores de los subespacios de reducción. Para $d > 1$, el artículo de B. Li, Zha, y Chiaromonte [2005] y el de B. Li y Wang [2007] usan la *distancia de proyección* (ver la Definición 9.2, que también se denomina *distancia de Gap*) entre el subespacio de reducción original y el estimado por el método de *contour regression* que proponen. K.-C. Li [1991a] toma el promedio de las correlaciones canónicas al cuadrado entre la variable proyectada al subespacio estimado $\widehat{\Delta}^{-1}\widehat{\Gamma}\mathbf{x}$ y las verdaderas proyecciones $\Delta_0^{-1}\Gamma_0\mathbf{x}$. H. Wang y Xia [2008] reportan el error de estimación promedio.

9.3.1. Vectores y ángulos principales

Sean $\mathcal{A}$ y $\mathcal{B}$ dos subespacios en $\mathbb{R}^p$. Introducimos la definición de ángulos principales entre un subespacio $k-$dimensional y uno $l-$dimensional. Para una tratamiento completo puede verse Eaton [1983], Proposición 1.48 y subsiguientes, o Watkins [1991], Sección 7.5.

Definición 9.1 *Sean $\mathcal{A}$ y $\mathcal{B}$ dos subespacios en $\mathbb{R}^p$, con $\dim(\mathcal{A}) = k, \dim(\mathcal{B}) = l, r = \min(k, l)$. Definimos el **i-ésimo par de vectores principales** $(\mathbf{a}_i, \mathbf{b}_i)$ para $i = 1, \ldots, r$ recursivamente como las soluciones al siguiente problema de optimización*

$$\begin{aligned} &\max \mathbf{a}^T\mathbf{b} \\ \textit{sujeto a:}\quad &\mathbf{a} \in \mathcal{A}, \mathbf{b} \in \mathcal{B} \\ &\|\mathbf{a}\| = \|\mathbf{b}\| = 1 \\ &\mathbf{a}^T\mathbf{a}_1 = \cdots = \mathbf{a}^T\mathbf{a}_{i-1} = 0 \\ &\mathbf{b}^T\mathbf{b}_1 = \cdots = \mathbf{b}^T\mathbf{b}_{i-1} = 0, \end{aligned}$$

(para $i = 1$ las condiciones de ortogonalidad son vacías). Los ángulos principales se definen por

$$\cos\zeta_i = \mathbf{a}_i^T\mathbf{b}_i, \qquad i = 1, \ldots, r.$$

Como el coseno es una función decreciente, resulta $0 \leq \zeta_1 \leq \cdots \leq \zeta_r \leq \frac{\pi}{2}$. *Notaremos* $\zeta_i(\mathcal{A}, \mathcal{B}) = \zeta_i$ *al* ***i-ésimo ángulo principal*** *entre* $\mathcal{A}$ *y* $\mathcal{B}$.

En otras palabras, el primer ángulo principal entre los subespacios $\mathcal{A}$ y $\mathcal{B}$ es el ángulo más pequeño que se puede armar entre un vector en $\mathcal{A}$ y un vector en $\mathcal{B}$, ya que el ángulo se minimiza cuando el coseno se maximiza. El par de vectores unitarios $\mathbf{a}_1 \in \mathcal{A}, \mathbf{b}_1 \in \mathcal{B}$ en los que se realiza dicho máximo son el primer par de vectores principales. A partir de él, se define el segundo ángulo principal ζ_2 como el menor ángulo que se puede formar entre un vector en $\mathcal{A}$ que sea ortogonal a $\mathbf{a}_1$ y un vector en $\mathcal{B}$ que sea ortogonal a $\mathbf{b}_1$. Los restantes ángulos se definen en forma similar.

Observación 9.1 *Si* $\zeta_1 = 0$ *entonces* $\dim(\mathcal{A} \cap \mathcal{B}) \geq 1$. *Más aún, si* $\zeta_1 = \cdots = \zeta_m = 0 < \zeta_{m+1}$, *entonces*

$$\mathcal{A} \cap \mathcal{B} = \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_m\} = \operatorname{span}\{\mathbf{b}_1, \ldots, \mathbf{b}_m\}.$$

Este resultado es el Teorema 6.4.2 de Golub y Van Loan [2013].

Si $\dim(\mathcal{A}) = \dim(\mathcal{B}) = k$, claramente $\mathcal{A} = \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ y $\mathcal{B} = \operatorname{span}\{\mathbf{b}_1, \ldots, \mathbf{b}_k\}$. Los vectores principales no quedan unívocamente determinados, pero sí lo están los ángulos principales. Por ejemplo, el par de vectores principales $\mathbf{a}_i$ y $\mathbf{b}_i$ se puede intercambiar con $-\mathbf{a}_i$ y $-\mathbf{b}_i$, pero si $\zeta_1 < \cdots < \zeta_r$, los vectores principales estarán unívocamente determinados, exceptuando el signo de la primer componente no nula del vector. El siguiente Teorema muestra como obtener los ángulos principales.

Teorema 9.1 *Sean* $\mathcal{A}$ *y* $\mathcal{B}$ *subespacios en* $\mathbb{R}^p$, *y sean* $P_{\mathcal{A}}$ *y* $P_{\mathcal{B}}$ *las matrices de las proyecciones ortogonales en* $\mathcal{A}$ *y* $\mathcal{B}$ *en la base canónica. Si* $k = \operatorname{rango}(P_{\mathcal{B}}P_{\mathcal{A}})$, *entonces existen conjuntos ortonormales* $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\} \subset \mathcal{A}$, $\{\mathbf{b}_1, \ldots, \mathbf{b}_k\} \subset \mathcal{B}$ *y números positivos* $\mu_1 \geq \cdots \geq \mu_k$ *tales que*

i. $P_B P_{\mathcal{A}} = \sum_{i=1}^k \mu_i \mathbf{b}_i \mathbf{a}_i^T = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_k \end{bmatrix} \operatorname{diag}(\mu_1, \ldots, \mu_k) \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_k^T \end{bmatrix}$.

ii. $P_{\mathcal{A}} P_B P_{\mathcal{A}} = \sum_{i=1}^k \mu_i^2 \mathbf{a}_i \mathbf{a}_i^T$.

iii. $P_B P_{\mathcal{A}} P_B = \sum_{i=1}^k \mu_i^2 \mathbf{b}_i \mathbf{b}_i^T$.

iv. $0 < \mu_j \leq 1$ *y* $\mathbf{a}_i^T \mathbf{b}_j = \delta_{ij}\mu_j$ *para* $i, j = 1, ..., k$.

v. *Si* $\mathbf{x} \in \mathcal{A} \cap (\operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_k\})^{\perp}$ *y* $\mathbf{w} \in \mathcal{B}$, *entonces* $\mathbf{x}^T\mathbf{w} = 0$.
Si $\mathbf{w} \in \mathcal{B} \cap (\operatorname{span}\{\mathbf{b}_1, \ldots, \mathbf{b}_k\})^{\perp}$ *y* $\mathbf{x} \in \mathcal{A}$, *entonces* $\mathbf{x}^T\mathbf{w} = 0$.

Teorema 9.2 *En la notación del Teorema previo, sea* $\mathcal{A}_1 = \mathcal{A}$, $\mathcal{B}_1 = \mathcal{B}$,

$$\mathcal{A}_i = \mathcal{A} \cap (\text{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_{i-1}\})^\perp$$
$$\mathcal{B}_i = \mathcal{B} \cap (\text{span}\{\mathbf{b}_1, \ldots, \mathbf{b}_{i-1}\})^\perp$$

para $i = 2, \ldots, k+1$. *Entonces*

$$\sup_{\substack{\mathbf{a}\in\mathcal{A}_i \\ \|\mathbf{a}\|=1}} \sup_{\substack{\mathbf{b}\in\mathcal{B}_i \\ \|\mathbf{b}\|=1}} \mathbf{a}^T\mathbf{b} = \mathbf{a}_i^T\mathbf{b}_i = \mu_i,$$

para $i = 1, \ldots, k$. *También,* $\mathcal{A}_{k+1} \perp \mathcal{B}$ *y* $\mathcal{A} \perp \mathcal{B}_{k+1}$. ■

Observación 9.2 *En virtud del Teorema 9.1 y el Teorema 9.2 y la definición de ángulos principales entre subespacios, tenemos que*

$$\cos\zeta_i = \mu_i, \qquad i = 1, \ldots, k.$$

En el Apéndice I.1 damos tres procedimientos distintos para calcular los ángulos principales, dependiendo de la información que tengamos de cada subespacio, o de cuán pequeños sean estos ángulos.

9.3.2. Distancias entre subespacios

Queremos definir una noción de distancia entre subespacios $k-$dimensionales de $\mathbb{R}^p$. Para llevar adelante esta tarea, podríamos valernos tanto de los ángulos principales como de las posibles bases ortonormales que se pueden tomar en ellos. El siguiente teorema justifica esta idea. Sea $\mathcal{O}(p)$ el conjunto de matrices $p \times p$ ortogonales. A continuación, se enuncia el Teorema 3 de Wong [1967].

Teorema 9.3 *Cualquier noción de distancia entre subespacios* k *dimensionales de* $\mathbb{R}^p$ *que convierte a la Grassmaniana* $\text{Grass}(k, \mathbb{R}^p)$ *en un espacio métrico que sea invariante por rotaciones en* $\mathcal{O}(p)$ *debe ser una función de los ángulos principales. Es decir, si* $d : \text{Grass}(k, \mathbb{R}^p) \times \text{Grass}(k, \mathbb{R}^p) \to [0, +\infty)$ *es una distancia que además cumple*

$$d(\text{span}(VA), \text{span}(VB)) = d(\text{span}(A), \text{span}(B)) \qquad \textit{para toda } V \in \mathcal{O}(p)$$

para toda $A, B \in \mathbb{R}^{p\times k}$ *con* $\text{rango}(A) = \text{rango}(B) = k$, *entonces* d *debe ser una función de los* $\zeta_i(\text{span}(A), \text{span}(B))$.

Se pueden definir varias nociones de distancia entre subespacios de la misma dimensión. A continuación presentamos varias. Puede consultarse, por ejemplo Hamm y Lee [2008] o Edelman, Arias, y Smith [1998], para una descripción más extensa. La más utilizada es la distancia de proyección.

Definición 9.2 ***(Distancia de proyección)*** *Sean $\mathcal{A}$ y $\mathcal{B}$ subespacios $k-$dimensionales de $\mathbb{R}^p$. Sean $A, B \in \mathbb{R}^{p\times k}$ con $\operatorname{rango}(A) = \operatorname{rango}(B) = k$, tales que $\mathcal{A} = \operatorname{span}(A), \mathcal{B} = \operatorname{span}(B)$. Definimos la* ***distancia de proyección*** *d_{proy} entre ellos por*

$$d_{\text{proy}}(\mathcal{A},\mathcal{B}) = \max_{\substack{\mathbf{a}\in\mathcal{A}\\ \|\mathbf{a}\|_2=1}} d(\mathbf{a},\mathcal{B}) \tag{98}$$

donde $\|\mathbf{a}\|_2 = \sqrt{\mathbf{a}^T\mathbf{a}}$ y

$$d(\mathbf{a},\mathcal{B}) = \min_{\mathbf{b}\in\mathcal{B}} \|\mathbf{a}-\mathbf{b}\|_2 . \tag{99}$$

Observación 9.3 *Algunos autores la denominan la distancia de Gap.*

El siguiente teorema permite calcular esta distancia.

Teorema 9.4 *Sean $\{\mathbf{q}_1,\ldots,\mathbf{q}_k\}$ y $\{\mathbf{q}_{k+1},\ldots,\mathbf{q}_d\}$ bases ortonormales para $\mathcal{A}$ y $\mathcal{A}^\perp$, respectivamente, y sean $\{\mathbf{r}_1,\ldots,\mathbf{r}_k\}$ y $\{\mathbf{r}_{k+1},\ldots,\mathbf{r}_d\}$ bases ortonormales para $\mathcal{B}$ y $\mathcal{B}^\perp$, respectivamente. Llamamos*

$$Q_1 = \begin{bmatrix}\mathbf{q}_1 & \cdots & \mathbf{q}_k\end{bmatrix} \in \mathbb{R}^{p\times k}, \qquad Q_2 = \begin{bmatrix}\mathbf{q}_{k+1} & \cdots & \mathbf{q}_p\end{bmatrix} \in \mathbb{R}^{p\times(p-k)}$$
$$R_1 = \begin{bmatrix}\mathbf{r}_1 & \cdots & \mathbf{r}_k\end{bmatrix} \in \mathbb{R}^{p\times k}, \qquad R_2 = \begin{bmatrix}\mathbf{r}_{k+1} & \cdots & \mathbf{r}_p\end{bmatrix} \in \mathbb{R}^{p\times(p-k)}.$$

Entonces se tiene

$$\begin{aligned} d_{\text{proy}}(\mathcal{A},\mathcal{B}) &= \|R_2^TQ_1\|_2 = \|R_1^TQ_2\|_2 = \|Q_1^TR_2\|_2 = \|Q_2^TR_1\|_2 \\ &= \|Q_1Q_1^T - R_1R_1^T\|_2 = \|P_{\mathcal{A}} - P_{\mathcal{B}}\|_2 \\ &= \operatorname{sin}\zeta_k(\mathcal{A},\mathcal{B}) \end{aligned}$$

donde para matrices, la norma $\|\cdot\|_2$ denota la norma de operadores lineales (o norma inducida o espectral ver la Sección B.1.2, es el mayor valor singular de la matriz, $\|Q\|_2 = \sqrt{\lambda_{\text{máx}}(Q^TQ)}$) y $P_{\mathcal{A}}$ denota la matriz asociada a la proyección ortogonal en el subespacio $\mathcal{A}$).

Para una demostración de este Teorema, ver Watkins [1991].

Definición 9.3 ***(Distancia Grassmaniana)*** *Sean $\mathcal{A}$ y $\mathcal{B}$ subespacios $k-$dimensionales de $\mathbb{R}^p$. Se define la* ***distancia Grassmaniana*** *entre ellos por*

$$d_{\text{Grass}(k,\mathbb{R}^p)}(\mathcal{A},\mathcal{B}) = \left(\sum_{i=1}^{k}\zeta_i^2(\mathcal{A},\mathcal{B})\right)^{1/2}.$$

Definición 9.4 ***(Distancia Cordal)*** *Sean $\mathcal{A}$ y $\mathcal{B}$ subespacios $k-$dimensionales de $\mathbb{R}^p$. Se define la* ***distancia cordal*** *entre ellos por*

$$d_{\text{cordal}}(\mathcal{A},\mathcal{B}) = \left(\sum_{i=1}^{k}\operatorname{sin}^2\zeta_i(\mathcal{A},\mathcal{B})\right)^{1/2} = \frac{1}{\sqrt{2}}\|P_{\mathcal{A}} - P_{\mathcal{B}}\|_F ,$$

donde $\|\cdot\|_F$ es la norma Frobenius para matrices, ver la Sección B.1.2.

Al valor $\sin\zeta_1(\mathcal{A},\mathcal{B})$ a veces se lo denomina la *distancia de máxima correlación,* pero no es una métrica ya que puede valer cero sin que los subespacios $\mathcal{A}$ y $\mathcal{B}$ sean iguales. Puede probarse que, equivalentemente, está basada en la mayor correlación canónica, en el sentido dado en la página 66.

En virtud de las descripciones anteriores, para comparar subespacios utilizaremos el ángulo principal (para $d=1$), o en general el d-ésimo ángulo principal entre el verdadero subespacio de reducción de la dimensión y el estimado, como medida de performance de los estimadores.

9.4. Simulación

Realizamos dos simulaciones que describimos a continuación, para el caso en el que $d=1$ ó $d=2$.

9.4.1. Caso $d=1$

El modelo simulado es

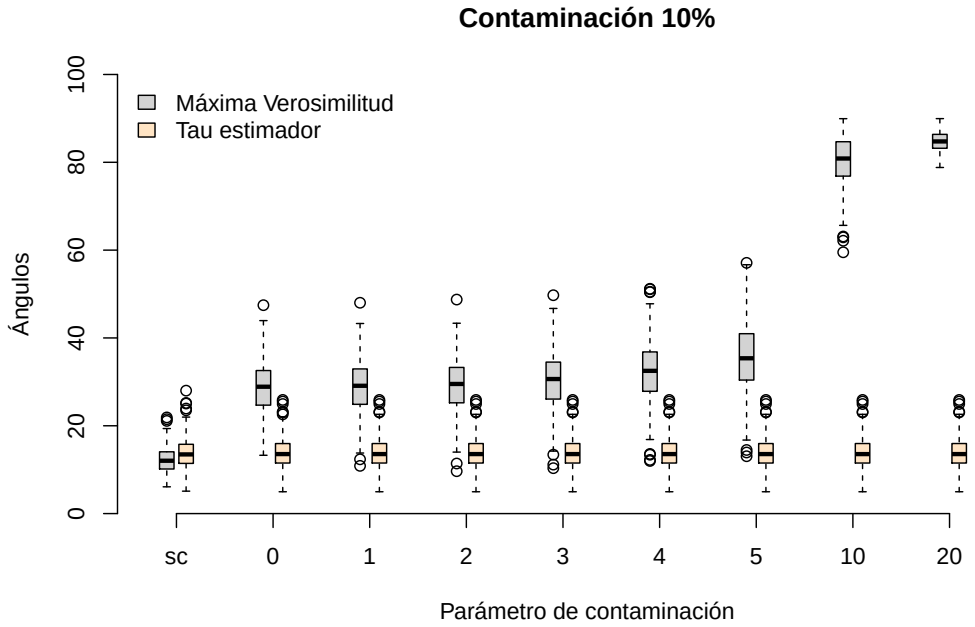
$$\mathbf{x} = \boldsymbol{\mu}_0 + \Gamma_0\beta_0\mathbf{f}(y) + \Delta_0^{1/2}\mathbf{u}$$

donde $\mathbf{x}\in\mathbb{R}^p$, con $p=10$, $d=1$, $r=2$, $\boldsymbol{\mu}_0=\mathbf{0}$, $\Gamma_0=(1,0,\ldots,0)^T\in\mathbb{R}^{p\times d}$, $\beta_0=(1,1)\in\mathbb{R}^{d\times r}$, $\mathbf{f}(y)=\left(y,y^2\right)^T\in\mathbb{R}^r$, $\Delta_0=I_p$, $\mathbf{u}\sim N_p(\mathbf{0},I_p)$, $y\sim U(0,4)$. Para evaluar el desempeño del estimador bajo contaminación de los datos, tomamos un 10 % de las observaciones $(\mathbf{f}(y),\mathbf{x})$ de la forma $(\mathbf{f}^*(y),\mathbf{x}^*)$, donde $\mathbf{f}^*(y)=\mathbf{f}(10)=\left(10,10^2\right)^T$ y $\mathbf{x}^*=\lambda(0,1,0,\ldots,0)^T$ con $\lambda=0,1,2,3,4,5,10,15,20$. En la Figura 8 mostramos los boxplots de los ángulos que se obtienen entre el verdadero subespacio de reducción $\left(\Delta^{-1}\mathcal{S}_\Gamma\right)$ y el subespacio de reducción estimado $\left(\widehat{\Delta}^{-1}\mathcal{S}_{\widehat{\Gamma}}\right)$ por máxima verosimilitud y por el $\tau-$estimador que proponemos, en los distintos escenarios: sin contaminación (SC) y para los distintos valores de λ considerados, a partir de $B=500$ muestras de tamaño $n=200$, con 10 % de datos contaminados. Las constantes descriptas en la Sección 9.2.3 resultan ser, para este caso

$$\begin{aligned} c_1 &= 5{,}210748 \\ \kappa_1 &= 0{,}5000047 \\ c_2 &= 4{,}841626 \\ \kappa_2 &= 0{,}569148. \end{aligned} \tag{100}$$

En la Figura 8 vemos que cuando no hay contaminación, el ángulo entre el subespacio verdadero y el estimado por ambos métodos es bastante pequeño, siendo más pequeño aún el de máxima verosimilitud (mediana menor a 15^o en ambos métodos). Sin embargo, apenas empezamos a contaminar, el estimador de máxima verosimilitud se rompe, dando ángulos alrededor de 35^o cuando $\lambda\leq 5$, para romperse abiertamente (estimando una dirección prácticamente ortogonal a la original) para $\lambda=10$ ó $\lambda=20$ (mediana superior a 80^o), mientras que el $\tau-$estimador prácticamente no se modifica, siempre la mediana del ángulo entre la dirección verdadera y la estimada da apenas menor a 15^o.

Figura 8: Boxplots de los ángulos entre el subespacio verdadero y el subespacio estimado por máxima verosimilitud (en gris) y usando el $\tau-$estimador propuesto (en salmón) para la simulación descripta en la Sección 9.4.1.



9.4.2. Caso $d = 2$

Primera simulación

Sin contaminar El modelo simulado es

$$\mathbf{x} = \boldsymbol{\mu}_0 + \Gamma_0 \beta_0 \mathbf{f}_0(y) + \Delta_0^{1/2}\mathbf{u}$$

donde $\mathbf{x} \in \mathbb{R}^p$, con $p = 10, d = 2, r = 2, \boldsymbol{\mu}_0 = \mathbf{0}, \Gamma_0 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \cdots & 0 & 0 \end{bmatrix}^T \in \mathbb{R}^{p\times d}$, $\beta_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{d\times r}$, $\mathbf{f}_0(y) = \left(y, y^2\right)^T \in \mathbb{R}^r$, $\mathbf{u} \sim N_p(\mathbf{0}, I_p)$, $y \sim U(0,4)$, $\Delta_0 = 0{,}1 \cdot I_p$.

Para evaluar el desempeño del estimador bajo contaminación de los datos, tomamos un 10 % de las observaciones contaminadas. Las observaciones contaminadas ($\mathbf{x}$,y) siguen el siguiente modelo

$$\mathbf{x} = \boldsymbol{\mu}_0 + \Gamma_{C_j} \beta_0 \mathbf{f}_{C_j}(y) + \Delta_{C_j}^{1/2}\mathbf{u},$$

con $1 \leq j \leq 4$. En cada uno de los siguientes cuatro escenarios, contaminamos de formas distintas.

Contaminación 1 Tomamos $\mathbf{f}_{C_1}(y) = (10,10)^T$, $\Delta_{C_1} = \Delta_0 = 0{,}1 \cdot I_p$ y

$$\Gamma_{C_1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T.$$

Observemos que en este escenario, el subespacio de reducción de las observaciones no contaminadas (span $\left(\Delta_0^{-1}\Gamma_0\right)$) y el de las contaminadas $\left(\text{span}\left(\Delta_0^{-1}\Gamma_{C_1}\right)\right)$ forman los siguientes ángulos principales: 0^o y 90^o, es decir, uno de los generadores de ambos subespacios coincide, el otro está a 90^o.

Contaminación 2 Tomamos $\mathbf{f}_{C_2}(y) = \mathbf{f}_0(10) = \left(10, 10^2\right)^T$, $\Delta_{C_2} = \Delta_0 = 0{,}1 \cdot I_p$ y $\Gamma_{C_2} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T$.

Contaminación 3 Tomamos $\mathbf{f}_{C_3}(y) = \mathbf{f}_0(10) = \left(10, 10^2\right)^T$, $\Delta_{C_3} = 0{,}5 \cdot I_p$ y $\Gamma_{C_3} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T$.

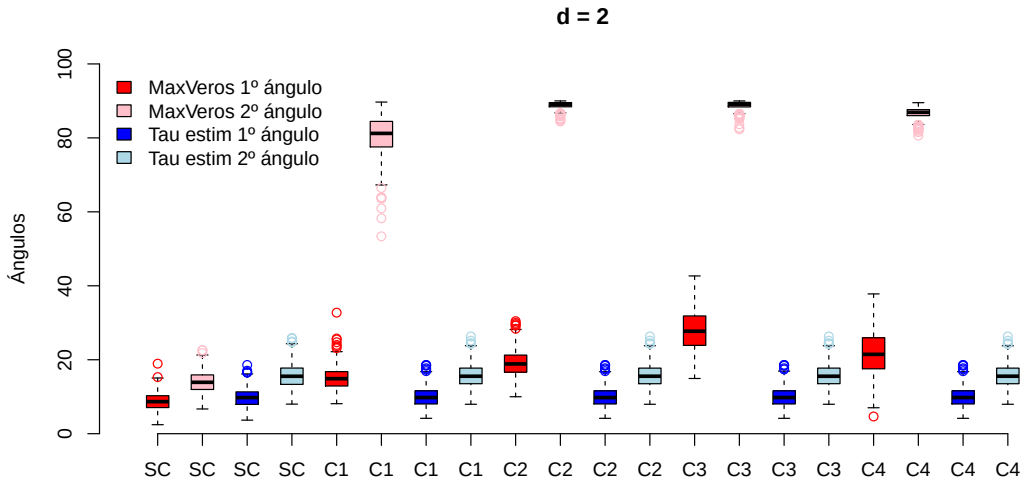
Contaminación 4 Tomamos $\mathbf{f}_{C_4}(y) = \mathbf{f}_0(10) = \left(10, 10^2\right)^T$, $\Delta_{C_4} = 0{,}5 \cdot I_p$ y $\Gamma_{C_4} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix}^T$. En esta situación, los ángulos principales son ambos iguales a 90^o, es decir, en este caso están contaminadas ambas direcciones principales.

Si bien la distancia de proyección entre dos subespacios de dimensión dos depende solamente del segundo ángulo principal entre el subespacio verdadero y el estimado, hemos resumido el comportamiento de ambos ángulos principales a lo largo de la simulación, para tener un mejor conocimiento del funcionamiento de ambos métodos bajo los distintos escenarios.

Las constantes $c_1, \kappa_1, c_2, \kappa_2$ las tomamos como en (100). En la Figura 9 graficamos los boxplots del primer y segundo ángulo principal entre el subespacio verdadero y el subespacio estimado por máxima verosimilitud (en color rojo y rosa, respectivamente) y lo mismo para el $\tau-$estimador propuesto (en color azul y celeste, respectivamente) para la simulación descripta en la Sección 9.4.2. Se generaron $n = 200$ datos sin contaminación. Luego, se reemplazaron 20 observaciones de cada muestra por otras 20 contaminadas bajo los cuatro escenarios descriptos en dicha sección.

En dicho gráfico vemos que, para los datos sin contaminación, ambos métodos estiman igualmente bien el subespacio de reducción de la dimensión (las medianas de los ángulos principales para máxima verosimilitud son 8.7^o y 13.9^o, respectivamente, mientras que para el $\tau-$estimador resultan ser 9.8^o y 15.5^o). Para las contaminaciones C1 a C4 vemos que el método de máxima verosimilitud erra en la estimación de la segunda dirección principal en todos los casos (medianas cercanas a 90^o, el método estima una dirección perpendicular a la que debería), y con respecto a la primera, la estima razonablemente bien (medianas de 14.9^o, 18.9^o, 27.7^o y 21.5^o respectivamente, en los ángulos entre el primer vector principal verdadero y estimado), mientras que los ángulos principales de

Figura 9: Boxplots del primer y segundo ángulo principal entre el subespacio verdadero y el subespacio estimado por máxima verosimilitud (en rojo y rosa, respectivamente) y usando el $\tau-$estimador propuesto (en azul y celeste, respectivamente) para la simulación descripta en la Sección 9.4.2. Se generaron $n = 200$ datos sin contaminación. Luego, se reemplazaron 20 observaciones de cada muestra por otras 20 contaminadas bajo los cuatro escenarios descriptos en dicha sección.

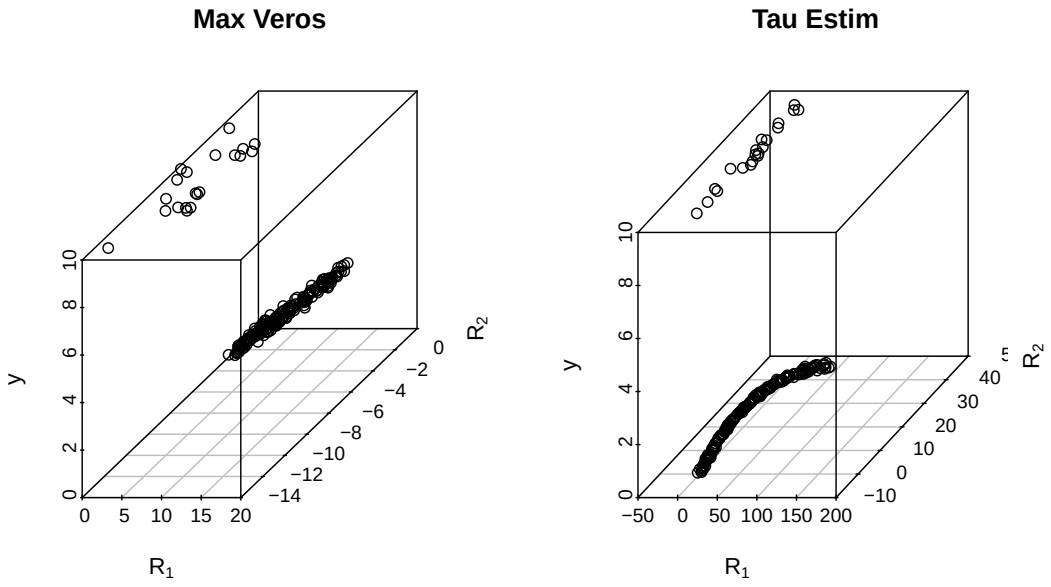


los $\tau-$estimadores casi no se modifican con la contaminación, mostrando su desempeño robusto en muestras finitas (medianas de 9.8^o y 15.5^o en todos los casos).

Para comparar ambos estimadores en una muestra, vemos en la Figura 10 un scatter plot de las 200 observaciones correspondientes a una de las muestras simuladas bajo la contaminación C3. En la imagen de la izquierda graficamos las $y's$ en función de las dos reducciones estimadas bajo máxima verosimilitud (es decir, los puntos $(R_1(\mathbf{x}_i), R_2(\mathbf{x}_i), y_i)$, $1 \leq i \leq n = 200$), y en la derecha, lo mismo para la estimación dada por los $\tau-$estimadores propuestos.

En ambos scatter plots vemos que, si bien los subespacios identificados son diferentes, ambos métodos apartan las observaciones contaminadas del resto. Para ver la bondad del ajuste obtenido, eliminamos dichas observaciones atípicas de ambos conjuntos de datos, y corremos un estimador no paramétrico de la regresión, es decir, de la esperanza condicional $E(y \mid (R_1(\mathbf{x}), R_2(\mathbf{x})))$ con la función `sm.regression` del paquete `sm` Bowman y Azzalini [2014] de R. En la Figura 11(a) puede verse el mismo conjunto de datos de la Figura 10 sin los outliers, con la función de regresión estimada por esta rutina no paramétrica. En el gráfico de la izquierda, la reducción utilizada es la de máxima verosimilitud. A la derecha, tenemos el mismo gráfico, esta vez para la reducción estimada vía el $\tau-$estimador. Para ambas situaciones también reportamos el estimador del desvío estándar del error que se obtiene junto con el ajuste no paramétrico. En este caso podemos

Figura 10: Scatter plot de las observaciones de una muestra generada bajo la contaminación C3. A la izquierda, las y's están ploteadas en función de las proyecciones de las $\mathbf{x}$ sobre el subespacio de reducción estimado por máxima verosimilitud. A la derecha, las mismas observaciones y están graficadas en función de su proyección sobre el subespacio de reducción estimado por el $\tau-$estimador.



usar un estimador clásico (no necesariamente robusto) ya que en el proceso de reducir la dimensión de los datos, los outliers fueron identificados y eliminados.

Para comparar estos ajustes, estimamos el desvío estándar asociado al error, estimado por la rutina `sm.regression`. En el ejemplo graficado en la Figura 11 (b) obtenemos que $\widehat{\sigma}_{MV} = 0{,}102$ y $\widehat{\sigma}_{TAU} = 0{,}045$. Para comparar ambos ajustes, en la Figura 11 exhibimos un boxplot de los desvíos estándares estimados para cada conjunto de datos simulado bajo la contaminación C3 (siempre tomando en cuenta las 180 observaciones no atípicas).

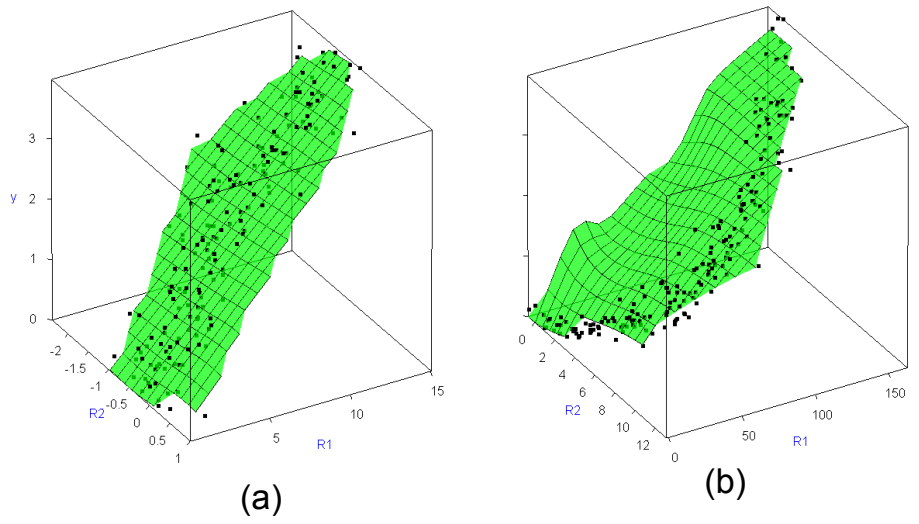
En dicho gráfico vemos que los desvíos estándares de los errores estimados por la regresión no paramétrica dan todos más chicos si el subespacio de reducción de la dimensión se estima a través de los $\tau-$estimadores, que si lo estimamos por máxima verosimilitud.

Segunda simulación El modelo simulado es

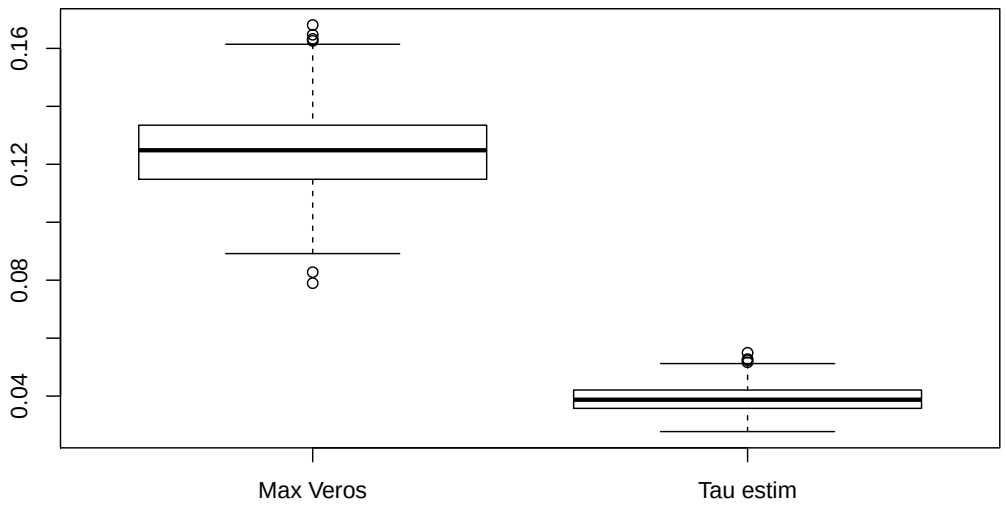
$$\mathbf{x} = \boldsymbol{\mu}_0 + \Gamma_0 \beta_0 \mathbf{f}_0 (y) + \Delta_0^{1/2} \mathbf{u}$$

Figura 11: Resúmenes gráficos del ajuste no paramétrico a una muestra.

(a) Scatter plot de las observaciones de una muestra generada bajo la contaminación C3, excluyendo los outliers. Las superficies fueron ajustadas usando la rutina de regresión no paramétrica sm.regression del paquete sm del R. La figura (a), a la izquierda, muestra las $y's$ en función de las proyecciones de las x sobre el subespacio de reducción estimado por máxima verosimilitud. A la derecha, en la figura (b) vemos las mismas observaciones y graficadas en función de su proyección sobre el subespacio de reducción estimado por el $\tau-$estimador.



(b) Boxplot de los desvíos estándares del error estimados por la rutina de regresión no paramétrica `sm.regression` del paquete `sm` del R, para las 500 muestras generadas bajo la contaminación C3, excluyendo los outliers, para Máxima verosimilitud y $\tau-$estimadores.



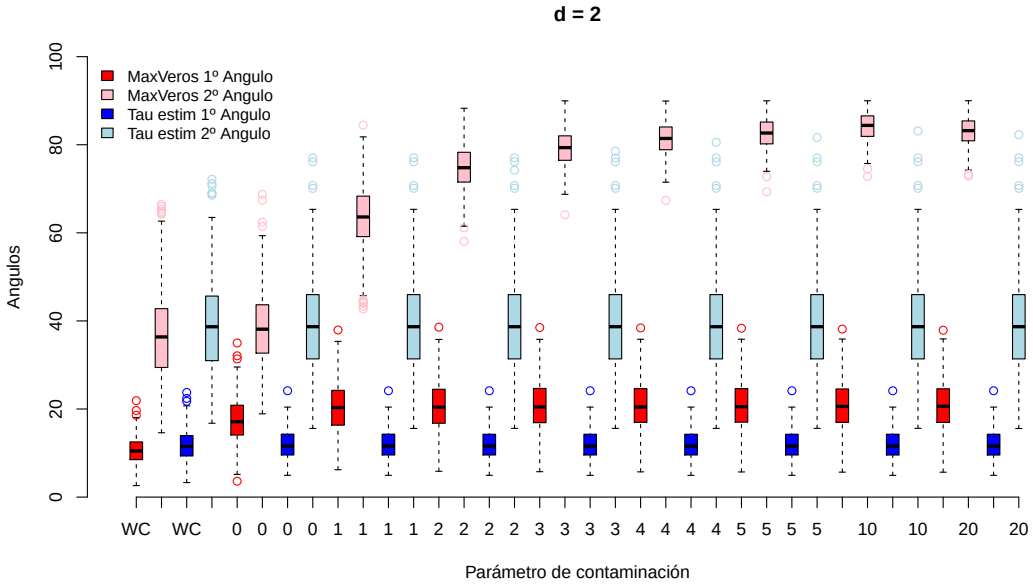
donde $\mathbf{x} \in \mathbb{R}^p$, con $p = 10$, $d = 2$, $r = 2$, $\boldsymbol{\mu}_0 = \mathbf{0}$, $\Gamma_0 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \cdots & 0 & 0 \end{bmatrix}^T \in \mathbb{R}^{p \times d}$, $\beta_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{d \times r}$, $\mathbf{f}(y) = \left(y, y^2\right)^T \in \mathbb{R}^r$, $\mathbf{u} \sim N_p(\mathbf{0}, I_p)$, $y \sim U(0, 4)$. Simulamos bajo dos escenarios: (a) $\Delta_0 = I_p$, y (b) $\Delta_0 = 0{,}5 I_p$.

Como en el caso simulado para $d = 1$, para evaluar el desempeño del estimador bajo contaminación de los datos, tomamos un 10 % de las observaciones $(\mathbf{x}, \mathbf{f}(y))$ de la forma $(\mathbf{x}^*, \mathbf{f}^*(y))$, donde $\mathbf{f}^*(y) = \mathbf{f}(10) = \left(10, 10^2\right)^T$ y $\mathbf{x}^* = \lambda (0, 0, \ldots, 0, 1)^T$ con $\lambda = 0, 1, 2, 3, 4, 5, 10, 15, 20$. En las Figuras 12a y 12b mostramos los boxplots de los ángulos que se obtienen entre el verdadero subespacio de reducción $\left(\Delta^{-1}\mathcal{S}_\Gamma\right)$ y el subespacio de reducción estimado $\left(\widehat{\Delta}^{-1}\mathcal{S}_{\widehat{\Gamma}}\right)$ por máxima verosimilitud (en rojo y rosa) y por el $\tau-$estimador que proponemos (en azul y celeste), con las matrices de covarianza del error descripta en (a) y (b) respectivamente, bajo los distintos escenarios: sin contaminación (SC) y para los distintos valores de λ considerados, a partir de $B = 500$ muestras de tamaño $n = 200$, con 10 % de datos contaminados. Las constantes son las descriptas en la Sección 9.2.3.

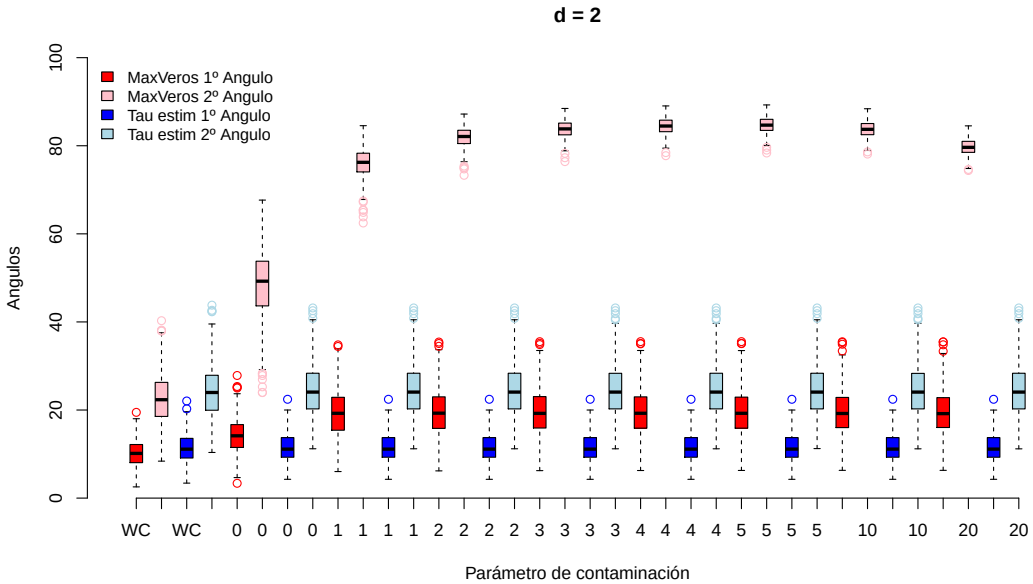
Tanto en la Figura 12a como en la 12b puede apreciarse el mismo fenómeno que en la 9: cuando no hay contaminación ambos métodos encuentran el subespacio correctamente. Sin embargo cuando contaminamos los datos, el método de máxima verosimilitud falla en detectar el subespacio, mientras que el $\tau-$estimador lo encuentra. Además, cuánto menor es la varianza del error, mejor funciona el $\tau-$estimador, lo cual era esperable, aunque no pasa lo mismo con el EMV.

Figura 12: Boxplots del primer y segundo ángulo principal entre el subespacio verdadero y el subespacio estimado por máxima verosimilitud (en rojo y rosa, respectivamente) y usando el $\tau-$estimador propuesto (en azul y celeste, respectivamente) para la simulación descripta en la Sección 9.4.2. Se generaron $n = 200$ datos sin contaminación. Luego, se reemplazaron 20 observaciones de cada muestra por otras 20 contaminadas según lo descripto en dicha sección, para distintos valores del parámetro λ.

(a) Matriz de covarianza del error $\Delta_0 = I_p$



(b) Matriz de covarianza del error $\Delta_0 = 0{,}5I_p$



9.5. Selección de d

9.5.1. Propuestas existentes

Hasta aquí, asumimos que d, la dimensión del subespacio de reducción suficiente, era conocida de antemano. En las aplicaciones prácticas, sin embargo, esta dimensión debe ser estimada a partir de los datos. Cook y Forzani [2008] proponen dos formas distintas de elegir la dimensión, usando el estimador de máxima verosimilitud bajo normalidad del modelo PFC. La primera, basada en el test del cociente de verosimilitud, para testear las hipótesis

$$H_0 : d = m$$
$$H_1 : d > m.$$

El estadístico correspondiente a este test es

$$\Lambda_m = 2\left[\ln L\left(\widehat{\boldsymbol{\theta}}_{\text{mín}(r,p)}\right) - \ln L\left(\widehat{\boldsymbol{\theta}}_m\right)\right],$$

donde el logaritmo de la función de verosimilitud L se puede encontrar en (18). Aquí $\widehat{\boldsymbol{\theta}}_m$ denota el estimador de máxima verosimilitud de los parámetros del modelo PFC que se calcula asumiendo que $d = m$. Bajo la hipótesis nula, Λ_m tiene distribución asintótica χ^2 con $(r-m)(p-m)$ grados de libertad. El test del cociente de verosimilitud (LRT; *likelihood ratio test*) se usa de forma secuencial en este contexto, comenzando con $m = 0$, se estima a d como el primer valor de dimensión hipotetizado que no sea rechazado.

El segundo enfoque que proponen Cook y Forzani [2008] es usar un criterio basado en la información, como el AIC o el BIC, nuevamente construido con los estimadores de máxima verosimilitud. Para $m \in \{0, ..., \text{mín}(r,p)\}$ se elige la dimensión d que minimice el

$$IC(m) = -2\ln L\left(\widehat{\boldsymbol{\theta}}_m\right) + h(n)g(m),$$

donde el logaritmo de la función de verosimilitud L fue definido en (18), $g(m)$ es el número de parámetros a ser estimados, en este caso $p(p+3)/2 + rm + m(p-m)$ y $h(n)$ depende de si se calcula el BIC (*Bayesian Information Criterion*) en cuyo caso se toma $h(n) = \log(n)$ o el AIC (*Akaike Information Criterion*) en cuyo caso $h(n) = 2$. Estas versiones son adaptaciones de lo que ocurre frecuentemente con el comportamiento asintótico para otros modelos. En las simulaciones observan que LRT y AIC sobreestiman el valor de d, mientras que el BIC lo subestima. Los autores enfatizan que, sin embargo, las consecuencias de ambas fallas no son igualmente peligrosas: pequeños niveles de sobreestimación son razonables en el contexto de la búsqueda de la reducción suficiente, ya que una estimación más fina se conseguirá luego al estimar el modelo directo (no ya inverso) de regresión no paramétrica. En cambio, si la dimensión es subestimada, $\mathbf{R}$ ya no será una reducción suficiente. Bura y Cook [2003] proponen un test asintótico para el rango de la matriz de coeficientes en un contexto de regresión con matrices de rango reducido, que no requiere supuestos sobre la distribución de la respuesta. Extienden el test para abarcar también el caso de matriz de covarianza de los errores no constante.

Los procedimientos conducen a tests asintóticos χ^2 ó χ^2 pesados. También Bura y Yang [2011] proponen dos tests para seleccionar el rango de la matriz en el contexto más general de reducción de la dimensión, basados en autovalores de una matriz, siendo uno de ellos asintóticamente χ^2 pesado y el otro, un test asintótico de tipo Wald.

Nuestra propuesta consiste en realizar un procedimiento de validación cruzada para seleccionar a d.

9.5.2. Validación cruzada

La validación cruzada (*cross-validation*) es una técnica estadística que se usa tanto para evaluar el comportamiento de un estimador como para seleccionar un modelo entre varios posibles. Consiste en la división aleatoria de la muestra en dos partes: una parte de *entrenamiento* y otra de *validación*. la primera se utiliza para ajustar el modelo, la segunda parte para validarlo. Mosteller y Tukey [1968] proponen tomar la muestra de validación de tamaño uno, de modo que el procedimiento se repite n veces, una para cada observación como muestra de validación. Finalmente se resume lo obtenido en las distintas submuestras, usualmente se toma el promedio de los estimadores obtenidos (o algún otro resumen) como criterio de cross-validación. Nuestra propuesta para la selección de la dimensión d a partir de una muestra está basada en una validación cruzada con cinco cortes o subgrupos (*5-fold cross-validation*), como describimos a continuación. Complementamos la elección de la dimensión con la *one-standard-deviation rule*, (regla de un desvío estándar) según la propuesta dada por Hastie, Tibshirani, y Friedman [2009].

Cross-validation basado en el $\tau-$estimador

1. Dividimos a la muestra en cinco partes (o *folds*) de igual tamaño, elegidas de forma aleatoria, $\bigcup_{k=1}^{5} B_k = \{1, 2, \ldots, n\}$, $B_k \cap B_l = \emptyset$ para todo $1 \leq k \neq l \leq 5$, $\sharp(B_k) = \left[\frac{n}{5}\right]$. Elegimos un $d_{\text{máx}} \leq \text{mín}\{p, r\}$, $d_{\text{máx}}$ es el máximo valor de d que evaluaremos por validación cruzada.

2. Para cada fold o parte, $k = 1, \ldots, 5$,

   a) **(Entrenamiento)** Para cada $d = 0, \ldots, d_{\text{máx}}$, calculamos los $\tau-$estimadores para PFC con todas las observaciones que no pertenecen a B_k. Sea

   $$\begin{aligned} \widehat{\boldsymbol{\theta}}_k^{(d)} &:= \widehat{\boldsymbol{\theta}}\left(\{(\mathbf{x}_i, y_i) : i \in B_k^c\}\right) \\ &= \left(\widehat{\boldsymbol{\mu}}_k^{(d)}, \widehat{\Gamma}_k^{(d)}, \widehat{\beta}_k^{(d)}, \widehat{\Delta}_k^{(d)}\right). \end{aligned} \tag{101}$$

   basado en (entrenado con) $4n/5$ observaciones.

   b) **(Validación o testeo)** Para cada observación $(\mathbf{x}_j, y_j)$ tal que $j \in B_k$, es decir para cada observación del grupo de testeo, calculamos la distancia de Mahalanobis de dicha observación tomada con los estimadores calculados en

(101), sea

$$d_j^{(d)} := d_j\left(\mathbf{x}_j, y_j, \widehat{\boldsymbol{\theta}}_k^{(d)}\right)$$
$$= \left\{\left(\mathbf{x}_j - \widehat{\boldsymbol{\mu}}_k^{(d)} - \widehat{\Gamma}_k^{(d)}\widehat{\beta}_k^{(d)}\mathbf{f}(y_j)\right)^T \left(\widehat{\Delta}_k^{(d)}\right)^{-1}\left(\mathbf{x}_j - \widehat{\boldsymbol{\mu}}_k^{(d)} - \widehat{\Gamma}_k^{(d)}\widehat{\beta}_k^{(d)}\mathbf{f}(y_j)\right)\right\}^{1/2}.$$

Observar que, en realidad, $k = k(j)$. A partir de ellas, calculamos $S_k(d)$, el M-estimador de escala basado en las distancias $\left\{d_j^{(d)} : j \in B_k\right\}$, tal que

$$\frac{1}{\left(\frac{1}{5}n\right)}\sum_{j\in B_k}\rho_1\left(\frac{d_j^{(d)}}{S_k(d)}\right) = \kappa_1.$$

Para cada $j \in B_k$, sea

$$L_j^{(d)} = \left|\widehat{\Delta}_k^{(d)}\right|^{1/p} S_k^2(d)\,\rho_2\left(\frac{d_j^{(d)}}{S_k(d)}\right).$$

3. (**Función objetivo**) Para cada $d = 0, \ldots, d_{\text{máx}}$, Calculamos el valor de la función objetivo robusta, promediando los $L_j^{(d)}$,

$$CV_{\text{tau}}(d) = \frac{1}{n}\sum_{j=1}^{n} L_j^{(d)} \tag{102}$$
$$= \frac{1}{5}\sum_{k=1}^{5}\left|\widehat{\Delta}_k^{(d)}\right|^{1/p} S_k^2(d)\left(\frac{5}{n}\sum_{j\in B_k}\rho_2\left(\frac{d_j^{(d)}}{S_k(d)}\right)\right),$$

y su desvío estándar muestral como

$$SD_{\text{tau}}(d) = sd\left(L_1^{(d)}, \ldots, L_n^{(d)}\right)$$
$$= \sqrt{Var\left(L_1^{(d)}, \ldots, L_n^{(d)}\right)}.$$

Finalmente, estimamos la desviación estándar de $CV_{\text{tau}}(d)$, que llamaremos el *error estándar de* $CV_{\text{tau}}(d)$, por

$$SE_{\text{tau}}(d) = \frac{SD_{\text{tau}}(d)}{\sqrt{n}}. \tag{103}$$

4. El criterio que suele utilizarse para elegir el valor de d por validación cruzada sería tomar el

$$\widehat{d} = \arg\min_{0\le d\le d_{\text{máx}}} CV_{\text{tau}}(d).$$

En la Figura 13, en el gráfico de la derecha vemos las curvas de cross validación para un ejemplo simulado, con $d = 4$ y $r = 5$ (detalles descriptos en la Sección 9.6.1, simulación sin contaminar). En dicho gráfico observamos que la curva es bastante plana para valores de d cerca del mínimo. Las barras verticales representan los errores estándares calculados por cross validación, como se describe en el punto 3. Teniendo esto en cuenta es que proponemos utilizar la regla de "un desvío estándar": elegir el modelo más parsimonioso cuya función de performance diste menos de un error estándar de la mínima. Esta regla, sugerida por Hastie et al. [2009] (Sección 7.10) toma en cuenta que la curva de CV_{tau} está estimada con error, y se basa en el hecho de que cuánto menor sea la cantidad de parámetros a estimar con los mismos datos, mejor será la calidad de la estimación obtenida (i.e. será menor la varianza de los mismos). El criterio que utilizaremos para elegir el d siguiendo *la regla de un desvío estándar* será

$$\widehat{d}_{cross} = \text{mín}\left\{d = 0, \ldots, d_{\text{máx}} : CV_{\text{tau}}\left(d\right) < CV_{\text{tau}}\left(\widehat{d}\right) + SE_{\text{tau}}\left(\widehat{d}\right)\right\}.$$

Para poder comparar la performance de este criterio robusto de selección de la dimensión que proponemos, aplicamos el mismo criterio de selección de la dimensión pero basado en el método de máxima verosimilitud. A continuación describimos los pasos en los que consiste la propuesta.

Cross-validation basado en el estimador de máxima verosimilitud El mecanismo es similar al descripto en la Sección para $\tau-$estimadores. Las diferencias son las siguientes.

2. Para cada fold o parte, $k = 1, \ldots, 5$,

   a) **(Entrenamiento)** En (101), calculamos los $\widehat{\boldsymbol{\theta}}_k^{(d)}$ usando los estimadores de máxima verosimilitud para PFC descriptos en el Capítulo 4, en vez de los $\tau-$estimadores.

   b) **(Validación o testeo)** Las distancias las calculamos usando los EMV $\widehat{\boldsymbol{\theta}}_k^{(d)}$, y luego definimos, para cada $j \in B_k$,

$$ML_j^{(d)} = \ln\left|\widehat{\Delta}_k^{(d)}\right| + d^2\left(\mathbf{x}_j, y_j, \widehat{\boldsymbol{\theta}}_k^{(d)}\right).$$

3. (**Función objetivo**) Para cada $d = 0, \ldots, d_{\text{máx}}$, Calculamos el valor de la función objetivo de máxima verosimilitud, promediando los $ML_j^{(d)}$,

$$\begin{aligned} CV_{\text{MV}}\left(d\right) &= \frac{1}{n}\sum_{j=1}^{n} ML_j^{(d)} \\ &= \frac{1}{5}\sum_{k=1}^{5} \ln\left|\widehat{\Delta}_k^{(d)}\right| + \frac{1}{n}\sum_{j=1}^{n} d^2\left(\mathbf{x}_j, y_j, \widehat{\boldsymbol{\theta}}_k^{(d)}\right), \end{aligned} \tag{104}$$

y su desvío estándar muestral como

$$SD_{\mathrm{MV}}(d) = sd\left(ML_1^{(d)}, \ldots, ML_n^{(d)}\right) = \sqrt{Var\left(ML_1^{(d)}, \ldots, ML_n^{(d)}\right)}.$$

Finalmente, estimamos la desviación estándar de $CV_{\mathrm{MV}}(d)$, que llamaremos el *error estándar de* $CV_{\mathrm{MV}}(d)$, por

$$SE_{\mathrm{MV}}(d) = \frac{SD_{\mathrm{MV}}(d)}{\sqrt{n}}. \tag{105}$$

El criterio que utilizaremos para elegir el d siguiendo *la regla de un desvío estándar* será

$$\widehat{d}_{cross}^{(ML)} = \mín\left\{d = 0, \ldots, d_{\mathrm{máx}} : CV_{\mathrm{MV}}(d) < CV_{\mathrm{MV}}\left(\widehat{d}^{(ML)}\right) + SE_{\mathrm{MV}}\left(\widehat{d}^{(ML)}\right)\right\},$$

siendo

$$\widehat{d}^{(ML)} = \arg \min_{0 \le d \le d_{\mathrm{máx}}} CV_{\mathrm{MV}}(d).$$

En la Figura 13 vemos las curvas de cross-validación para máxima verosimilitud a la izquierda y la correspondiente a $\tau-$estimadores a la derecha, para la misma muestra considerada simulada, con $d = 4$ y $r = 5$ (detalles descriptos en la Sección 9.6.1, simulación sin contaminar). Las barras verticales tienen la amplitud de un error estándar a cada lado, y con un punto más grande indicamos la dimensión seleccionada. La muestra sigue el modelo PFC y no contiene outliers.

9.6. Simulaciones de Validación Cruzada para seleccionar a d

9.6.1. Caso $d = 2$, $r = 2$

Simulamos según lo descripto en la Sección 9.4.2 los siguientes cinco escenarios. Para cada caso, 500 replicaciones. El tamaño de muestra es siempre $n = 200$.

Sin contaminar El modelo simulado es

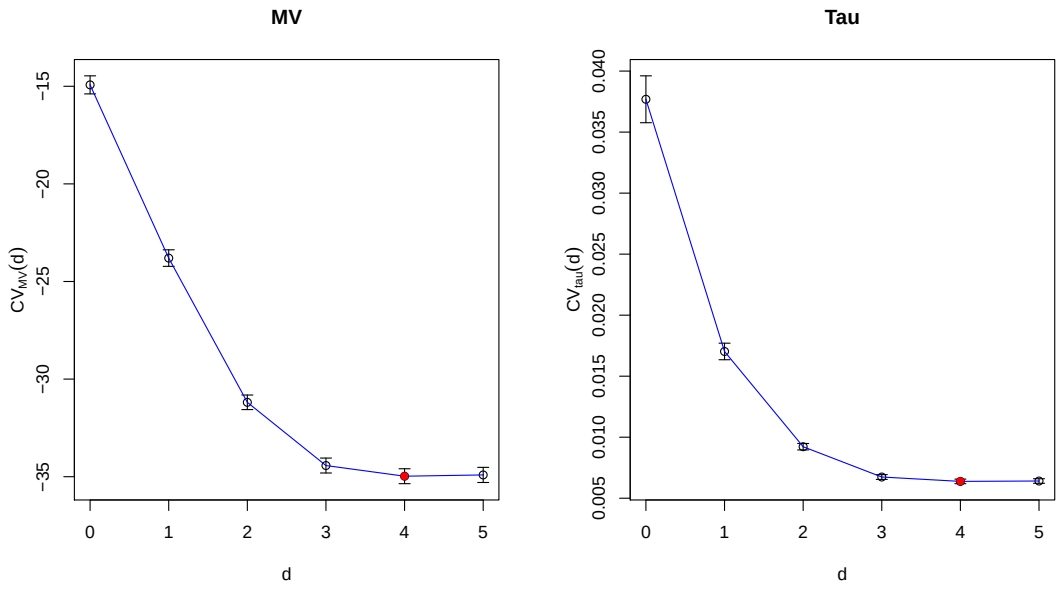
$$\mathbf{x} = \boldsymbol{\mu}_0 + \Gamma_0 \beta_0 \mathbf{f}_0(y) + \Delta_0^{1/2} \mathbf{u}$$

donde $\mathbf{x} \in \mathbb{R}^p$, con $p = 10$, $d = 2$, $r = 2$, $\boldsymbol{\mu}_0 = \mathbf{0}$, Γ_0 es una matriz de $p \times d$ de la siguiente forma

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \cdots & 0 & 0 \end{bmatrix}^T,$$

β_0 is I_2, $\mathbf{f}_0(y) = \left(y, y^2\right)^T$, $\mathbf{u} \sim N_p(\mathbf{0}, I_{10})$, $y \sim U(0,4)$, $\Delta_0 = 0{,}1 I_{10}$. El tamaño de muestra es $n = 200$.

Figura 13: Función objetivo de la validación cruzada dada en y (104) para máxima verosimilitud a la izquierda, y (102) para $\tau-$estimador de PFC a la derecha. Las barras verticales indican los límites dados por un error estándar, según (103) y (105), respectivamente. Los datos fueron generados con $d = 4$ y $r = 5$ (detalles descriptos en la Sección 9.6.1, simulación sin contaminar). Indicamos con un punto la dimensión elegida por el criterio de un desvío estándar.



Para evaluar el desempeño del estimador bajo contaminación de los datos, tomamos un 10 % de las observaciones contaminadas. Consideramos 4 escenarios de contaminación. Las observaciones contaminadas ($\mathbf{x}$,y) en el escenario de contaminación j siguen el siguiente modelo

$$\mathbf{x} = \boldsymbol{\mu}_0 + \Gamma_{C_j}\beta_0\mathbf{f}_{C_j}(y) + \Delta_{C_j}^{1/2}\mathbf{u},$$

con $1 \leq j \leq 4$. Luego, además del escenario sin contaminar, agregamos las siguientes cuatro contaminaciones.

Contaminación 1 Tomamos $\mathbf{f}_{C_1}(y) = (10,10)^T$, $\Delta_{C_1} = \Delta_0 = 0{,}1I_{10}, n = 200$ y

$$\Gamma_{C_1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T.$$

Observemos que en este escenario, el subespacio de reducción de las observaciones no contaminadas (span $\left(\Delta_0^{-1}\Gamma_0\right)$) y el de las contaminadas (span $\left(\Delta_0^{-1}\Gamma_{C_1}\right)$) forman los siguientes ángulos principales: 0^o y 90^o, es decir, uno de los generadores de ambos subespacios coincide, el otro está a 90^o.

Contaminación 2 Tomamos $\mathbf{f}_{C_2}(y) = \mathbf{f}_0(10) = \left(10, 10^2\right)^T$, $\Delta_{C_2} = \Delta_0 = 0{,}1 \cdot I_p, n =$ 200 y $\Gamma_{C_2} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T$.

Contaminación 3 Tomamos $\mathbf{f}_{C_3}(y) = \mathbf{f}_0(10) = \left(10, 10^2\right)^T$, $\Delta_{C_3} = 0{,}5 \cdot I_p, n = 200$ y $\Gamma_{C_3} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T$.

Tomamos $\mathbf{f}_{C_4}(y) = \mathbf{f}_0(10) = \left(10, 10^2\right)^T$, $\Delta_{C_4} = 0{,}5 \cdot I_p, n = 200$ y

$$\Gamma_{C_4} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix}^T$$

En este caso, los ángulos principales son ambos iguales a 90^o, es decir, están contaminadas ambas direcciones principales.

Las constantes $c_1, \kappa_1, c_2, \kappa_2$ las tomamos como en (100). Para cada uno de los cinco escenarios (sin contaminar, y C1 a C4) generamos $B = 500$ muestras de $n = 200$ observaciones cada una. Resumimos la performance de ambas propuestas de estimar la dimensión del subespacio de reducción por validación cruzada junto al criterio de un desvío estándar, de dos formas distintas. La primera es una tabla de valores de d estimados por validación cruzada utilizando máxima verosimilitud y $\tau-$estimadores, obtenidos en las 500 muestras simuladas bajo cada escenario. La segunda es una gráfica de las funciones objetivo de la validación cruzada dadas en (102) y (104), una curva para cada muestra, donde indicamos con un punto la dimensión elegida por el criterio de un desvío estándar descripto en la Sección 9.5.2. En cada caso hacemos un gráfico con sólo 20 curvas, en vez de las 200 calculadas en cada situación, ya que al ser unas pocas, pueda apreciarse

la forma de cada una. En cada caso, elegimos las primeras 20 curvas simuladas. A la izquierda siempre el gráfico correspondiente a máxima verosimilitud, a la derecha el de los $\tau-$estimadores.

Tabla 1: Porcentajes de los valores de d seleccionados por validación cruzada con máxima verosimilitud (a la izquierda) y con el $\tau-$estimador para el modelo PFC (a la derecha), en 500 muestras de tamaño 200 cada una, bajo los escenarios previamente descriptos, todos generados con $d = 2$, $r = 2$.

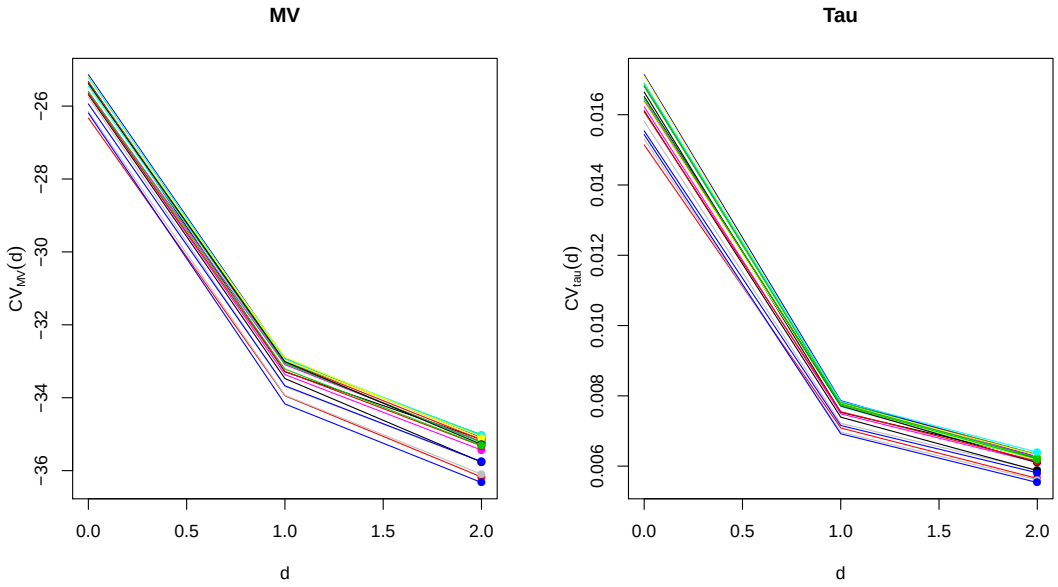
	d seleccionado por MV			d seleccionado por $\tau-$est		
Escenario	0	1	2	0	1	2
SC	0	0	100 %	0	0	100 %
C1	0.2 %	64.4 %	35.4 %	0	0	100 %
C2	0.6 %	68.8 %	30.6 %	0	0	100 %
C3	68.6 %	31.4 %	0	0	0	100 %
C4	41.8 %	58.2 %	0	0	0	100 %

Podemos ver que para este caso, ambos procedimientos de cross-validation eligen bien la dimensión del subespacio de reducción de dimensión cuando no hay contaminación. Sin embargo, con la aparición de contaminación, el método de máxima verosimilitud es sensible a su presencia y ya no selecciona la dimensión de manera apropiada. En cambio, la cross-validación basada en $\tau-$estimadores resiste la presencia de outliers y consigue estimar bien la dimensión.

Para completar el estudio, también aplicamos sobre las mismas muestras los tres métodos existentes para seleccionar la dimensión de la reducción, introducidas en la Sección 9.5.1. Fueron corridas con el paquete `LDR` Cook, Forzani, y Tomassi [2011] programado en `MATLAB`. Para todas las muestras generadas, tanto las contaminadas como las sin contaminar, los tres métodos eligieron $d = 2$ correctamente. Una contra que tiene este paquete es que sólo prueba valores de $d \geq 1$. En estas simulaciones con $d = r = 2$ esto quiere decir que sólo prueba $d = 1$ y $d = 2$. Cook y Forzani [2008] remarcan que además estos métodos tienden a sobreestimar el valor de d, por lo cual no parece tan sorprendente el resultado obtenido.

Figura 14: Funciones objetivo de la validación cruzada dadas en (102) y (104), una curva para cada muestra simulada en los distintos escenarios, con $d = 2, r = 2$. Indicamos con un punto la dimensión elegida por el criterio de un desvío estándar. A la izquierda el gráfico correspondiente a máxima verosimilitud, a la derecha el de los $\tau-$estimadores.

(a) Sin Contaminación



(b) Contaminación 1

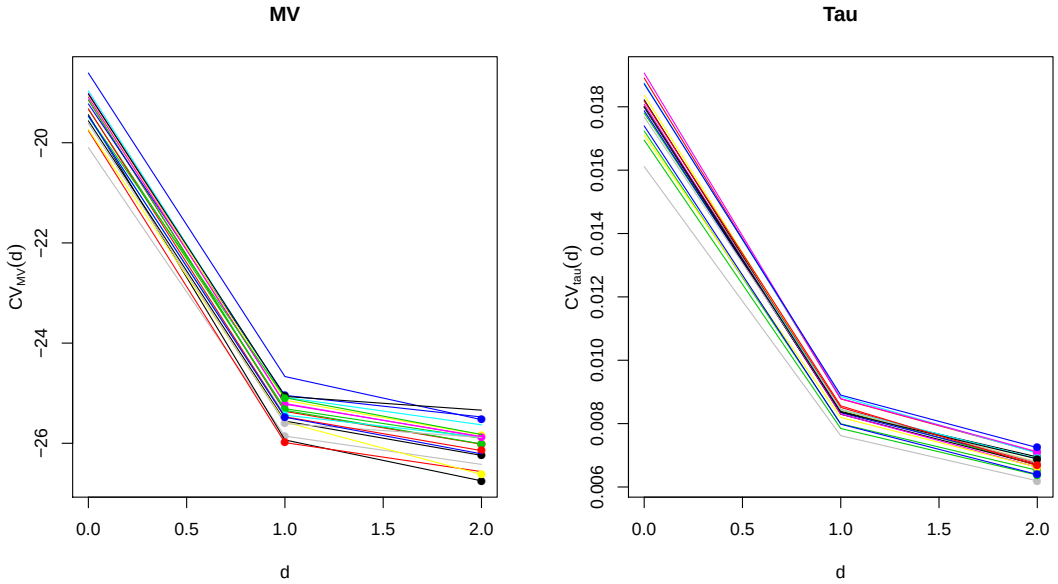
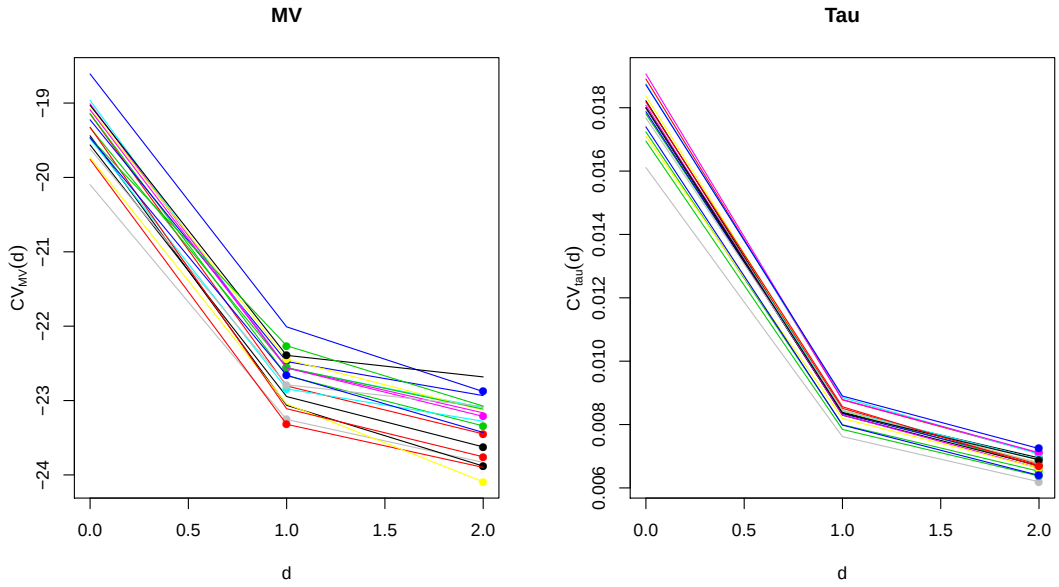


Figura 15: Contaminaciones 3 y 4.

(c) Contaminación 2



(d) Contaminación 3

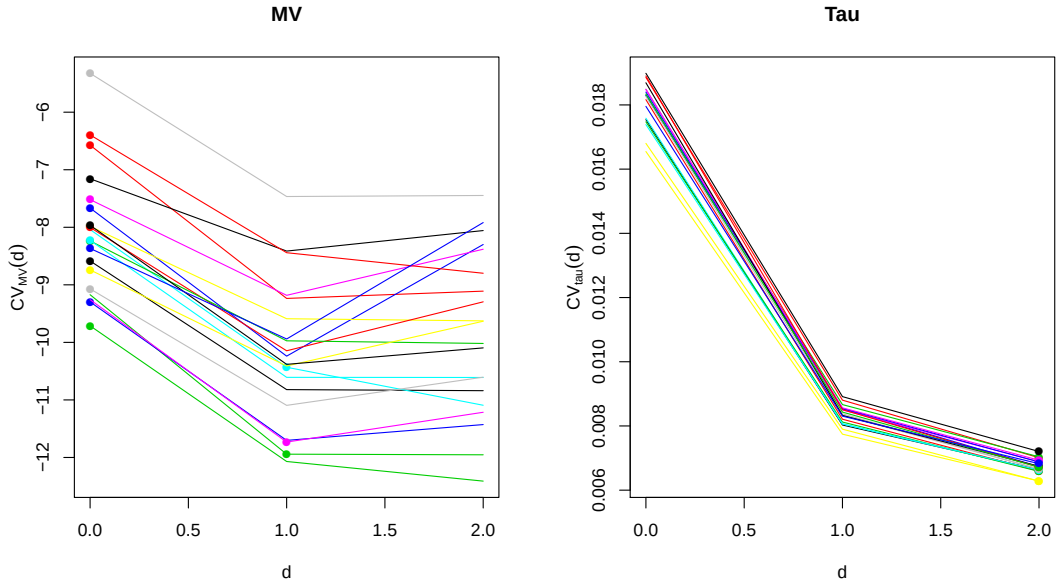
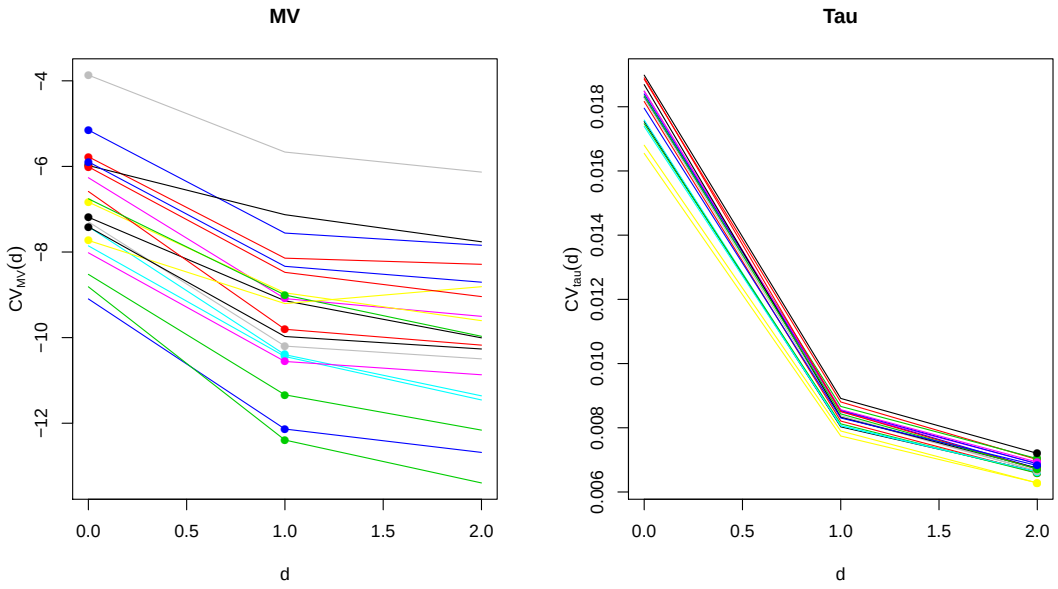


Figura 15: Contaminación 5.

(e) Contaminación 4



9.6.2. Caso $d = 2$, $r = 5$

Simulamos los siguientes cinco escenarios, que son similares a los cinco escenarios del caso anterior.

Sin contaminar El modelo simulado es

$$\mathbf{x} = \boldsymbol{\mu}_0 + \Gamma_0 \beta_0 \mathbf{f}_0(y) + \Delta_0^{1/2} \mathbf{u}$$

donde $\mathbf{x} \in \mathbb{R}^p$, con $p = 10, d = 2, r = 5, \boldsymbol{\mu}_0 = \mathbf{0}, \Gamma_0 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \cdots & 0 & 0 \end{bmatrix}^T \in$

$\mathbb{R}^{p\times d}$, $\beta_0 = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix} \in \mathbb{R}^{d\times r}$, $\mathbf{f}_0(y) = \begin{bmatrix} \left(\frac{y-2}{2/\sqrt{3}}\right) \\ \left(\frac{y-2}{2/\sqrt{3}}\right)^2 \\ \left(\frac{y-2}{2/\sqrt{3}}\right)^3 \\ \left(\frac{y-2}{2/\sqrt{3}}\right)^4 \\ \left(\frac{y-2}{2/\sqrt{3}}\right)^5 \end{bmatrix} \in \mathbb{R}^r$, $\mathbf{u} \sim N_p(\mathbf{0}, I_p)$,

$y \sim U(0,4)$, $\Delta_0 = 0{,}1 \cdot I_p$.

Observemos que la función $\mathbf{f}_0(y)$ corresponde a los polinomios mónicos de grado 1 a 5, para la y estandarizada, ya que si $y \sim U(0,4)$ entonces, $E(y) = 2$ y $Var(y) = 4/3$. Para evaluar el desempeño del estimador bajo contaminación de los datos, tomamos un 10 % de las observaciones contaminadas. Las observaciones contaminadas ($\mathbf{x}$,y) siguen el siguiente modelo

$$\mathbf{x} = \boldsymbol{\mu}_0 + \Gamma_{C_j} \beta_0 \mathbf{f}_{C_j}(y) + \Delta_{C_j}^{1/2} \mathbf{u},$$

con $1 \leq j \leq 4$. Luego, además del escenario sin contaminar, agregamos las siguientes cuatro contaminaciones.

Contaminación 1 Tomamos $\mathbf{f}_{C_1}(y) = (10, 10, 10, 10, 10)^T$, $\Delta_{C_1} = \Delta_0 = 0{,}1 \cdot I_p$ y

$$\Gamma_{C_1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T.$$

Nuevamente, $(\text{span}\left(\Delta_0^{-1}\Gamma_0\right))$ y $\left(\text{span}\left(\Delta_0^{-1}\Gamma_{C_1}\right)\right)$ forman los siguientes ángulos principales: 0^o y 90^o, es decir, uno de los generadores de ambos subespacios coincide, el otro está a 90^o.

Contaminación 2 Tomamos $\mathbf{f}_{C_2}(y) = \mathbf{f}_0(10) = \left[6.928\,2 \quad 48{,}0 \quad 332.55 \quad 2304{,}0 \quad 15963\right]^T$,

$\Delta_{C_2} = \Delta_0 = 0{,}1 \cdot I_p$ y $\Gamma_{C_2} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T$.

Contaminación 3 Tomamos $\mathbf{f}_{C_3}(y) = \mathbf{f}_0(10)$, $\Delta_{C_3} = 0{,}5 \cdot I_p$ y $\Gamma_{C_3} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T$.

Tabla 2: Porcentaje de veces en el que el valor d fue seleccionado por validación cruzada con máxima verosimilitud (a la izquierda) y con el $\tau-$estimador para el modelo PFC (a la derecha), en 500 muestras de tamaño 200 cada una, bajo los escenarios previamente descriptos, todos generados con $d = 2$, $r = 5$.

	d seleccionado por MV				d seleccionado por $\tau-$est			
Escenario	0	1	2	3	0	1	2	3
SC	0	0	100 %	0			100 %	
C1	38.4 %	17.8 %	43.4 %	0.4 %			100 %	
C2	9.8 %	48 %	41.8 %	0.4 %			100 %	
C3	21.6 %	75 %	3.4 %	0			100 %	
C4	18.8 %	27 %	54.2 %	0			100 %	

Contaminación 4 Tomamos $\mathbf{f}_{C_4}(y) = \mathbf{f}_0(10)$, $\Delta_{C_4} = 0{,}5 \cdot I_p$ y $\Gamma_{C_4} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix}^T$.

Ahora, los ángulos principales son ambos iguales a 90^o, es decir, están contaminadas ambas direcciones principales.

Para cada uno de los cinco escenarios (sin contaminar, y C1 a C4) generamos $B = 500$ muestras de $n = 200$ observaciones cada una. Resumimos la performance de ambas propuestas de estimar la dimensión del subespacio de reducción de la misma forma que describimos en la Sección 9.6.1.

Nuevamente, en este caso, la validación cruzada tanto por máxima verosimilitud como con $\tau-$estimadores funciona bien para la selección de d cuando no hay contaminación. Sin embargo, con la aparición de contaminación, el método de máxima verosimilitud es sensible a su presencia y ya no selecciona la dimensión de manera apropiada. En cambio, la cross-validación basada en $\tau-$estimadores resiste la presencia de outliers y consigue estimar bien la dimensión. También repetimos esta simulación completa (sin contaminar y con las cuatro contaminaciones) tomando a $\mathbf{f}_0(y) = \left(y, y^2, y^3, y^4, y^5\right)^T$ y los resultados de la validación cruzada tanto bajo MV como usando el $\tau-$estimador fueron similares a los obtenidos en este caso, por eso no los reportamos.

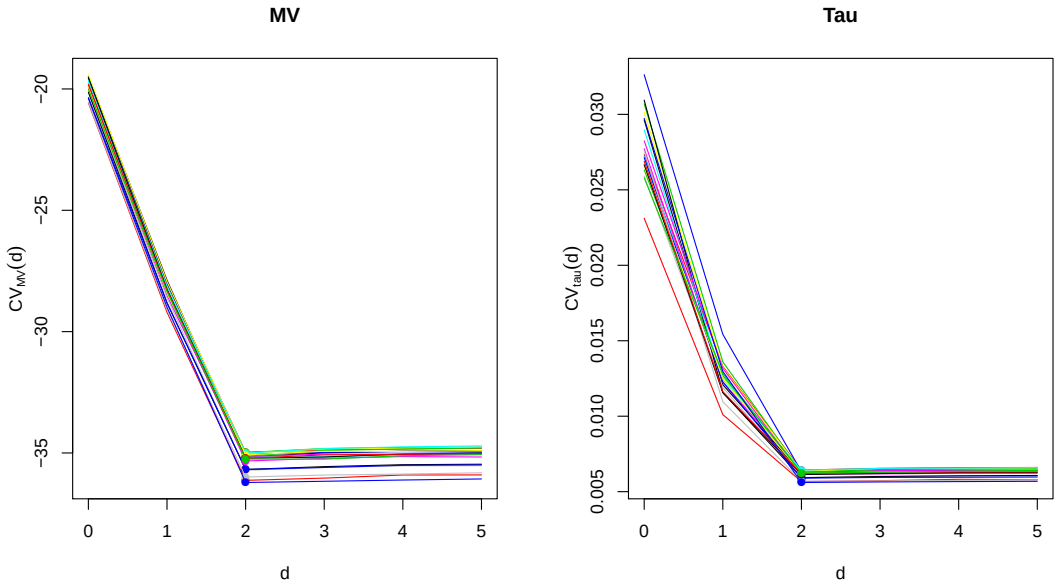
Para permitir la comparación con métodos publicados, también aplicamos para estas mismas muestras los tres métodos existentes para seleccionar la dimensión de la reducción, introducidas en la Sección 9.5.1, nuevamente con el paquete `LDR` Cook et al. [2011] programado en `MATLAB`. Recordemos que estas propuestas sólo prueban valores de $d \geq 1$. Los resultados obtenidos figuran en la Tabla 3. En ella puede verse que cuando no hay contaminación, los tres métodos seleccionan bien la dimensión d. Sin embargo, en presencia de la contaminación los tres métodos eligen $d = 3$ mayoritariamente, sobreestimando el valor correcto.

Tabla 3: Porcentaje de veces en el que el valor d fue seleccionado por AIC, BIC y LRT, en las 500 muestras correspondientes a la Tabla 2 tamaño 200 cada una, bajo los escenarios previamente descriptos, todos generados con $d = 2$, $r = 5$.

		Porcentaje de veces que es elegido		
Escenario	Valor de d	AIC	BIC	LRT
SC	1	0	0	0
	2	88.4 %	100 %	95.2 %
	3	11.4 %	0	4.4 %
	4	0.2 %	0	0.4 %
C1	1	0	0	0
	2	0	0.8 %	0
	3	88.2 %	99.2 %	94 %
	4	11.6 %	0	5.8 %
	5	0.2 %	0	0.2 %
C2	1	0	0	0
	2	0	0.8 %	0
	3	89.4 %	99.2 %	93.4 %
	4	10.6 %	0	6.6 %
C3	1	0	0	0
	2	0	0	0
	3	100 %	100 %	100 %
	4	0	0	0
C4	1	0	0	0
	2	0	0	0
	3	100 %	100 %	100 %
	4	0	0	0

Figura 15: Funciones objetivo de la validación cruzada dadas en (102) y (104), una curva para cada muestra simulada en los distintos escenarios, con $d = 2, r = 5$. Indicamos con un punto la dimensión elegida por el criterio de un desvío estándar. A la izquierda el gráfico correspondiente a máxima verosimilitud, a la derecha el de los $\tau-$estimadores.

(a) Sin Contaminación



(b) Contaminación 1

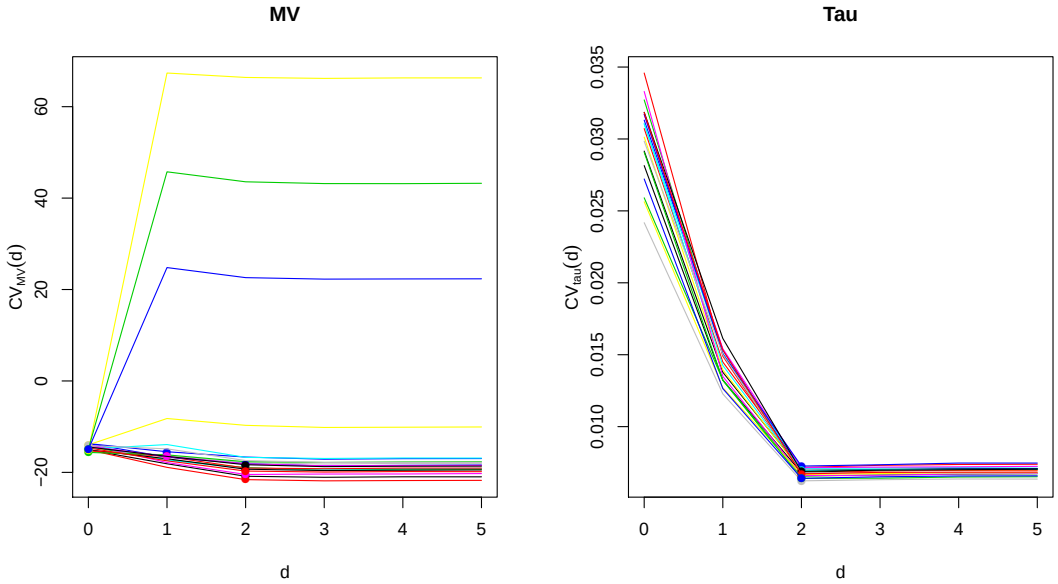
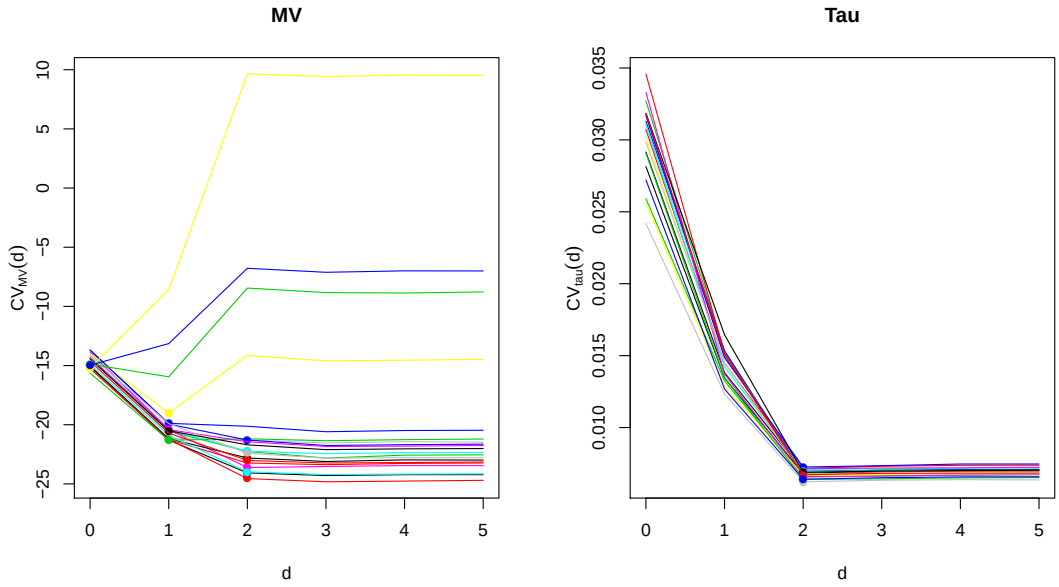


Figura 16: Contaminaciones 2 y 3.

(c) Contaminación 2



(d) Contaminación 3

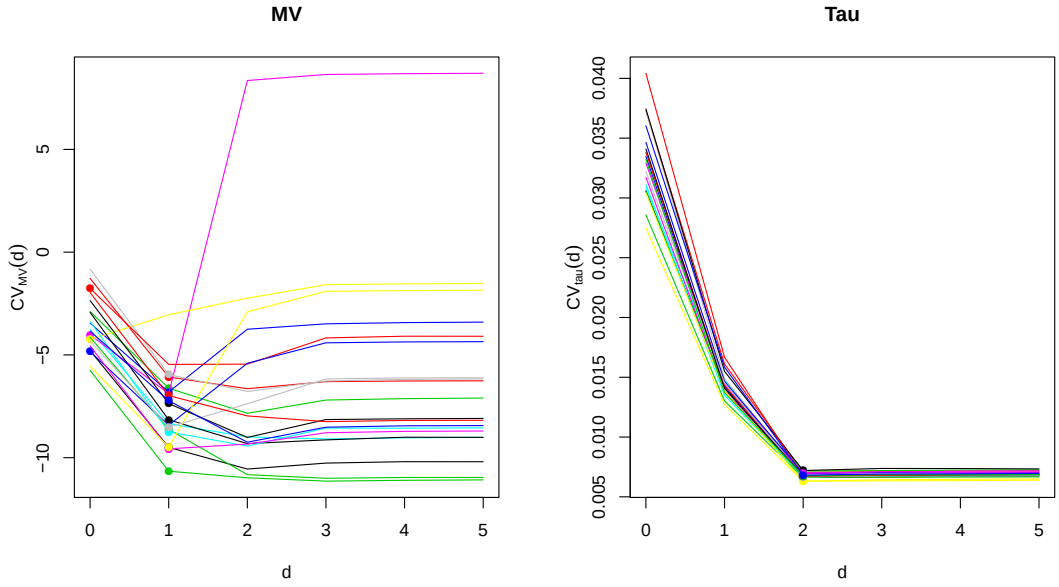
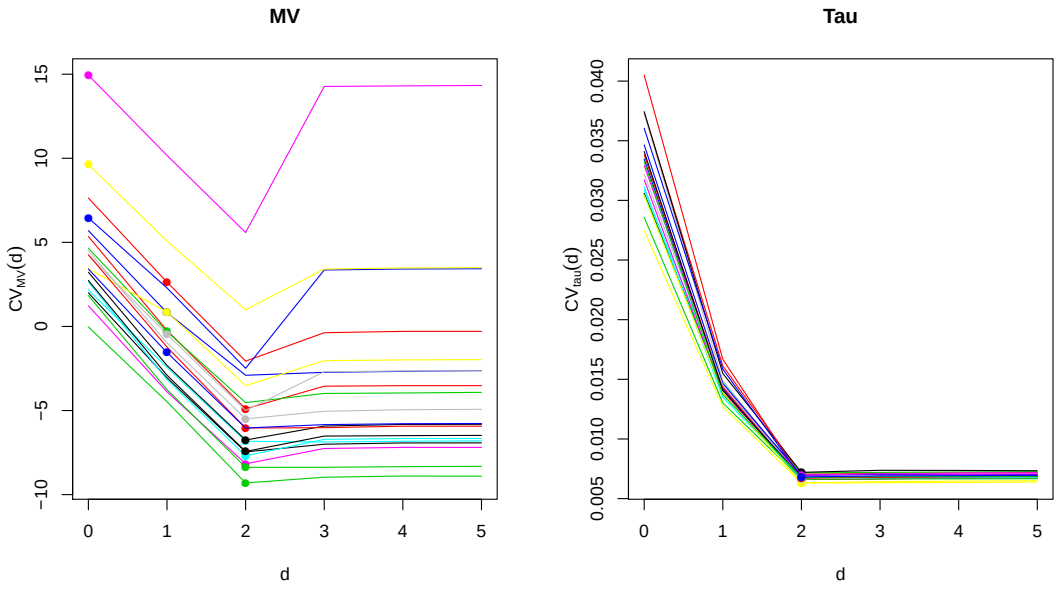


Figura 16: Contaminación 4.

(e) Contaminación 4



9.6.3. Caso $d = 4,\ r = 5$

Simulamos los siguientes cinco escenarios.

Sin contaminar El modelo simulado es

$$\mathbf{x} = \boldsymbol{\mu}_0 + \Gamma_0 \beta_0 \mathbf{f}_0(y) + \Delta_0^{1/2}\mathbf{u}$$

donde $\mathbf{x} \in \mathbb{R}^p$, con $p = 10, d = 4, r = 5$, $\boldsymbol{\mu}_0 = \mathbf{0}$, $\Gamma_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 & \cdots & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 & 0 \end{bmatrix}^T \in$

$\mathbb{R}^{p\times d}$, $\beta_0 = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \in \mathbb{R}^{d\times r}$, $\mathbf{f}_0(y) = \begin{bmatrix} \left(\frac{y-2}{2/\sqrt{3}}\right) \\ \left(\frac{y-2}{2/\sqrt{3}}\right)^2 \\ \left(\frac{y-2}{2/\sqrt{3}}\right)^3 \\ \left(\frac{y-2}{2/\sqrt{3}}\right)^4 \\ \left(\frac{y-2}{2/\sqrt{3}}\right)^5 \end{bmatrix} \in \mathbb{R}^r$, $\mathbf{u} \sim N_p(\mathbf{0}, I_p)$,

$y \sim U(0,4)$, $\Delta_0 = 0{,}1 \cdot I_p$.

Para evaluar el desempeño del estimador bajo contaminación de los datos, tomamos un 10 % de las observaciones contaminadas. Las observaciones contaminadas ($\mathbf{x}$,y) siguen el siguiente modelo

$$\mathbf{x} = \boldsymbol{\mu}_0 + \Gamma_{C_j} \beta_0 \mathbf{f}_{C_j}(y) + \Delta_{C_j}^{1/2}\mathbf{u},$$

con $1 \leq j \leq 4$. Luego, además del escenario sin contaminar, agregamos las siguientes cuatro contaminaciones.

Contaminación 1 Tomamos $\mathbf{f}_{C_1}(y) = (10, 10, 10, 10, 10)^T$, $\Delta_{C_1} = \Delta_0 = 0{,}1 \cdot I_p$ y

$$\Gamma_{C_1} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T.$$

Nuevamente, $(\text{span}\left(\Delta_0^{-1}\Gamma_0\right))$ y $\left(\text{span}\left(\Delta_0^{-1}\Gamma_{C_1}\right)\right)$ forman los siguientes ángulos principales: $0^o, 0^o, 90^o$ y 90^o, es decir, dos de los generadores de ambos subespacios coinciden, los otros dos están a 90^o.

Contaminación 2 Tomamos $\mathbf{f}_{C_2}(y) = \mathbf{f}_0(10) = \begin{bmatrix} 6.928\,2 & 48{,}0 & 332.55 & 2304{,}0 & 15963 \end{bmatrix}^T$, $\Delta_{C_2} = \Delta_0 = 0{,}1 \cdot I_p$ y $\Gamma_{C_2} = \Gamma_{C_1}$.

Contaminación 3 Tomamos $\mathbf{f}_{C_3}(y) = \mathbf{f}_0(10)$, $\Delta_{C_3} = 0{,}5 \cdot I_p$ y $\Gamma_{C_3} = \Gamma_{C_1}$.

Contaminación 4 Tomamos $\mathbf{f}_{C_4}(y) = \mathbf{f}_0(10)$, $\Delta_{C_4} = 0{,}5 \cdot I_p$ y

$$\Gamma_{C_4} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 1 \end{bmatrix}^T .$$

Ahora, los cuatro ángulos principales son iguales a 90^o, es decir, están contaminadas las 4 direcciones principales.

Tabla 4: Porcentaje de veces en el que el valor d fue seleccionado por validación cruzada con máxima verosimilitud (a la izquierda) y con el $\tau-$estimador para el modelo PFC (a la derecha), en 500 muestras de tamaño 200 cada una, bajo los escenarios previamente descriptos, todos generados con $d = 4, r = 5$.

	d seleccionado por MV					d seleccionado por $\tau-$est	
Escenario	0	1	2	3	4	3	4
SC	0	0	0	1 %	99 %	0	100 %
C1	28.8 %	14.4 %	53.4 %	4 %	0	0	100 %
C2	22.2 %	25.4 %	36.4 %	15.8 %	0.2 %	0.2 %	99.8 %
C3	41.8 %	45.4 %	12.8 %	0	0	0	100 %
C4	26.8 %	24.4 %	48.8 %	0	0	0	100 %

En la Tabla 4 se ven los resultados obtenidos en la simulación. Allí puede apreciarse el mismo fenómeno que ocurría con los casos anteriores: cuando no hay contaminación, ambos procesos eligen la dimensión correcta; sin embargo, cuando las muestras están contaminadas, el método de seleccionar la dimensión basado en máxima verosimilitud se corrompe, mientras que el basado en $\tau-$estimadores resiste y selecciona adecuadamente.

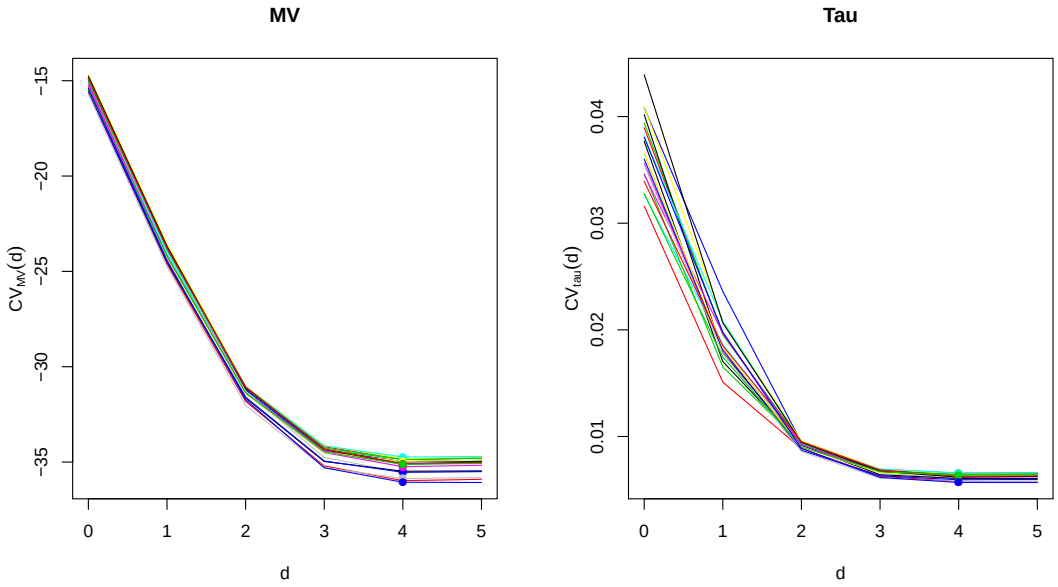
Nuevamente, para comparar con los métodos de selección de dimensión existente, en la Tabla 5 exhibimos los resultados de correr los métodos AIC, BIC y LRT para las mismas muestras. En ellas vemos que los tres métodos seleccionan mayoritariamente bien la dimensión correcta en el escenario sin contaminar. Sin embargo, en las contaminaciones el AIC selecciona bastante mal la dimensión (excepto para C4), lo mismo que LRT. En cuanto al BIC, este método selecciona bien la dimensión para tres de los cuatro escenarios contaminados, sobreestimando la dimensión sólo bajo C1, mostrando una alta capacidad de detección, que ya había sido señalada en la literatura.

Tabla 5: Porcentaje de veces en el que el valor d fue seleccionado por por AIC, BIC y LRT, en las 500 muestras correspondientes a la Tabla 4 tamaño 200 cada una, bajo los escenarios previamente descriptos, todos generados con $d = 4$, $r = 5$. Resultados corridos con el paquete `LDR` Cook et al. [2011] programado en `MATLAB`. Cabe recordar que este paquete sólo prueba valores de d mayores o iguales a uno.

		Porcentaje de veces que es elegido		
Escenario	Valor de d	AIC	BIC	LRT
SC	4	92.6 %	100 %	93.6 %
	5	7.4 %	0	6.4 %
C1	3	0	1.6 %	0
	4	0	20.2 %	0
	5	100 %	78.2 %	100 %
C2	3	0	2.8 %	0
	4	23.6 %	89 %	26.2 %
	5	76.4 %	8.2 %	73.8 %
C3	3	0	3.2 %	0
	4	80 %	96.8 %	84 %
	5	20 %	0	16 %
C4	4	58.2 %	99.4 %	60.4 %
	5	41.8 %	0.6 %	39.6 %

Figura 16: Funciones objetivo de la validación cruzada dadas en (102) y (104), una curva para cada muestra simulada en los distintos escenarios, con $d = 2, r = 5$. Indicamos con un punto la dimensión elegida por el criterio de un desvío estándar. A la izquierda el gráfico correspondiente a máxima verosimilitud, a la derecha el de los $\tau-$estimadores.

(a) Sin Contaminación



(b) Contaminación 1

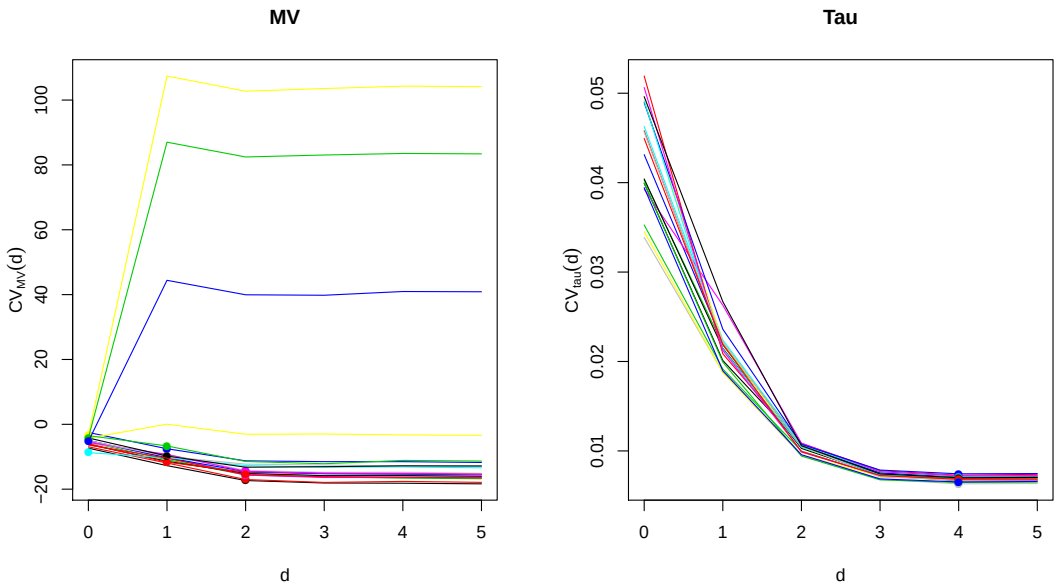
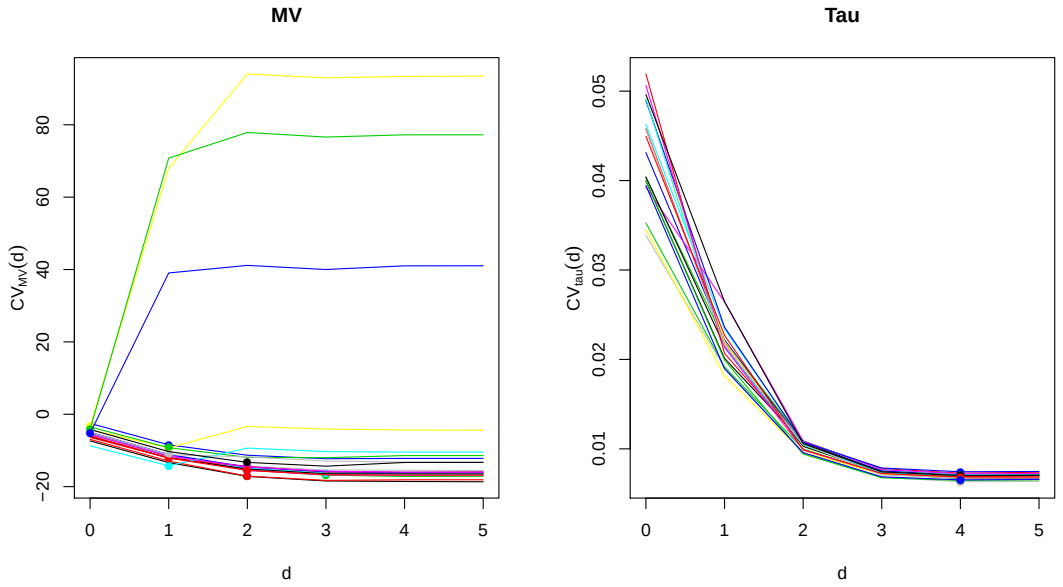


Figura 17: Contaminaciones 2 y 3.

(c) Contaminación 2



(d) Contaminación 3

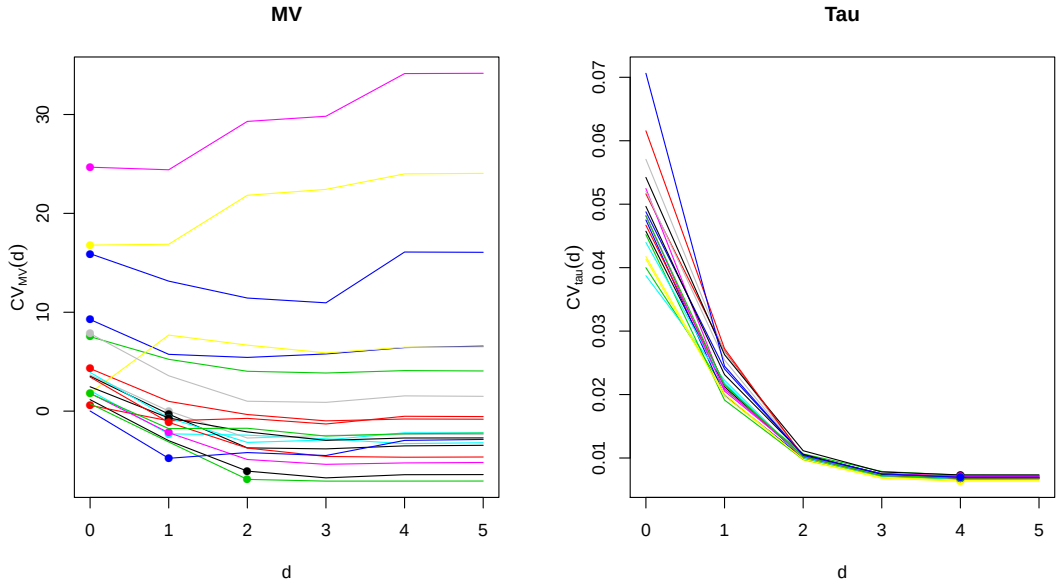
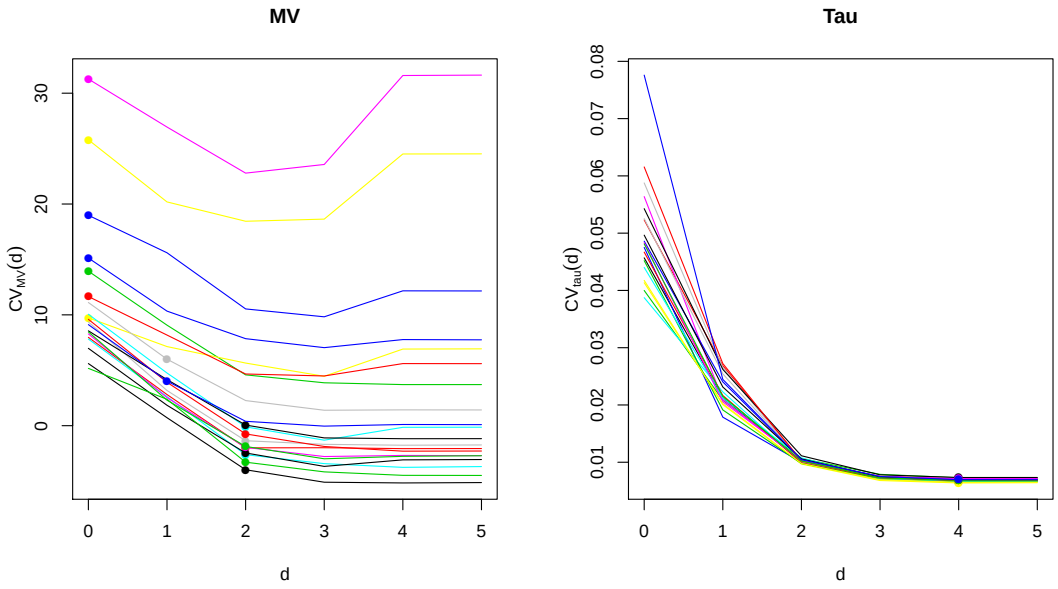


Figura 17: Contaminación 4.

(e) Contaminación 4



9.7. Ejemplos de datos reales

Finalmente, para ilustrar el desempeño de nuestras propuestas, aplicamos los $\tau-$estimadores de reducción suficiente propuestos a dos conjuntos de datos.

9.7.1. Datos ais

Utilizamos los datos `ais` ya analizados por Cook [1998] y por Bura, Duarte, y Forzani [2016] que se encuentran disponibles en el paquete `DAAG` de R. Estos datos fueron recolectados en un estudio para analizar determinadas características en el tamaño corporal y en la sangre de atletas. Para una muestra de $n = 202$ individuos que entrenan en el *Australian Institute of Sport* tomamos como variable respuesta a $y =$ masa corporal magra (ver Bura et al. [2016] para su definición) y como variables predictoras la cantidad de glóbulos rojos (`rcc`), cantidad de glóbulos blancos (`wcc`), concentración de hemoglobina en g/dl (`hg`), altura en cm (`ht`) y peso en kg (`wt`).

Para estos datos ajustamos el modelo

$$\mathbf{x}_i = \Gamma\beta\mathbf{f}\left(y_i\right) + \boldsymbol{\mu} + \Delta^{1/2}\mathbf{u}_i,$$

con $i = 1, \ldots, n = 202$, donde $\mathbf{x}_i$ es un vector de dimensión $p = 5$ que contiene las variables explicativas para la i-ésima observación, y $f(y) = (y^*, y^{*2}, y^{*3})^T$ con $y^* = (y - \overline{y})/sd(y)$.

Para la elección de la dimensión d del espacio de reducción utilizamos el criterio de validación cruzada descripto en la Sección 9.5.2. Aplicamos dicho método usando el estimador de máxima verosimilitud y el estimador τ, en ambos casos encontramos que la dimensión elegida resulta $d = 1$. Para la comparación de los subespacios generados a partir de estas dos estimaciones, calculamos el ángulo entre ellos que resulta de 35,36°.

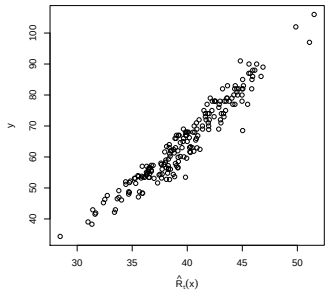
Sin embargo, detectamos que los datos utilizados presentan observaciones atípicas. Para identificarlas, nos basamos en las distancias de Mahalanobis (28) calculadas con el ajuste utilizando el estimador τ. Identificamos como atípicas todas las observaciones cuya distancia de Mahalanobis al cuadrado supere el cuantil 0,975 de la distribución χ^2 con 5 grados de libertad. Dicho cuantil es igual a 12,8325. Este criterio selecciona 14 observaciones de la muestra, las correspondientes a los índices $i = 11, 36, 56, 75, 113, 160, 161, 163, 166, 178, 181, 186, 187, 194$. Luego, realizamos el ajuste utilizando el estimador de máxima verosimilitud pero eliminando estas observaciones. El ángulo entre el subespacio de reducción estimado por MV con estas 188 observaciones y el obtenido a utilizando los $\tau-$estimadores basados en toda la muestra se reduce a $2{,}31^o$.

La Figura 17 muestra la relación entre y y $\widehat{R}_\tau(x) = \widehat{\Delta}^{-1}\widehat{\Gamma}\mathbf{x}$ para los parámetros Δ y Γ estimados con el estimador τ.

9.7.2. Datos eólicos

El segundo ejemplo de datos reales corresponde a mediciones de viento y producción eólica realizados en el Parque Eólico de Rawson, entre el 31 de marzo y el 7 de noviembre de 2012. El parque, que se halla cercano a la ciudad de Rawson, provincia de Chubut,

Figura 17: Scatter plot de y versus $\widehat{R}_\tau(\mathbf{x})$ para el ejemplo `ais`.



cuenta con 43 generadores Vestas V90, con capacidad de generación de 1.8 MW de potencia cada uno. Tomamos un aerogenerador del parque. La base de datos cuenta con 29771 observaciones, sin embargo, solamente 18342 tienen todas las variables observadas (sin datos faltantes). Cada observación es un promedio de lo que sucede en un periodo de diez minutos. Se relevaron las siguientes cuatro variables generales:

1. `vw80m`: es la velocidad del viento en metros por segundo, medido por un anemómetro en una torre del parque, a 80 metros de altura.

2. `vw35m`: es la velocidad del viento, en m/s, medido en una torre del parque a 35 metros de altura.

3. `dw79m`: la dirección del viento en grados, medida en la torre a 79 metros de altura.

4. `pre78m`: es la presión en milibares, medido a 78 metros de altura.

Además, con el sensor del aerogenerador ubicado a 85 metros de altura se mide:

5. `vwg`: la velocidad del viento que impacta en el aerogenerador, en m/s.

6. `dwg`: la dirección del viento en grados.

7. `pow`: la potencia, enKW, producida por el aerogenerador.

El objetivo del estudio es predecir la potencia generada (`pow`) de la mejor manera posible a partir de las seis variables explicativas disponibles.

Como la base de datos dispone de un gran número de observaciones, antes de hacer ningún análisis la dividimos en dos partes de tamaño 9171 cada una, que denominamos la *muestra de entrenamiento* y la *muestra de validación.*

Para las observaciones de la muestra de entrenamiento, ajustamos el modelo PFC

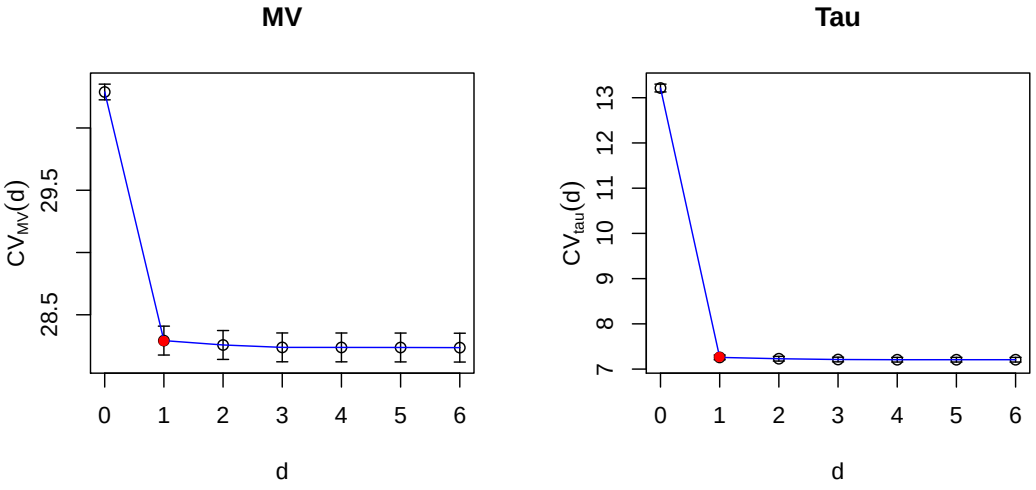
$$\mathbf{x}_i = \Gamma\beta\mathbf{f}(y_i) + \boldsymbol{\mu} + \Delta^{1/2}\mathbf{u}_i, \tag{106}$$

donde $\mathbf{x}_i \in \mathbb{R}^6$ contiene a las variables explicativas de la $i-$ésima observación antes descriptas (4 variables generales, y 2 medidas directamente en el aerogenerador), y_i es la potencia (`pow`) en la iésima observación, y $\mathbf{f}(y) = (y^*, y^{*2}, y^{*3}, y^{*4}, y^{*5}, y^{*6})^T$ con $y^* = (y - \overline{y})/sd(y)$, y $1 \leq i \leq n = 9171$.

Para este modelo ajustamos el estimador τ robusto, el EMV, y también calculamos las componentes principales robustas para las $(\mathbf{x}_i)_{1 \leq i \leq n}$ basadas tanto en la matriz de covarianza como en la matriz de correlación. Para calcular estas matrices usamos un S-estimador de covarianza calculado con la función bicuadrada de Tukey usando un algoritmo rápido para su cómputo, que se encuentra programado en el software R Core Team [2015], bajo la función `CovSest`, (*method bisquare*), de la librería `rrcov`, desarrollada por Todorov y Filzmoser [2009], como describimos en la Sección 9.2.1.

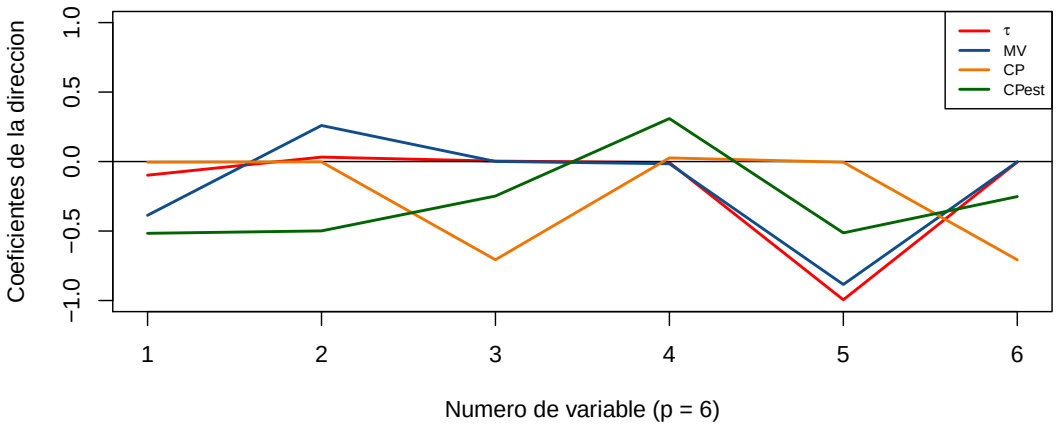
Para los dos métodos de estimación para PFC seleccionamos la dimensión d del subespacio de reducción suficiente mediante validación cruzada, según lo propuesto en la Sección 9.5.2. Tanto para los $\tau-$estimadores como para los EMV, la dimensión seleccionada resultó ser $d = 1$, como puede verse en la Figura 18.

Figura 18: Funciones de validación cruzada para máxima verosimilitud (a la izquierda) y $\tau-$estimadores del modelo PFC (a la derecha) para el ejemplo de datos eólicos, con $p = r = 6$. En ambos gráficos se resalta la dimensión elegida por el criterio de un desvío estándar.



En la Figura 19 graficamos los coeficientes que definen las direcciones estimadas tomando $d = 1$, en el espacio $\mathbb{R}^6$, para los cuatro métodos empleados. El orden de las variables es el que aparece en la descripción de las $\mathbf{x}_i$ luego de la ecuación (106). Cada método está representado por una curva, de un color en particular: en rojo para el τ, azul el EMV, naranja las componentes principales robustas basadas en los datos crudos y verde para las componentess principales robustas basadas en las variables estandarizadas. La primer componente principal explica el 99,1 % de la variabilidad de las observaciones

Figura 19: Coeficientes que definen la reducción suficiente correspondientes al $\tau-$estimador (rojo), a máxima verosimilitud (azul), la primera direcciones principales calculadas con los datos crudos (naranja) y sobre los datos estandarizados (verde).



cuando son calculadas en base a las variables crudas y 53,5 % cuando las calculamos en base a las variables estandarizadas.

Las constantes utilizadas para el estimador τ resultaron en este caso:

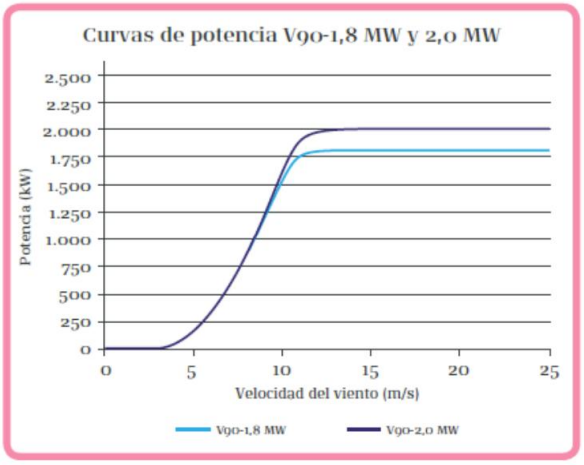
$$c_1 = 3{,}983, \quad k_1 = 0{,}499, \quad c_2 = 4{,}174, \quad k_2 = 0{,}461,$$

para obtener una eficiencia asintótica de 0,90, como se describe en la Sección 9.2.3.

En la Figura 19 vemos que el estimador τ selecciona esencialmente a la velocidad del viento medida en el aerogenerador como la variable que más interesa para predecir a la variable respuesta y, en consonancia con el modelo físico que se utiliza para modelar la potencia producida por la fuerza eólica, como puede apreciarse en la Figura 20. El EMV para PFC selecciona un contraste entre dicha velocidad y la velocidad medida en el parque a 35 metros de altura como principal responsable de la potencia. La primera dirección principal robusta, medida sobre las variables crudas, selecciona el promedio entre la dirección del viento medido en la torre del parque y la dirección del viento medida en el aerogenerador. Vemos que el método de componentes principales que también reduce la dimensión aunque de forma no supervisada por una variable respuesta, no detecta como variables relevantes a las velocidades del viento, a pesar de explicar una alta proporción de la variabilidad original de las covariables. Esto puede deberse a las importantes diferencias de escala que tienen la covariables. Por esta razón también calculamos las direcciones principales robustas para las covariables estandarizadas. Finalmente, en la Figura 19 vemos que la primer dirección depende de las seis covariables medidas, siendo esencialmente un contraste entre la presión y el promedio de las restantes. Concluimos

que la estandarización de las variables tampoco permite la detección adecuada del fenómeno que regula la producción eólica. En síntesis, vemos que la reducción dada por los τ estimadores basados en la muestra entrenamiento es la única reducción que logra recuperar la información físicamente relevante del problema.

Figura 20: Curva teórica de potencia (en kW) en función de la velocidad del viento (en m/s) para el aerogenerador Vestas V90, 1.8 MW y 2.0 MW.



Los ángulos principales entre los subespacios estimados figuran en la Tabla 6 que sigue. En ella vemos que las dos reducciones (MV y la robusta) estiman subespacios distintos, distintos también de los subespacios de dimensión uno generados por las primeras direcciones principales tanto de las variables crudas como de las variables estandarizadas.

Tabla 6: Ángulo entre subespacios estimados por los cuatro métodos presentados: PFC ajustado por un $\tau-$estimador, PFC ajustado por EMV, componentes principales robustas basadas en los datos crudos, componentes principales robustas basadas en los datos estandarizados.

Ángulos	Reducción Máx. Veros	Comp. Principales Robustas	Comp. Princ. Rob Estandariz.
Reducción Tau	$22{,}1^o$	$89{,}7^o$	$57{,}1^o$
Reducción MV	·	$89{,}7^o$	$58{,}7^o$
CP Robustas		·	$68{,}4^o$

Para comparar el poder explicativo de los cuatro subespacios de reducción estimados, usamos un estimador no paramétrico de regresión robusto: tomando como respuesta a y, y como variables explicativas las correspodientes proyecciones a los subespacios. Es decir, ajustamos una curva de regresión no paramétrica de y explicada por:

- $R_\tau(\mathbf{x}) = \widehat{\Delta}_\tau^{-1}\widehat{\Gamma}_\tau\mathbf{x}$, donde $\widehat{\Delta}_\tau$ y $\widehat{\Gamma}_\tau$ fueron estimados por los $\tau-$estimadores calculados con la muestra entrenamiento,

- $R_{\text{MV}}(\mathbf{x}) = \widehat{\Delta}_{\text{MV}}^{-1}\widehat{\Gamma}_{\text{MV}}\mathbf{x}$, donde $\widehat{\Delta}_{\text{MV}}$ y $\widehat{\Gamma}_{\text{MV}}$ fueron estimados por los EMV calculados con la muestra entrenamiento,
- $\mathbf{v}_1^T\mathbf{x}$, donde $\mathbf{v}_1$ es la primer dirección principal robusta, es decir el autovector asociado al mayor autovalor del S-estimador de covarianza calculado con los datos crudos de la muestra entrenamiento, y
- $\mathbf{w}_1^T\mathbf{x}$, donde $\mathbf{w}_1$ es la primer dirección principal robusta, es decir el autovector asociado al mayor autovalor del S-estimador de covarianza calculado con los datos estandarizados de la muestra entrenamiento.

En los cuatro casos, para calcular la regresión no paramétrica, utilizamos el M-estimador no paramétrico basado en polinomios locales de orden 1 introducido en Boente y Martínez [2017], ajustada con cada uno de los cuatro conjuntos de datos de la muestra entrenamiento. Este estimador no paramétrico de la curva de regresión es robusto. Representa una modificación del propuesto por Boente y Fraiman [1989]. De este modo obtuvimos cuatro estimadores no paramétricos de la curva de regresión, que nos permiten explicar a y utilizando los cuatro subespacios hallados.

Luego, para cada una de las 9171 observaciones de la muestra de validación predijimos el valor de y utilizando cada uno de los estimadores no paramétricos ajustados, y calculamos los residuos de cada una de las observaciones como la diferencia entre el valor observado de la producción total y el valor estimado por la regresión robusta (basada en la muestra de entrenamiento). Esto nos permitió obtener cuatro conjuntos de residuos. Finalmente calculamos la escala de cada conjunto de residuos, tomando varias de las posibilidades descriptas en la Sección 4.3. Usamos la escala de los residuos como una medida de la bondad de las distintas reducciones. En la Tabla 7 aparecen los valores obtenidos. Consideramos distintas escalas: M-escala, mediana de los valores absolutos, escala L_1 y la raíz cuadrada de la escala cuadrática.

Tabla 7: Distintas escalas de los residuos dados por el ajuste no paramétrico robusto de regresión basado en los cuatro subespacios estimados ($\tau-$estimadores, EMV, componentes principales robustas crudas y estandarizadas) usando la muestra de entrenamiento para seleccionar la dimensión, estimar los subespacios y la curva no paramétrica de regresión, y la muestra de validación para predecir y calcular los residuos. Escalas consideradas: M-escala, mediana de los valores absolutos, escala L_1 y la raíz cuadrada de la escala cuadrática.

	M-escala	Mediana de los valores abs	Escala L_1	Raíz cuad de esc cuadrática
$\tau-$est	151.25	38.97	72.43	123.22
MV	160.23	51.86	91.53	150.14
CP Rob	816.72	562.74	552.30	627.28
CP Rob Est	803.66	551.06	548.11	622.49

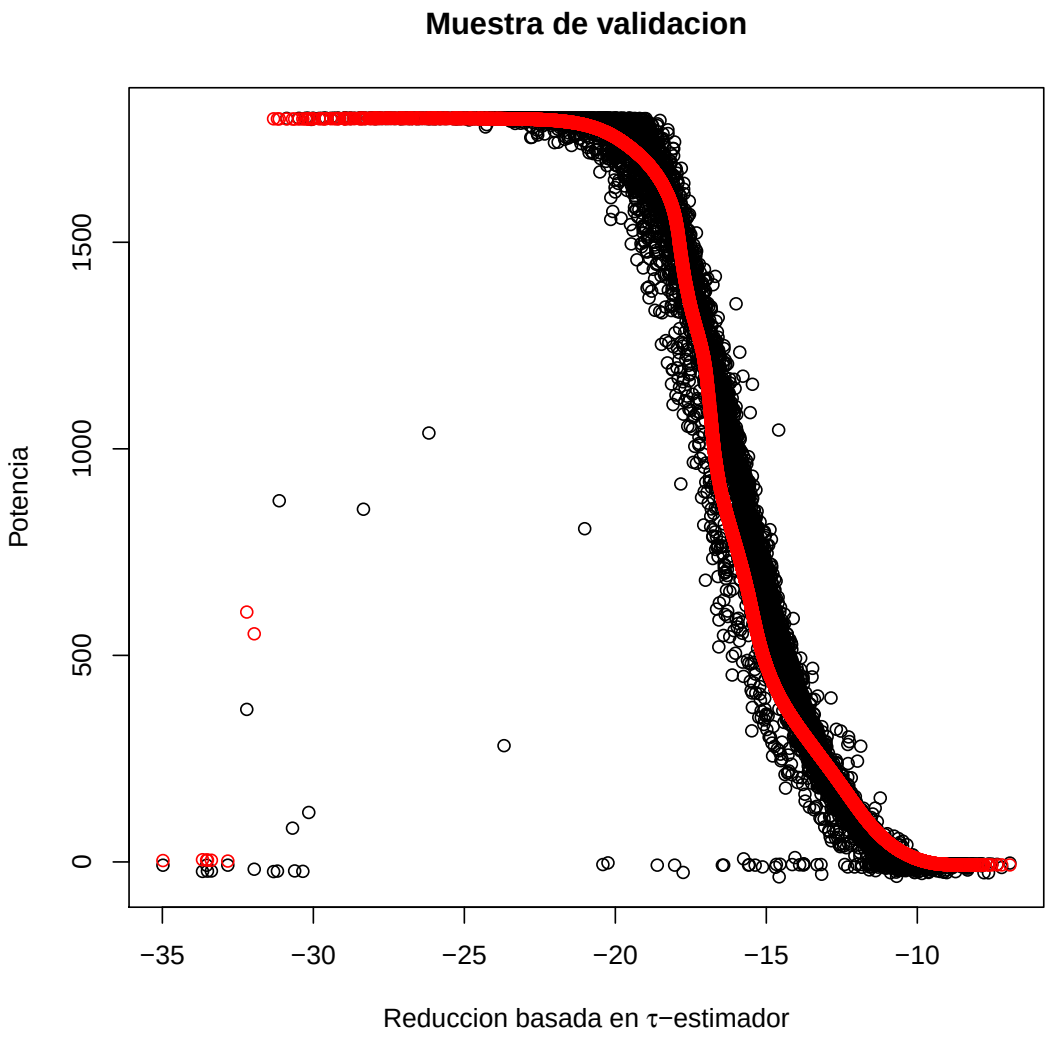
En la tabla, vemos que para todas las escalas consideradas, el menor valor calculado

sobre los residuos siempre se alcanza en la reducción estimada por los $\tau-$estimadores propuestos. El hecho de evaluar el ajuste obtenido con la muestra de validación que no fue utilizada para estimar nos permite evitar los efectos distorsionantes del sobreajuste. Podemos concluir que la reducción que mejor resume el vínculo entre las covariables y la respuesta (la potencia generada en todo el parque) está dada por la encontrada por los estimadores τ propuestos, ya que al predecir con ella los valores de potencia en una muestra independiente, el error cometido (resumido a través de una medida de escala) es menor.

En la Figura 21 presentamos un gráfico de los 9171 observaciones de la muestra de validación. Cada observación representa la variable respuesta (`pow`) versus la primera reducción estimada por τ (basada en la muestra de estimación) en negro. En rojo presentamos la predicción de la curva de regresión no paramétrica descripta antes, basada en la muestra de entrenamiento, calculada para cada punto de la muestra de validación. Vemos que en este caso, a diferencia del ejemplo AIS, el vínculo entre la variable respuesta y la reducción suficiente resulta (fuertemente) no lineal. A pesar de ello, el modelo de PFC ajustado con el estimador robusto propuesto logra recuperar claramente la reducción responsable de la variabilidad de la potencia.

Como una última comparación entre el ajuste dado por los EMV y los estimadores τ, siguiendo lo realizado en el ejemplo de las atletas, miramos las distancias de Mahalanobis estimadas por el $\tau-$estimador del modelo PFC a los datos de la muestra de entrenamiento. Podemos considerar outliers a las observaciones con mayores distancias de Mahalanobis estimadas, y comparar el subespacio de reducción estimado por el $\tau-$estimador basado en todos los datos (de la muestra de entrenamiento) y el subespacio de reducción estimado en base a las observaciones de la muestra de entrenamiento cuya distancia de Mahalanobis al cuadrado estimada sea menor al cuantil 0,999 de la distribución χ^2 con 6 grados de libertad, que es 22,458. Eliminamos dichas observaciones (que resultaron ser 1258), y calculamos los EMV sobre las 7913 restantes. El subespacio de reducción suficiente estimado por máxima verosimilitud basado en esta muestra más pequeña forma un ángulo principal de $1{,}29^o$ con el subespacio estimado por los estimadores τ en vez de $22{,}13^o$ que formaban ambos subespacios de reducción estimados, basados en todas las observaciones. Esto indica que también el método robusto para PFC propuesto es capaz de detectar las observaciones atípicas para el modelo, ya que al calcular los estimadores clásicos sin tenerlas en cuenta llegamos esencialmente al mismo estimador que nos proporciona el método τ.

Figura 21: Cada punto negro corresponde a una observación de la muestra de validación: graficamos la varible respuesta (`pow`) versus la reducción suficiente correspondiente estimada por el $\tau-$estimador entrenado con la muestra de entrenamiento. En rojo, para cada observación de la muestra de validación, graficamos la variable y predicha con el estimador no paramétrico de regresión estimado robustamente con un estimador entrenado con la muestra de entrenamiento, versus la reducción correspondiente.



A. Apéndice del Capítulo 1

A.1. Independencia condicional

Sea $(\Omega, \mathcal{F}_0, P)$ un espacio de probabilidad. Todas las variables y vectores aleatorios que consideremos a continuación estarán definidas en él. A la menor sub$-\sigma-$álgebra que contiene a los eventos nulos (i.e. $A \in \mathcal{F}_0$ tales que $\mathrm{P}(A) = 0$) de $\mathcal{F}_0$ la denotaremos por $null(\mathcal{F}_0)$. Decimos que una sub$-\sigma-$álgebra de $\mathcal{F}_0$ es completa si contiene a $null(\mathcal{F}_0)$. Para toda sub$-\sigma-$álgebra $\mathcal{F}'$ de $\mathcal{F}_0$ su completación se define por $\sigma(\mathcal{F}', null(\mathcal{F}_0))$, es decir, la menor sub$-\sigma-$álgebra que contiene a ambas $\sigma-$álgebras. La denotaremos $\overline{\mathcal{F}'}$.

Recordemos la definición de esperanza condicional.

Definición A.1 *Dados un espacio de probabilidad $(\Omega, \mathcal{F}_0, P)$, una $\sigma-$ álgebra $\mathcal{F} \subset \mathcal{F}_0$, y una variable aleatoria X que es $\mathcal{F}_0-$ medible con $E|X| < \infty$, definimos la* ***esperanza condicional de X dado $\mathcal{F}$****, $E[X \mid \mathcal{F}]$, a cualquier variable aleatoria que satisface:*

(EC i) $E[X \mid \mathcal{F}] \in \mathcal{F}$, *i.e. es $\mathcal{F}$ medible.*

(EC ii) Para todo $A \in \mathcal{F}, \int_A X dP = \int_A E[X \mid \mathcal{F}] dP$.

La condición (EC ii) puede ser escrita de la siguiente forma: $E(X1_A) = E(E[X \mid \mathcal{F}] 1_A)$.

Si Y es otra variable aleatoria y $\mathcal{F} = \sigma(Y)$ es la menor $\sigma-$ álgebra que hace a Y medible, entonces la esperanza condicional $E[X \mid \sigma(Y)]$ se escribe $E[X \mid Y]$. Con esta notación, $W \in \sigma(Y)$ significa $\sigma(W) \subset \sigma(Y)$, o equivalentemente, que existe una función medible $s : \mathbb{R} \to \mathbb{R}$ tal que $W = s(Y)$.

Cualquier variable aleatoria que cumple (EC i) y (EC ii) de la definición previa se dice que es una **versión** de $E[X \mid \mathcal{F}]$.

Teorema A.1 *$E[X \mid \mathcal{F}]$ existe y es única (c.s.). Tiene esperanza finita (es integrable).*

La prueba de la existencia se basa en el Teorema de Radon-Nikodym. Ver, por ejemplo, Durrett [2010], Sección 5.1.

Intuitivamente, pensamos que $\mathcal{F}$ describe la información que tenemos a nuestra disposición en el sentido de que para cada $A \in \mathcal{F}$, sabemos si A ha ocurrido o no. Ese es el motivo por el cual la esperanza condicional debe satisfacer la condición (EC i) de la definición, es decir, la esperanza condicional sólo debe depender de la información disponible. Por otro lado, $E[X \mid \mathcal{F}]$ es nuestra "mejor conjetura" del valor de X dada la información que tenemos, para cumplir (EC ii) tiene que tomar el mismo valor medio que X en cada subconjunto de $\mathcal{F}$.

Teorema A.2 *(Propiedades de la esperanza condicional)*

i. (Linealidad) $E[aX + Y \mid \mathcal{F}] = aE[X \mid \mathcal{F}] + E[Y \mid \mathcal{F}]$.

ii. (Monotonía) Si $X \leq Y$ *entonces* $E[X \mid \mathcal{F}] \leq E[Y \mid \mathcal{F}]$

iii. $E\left(E\left[Y \mid \mathcal{F}\right]\right) = E\left(Y\right).$

La demostración de estas propiedades puede verse, por ejemplo, en Durrett [2010], Teorema 5.1.2 e igualdad (5.1.5).

Lema A.1 *Si* $\mathcal{F}_1 \subset \mathcal{F}_2$ *entonces*

i. $E\left[E\left[X \mid \mathcal{F}_1\right] \mid \mathcal{F}_2\right] = E\left[X \mid \mathcal{F}_1\right].$

ii. $E\left[E\left[X \mid \mathcal{F}_2\right] \mid \mathcal{F}_1\right] = E\left[X \mid \mathcal{F}_1\right].$

En palabras, la menor $\sigma-$álgebra gana. La primer igualdad es trivial. La segunda es una poderosa herramienta para calcular esperanzas condicionales. Una demostración puede encontrarse en Durrett [2010], Teorema 5.1.6.

Teorema A.3 *Si* $X \in \mathcal{F}$ *y* $E\left|X\right| < \infty, E\left|XY\right| < \infty$, *entonces* $E\left[XY \mid \mathcal{F}\right] = XE\left[Y \mid \mathcal{F}\right].$

Una demostración puede encontrarse en Durrett [2010], Teorema 5.1.7.

Lema A.2 *Sean* W_1, W_2, X *variables aleatorias y* $h \in L^{\infty}\left(\Omega\right)$, *entonces*

$$\begin{aligned} E\left(E\left[W_1 \mid X\right] W_2 h\left(X\right)\right) &= E\left(W_1 E\left[W_2 \mid X\right] h\left(X\right)\right) \\ &= E\left(E\left[W_1 \mid X\right] E\left[W_2 \mid X\right] h\left(X\right)\right). \end{aligned}$$

Demostración. El Teorema A.3 justifica la segunda igualdad,

$$\begin{aligned} E\left(E\left[W_1 \mid X\right] W_2 h\left(X\right)\right) &= E\left(E\left[E\left[W_1 \mid X\right] W_2 h\left(X\right) \mid X\right]\right) \\ &= E\left(E\left[W_1 \mid X\right] h\left(X\right) E\left[W_2 \mid X\right]\right) \end{aligned}$$

lo cual prueba que la primera expresión es igual a la última. La simetría en los roles de W_1 y W_2 prueba la otra. ■

La definición y teorema que siguen se deben a Dynkin (en el mismo espíritu del Teorema $\pi - \lambda$), son muy importantes y nos permitirán probar el Teorema A.5.

Definición A.2 *Sea* $\mathcal{D}$ *una clase de subconjuntos de* Ω. *Diremos que* $\mathcal{D}$ *es un* ***D-sistema*** *si se satisfacen las siguientes condiciones:*

i. $\Omega \in \mathcal{D}$.

ii. *Si* $A, B \in \mathcal{D}, A \subset B$ *entonces* $B - A \in \mathcal{D}$.

iii. *Si* $\left(A_n\right)_{n \geq 1} \subset \mathcal{D}$ *y* $A_n \nearrow A$ *entonces* $A \in \mathcal{D}$.

Teorema A.4 *Sea $\mathcal{C}$ una clase de subconjuntos de Ω y asumamos que $\mathcal{C}$ es cerrada bajo intersecciones finitas. Si $\mathcal{D}$ es un D-sistema tal que $\mathcal{C} \subset \mathcal{D}$ entonces $\sigma(\mathcal{C}) \subset \sigma(\mathcal{D})$.*

Para una prueba de este resultado, ver Ash [1972] p. 168-169.

La siguiente exposición de independencia condicional sigue el artículo de Basu y Pereira [1983].

Definición A.3 *Los vectores aleatorios* $\mathbf{x}$ *e* $\mathbf{y}$ ***son condicionalmente independientes dado*** $\mathbf{z}$ *si para toda f, g tales que $f(\mathbf{x}), g(\mathbf{y}) \in L^{\infty}(\Omega)$ resulta*

$$E[f(\mathbf{x}) g(\mathbf{y}) \mid \mathbf{z}] = E[f(\mathbf{x}) \mid \mathbf{z}] E[g(\mathbf{y}) \mid \mathbf{z}].$$

Lo notaremos $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$.

Teorema A.5 $\mathbf{x}$ *e* $\mathbf{y}$ *son condicionalmente independientes dado* $\mathbf{z}$ *si y sólo si*

$$E[f(\mathbf{x}) \mid (\mathbf{y}, \mathbf{z})] = E[f(\mathbf{x}) \mid \mathbf{z}] \text{ para toda } f \text{ tal que } f(\mathbf{x}) \in L^{\infty}(\Omega). \tag{107}$$

La notación usual para (107) es $\mathbf{x} \mid (\mathbf{y}, \mathbf{z}) \sim \mathbf{x} \mid \mathbf{z}$.

Entonces, decir que $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$ es equivalente a decir que $\mathbf{x} \mid (\mathbf{y}, \mathbf{z})$ tiene la misma distribución condicional que $\mathbf{x} \mid \mathbf{z}$, o, en palabras, qeu la distribución de $\mathbf{x}$, dados $\mathbf{y}$ y $\mathbf{z}$ está de hecho, completamente determinada solamente por el valor de $\mathbf{z}$, $\mathbf{y}$ se vuelve superflua una vez que uno conoce a $\mathbf{z}$. Observemos que en la definición de esperanza condicional lo roles de $\mathbf{x}$ e $\mathbf{y}$ son claramente simétricos, esta simetría no es tan clara en (107).

Demostración. Veamos que (107) implica la independencia condicional. Por el Lema A.1 (ii), obtenemos la primera igualdad:

$$\begin{aligned} E[f(\mathbf{x}) g(\mathbf{y}) \mid \mathbf{z}] &= E[E[f(\mathbf{x}) g(\mathbf{y}) \mid (\mathbf{y}, \mathbf{z})] \mid \mathbf{z}] \\ &= E[g(\mathbf{y}) E[f(\mathbf{x}) \mid (\mathbf{y}, \mathbf{z})] \mid \mathbf{z}] \\ &= E[g(\mathbf{y}) E[f(\mathbf{x}) \mid \mathbf{z}] \mid \mathbf{z}] \\ &= E[f(\mathbf{x}) \mid \mathbf{z}] E[g(\mathbf{y}) \mid \mathbf{z}]. \end{aligned}$$

Hemos usado la hipótesis en la tercera igualdad, y el Teorema A.3 en la segunda y cuarta, probando la impicación es verdadera.

Recíprocamente, queremos probar que si para toda f y g funciones medibles tales que $f(\mathbf{x})$ y $g(\mathbf{y})$ son acotadas, se cumple,

$$E[f(\mathbf{x}) g(\mathbf{y}) \mid \mathbf{z}] = E[f(\mathbf{x}) \mid \mathbf{z}] E[g(\mathbf{y}) \mid \mathbf{z}].$$

entonces, para toda f tal que $f(\mathbf{x}) \in L^{\infty}(\Omega)$, se tendrá

$$E[f(\mathbf{x}) \mid (\mathbf{y}, \mathbf{z})] = E[f(\mathbf{x}) \mid \mathbf{z}].$$

Sea $\mathcal{E}$ la clase de subconjuntos definida por

$$\mathcal{E} = \left\{ \begin{array}{c} E : E \in \sigma(\mathbf{y}, \mathbf{z}, null(\mathcal{F}_0)) = \overline{\sigma(\mathbf{y}, \mathbf{z})} \\ \text{y } \int_E f(\mathbf{x})\, dP = \int_E E[f(\mathbf{x}) \mid \mathbf{z}]\, dP \ \forall f \text{ tal que } f(\mathbf{x}) \in L^\infty(\Omega) \end{array} \right\}.$$

$\mathcal{E}$ es un D-sistema pues

1. $\Omega \in \mathcal{E}$,
2. Para $A, B \in \mathcal{E},\ A \subset B$ tenemos $B - A \in \mathcal{E}$.
3. $\mathcal{E}$ es una clase monótona: Sea $(A_n)_{n\geq 1} \subset \mathcal{E}$ y $A_n \nearrow A$ entonces queremos probar que $A \in \mathcal{E}$.

   - Para $f \geq 0$, tenemos $1_{A_n} \nearrow 1_A$, luego por el teorema de la convergencia monótona tenemos, tenemos

$$\begin{aligned} \int_A f(\mathbf{x})\, dP &= \int f(\mathbf{x})\, 1_A dP = \lim_{n\to\infty} \int f(\mathbf{x})\, 1_{A_n} dP = \lim_{n\to\infty} \int_{A_n} f(\mathbf{x})\, dP \\ &= \lim_{n\to\infty} \int_{A_n} f(\mathbf{x})\, dP = \lim_{n\to\infty} \int_{A_n} E[f(\mathbf{x}) \mid \mathbf{z}]\, dP \qquad (108) \\ &= \lim_{n\to\infty} \int E[f(\mathbf{x}) \mid \mathbf{z}]\, 1_{A_n} dP = \int E[f(\mathbf{x}) \mid \mathbf{z}]\, 1_A dP \qquad (109) \\ &= \int_A E[f(\mathbf{x}) \mid \mathbf{z}]\, dP. \end{aligned}$$

     La segunda igualdad en (108) se debe a que $A_n \in \mathcal{E}$. Y la segunda igualdad en (109) deriva nuevamente de la convergencia monótona y del hecho de que $E[f(\mathbf{x}) \mid \mathbf{z}] \geq 0$ cuando f es positiva.
   - Para f general, escribimos $f = f^+ - f^-$ y usamos la linealidad de la esperanza condicional.

Queremos mostrar que $\mathcal{E}$ contiene a $\overline{\sigma(\mathbf{y}, \mathbf{z})}$. Sean C, D dos conjuntos que cumplen $C \in \overline{\sigma(\mathbf{y})}$, $D \in \overline{\sigma(\mathbf{z})}$. Claramente, $C \cap D \in \overline{\sigma(\mathbf{y}, \mathbf{z})}$ y como $1_C \in \overline{\sigma(\mathbf{y})}$ si y sólo si existe una función s medible tal que $1_C = s(\mathbf{y})$. Análogamente, $1_D = h(\mathbf{z})$, entonces

$$\begin{aligned} \int_{C\cap D} f(\mathbf{x})\, dP &= E(1_C 1_D f(\mathbf{x})) = E(E[1_C 1_D f(\mathbf{x}) \mid \mathbf{z}]) \\ &= E(1_D E[1_C f(\mathbf{x}) \mid \mathbf{z}]) \\ &= E(1_D E[1_C \mid \mathbf{z}]\, E[f(\mathbf{x}) \mid \mathbf{z}]) \\ &= E(1_D 1_C E[f(\mathbf{x}) \mid \mathbf{z}]) = \int_{C\cap D} E[f(\mathbf{x}) \mid \mathbf{z}]\, dP \end{aligned}$$

donde hemos usado las hipótesis para la cuarta ecuación y el Lema A.2 para la quinta. Luego $\mathcal{E}' \subset \mathcal{E}$ donde

$$\mathcal{E}' = \left\{ C \cap D : C \in \overline{\sigma(\mathbf{y})} \text{ y } D \in \overline{\sigma(\mathbf{z})} \right\}.$$

Como $\mathcal{E}'$ es cerrado bajo intersecciones finitas, y $\sigma(\mathcal{E}') = \overline{\sigma(\mathbf{y},\mathbf{z})}$, por el Teorema A.4, concluimos que $\overline{\sigma(\mathbf{y},\mathbf{z})} \subset \mathcal{E}$, esto es $E[f(\mathbf{x}) \mid (\mathbf{y},\mathbf{z})] = E[f(\mathbf{x}) \mid \mathbf{z}]$ para toda f tal que $f(\mathbf{x}) \in L^{\infty}(\Omega)$. ■

Citamos a Basu y Pereira [1983]. El concepto de independencia condicional da lugar a muchas preguntas. Entre ellas están las que involucran a la preservación de la misma al agregar o quitar variables. Supongamos que $\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{w},\mathbf{x}_1,\mathbf{z}_1$, son vectores aleatorios tales que $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$, $\mathbf{x}_1 \in \sigma(\mathbf{x}), \mathbf{z}_1 \in \sigma(\mathbf{z})$. ¿Qué se puede decir de la relación $\perp\!\!\!\perp$ si $\mathbf{x}_1$ sustituye a $\mathbf{x}$, $\mathbf{z}_1$ sustituye a $\mathbf{z}$, cambiamos a $\mathbf{y}$ por $(\mathbf{y},\mathbf{w})$ o a $\mathbf{z}$ por $(\mathbf{z},\mathbf{w})$? En otras palabras, podemos reducir o aumentar a $\sigma(\mathbf{x}), \sigma(\mathbf{y})$ ó $\sigma(\mathbf{z})$ sin destruir la independencia condicional? En general, la respuesta es no. Sin embargo, para ciertas clases de aumentos o disminuciones la relación puede ser preservada. La esencia de los principios de Drop/Add para independencia condicional está contenida en el siguiente resultado.

Lema A.3 ***(add and drop)*** *Si* $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$, *entonces para cada* $\mathbf{x}' \in \sigma(\mathbf{x})$ *se tiene:*

i. $\mathbf{x}' \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$

ii. $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid (\mathbf{x}',\mathbf{z})$

iii. $(\mathbf{x},\mathbf{z}) \perp\!\!\!\perp (\mathbf{y},\mathbf{z}) \mid \mathbf{z}$

Entonces, si $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$, la relación $\perp\!\!\!\perp$ es preservada cuando (i) $\mathbf{x}$ ó $\mathbf{y}$ son arbitrariamente reducidas (*drop*), (ii) $\mathbf{z}$ es aumentada (*add*) en cualquier parte esencial de $\mathbf{x}$ ó $\mathbf{y}$, o bien (iii) $\mathbf{x}$ ó $\mathbf{y}$ son aumentadas (*add*) por cualquier parte esencial de $\mathbf{z}$.

Demostración.

i. Como $\mathbf{x}' \in \sigma(\mathbf{x})$ existe una función medible s tal que $\mathbf{x}' = s(\mathbf{x})$, luego puesto que $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$, entonces para toda f tal que $f(\mathbf{x}') \in L^{\infty}(\Omega)$, tendremos $f(\mathbf{x}') = f(s(\mathbf{x})) \in L^{\infty}(\Omega)$, y

$$E\left[f\left(\mathbf{x}'\right) \mid (\mathbf{y},\mathbf{z})\right] = E\left[f\left(s\left(\mathbf{x}\right)\right) \mid (\mathbf{y},\mathbf{z})\right] = E\left[f\left(s\left(\mathbf{x}\right)\right) \mid \mathbf{z}\right] = E\left[f\left(\mathbf{x}'\right) \mid \mathbf{z}\right].$$

que es equivalente a (i).

ii. Por el Teorema A.5, si mostramos que para cada g tal que $g(\mathbf{y}) \in L^{\infty}(\Omega)$, $E[g(\mathbf{y}) \mid (\mathbf{x},\mathbf{x}',\mathbf{z})] = E[g(\mathbf{y}) \mid (\mathbf{x}',\mathbf{z})]$, habremos completado la prueba de (ii).

$$E\left[g\left(\mathbf{y}\right) \mid \left(\mathbf{x},\mathbf{x}',\mathbf{z}\right)\right] = E\left[g\left(\mathbf{y}\right) \mid (\mathbf{x},\mathbf{z})\right] = E\left[g\left(\mathbf{y}\right) \mid \mathbf{z}\right] \tag{110}$$

donde la última igualdad sigue de la equivalencia a $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$ establecida en el Teorema A.5.

Por otro lado, por el Lema A.1, tenemos

$$\begin{aligned} E\left[g\left(\mathbf{y}\right) \mid \left(\mathbf{x}',\mathbf{z}\right)\right] &= E\left[E\left[g\left(\mathbf{y}\right) \mid \left(\mathbf{x},\mathbf{x}',\mathbf{z}\right)\right] \mid \left(\mathbf{x}',\mathbf{z}\right)\right] \\ &= E\left[E\left[g\left(\mathbf{y}\right) \mid \mathbf{z}\right] \mid \left(\mathbf{x}',\mathbf{z}\right)\right] = E\left[g\left(\mathbf{y}\right) \mid \mathbf{z}\right]. \end{aligned} \tag{111}$$

Finalmente, a partir de (110) y (111) puede obtenerse el resultado buscado.

iii. Basta mostrar que $(\mathbf{x}, \mathbf{z}) \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$ puesto que entonces (iii) sigue por simetría. Por el Teorema A.5, queremos probar qeu para cada g tal que $g(\mathbf{y}) \in L^{\infty}(\Omega)$, $E[g(\mathbf{y}) \mid (\mathbf{x}, \mathbf{z})] = E[g(\mathbf{y}) \mid \mathbf{z}]$, lo cual se deduce de la independencia condicional de $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$. ■

B. Apéndice del Capítulo 2

Comenzamos este apéndice resumiendo algunas cuestiones de álgebra lineal que pueden ser de interés para el tema que estudiamos. Comencemos por recordar la descomposición en valores singulares de una matriz, y algunas propiedades de interés, para luego resumir las distintas nociones de normas en espacios matriciales.

B.1. Descomposición en Valores Singulares (SVD)

B.1.1. Definición y construcción

En álgebra lineal, la descomposición en valores singulares (en inglés, singular value decomposition (SVD)) es una factorización de una matriz real o compleja, con muchas aplicaciones en distintas áreas de la matemática, desde el procesamiento de señales hasta la estadística. Formalmente

Teorema B.1 *Sea M una matriz $m \times n$ cuyos coeficientes son números reales o complejos. Entonces existe una factorización de la forma*

$$M = U\Sigma\overline{V}^T, \tag{112}$$

donde U es una matriz unitaria $m \times m$ real o compleja (unitaria significa que $U\overline{U}^T = \overline{U}^T U = I$, donde $\overline{U}$ es la matriz conjugada de U. En el caso en el que U es real, decimos que es una matriz ortogonal), Σ es una matriz rectangular diagonal $m \times n$ con números reales no negativos en la diagonal, y la matriz unitaria $n \times n$ $\overline{V}^T$ denota la traspuesta conjugada de la matriz unitaria V de dimensión $n \times n$. Tal factorización se denomina la ***descomposición en valores singulares*** *de M.*

Los elementos σ_i de la diagonal de Σ se denominan los valores singulares de M. Una convención habitual es listarlos en orden descendiente. En este caso, la matriz diagonal Σ queda unívocamente determinada por M (aunque las matrices U y V no lo están). Las m columnas de U y las n columnas de V se denominan los vectores singulares a izquierda de M y a derecha de M, respectivamente.

La descomposición en valores singulares y la diagonalización están muy relacionadas. En efecto:

- Los vectores singulares a izquierda de M son autovectores de $M\overline{M}^T$.
- Los vectores singulares a derecha de M son autovectores de $\overline{M}^T M$.
- Los valores singulares no nulos de M (que se encuentran en la diagonal de Σ) son las raíces cuadradas de los autovalores no nulos tanto de $M\overline{M}^T$ como de $\overline{M}^T M$.

Para una demostración de este resultado puede verse Eaton [1983], Teorema 1.3.

¿Es única esta descomposición? Si asumimos que $m \leq n$, Σ será única, puesto que los autovalores de $\overline{M}^T M$ lo son. Sin embargo, los autovectores que componen a U y V no serán únicos salvo que los autovalores sean todos distintos, y se utilice una convención apropiada para los autovectores (por ejemplo, que la primer coordenada no nula de cada uno sea positiva).

Observación B.1 *Si escribimos*

$$U = \left[\begin{array}{cccc} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_m \end{array} \right]$$

donde los $\mathbf{u}_i$ *son las columnas de* U*, y*

$$V = \left[\begin{array}{cccc} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{array} \right]$$

donde los $\mathbf{v}_i$ *son las columnas de* V*, y* $D^{1/2} = diag\left(\sigma_1, \ldots, \sigma_{q_1}\right)$*, donde* $\sigma_1 \geq \cdots \geq \sigma_{q_1}$*, entonces una forma alternativa de escribir la ecuación (112) es*

$$M = \sum_{i=1}^{q_1} \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

B.1.2. Normas de vectores y matrices

Norma de vectores Para vectores $\mathbf{x} \in \mathbb{R}^n$ ó $\mathbb{C}^n, \mathbf{x} = (x_1, \ldots, x_n)^T$, para $p \geq 1$ la $p-$**norma** está dada por

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

Si $p = 2$, la **2−norma** recibe el nombre de **norma euclídea**.

Norma de matrices

Norma inducida A partir de la norma para vectores de $\mathbb{R}^n$ ó $\mathbb{C}^n$, se pueden definir la correspondiente **norma inducida** o **norma de operadores** en el espacio de matrices m-por-n de la siguiente forma,

$$\|A\| = \text{máx}\left\{ \|A\mathbf{x}\|_m : \mathbf{x} \in \mathbb{R}^n \text{ con } \|\mathbf{x}\|_n = 1 \right\} = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|_m}{\|\mathbf{x}\|_n}.$$

Si $m = n$ y se utiliza la misma norma en el dominio y en el rango, entonces la norma inducida resulta una norma matricial sub-multiplicativa. La norma de operadores correspondiente a la $p-$norma de vectores es

$$\|A\|_p = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p}.$$

Estas normas son diferentes de las $p-$normas de coordenadas de matrices que presentamos más adelante, aunque también son usualmente notadas por $\|A\|_p$.

Norma espectral La norma espectral de una matriz A es el mayor valor singular de A, es decir, la raíz cuadrada del mayor autovalor de la matriz definida positiva $\overline{A}^T A$

$$\|A\|_2 = \sqrt{\lambda_{\text{máx}}\left(\overline{A}^T A\right)} = \sigma_{\text{máx}}\left(A\right).$$

En el caso especial de $p = 2$ (la norma Euclídea) y matrices cuadradas $m = n$, la norma inducida es la norma espectral.

Normas componente a componente o Entrywise Estas normas toman a la matriz $m \times n$ como si fuera un vector de dimensión mn, y utilizan alguna norma de vectores conocida.

Por ejemplo, usando la $p-$norma para vectores, se obtiene

$$\|A\|_p = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^p\right)^{1/p}.$$

Una expresión alternativa se obtiene si escribimos a la matriz en términos de los vectores que componen sus columnas, es decir, si

$$A = (a_{ij})_{1\le i\le m, 1\le j\le n} = \begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_n\end{pmatrix}$$

donde $\mathbf{a}_j^T = (a_{1j}, \ldots, a_{mj})$ son los vectores columna de la matriz A, entonces

$$\|A\|_p = \left\| \left(\|\mathbf{a}_1\|_p, \ldots, \|\mathbf{a}_n\|_p \right)^T \right\|_p.$$

Para $p = 2$, se denomina la **norma de Frobenius** o norma de Hilbert–Schmidt. Hay varias formas de definirla,

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2} = \sqrt{\text{traza}(\overline{A}^T A)} = \sqrt{\sum_{i=1}^{\text{mín}\{m,n\}} \sigma_i^2}$$

donde, como antes, $\overline{A}^T$ denota la conjugada traspuesta de A, y (σ_i) son los valores singulares de A. La norma de Frobenius es muy parecida a la norma Euclídea en $\mathbb{R}^n$ o $\mathbb{C}^n$.

La norma de Frobenius es sub-multiplicativa y muy utilizada en el álgebra lineal numérica. En general es más fácil de calcular que las normas inducidas y tiene la propiedad de ser invariante bajo transformaciones unitarias (ortogonales en los reales). Veámoslo en el caso real, sea C una matriz ortogonal, luego

$$\|CA\|_F^2 = \text{traza}\left((CA)^T \cdot (CA)\right) = \text{traza}(A^T C^T C A) = \text{traza}(A^T A) = \|A\|_F^2.$$

B.1.3. Aproximación por matrices de menor rango

Un problema frecuentemente tratado en diversas aplicaciones es el siguiente. Dada una matriz M se la quiere aproximar por otra matriz $\widetilde{M}$, que se denomina truncada, que tiene un rango específico r. Esta área se conoce como *low-rank matrix approximation*. Antes de presentar el resuEn el caso en el que la aproximación se base en minimizar la norma de Frobenius de la diferencian entre M y $\widetilde{M}$ bajo la restricción de que rango$(\widetilde{M}) = r$ resulta ser que la solución está dada por la SVD de M, más específicamente

$$\widetilde{M} = U\tilde{\Sigma}\overline{V}^T \tag{113}$$

donde $\tilde{\Sigma}$ es la misma matriz que Σ excepto que sólo contiene los r mayores valores singulares (los otros valores singulares se reemplazan por cero). Esto se conoce como el **Teorema de Eckart–Young theorem**, ya que fue probado por estos dos autores en 1936 (aunque más tarde se descubrió que era un resultado conocido por autores anteriores (Stewart 1993). Otra manera de expresar este resultado es la siguiente

$$\widetilde{M} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

Puede verse, por ejemplo, Eaton [1983], página 458, o Seber [1984], página 544.

B.2. Variedades de Stiefel y de Grassmann

En esta sección presentamos las definiciones y los resultados básicos que involucran a las variedades de Stiefel y de Grassmann. Puede verse Absil et al. [2009], Capítulo 3, para una presentación completa. Algunos resultados fueron extraidos de Ferrer, García, y Puerta [1994] y de Edelman et al. [1998].

Sea $d \leq p$. Denotamos por Grass $(d, \mathbb{R}^p)$ al conjunto de todos los subespacios d-dimensionales de $\mathbb{R}^p$. Es decir,

$$\text{Grass}\,(d, \mathbb{R}^p) := \{\mathcal{V} \subset \mathbb{R}^p : \mathcal{V} \text{ es subespacio de } \mathbb{R}^p,\ \dim(\mathcal{V}) = d\}.$$

A este conjunto se lo puede dotar con una estructura de variedad diferencial. Se la denomina *variedad de Grassmann o Grassmanniana de dimensión d en* $\mathbb{R}^p$.

Sea $A \in \mathbb{R}^{p\times d}, A = (a_{ij})_{1\leq i\leq p, 1\leq j\leq d} = \begin{bmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_d\end{bmatrix}$ donde $\mathbf{a}_j^T = (a_{1j}, \ldots, a_{pj})$ son los vectores columna de la matriz A. Consideraremos la norma en $\mathbb{R}^{p\times d}$ inducida por el producto interno de Frobenius (ver la Sección B.1.2), es decir

$$\|A\|_F = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p} |a_{ij}|^2} = \sqrt{\text{traza}(A^T A)}.$$

Definimos la *variedad no compacta de Stiefel* por

$$\begin{aligned}\text{St}\,(d, \mathbb{R}^p) &:= \left\{A \in \mathbb{R}^{p\times d} : \text{rango}\,(A) = d\right\} \\ &= \left\{A = \begin{bmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_d\end{bmatrix} \in \mathbb{R}^{p\times d} : \mathbf{a}_1, \ldots, \mathbf{a}_d \text{ son linealmente independientes}\right\},\end{aligned}$$

es decir, $\mathrm{St}(d, \mathbb{R}^p)$ es el conjunto de matrices $p \times d$ cuyas columnas son linealmente independientes. Observemos que este conjunto es abierto en $\mathbb{R}^{p \times d}$ puesto que,

$$\mathrm{St}(d, \mathbb{R}^p) := \left\{ A \in \mathbb{R}^{p \times d} : \det\left(A^T A\right) \neq 0 \right\}$$

es la preimagen de un conjunto abierto, $\mathbb{R} - \{0\}$, vía una función continua $\left(A \to \det\left(A^T A\right)\right)$. En consecuencia, admite una estructura diferenciable de una variedad abierta de $\mathbb{R}^{p \times d}$. Esta estructura diferenciable está generada por la aplicación $\mathrm{vec} : \mathbb{R}^{p \times d} \to \mathbb{R}^{pd}$, donde $\mathrm{vec}(A)$ denota el vector de $\mathbb{R}^{pd}$ que se obtiene al apilar las columnas de A una debajo de la otra. Cabe resaltar que la aplicación vec no es muy adecuada, puesto que destruye la estructura matricial de su argumento, ya que en particular $\mathrm{vec}(AB)$ no puede escribirse de manera sencilla a partir de $\mathrm{vec}(A)$ y de $\mathrm{vec}(B)$.

También consideramos la *variedad compacta de Stiefel* dada por

$$\mathrm{St}^*(d, \mathbb{R}^p) := \left\{ A \in \mathbb{R}^{p \times d} : A^T A = I_d \right\},$$

es decir, el conjunto de matrices $p \times d$ con columnas ortonormales. Claramente, $\mathrm{St}^*(d, \mathbb{R}^p) \subset \mathrm{St}(d, \mathbb{R}^p)$. Este conjunto resulta acotado (por $\sqrt{d}$) y cerrado (por ser la preimagen del conjunto cerrado $\{I_d\}$ vía la función continua $A \to A^T A$), y por lo tanto es un subconjunto compacto de $\mathrm{St}(d, \mathbb{R}^p)$.

Observemos que si utilizamos otro producto interno en $\mathbb{R}^p$ y consideramos la noción de perpendicularidad inducida por él, podríamos generalizar la definición de la variedad compacta de Stiefel. Más específicamente, sea $\Delta \in PDS(p)$, se puede definir un producto interno en $\mathbb{R}^p$ inducido por Δ, dado por

$$\langle \mathbf{x}, \mathbf{y} \rangle_\Delta = \mathbf{x}^T \Delta^{-1} \mathbf{y},$$

y la norma dada por

$$\|\mathbf{x}\|_\Delta = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_\Delta}.$$

Luego, se podría generalizar la variedad compacta de Stiefel de la siguiente forma

$$\mathrm{St}^*_\Delta(d, \mathbb{R}^p) := \left\{ A \in \mathbb{R}^{p \times d} : A^T \Delta^{-1} A = I_d \right\},$$

es decir, pidiendo que las columnas de A sean ortonormales con la norma (y producto interno) inducida por Δ.

Veamos cómo se relacionan estos tres conjuntos entre sí. Consideremos la siguiente función,

$$\begin{aligned} \mathrm{span} : \mathrm{St}(d, \mathbb{R}^p) &\to \mathrm{Grass}(d, \mathbb{R}^p), \\ A &\to \mathrm{span}(A) \end{aligned}$$

donde $\mathrm{span}(A)$ denota al subespacio generado por las columnas de A, es decir, $\mathrm{span}(A) = \left\{ A\alpha : \alpha \in \mathbb{R}^d \right\}$. En el espacio de las Grassmannianas $(\mathrm{Grass}(d, \mathbb{R}^p))$ consideramos la topología final respecto de π, es decir, $U \subseteq \mathrm{Grass}(d, \mathbb{R}^p)$ es abierto si y sólo si $\mathrm{span}^{-1}(U)$

es abierto en $\mathrm{St}(d, \mathbb{R}^p)$. Siguiendo a Absil et al. [2009], nos referiremos a esta topología como la topología de Grassmann. Observemos que la aplicación span permite definir la siguiente relación de equivalencia en $\mathrm{St}(d, \mathbb{R}^p)$. Usaremos $\sim$ para denotarla. Diremos que

$$A \sim B \text{ si y sólo si } \mathrm{span}(A) = \mathrm{span}(B).$$

La clase de equivalencia de A en $\mathrm{St}(d, \mathbb{R}^p)$ inducida por esta relación está dada por

$$[A] = \left\{ B \in \mathbb{R}^{p \times d} : \mathrm{span}(A) = \mathrm{span}(B) \right\},$$

es decir, todas las matrices que tienen por columnas una base del subespacio generado por A. El conjunto $\mathrm{St}(d, \mathbb{R}^p) / \sim$ definido por

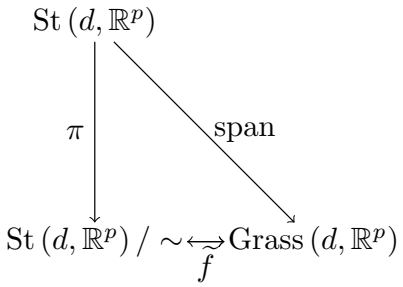
$$\mathrm{St}(d, \mathbb{R}^p) / \sim := \{[A] : A \in \mathrm{St}(d, \mathbb{R}^p)\}$$

de todas las clases de equivalencia de $\sim$ en $\mathrm{St}(d, \mathbb{R}^p)$ es el cociente de $\mathrm{St}(d, \mathbb{R}^p)$ por $\sim$. Cada punto del espacio cociente es un subconjunto de $\mathrm{St}(d, \mathbb{R}^p)$. La función

$$\begin{aligned} \pi : \mathrm{St}(d, \mathbb{R}^p) &\to \mathrm{St}(d, \mathbb{R}^p) / \sim \\ A &\to [A] \end{aligned}$$

es la proyección canónica. Como los conjuntos de nivel de $\mathrm{span}(\cdot)$ son las clases de equivalencia de la relación $\sim$ y como además la aplicación $\mathrm{span}(\cdot)$ es sobreyectiva sobre $\mathrm{Grass}(d, \mathbb{R}^p)$, resulta que $\mathrm{span}(\cdot)$ induce una correspondencia biyectiva entre $\mathrm{Grass}(d, \mathbb{R}^p)$ y $\mathrm{St}(d, \mathbb{R}^p) / \sim$. Esta biyección se ilustra en la Figura 22.

Figura 22: Diagrama de correspondencia entre la Grassmaniana y una subvariedad en el espacio de matrices.



En Absil et al. [2009], Proposición 3.4.6, se muestra que esta relación de equivalencia admite una (única) estructura de variedad cociente. Si una matriz A y un subespacio $\mathcal{S}_A$ satisfacen $\mathcal{S}_A = \mathrm{span}(A)$, decimos que A es la representación matricial de $\mathcal{S}_A$. El

conjunto de todas las representaciones matriciales de span(A) es la clase de equivalencia $\pi^{-1}(\pi(A))$. Se tiene que

$$\pi^{-1}(\pi(A)) = \{AM : M \in GL_d\},$$

donde GL_d es el grupo lineal general, es decir, el conjunto de todas las matrices $d \times d$ inversibles. De hecho, la operación $A \to AM$, $M \in GL_d$, corresponde a todos los posibles cambios de base para el span(A). Luego, suele usarse la notación $\mathrm{St}(d, \mathbb{R}^p)/GL_d$ para $\mathrm{St}(d, \mathbb{R}^p)/\sim$. Por lo tanto, tenemos la identificación

$$\mathrm{Grass}(d, \mathbb{R}^p) \simeq \mathrm{St}(d, \mathbb{R}^p)/GL_d.$$

Esta descripción de $\mathrm{Grass}(d, \mathbb{R}^p)$ como variedad cociente permite probar que

$$\dim(\mathrm{Grass}(d, \mathbb{R}^p)) = d(p-d),$$

y dotar a $\mathrm{Grass}(d, \mathbb{R}^p)$ de una topología y una estructura diferenciable. Ver Absil et al. [2009], Sección 3.4 para una demostración del cálculo de la dimensión y la construcción de las mismas. También resulta que

$$\dim(\mathrm{St}^*(d, \mathbb{R}^p)) = pd - \frac{d(d+1)}{2} = d(p-d) + \frac{d(d-1)}{2}.$$

Otras representaciones de la Grassmanniana como cociente están dadas por

$$\mathrm{Grass}(d, \mathbb{R}^p) \simeq \mathrm{St}^*(d, \mathbb{R}^p)/O(d) \simeq O(p)/(O(d) \times O(p-d)),$$

(ver Edelman et al. [1998]), donde

$$O(p) = \left\{A \in \mathbb{R}^{p \times p} : A^T A = I_p\right\} = \mathrm{St}^*(p, \mathbb{R}^p)$$

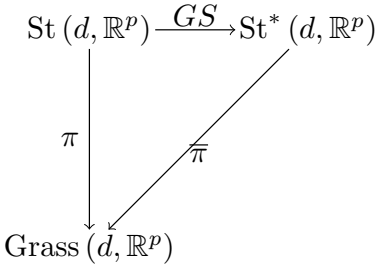
es el grupo de matrices ortogonales de $\mathbb{R}^{p \times p}$. De lo anterior, tenemos $\dim(O(p)) = \frac{p(p-1)}{2}$.

Finalmente, consideremos la restricción de π a $\mathrm{St}^*(d, \mathbb{R}^p)$, es decir, $\overline{\pi} := \pi\,|_{\mathrm{St}^*(d,\mathbb{R}^p)}$ y definamos la función dada por el *proceso de ortonormalización de Gram-Schmidt*, es decir, la función

$$\begin{aligned} GS : \mathrm{St}(d, \mathbb{R}^p) &\to \mathrm{St}^*(d, \mathbb{R}^p) \\ A &\to GS(A) \end{aligned}$$

$GS(A)$ es la matriz que se obtiene al aplicar el proceso de ortonormalización de Gram-Schmidt a las columnas $\mathbf{a}_1, \ldots, \mathbf{a}_d$ de A. Es un resultado conocido que el método de Gram-Schmidt define una función continua. Luego $\overline{\pi} \circ GS = \pi$, como puede verse en la Figura 23, que vincula los tres espacios que definimos. Se prueba que $\mathrm{Grass}(d, \mathbb{R}^p)$ es un espacio compacto respecto de la topología de Grassmann.

Figura 23: Diagrama conmutativo que permite definir la topología Grassmann de forma directa en la Grassmanniana.



B.3. Caracterización del espacio Ω

Nuestro objetivo es caracterizar el espacio

$$\Omega = \left\{B \in \mathbb{R}^{p\times r} : \text{rango}\,(B) \leq d\right\},$$

involucrado en la definición del espacio de parámetros del modelo PFC. Dado $B \in \Omega$, por la Proposición 2.1, igualdad (4), existen $U \in \text{St}^*\,(d, \mathbb{R}^p)$, $\Sigma_1 = \text{diag}\,(\sigma_1, \ldots, \sigma_d) \in \mathbb{R}^{d\times d}$ con $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$ y $V \in \text{St}^*\,(d, \mathbb{R}^r)$ tales que

$$B = U\Sigma_1 V^T. \tag{114}$$

Como comentamos luego del Teorema B.1 de descomposición en valores singulares, esta descomposción es única si asumimos que los valores singulares $\sigma_1, \ldots, \sigma_d$ de B son todos distintos y se adopta la convención de que la primer coordenada no nula de cada columna de U y V sea positiva. Luego, la función dada por

$$\begin{aligned} \phi : \Omega &\to \text{St}^*\,(d, \mathbb{R}^p) \times (0, +\infty)^d \times \text{St}^*\,(d, \mathbb{R}^r) \\ \phi\,(B) &= (U, \text{diag}\,(\Sigma_1), V) \end{aligned} \tag{115}$$

está bien definida y es biyectiva al restringirla al conjunto de matrices de Ω con todos sus valores singulares distintos. De acuerdo a lo anterior, resulta que la dimensión algebraica de Ω es

$$\begin{aligned} \dim(\Omega) &= \dim\left(\text{St}^*\,(d, \mathbb{R}^p)\right) + \dim\left((0, +\infty)^d\right) + \dim\left(\text{St}^*\,(d, \mathbb{R}^r)\right) \\ &= d\,(p-d) + \frac{d\,(d-1)}{2} + d + d\,(r-d) + \frac{d\,(d-1)}{2} \\ &= d\,(p+r-d)\,. \end{aligned} \tag{116}$$

Recordemos que si tomamos $\Gamma = U$ y $\beta = \Sigma_1 V^T$ se cumplen para estas matrices las restricciones impuestas y se recupera a B. Luego (115) es una caracterización de Ω.

Otra posibilidad es la siguiente. Para cada $B \in \Omega$, $V = \text{span}(B)$ es un elemento de $\text{Grass}(d, \mathbb{R}^p)$. Entonces existe una matriz $A \in \mathbb{R}^{p\times d}$ tal que $[A] = V = \mathcal{S}_A = \text{span}(A)$. Consideremos $P_{[A]} = A\left(A^T A\right)^{-1} A^T$ la proyección ortogonal (con la norma usual) en $\text{span}(A)$. La proyección no depende de los generadores de $\text{span}(A)$ elegidos, ya que tiene por columnas a $\left\{P_{[A]}\mathbf{e}_1, \ldots, P_{[A]}\mathbf{e}_p\right\}$. Sea

$$A_0 := \begin{bmatrix} P_{[A]}\mathbf{e}_{i_1} & \cdots & P_{[A]}\mathbf{e}_{i_d} \end{bmatrix}$$

donde $1 = i_1 < \cdots < i_d$ elegidos de modo tal que $\text{rango}(A_0) = d$ y si

$$A_1 := \begin{bmatrix} P_{[A]}\mathbf{e}_{j_1} & \cdots & P_{[A]}\mathbf{e}_{j_d} \end{bmatrix}$$

cumple $\text{rango}(A_1) = d$, entonces hay al menos un k entre 1 y d para el cual $j_k > i_k$. Es decir, en A_0 ponemos por columnas d vectores linealmente independientes que generan a $\mathcal{S}_A$, elegidos de modo de poner las primeras d columnas si fuera posible y sino (si estas fueran linealmente dependientes) comenzamos por $P_{[A]}\mathbf{e}_1$ y tomamos el menor l tal que el conjunto $\left\{P_{[A]}\mathbf{e}_1, P_{[A]}\mathbf{e}_l\right\}$ es linealmente independiente. Tomamos $i_2 = l$. Si $d > 2$, buscamos ahora el menor s con $i_2 < s$ de modo que $\left\{P_{[A]}\mathbf{e}_1, P_{[A]}\mathbf{e}_{i_2}, P_{[A]}\mathbf{e}_s\right\}$ sea linealmente independiente. Tomamos $i_3 = s$. Seguimos hasta definir i_d, de modo que $\text{rango}(A_0) = d$.

Esto define la siguiente función, que a cada subespacio le asigna una matriz que tiene por columnas a una base de dicho subespacio,

$$\begin{aligned} Gen &: \text{Grass}(d, \mathbb{R}^p) \to \text{St}(d, \mathbb{R}^p) \\ Gen(\mathcal{V}) &= Gen(\mathcal{S}_A) = Gen(\text{span}(A_0)) = A_0. \end{aligned}$$

Esto nos permite tener otra representación de Ω.

Como $A \in \mathbb{R}^{p\times d}$ y $A_0 := Gen(\mathcal{S}_A) \in \mathbb{R}^{p\times d}$ generan el mismo espacio, existe una matriz $B_0 \in \mathbb{R}^{d\times r}$ tal que $B = A_0 B_0$, luego podemos definir

$$\begin{aligned} \varphi &: \Omega \to \text{Grass}(d, \mathbb{R}^p) \times \mathbb{R}^{d\times r} \\ \varphi(B) &= (\text{span}(B), B_0) \end{aligned} \tag{117}$$

Por definición, esta aplicación resulta biyectiva, y es (una posible formalización de) la descripción introducida en Cook y Forzani [2008], por lo que (117) resulta ser otra caracterización alternativa de Ω. Con ella, podemos, por supuesto, repetir el cálculo de dimensiones para Ω :

$$\begin{aligned} \dim(\Omega) &= \dim(\text{Grass}(d, \mathbb{R}^p)) + \dim\left(\mathbb{R}^{d\times r}\right) \\ &= d(p-d) + dr \\ &= d(p+r-d). \end{aligned}$$

B.4. Caracterización alternativa de $\boldsymbol{\Theta}$

A partir de la descomposición de B dada en (114) si tomamos $\Gamma = U$ y $\beta = \Sigma_1 V^T$, entonces

$$\begin{aligned} &\Gamma \in \mathbb{R}^{p\times d}, \quad \Gamma^T\Gamma = I_d \\ &\beta \in \mathbb{R}^{d\times r}, \quad \beta\beta^T = \Sigma_1 V^T V \Sigma_1 = \Sigma_1\Sigma_1 = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_d^2\right), \\ &\text{con } \sigma_1^2 \geq \cdots \geq \sigma_d^2. \end{aligned}$$

Con estas herramientas, podemos construir la siguiente representación del espacio de parámetros del modelo PFC.

$$\boldsymbol{\Theta}_3 = \left\{ \begin{array}{c} (\boldsymbol{\mu}, \Gamma, \Sigma_1, V, \Delta) : \boldsymbol{\mu} \in \mathbb{R}^p, \Gamma \in \mathrm{St}^*(d, \mathbb{R}^p), V \in \mathrm{St}^*(d, \mathbb{R}^r), \Delta \in PDS(p), \\ \Sigma_1 \in \mathbb{R}^{d\times d} \text{ diagonal, con diagonal decreciente} \end{array} \right\}, \tag{118}$$

o, también

$$\boldsymbol{\Theta}^* = \left\{ \begin{array}{c} (\boldsymbol{\mu}, \Gamma, \beta, \Delta) : \boldsymbol{\mu} \in \mathbb{R}^p, \Gamma \in \mathbb{R}^{p\times d}, \beta \in \mathbb{R}^{d\times r}, \Delta \in PDS(p) : \Gamma^T\Gamma = I_d, \\ \beta\beta^T \text{ es diagonal, con diagonal decreciente} \end{array} \right\}, \tag{119}$$

ya que dado $\boldsymbol{\theta} \in \boldsymbol{\Theta}_3$ se puede asociar un único $\widetilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}^*$ y viceversa. Como comentamos en la Sección B.2, podríamos imponer otras condiciones de normalización en la definición de $\boldsymbol{\Theta}_3$ y $\boldsymbol{\Theta}^*$. En la Sección 8.2, cuando calculamos el funcional propuesto en este trabajo, las condiciones de normalización impuestas surgirán del criterio de estimación elegido. A esta altura nos basta notar que responden a una restricción de ortogonalidad sobre Γ y otra de diagonalización sobre β.

Otra posibilidad es describir a $\boldsymbol{\Theta}$ a través de una biyección entre Ω y el producto cartesiano $\mathrm{Grass}(d, \mathbb{R}^p) \times \mathbb{R}^{d\times r}$, que puede darse a través de la función φ definida en (117) o a través de cualquier otra biyección que se pueda definir entre estos espacios, dando lugar a la siguiente notación

$$\begin{aligned} \boldsymbol{\Theta}_5 &= \left\{ (\boldsymbol{\mu}, \mathcal{S}_\Gamma, \beta, \Delta) : \boldsymbol{\mu} \in \mathbb{R}^p, \mathcal{S}_\Gamma \in \mathrm{Grass}(d, \mathbb{R}^p), \beta \in \mathbb{R}^{d\times r}, \Delta \in PDS(p) \right\} \\ &= \left\{ (\boldsymbol{\mu}, \Gamma, \beta, \Delta) : \boldsymbol{\mu} \in \mathbb{R}^p, \operatorname{span}(\Gamma) \in \mathrm{Grass}(d, \mathbb{R}^p), \beta \in \mathbb{R}^{d\times r}, \Delta \in PDS(p) \right\}. \end{aligned}$$

Finalmente, podemos calcular la dimensión del espacio de parámetros $\boldsymbol{\Theta}$, a través de cualquiera de las caracterizaciones presentadas. En particular,

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta) \in \mathbb{R}^p \times \Omega \times PDS(p)\}.$$

luego su dimensión será

$$\begin{aligned} \dim(\boldsymbol{\Theta}) &= \dim(\mathbb{R}^p) + \dim(\Omega) + \dim(PDS(p)) \\ &= p + d(p + r - d) + \dim(PDS(p)) \end{aligned}$$

ya que la dimensión de Ω la calculamos en (116). Recordando que $A \in PDS(p)$ si y sólo si

$$A = U \operatorname{diag}(\lambda_1(A), \ldots, \lambda_p(A)) U^T$$

con $U \in \mathrm{St}^*(p, \mathbb{R}^p)$ y $\lambda_1(A) \geq \cdots \geq \lambda_p(A) > 0$ los autovalores (reales) de A. Luego su dimensión algebraica resulta ser

$$\dim(PDS(p)) = \dim(\mathrm{St}^*(p, \mathbb{R}^p)) + p = \frac{p(p+1)}{2}$$

y finalmente, $\dim(\boldsymbol{\Theta}) = \frac{p(p+3)}{2} + d(p + r - d)$.

También es de interés para nuestro problema definir una distancia entre los elementos del espacio de parámetros. Puede verse el Apéndice 9.3 para una discusión del tema.

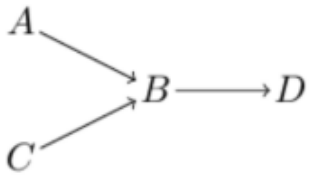
C. Apéndice del Capítulo 3

C.1. DAGs

Los grafos son una herramienta muy poderosa para descubrir, expresar y probar relaciones de independencia condicional entre las variables. Los introducimos en esta sección, siguiendo la presentación dada en Pearl [2009], Wasserman [2013] y Pearl [1988].

Un grafo dirigido $\mathcal{G}$ consiste en un conjunto de vértices V y un conjunto de aristas E, es decir, un conjunto de pares ordenados de vértices, $\mathcal{G} = \{V, E\}$. Para nuestros propósitos, cada vértice o nodo corresponderá a una variable aleatoria o vector. Si $(A, B) \in E$, entonces en el grafo hay una flecha que apunta desde A a B. Si hay una flecha de A y B entonces se dice que A es un **padre** de B y B es un **hijo** de A. Un **camino dirigido** entre dos variables es un conjunto de flechas, cada una apuntando en la misma dirección a lo largo del conjunto, que une una variable con la otra. Una secuencia de vértices adyacentes que comienza en A y termina en B ignorando la dirección de las flechas se llama **camino no dirigido**,o simplemente un **camino**. A es un **ancestro** de B si hay un camino dirigido de A a B. En ese caso también decimos que B es un **descendiente** de A. Una variable que aparece en el grafo con dos flechas apuntándola, como ocurre con B en el grafo $A \rightarrow B \leftarrow C$ se denomina un **colisionador** (*colider*). Ser (o no) colisionador es una propiedad del camino. En la figura 24, B es un colisionador para el camino $A \longrightarrow B \longleftarrow C$, pero B no es un colisionador para el camino de la ruta $A \longrightarrow B \longrightarrow D$.

Figura 24: DAG con un colisionador. En el camino $A \longrightarrow B \longleftarrow C$, B es un colisionador. En el camino $A \longrightarrow B \longrightarrow C$, B no es un colisionador.



Un camino dirigido que comienza y termina en la misma variable se denomina un ciclo. Un grafo dirigido se dice acíclico si no tiene ciclos. En este caso decimos que el grafo es un **grafo acíclico dirigido** (*directed acyclic graph*) ó DAG .

C.2. Probabilidad codificada por un DAG

Una flecha entre dos nodos, más específicamente, una flecha de A a B indica la posibilidad de conexión causal entre ellos: se lee "A causa o provoca a B". En términos estadísticos, una flecha $A \longrightarrow B$ codifica cualquier distribución de probabilidad conjunta que sea compatible con la existencia de una función f_0 tal que podemos escribir

$$B = f_0(A, \varepsilon_B)$$

donde ε_B es una variable error independiente de A. Por supuesto, si f_0 es una función que no depende de su primer argumento, entonces, de hecho, B no depende de A. Esta es la razón por la cual, en los DAGs, los supuestos causales quedan codificados no en las flechas dibujadas, sino en las flechas ausentes. Una flecha indica meramente la posibilidad de una conexión causal, una flecha ausente representa la afirmación de no influencia. Esto conduce a la siguiente definición.

Definición C.1 *Sea $\mathcal{G}$ un DAG con vértices $V = (X_1, \ldots, X_k)$. Si P es una distribución de probabilidad para V, decimos que es* ***compatible con*** *$\mathcal{G}$, o que $\mathcal{G}$ representa a P si*

$$P\left(\bigcap_{i=1}^{k}\{X_i \in A_i\}\right) = \prod_{i=1}^{k} P\left(X_i \in A_i \mid \bigcap_{j:X_j \in \pi(X_i)} \{X_j \in A_j\}\right)$$

para todos los conjuntos borelianos $A_1, \ldots, A_k$, donde $\pi(X_i)$ son los padres de X_i. Llamamos $P(\mathcal{G})$ al conjunto de todas las distribuciones de probabilidad compatibles con el grafo $\mathcal{G}$.

Esto significa que una variable en el grafo es independiente del "pasado" dados sus padres. Esto es usualmente denominada la **condición de Markov** de P.

Por ejemplo, para el grafo de la Figura 24, una densidad conjunta para este grafo debe factorizarse de la siguiente forma

$$f_{ABCD}(a, b, c, d) = f_A(a)\, f_C(c)\, f_{B|AC}(b \mid a, c)\, f_{D|B}(d \mid b).$$

La probabilidad que este grafo codifica también puede ser descripta a través del siguiente sistema de ecuaciones (que se denominan *ecuaciones estructurales*):

$$B = f_1(A, C, \varepsilon_B)$$
$$D = f_2(B, \varepsilon_D)$$

Aquí las variables ε_B y ε_D se denominan "exógenas". Representan factores de fondo observados o no observados que el modelador decide mantener no explicados, es decir, los factores que influyen, pero no están influenciados por otras variables (denominadas "endógenas") en el modelo. Las variables exógenas no observadas a veces se llaman "perturbaciones" o "errores", representan hechos omitidos por el modelo, pero que se juzgan relevantes para explicar el comportamiento de las variables incluídas en el modelo. Por

lo general son no observadas, y el modelador se resigna a reconocer su existencia y a asignarle un rol cualitativo a su vínculo con otras variables en el sistema. Se asume que las variables exógenas asociadas a cada variable del modelo son conjuntamente independientes.

La potencia de los modelos gráficos radica en que permiten expresar en forma simple las suposiciones hechas por el modelador acerca de las relaciones causales entre las variables. Y también le permite al analista deducir otras (in)dependencias implícitas en las suposiciones. Estas se puede resolver gráficamente a traves del criterio de $d-$separación, probado por Pearl [1988]. La d significa "directa".

Definición C.2 ($d-$separación) *Sean X e Y dos nodos de un DAG $\mathcal{G}$, y sea S un subconjunto de nodos que no contiene ni a X ni a Y. Decimos que S* ***bloquea (o $d-$separa) un camino*** *p que conecta a X con Y si*

i. p contiene al menos un nodo que emite una flecha que está en S, o bien

ii. p contiene al menos un colisionador que está fuera de S y que no tiene descendientes en S.

Si S bloquea ***todos*** *los caminos de X a Y, decimos que $d-$**separa a** X **e** Y.*

Sean A, B y S tres subconjuntos distintos de nodos de $\mathcal{G}$, con A y B no vacíos. Diremos que A y B están $d-$separados por S si para todo $X \in A$ e $Y \in B$, X e Y están $d-$separados por S.

Citemos a Pearl, "la interpretación del bloqueo es que detiene el flujo de la información (o dependencia) entre las variables que están conectadas por dichos caminos". El teorema que sigue, debido a Verma y Pearl [1988], y Verma y Pearl [1990] conecta la $d-$separación con la independencia condicional.

Teorema C.1 (Implicaciones probabilísticas de la $d-$separación) *Sean A, B y S conjuntos disjuntos de vértices en un grafo $\mathcal{G}$. Tenemos la siguiente equivalencia: $A \perp\!\!\!\perp B \mid S$ para toda $P \in P(\mathcal{G})$ si y sólo si A y B están $d-$separados por S en el DAG $\mathcal{G}$.*

En la Figura 24, $S = \emptyset$ bloquea el camino $A \longrightarrow B \longleftarrow C$ porque B es un colisionador en el camino pero no está en S y su descendiente D no está en S. Entonces $A \perp\!\!\!\perp C$. $S = \{B\}$ bloquea el camino $A \longrightarrow B \longrightarrow D$ porque B emite una flecha a lo largo del camino. A y D son variables dependientes pero se vuelven independientes (i. e. bloqueadas) cuando condicionamos a (i. e. conocemos el valor de) B. En simbolos, $A \perp\!\!\!\perp D \mid B$. Pero $S = \{B\}$ no bloquea el camino $A \longrightarrow B \longleftarrow C$. Este camino representa la situación en que dos causas independientes, A y C tienen un efecto común. Cuando condicionamos en (i. e. conocemos el valor de) este efecto común B (o cualquiera de sus descendientes), estas dos variables se vuelven dependientes: conocer el valor de una consecuencia común usualmente hace que las variables originales se vuelvan (condicionalmente) dependientes. Entonces, en general, A y C no serán independientes, cuando condicionemos en B. Condicionar en un

colisionador, abre un camino entre los ancestros. Lo mismo ocurre si condicionamos en un descendiente de un colisionador: en la Figura 24, el camino $A \longrightarrow B \longleftarrow C$ no está bloqueado por $S = \{D\}$. Entonces, en general, A y C no serán independientes cuando condicionemos en D.

Demostración del Teorema 3.1. Bajo el modelo PFC tenemos

$$\mathbf{x} = \boldsymbol{\mu}_0 + \Gamma_0 \boldsymbol{\beta}_0 \mathbf{f}(y) + \Delta_0^{1/2} \mathbf{u}.$$

Sea $\Gamma_1 \in \mp_1$, luego, por (10) vale que

$$\begin{aligned} p_{\Delta_0}(\mathbf{x}, \operatorname{span}(\Gamma_0)) &= \Gamma_0 \Gamma_0^T \Delta_0^{-1} \mathbf{x} = \Gamma_0 \mathbf{R}(\mathbf{x}) \\ p_{\Delta_0}(\mathbf{x}, \operatorname{span}(\Gamma_1)) &= \Gamma_1 \Gamma_1^T \Delta_0^{-1} \mathbf{x}, \\ \mathbf{x} &= p_{\Delta_0}(\mathbf{x}, \operatorname{span}(\Gamma_0)) + p_{\Delta_0}(\mathbf{x}, \operatorname{span}(\Gamma_1)) \end{aligned}$$

$$\begin{aligned} \mathbf{R}(\mathbf{x}) = \Gamma_0^T \Delta_0^{-1} \mathbf{x} &= \Gamma_0^T \Delta_0^{-1} \boldsymbol{\mu}_0 + \Gamma_0^T \Delta_0^{-1} \Gamma_0 \boldsymbol{\beta}_0 \mathbf{f}(y) + \Gamma_0^T \Delta_0^{-1/2} \mathbf{u} \\ &= \boldsymbol{\beta}_0 \mathbf{f}(y) + \Gamma_0^T \Delta_0^{-1} \left(\boldsymbol{\mu}_0 + \Delta_0^{1/2} \mathbf{u} \right) \end{aligned} \tag{120}$$

de modo que

$$\begin{aligned} \mathbf{x} &= p_{\Delta_0}(\mathbf{x}, \operatorname{span}(\Gamma_0)) + p_{\Delta_0}(\mathbf{x}, \operatorname{span}(\Gamma_1)) \\ &= \Gamma_0 \Gamma_0^T \Delta_0^{-1} \mathbf{x} + \Gamma_1 \Gamma_1^T \Delta_0^{-1} \mathbf{x} \\ &= \Gamma_0 \mathbf{R}(\mathbf{x}) + \Gamma_1 \Gamma_1^T \Delta_0^{-1} \left(\boldsymbol{\mu}_0 + \Gamma_0 \boldsymbol{\beta}_0 \mathbf{f}(y) + \Delta_0^{1/2} \mathbf{u} \right) \\ &= \Gamma_0 \mathbf{R}(\mathbf{x}) + \Gamma_1 \Gamma_1^T \Delta_0^{-1} \left(\boldsymbol{\mu}_0 + \Delta_0^{1/2} \mathbf{u} \right) \\ &= \Gamma_0 \boldsymbol{\beta}_0 \mathbf{f}(y) + \Gamma_0 \Gamma_0^T \Delta_0^{-1} \left(\boldsymbol{\mu}_0 + \Delta_0^{1/2} \mathbf{u} \right) + \Gamma_1 \Gamma_1^T \Delta_0^{-1} \left(\boldsymbol{\mu}_0 + \Delta_0^{1/2} \mathbf{u} \right) \end{aligned} \tag{121}$$

Recordemos que $\mathbf{u}$ e y son independientes. La descomposición de $\mathbf{x}$ establecida en (121) y (120) nos permite dibujar un DAG (ver la Sección C.1 para una presentación de los DAGs en relación con los modelos estocásticos) para describir la dependecia entre las variables involucradas en el modelo PFC. En la Figura 25, mostramos el DAG sin ningun supuesto extra. En este gráfico, $\mathbf{u}$ juega el rol de una variable exógena.

Asumimos que

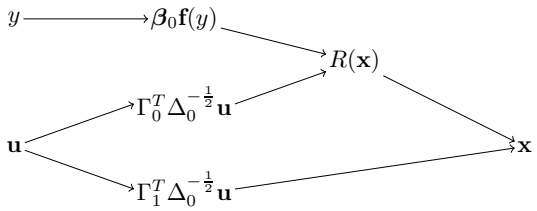
$$p_{\Delta_0}\left(\Delta_0^{1/2} \mathbf{u}, \operatorname{span}(\Gamma_0) \right) = \Gamma_0 \Gamma_0^T \Delta_0^{-1/2} \mathbf{u}$$

y

$$p_{\Delta_0}\left(\Delta_0^{1/2} \mathbf{u}, \operatorname{span}(\Gamma_1) \right) = \Gamma_1 \Gamma_1^T \Delta_1^{1/2} \mathbf{u}$$

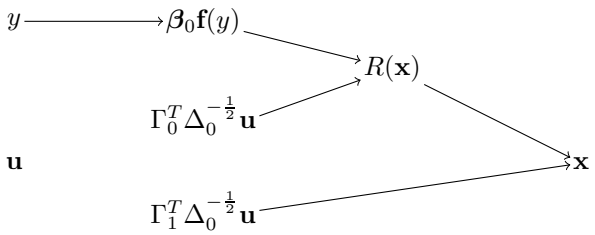
son independientes. Equivalentemente, $\Gamma_0^T \Delta_0^{-1/2} \mathbf{u}$ y $\Gamma_1^T \Delta_0^{-1/2} \mathbf{u}$ son independientes. Traducimos esta información a un DAG en el que borramos las flechas que emanan de $\mathbf{u}$ en la Figura 25. El nuevo grafo se puede observar en la Figura 26.

Figura 25: DAG que describe la relación de dependencia entre las variables del modelo PFC.



En este grafo, $S = \{\mathbf{R}(\mathbf{x})\}$ bloquea el único camino que conecta a y con $\mathbf{x} : y \longrightarrow \boldsymbol{\beta}_0\mathbf{f}(y) \longrightarrow \mathbf{R}(\mathbf{x}) \longrightarrow \mathbf{x}$, luego, por el Criterio de $d-$separación (Teorema C.1), $\mathbf{x} \perp\!\!\!\perp y \mid \mathbf{R}(\mathbf{x})$, y por lo tanto, $\mathbf{R}(\mathbf{x})$ es una reducción suficiente. ■

Figura 26: DAG que describe la dependencia entre las variables aleatorias involucradas en el modelo PFC asumiendo independencia entre $p_{\Delta_0}\left(\Delta_0^{1/2}\mathbf{u}, \text{span}\left(\Gamma_0\right)\right)$ y $p_{\Delta_0}\left(\Delta_0^{1/2}\mathbf{u}, \text{span}\left(\Gamma_1\right)\right)$.



D. Apéndice del Capítulo 4

D.1. Demostraciones de máxima verosimilitud

Demostración del Teorema 4.1 (d). Observemos que dados $\widehat{\boldsymbol{\mu}}_i = \widehat{\boldsymbol{\mu}} + \widehat{\Gamma}\widehat{\beta}\mathbf{f}(y_i)$ (definimos a $\boldsymbol{\mu}_i$ en la ecuación (17)) para $1 \leq i \leq n$, el estimador de máxima verosimilitud de Δ es el estimador de máxima verosimilitud de Δ cuando queremos ajustar una distribución normal $\mathrm{N}_p(\mathbf{0},\Delta)$ a las observaciones $\mathbf{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{\Gamma}\widehat{\beta}\mathbf{f}(y_i)$, lo cual justifica la expresión (19).

Finalmente, para probar (20) observemos que

$$\begin{aligned}
p &= \operatorname{traza}(I_p) = \operatorname{traza}\left(\widehat{\Delta}\widehat{\Delta}^{-1}\right) \\
&= \operatorname{traza}\left(\left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{\Gamma}\widehat{\beta}\mathbf{f}(y_i)\right)\left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{\Gamma}\widehat{\beta}\mathbf{f}(y_i)\right)^T\right)\widehat{\Delta}^{-1}\right) \\
&= \operatorname{traza}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{\Gamma}\widehat{\beta}\mathbf{f}(y_i)\right)\left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{\Gamma}\widehat{\beta}\mathbf{f}(y_i)\right)^T\widehat{\Delta}^{-1}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\operatorname{traza}\left(\left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{\Gamma}\widehat{\beta}\mathbf{f}(y_i)\right)\left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{\Gamma}\widehat{\beta}\mathbf{f}(y_i)\right)^T\widehat{\Delta}^{-1}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\operatorname{traza}\left(\left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{\Gamma}\widehat{\beta}\mathbf{f}(y_i)\right)^T\widehat{\Delta}^{-1}\left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{\Gamma}\widehat{\beta}\mathbf{f}(y_i)\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{\Gamma}\widehat{\beta}\mathbf{f}(y_i)\right)^T\widehat{\Delta}^{-1}\left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{\Gamma}\widehat{\beta}\mathbf{f}(y_i)\right),
\end{aligned}$$

lo cual prueba (d). ■

Demostración del Lema 4.1. Debemos mostrar que el mínimo de ambos problemas se alcanza en el mismo valor. Sea $\widehat{\boldsymbol{\theta}}_c = \left(\widehat{\boldsymbol{\mu}}_c, \widehat{\Gamma}_c, \widehat{\beta}_c, \widehat{\Delta}_c\right)$ el mínimo de $g(\boldsymbol{\theta})$ sujeto a $h(\boldsymbol{\theta}) = c$. Esto implica que

$$l\left(\widehat{\boldsymbol{\theta}}_c\right) = g\left(\widehat{\boldsymbol{\theta}}_c\right) + c.$$

Probemos que $\widehat{\boldsymbol{\theta}}_p = \left(\widehat{\boldsymbol{\mu}}_p, \widehat{\Gamma}_p, \widehat{\beta}_p, \widehat{\Delta}_p\right)$ es el EMV.

Valen las siguientes dos propiedades para $k > 0$, que se derivan de las definiciones de las funciones g y h.

(a) $g(\boldsymbol{\mu}, \Gamma, \beta, k\Delta) = p\log k + g(\boldsymbol{\mu}, \Gamma, \beta, \Delta)$

(b) $h(\boldsymbol{\mu}, \Gamma, \beta, k\Delta) = \dfrac{1}{k}h(\boldsymbol{\mu}, \Gamma, \beta, \Delta)$.

Entonces

$$l\left(\widehat{\boldsymbol{\theta}}_p\right) = g\left(\widehat{\boldsymbol{\theta}}_p\right) + p \leq g\left(\widehat{\boldsymbol{\mu}}_c, \widehat{\Gamma}_c, \widehat{\beta}_c, \frac{c}{p}\widehat{\Delta}_c\right) + p \tag{122}$$

puesto que $h\left(\widehat{\boldsymbol{\mu}}_p, \widehat{\Gamma}_p, \widehat{\beta}_p, \widehat{\Delta}_p\right) = p$ y también

$$h\left(\widehat{\boldsymbol{\mu}}_c, \widehat{\Gamma}_c, \widehat{\beta}_c, \frac{c}{p}\widehat{\Delta}_c\right) = \frac{p}{c} h\left(\widehat{\boldsymbol{\mu}}_c, \widehat{\Gamma}_c, \widehat{\beta}_c, \widehat{\Delta}_c\right) = p,$$

por la propiedad (a) enunciada más arriba. Entonces (122) sigue de la definición de $\left(\widehat{\boldsymbol{\mu}}_p, \widehat{\Gamma}_p, \widehat{\beta}_p, \widehat{\Delta}_p\right)$. También

$$g\left(\widehat{\boldsymbol{\mu}}_c, \widehat{\Gamma}_c, \widehat{\beta}_c, \frac{c}{p}\widehat{\Delta}_c\right) + p = p \log \frac{c}{p} + g\left(\widehat{\boldsymbol{\theta}}_c\right) + p.$$

Pero $r(p) = p \log c - p \log p + p$ alcanza un máximo para $p = c$, donde toma el valor c, por lo que

$$g\left(\widehat{\boldsymbol{\mu}}_c, \widehat{\Gamma}_c, \widehat{\beta}_c, \frac{c}{p}\widehat{\Delta}_c\right) + p = g\left(\widehat{\boldsymbol{\theta}}_c\right) + r\left(p\right) \leq g\left(\widehat{\boldsymbol{\theta}}_c\right) + c.$$

Luego,

$$\begin{aligned} l\left(\widehat{\boldsymbol{\theta}}_p\right) &= g(\widehat{\boldsymbol{\theta}}_p) + h\left(\widehat{\boldsymbol{\theta}}_p\right) \\ &\leq g(\widehat{\boldsymbol{\theta}}_c) + h\left(\widehat{\boldsymbol{\theta}}_c\right) \\ &= l\left(\widehat{\boldsymbol{\theta}}_c\right) \end{aligned} \tag{123}$$

para todo $c > 0$. Sea $\boldsymbol{\theta} \in \Theta_5$ cualquiera. Con probabilidad uno, $h\left(\boldsymbol{\theta}\right) \in (0, +\infty)$. Llamemos c_1 al valor de $h\left(\boldsymbol{\theta}\right)$ (que es aleatorio). Por un lado, tenemos que

$$l\left(\widehat{\boldsymbol{\theta}}_{c_1}\right) \leq l\left(\boldsymbol{\theta}\right)$$

y por otro, vale (123), luego

$$l\left(\widehat{\boldsymbol{\theta}}_p\right) \leq l\left(\boldsymbol{\theta}\right),$$

probando que $\widehat{\boldsymbol{\theta}}_p$ es el EMV.

Recíprocamente, sea $\widehat{\boldsymbol{\theta}}_{MV}$ el EMV. Por el Teorema 4.1 (d) sabemos que $h\left(\widehat{\boldsymbol{\theta}}_{MV}\right) = p$. Por ser el EMV para el problema PFC sabemos que para todo $\boldsymbol{\theta} \in \Theta_5$ vale

$$l\left(\widehat{\boldsymbol{\theta}}_{MV}\right) \leq l\left(\boldsymbol{\theta}\right).$$

Sea $\boldsymbol{\theta} \in \Theta_5$ tal que $h\left(\boldsymbol{\theta}\right) = p$. Luego, tenemos

$$\begin{aligned} l\left(\widehat{\boldsymbol{\theta}}_{MV}\right) &= g\left(\widehat{\boldsymbol{\theta}}_{MV}\right) + h\left(\widehat{\boldsymbol{\theta}}_{MV}\right) = g\left(\widehat{\boldsymbol{\theta}}_{MV}\right) + p \\ &\leq l\left(\boldsymbol{\theta}\right) = g\left(\boldsymbol{\theta}\right) + h\left(\boldsymbol{\theta}\right) = g\left(\boldsymbol{\theta}\right) + p, \end{aligned}$$

luego $g\left(\widehat{\boldsymbol{\theta}}_{MV}\right) \leq g\left(\boldsymbol{\theta}\right)$, probando el resultado. ■

E. Apéndice del Capítulo 5

E.1. Demostraciones de equivalencias entre funcionales

Demostración del Lema 5.1.

i. y ii. Por definición de S_M, tenemos

$$\begin{aligned}
\kappa_1 &= E_{H_{P,\boldsymbol{\theta}}}\left(\rho_1\left(\frac{v}{S_M\left(H_{P,\boldsymbol{\theta}}\right)}\right)\right) \\
&= E_P\left(\rho_1\left(\frac{\left\{\left(\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y)\right)^T\Delta^{-1}\left(\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y)\right)\right\}^{1/2}}{S_M\left(H_{P,\boldsymbol{\theta}}\right)}\right)\right) \\
&= E_P\left(\rho_1\left(\frac{\left\{\left(\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y)\right)^T\left(\Delta\lambda\right)^{-1}\left(\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y)\right)\right\}^{1/2}}{\frac{1}{\sqrt{\lambda}}S_M\left(H_{P,\boldsymbol{\theta}}\right)}\right)\right) \\
&= E_{H_{P,\boldsymbol{\theta}_\lambda}}\left(\rho_1\left(\frac{v}{\frac{1}{\sqrt{\lambda}}S_M\left(H_{P,\boldsymbol{\theta}}\right)}\right)\right).
\end{aligned}$$

Luego, tomando $\sqrt{\lambda} = S_M\left(H_{P,\boldsymbol{\theta}}\right)$ resulta que $\boldsymbol{\theta}_\lambda$ cumple la condición (40), probando (i). Además, resulta

$$\frac{1}{\sqrt{\lambda}}S_M\left(H_{P,\boldsymbol{\theta}}\right) = S_M\left(H_{P,\boldsymbol{\theta_\lambda}}\right)$$

o, equivalentemente

$$S_M^2\left(H_{P,\boldsymbol{\theta}}\right) = \lambda S_M^2\left(H_{P,\boldsymbol{\theta_\lambda}}\right),$$

probando (ii).

iii.

$$\begin{aligned}
&S_\tau^2\left(H_{P,\boldsymbol{\theta_\lambda}}\right) \\
&= S_M^2\left(H_{P,\boldsymbol{\theta_\lambda}}\right)E_{H_{P,\boldsymbol{\theta_\lambda}}}\left(\rho_2\left(\frac{v}{S_M\left(H_{P,\boldsymbol{\theta_\lambda}}\right)}\right)\right) \\
&= \frac{1}{\lambda}S_M^2\left(H_{P,\boldsymbol{\theta}}\right) \\
&\cdot E_P\left(\rho_2\left(\frac{\left\{\left(\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y)\right)^T\left(\lambda\Delta\right)^{-1}\left(\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y)\right)\right\}^{1/2}}{S_M\left(H_{P,\boldsymbol{\theta_\lambda}}\right)}\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\lambda} S_M^2 \left(H_{P,\boldsymbol{\theta}} \right) \\
&\cdot E_P \left(\rho_2 \left(\frac{\left\{ (\mathbf{x} - \boldsymbol{\mu} - B\mathbf{f}(y))^T \Delta^{-1} (\mathbf{x} - \boldsymbol{\mu} - B\mathbf{f}(y)) \right\}^{1/2}}{\sqrt{\lambda} S_M \left(H_{P,\boldsymbol{\theta}_{\boldsymbol{\lambda}}} \right)} \right) \right) \\
&= \frac{1}{\lambda} S_M^2 \left(H_{P,\boldsymbol{\theta}} \right) E_{H_{P,\boldsymbol{\theta}}} \left(\rho_2 \left(\frac{v}{\sqrt{\lambda} S_M \left(H_{P,\boldsymbol{\theta}_{\boldsymbol{\lambda}}} \right)} \right) \right) \\
&= \frac{1}{\lambda} S_M^2 \left(H_{P,\boldsymbol{\theta}} \right) E_{H_{P,\boldsymbol{\theta}}} \left(\rho_2 \left(\frac{v}{S_M \left(H_{P,\boldsymbol{\theta}} \right)} \right) \right) \\
&= \frac{1}{\lambda} S_\tau^2 \left(H_{P,\boldsymbol{\theta}} \right). \ \blacksquare
\end{aligned}
$$

Demostración del Lema 5.2. Primero veamos que $\boldsymbol{\theta}_A(P) = (\boldsymbol{\mu}_A(P), B_A(P), \Delta_A(P))$ minimiza a

$$\Phi_{B,P}(\boldsymbol{\theta}) = |\Delta| \left[S_\tau^2 \left(H_{P,\boldsymbol{\theta}} \right) \right]^p.$$

Por definición sabemos que satisface (39). Luego, para todo $\boldsymbol{\theta}$ que cumpla (39) tenemos

$$\Phi_{B,P}(\boldsymbol{\theta}_A(P)) = |\Delta_A(P)| \left[S_\tau^2 \left(H_{P,\boldsymbol{\theta}_A(P)} \right) \right]^p = |\Delta_A(P)| [\kappa_2]^p \leq |\Delta| [\kappa_2]^p = \Phi_{B,P}(\boldsymbol{\theta}).$$

En segundo lugar, veamos que un múltiplo de $\boldsymbol{\theta}_B(P)$ resuelve el problema que define a $\widetilde{\boldsymbol{\theta}}_C(P)$. Sea $a = S_M^2 \left(H_{P,\boldsymbol{\theta}_B(P)} \right)$. Entonces, como consecuencia del Lema 5.1(i) y (ii), $(\boldsymbol{\theta}_B(P))_a$ cumple la restricción (43). Además, para todo $\boldsymbol{\theta}$ con $S_M^2(H_{P,\boldsymbol{\theta}}) = 1$ se tiene

$$
\begin{aligned}
\Phi_{C,P}(\boldsymbol{\theta}) &= |\Delta| \left[E_{H_{P,\boldsymbol{\theta}}} (\rho_2(v)) \right]^p = |\Delta| \left[S_M^2 \left(H_{P,\boldsymbol{\theta}} \right) \right]^p \left[E_{H_{P,\boldsymbol{\theta}}} \left(\rho_2 \left(\frac{v}{S_M^2 \left(H_{P,\boldsymbol{\theta}} \right)} \right) \right) \right]^p \\
&= \Phi_{B,P}(\boldsymbol{\theta}) = \Phi_{B,P}(\boldsymbol{\theta}_\lambda),
\end{aligned}
\tag{124}
$$

donde la última igualdad vale para todo $\lambda > 0$. Luego, para todo $\boldsymbol{\theta}$ que cumple (43) se tiene

$$\Phi_{C,P}(\boldsymbol{\theta}) = \Phi_{B,P}(\boldsymbol{\theta}) \geq \Phi_{B,P}(\boldsymbol{\theta}_B(P)) = \Phi_{B,P}((\boldsymbol{\theta}_B(P))_a) = \Phi_{C,P}((\boldsymbol{\theta}_B(P))_a),$$

probando lo requerido. Entonces, multiplicando a $\Delta_B(P)$ por una constante adecuada se conseguirá a $\boldsymbol{\theta}_C(P)$.

En tercer lugar, veamos que un múltiplo de $\boldsymbol{\theta}_C(P) = (\boldsymbol{\mu}_C(P), B_C(P), \Delta_C(P))$ resuelve el problema que define a $\widetilde{\boldsymbol{\theta}}_D(P)$. Observemos que $(\boldsymbol{\theta}_C(P))_{\frac{1}{|\Delta_C(P)|^{1/p}}}$ satisface (44). Sea $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ que también cumple (44). Queremos probar que

$$S_\tau^2 \left(H_{P,(\boldsymbol{\theta}_C(P))_{\frac{1}{|\Delta_C(P)|^{1/p}}}} \right) \leq S_\tau^2 \left(H_{P,\boldsymbol{\theta}} \right). \tag{125}$$

Sean $\lambda = S_M^2\left(H_{P,\boldsymbol{\theta}_C(P)}\right)$ y $w = S_M^2\left(H_{P,\boldsymbol{\theta}}\right)$, sabemos que tanto $\left(\boldsymbol{\theta}_C\left(P\right)\right)_\lambda$ como $\left(\boldsymbol{\theta}\right)_w$ cumplen (43). Por lo tanto

$$\Phi_{C,P}\left(\left(\boldsymbol{\theta}_C\left(P\right)\right)_\lambda\right) \leq \Phi_{C,P}\left(\left(\boldsymbol{\theta}\right)_w\right). \tag{126}$$

Pero, por un lado,

$$\begin{aligned}
&\Phi_{C,P}\left(\left(\boldsymbol{\theta}_C\left(P\right)\right)_\lambda\right) \\
&= \left|\lambda\Delta_C\left(P\right)\right|\left[E_{H_{P,\left(\boldsymbol{\theta}_C(P)\right)_\lambda}}\left(\rho_2\left(v\right)\right)\right]^p \\
&= \left|\Delta_C\left(P\right)\right|\left[S_M^2\left(H_{P,\boldsymbol{\theta}_C(P)}\right)\right]^p \\
&\cdot\left[E_P\left(\rho_2\left(\frac{\left\{\left(\mathbf{x}-\boldsymbol{\mu}_C\left(P\right)-B_C\left(P\right)\mathbf{f}(y)\right)^T\left[\Delta\left(P\right)\right]^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_C\left(P\right)-B_C\left(P\right)\mathbf{f}(y)\right)\right\}^{1/2}}{S_M\left(H_{P,\boldsymbol{\theta}_C(P)}\right)}\right)\right)\right]^p \\
&= \left|\Delta_C\left(P\right)\right|\left[S_\tau^2\left(H_{P,\left(\boldsymbol{\theta}_C(P)\right)}\right)\right]^p \\
&= \left[S_\tau^2\left(H_{P,\left(\boldsymbol{\theta}_C(P)\right)\frac{1}{\left|\Delta_C(P)\right|^{1/p}}}\right)\right]^p
\end{aligned} \tag{127}$$

y por el otro

$$\begin{aligned}
\Phi_{C,P}\left(\left(\boldsymbol{\theta}\right)_w\right) &= \left|w\Delta\right|\left[E_{H_{P,\boldsymbol{\theta}_w}}\left(\rho_2\left(v\right)\right)\right]^p \\
&= \left|\Delta\right|\left[S_M^2\left(H_{P,\boldsymbol{\theta}}\right)\right]^p\left[E_P\left(\rho_2\left(\frac{\left\{\left(\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y)\right)^T\Delta^{-1}\left(\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y)\right)\right\}^{1/2}}{S_M\left(H_{P,\boldsymbol{\theta}}\right)}\right)\right)\right]^p \\
&= \left[S_\tau^2\left(H_{P,\boldsymbol{\theta}}\right)\right]^p
\end{aligned} \tag{128}$$

A partir de (126), (127) y (128) se obtiene (125), probando el resultado.

Finalmente, queremos ver que $\boldsymbol{\theta}_D\left(P\right) = \left(\boldsymbol{\mu}_D\left(P\right), B_D\left(P\right), \Delta_D\left(P\right)\right)$ resuelve el problema que define a $\boldsymbol{\theta}_A\left(P\right)$. Sea $\boldsymbol{\theta}\in\boldsymbol{\Theta}$. Luego, $\left(\boldsymbol{\theta}_D\left(P\right)\right)_{\frac{1}{\left|\Delta_D(P)\right|^{1/p}}}$ y $\left(\boldsymbol{\theta}\right)_{\frac{1}{\left|\Delta\right|^{1/p}}}$ cumplen (44) y por lo tanto,

$$S_\tau^2\left(H_{P,\left(\boldsymbol{\theta}_D(P)\right)\frac{1}{\left|\Delta_D(P)\right|^{1/p}}}\right) \leq S_\tau^2\left(H_{P,\left(\boldsymbol{\theta}\right)\frac{1}{\left|\Delta\right|^{1/p}}}\right).$$

Por el Lema 5.1(iii) esto equivale a

$$\left|\Delta_D\left(P\right)\right|^{1/p} S_\tau^2\left(H_{P,\boldsymbol{\theta}_D(P)}\right) \leq \left|\Delta\right|^{1/p} S_\tau^2\left(H_{P,\boldsymbol{\theta}}\right).$$

Si $\boldsymbol{\theta}$ satisface (39), es decir, $S_\tau^2\left(H_{P,\boldsymbol{\theta}}\right) = \kappa_2$, entonces, como $\boldsymbol{\theta}_D\left(P\right)$ también satisface (39), la última desigualdad equivale a

$$\left|\Delta_D\left(P\right)\right| \leq \left|\Delta\right|,$$

lo cual prueba que $\boldsymbol{\theta}_D\left(P\right)$ resuelve el problema que define a $\boldsymbol{\theta}_A\left(P\right)$. ■

E.2. Medida inducida

Sean $(\Omega, \mathcal{B})$ y $(\Omega', \mathcal{B}')$ espacios de probabilidad, $\mathcal{B}$ y $\mathcal{B}'$ son $\sigma-$álgebras de conjuntos de Ω y Ω', respectivamente. Se dice que una función $h : \Omega \to \Omega'$ es medible $(\mathcal{B}, \mathcal{B}')$ si la preimagen $h^{-1}(M')$ pertenece a $\mathcal{B}$ para cada M' de $\mathcal{B}'$. Sea P una medida de probabilidad definida en $(\Omega, \mathcal{B})$ y h una función $(\mathcal{B}, \mathcal{B}')$ medible. $P \circ h^{-1} (= Ph^{-1})$ denota la medida de probabilidad definida en $(\Omega', \mathcal{B}')$ dada por $\left(P \circ h^{-1}\right)(M') = P\left(h^{-1}(M')\right)$ para todo $M' \in \mathcal{B}'$. Sea $h : \Omega' \to \mathbb{R}$ medible $\mathcal{B}'$, entonces la función real $f \circ h$ definida en Ω es medible $\mathcal{B}$. Ver el Billingsley [1968] para una prueba del Teorema de Cambio de Variables, que recordamos a continuación.

Teorema E.1 (Teorema de Cambio de Variables) *La función f es integrable con respecto a la medida $P \circ h^{-1}$ si y sólo si la función real $f \circ h$ es integrable con respecto a P, en cuyo caso tenemos*

$$\int_{h^{-1}(M')} f(h(w))\, dP(w) = \int_{M'} f\left(w'\right) d\left(P \circ h^{-1}\right)\left(w'\right)$$

para todo $M' \in \mathcal{B}'$.

Demostración del Teorema 5.1. Consideremos la función $h : \mathbb{R}^{p+1} \to \mathbb{R}^{p+1}$, dada por $h(\mathbf{x},y) = (A\mathbf{x} + \mathbf{b}, y) = (\mathbf{x}^*, y^*)$ y la transformación en el espacio de parámetros dada por $W : \boldsymbol{\Theta} \to \boldsymbol{\Theta}$, $W(\boldsymbol{\theta}) = \left(A\boldsymbol{\mu} + \mathbf{b}, AB, A\Delta A^T\right)$. Luego, para toda f función real integrable tenemos

$$\begin{aligned}
&E_{H_{P^*,\boldsymbol{\theta}^*}}(f(v)) \\
&= E_{P^*}(f(d((\mathbf{x}^*, y^*), \boldsymbol{\theta}^*))) \\
&= \int f\left(\left\{(\mathbf{x}^* - \boldsymbol{\mu}^* - B^*\mathbf{f}(y^*))^T (\Delta^*)^{-1} (\mathbf{x}^* - \boldsymbol{\mu}^* - B^*\mathbf{f}(y^*))\right\}^{1/2}\right) dP^*(\mathbf{x}^*, y^*) \\
&= \int f\left(\left\{(\mathbf{x}^* - \boldsymbol{\mu}^* - B^*\mathbf{f}(y^*))^T (\Delta^*)^{-1} (\mathbf{x}^* - \boldsymbol{\mu}^* - B^*\mathbf{f}(y^*))\right\}^{1/2}\right) d\left(P \circ h^{-1}\right)(\mathbf{x}^*, y^*) \\
&= \int f\left(\left\{(A\mathbf{x} + \mathbf{b} - \boldsymbol{\mu}^* - B^*\mathbf{f}(y))^T (\Delta^*)^{-1} (A\mathbf{x} + \mathbf{b} - \boldsymbol{\mu}^* - B^*\mathbf{f}(y))\right\}^{1/2}\right) dP(\mathbf{x}, y) \\
&= \int f\left(\left\{\left(\mathbf{x} - A^{-1}(\boldsymbol{\mu}^* - \mathbf{b}) - A^{-1}B^*\mathbf{f}(y)\right)^T A^T (\Delta^*)^{-1} A\right.\right. \\
&\left.\left.\left(\mathbf{x} - A^{-1}(\boldsymbol{\mu}^* - \mathbf{b}) - A^{-1}B^*\mathbf{f}(y)\right)\right\}^{1/2}\right) dP(\mathbf{x}, y) \\
&= E_P\left(f\left(d\left((\mathbf{x}, y), W^{-1}(\boldsymbol{\theta}^*)\right)\right)\right) \\
&= E_{H_{P,W^{-1}(\boldsymbol{\theta}^*)}}(f(v)).
\end{aligned}$$

donde $W^{-1}(\boldsymbol{\theta}) = \left(A^{-1}(\boldsymbol{\mu} - \mathbf{b}), A^{-1}B, A^{-1}\Delta\left(A^{-1}\right)^T\right)$, por el Teorema de cambio de variables (ver el Apéndice E, página 154). En particular, esta igualdad para toda función

f integrable, implica que las dos distribuciones coinciden, es decir, $H_{P^*,\boldsymbol{\theta}^*} = H_{P,W^{-1}(\boldsymbol{\theta}^*)}$. Luego, $S_M^2\left(H_{P^*,\boldsymbol{\theta}^*}\right) = S_M^2\left(H_{P,W^{-1}(\boldsymbol{\theta}^*)}\right)$ y $S_\tau^2\left(H_{P^*,\boldsymbol{\theta}^*}\right) = S_\tau^2\left(H_{P,W^{-1}(\boldsymbol{\theta}^*)}\right)$. Luego

$$\begin{aligned}
\Phi_{B,P^*}\left(\boldsymbol{\theta}^*\right) &= \left|\Delta^*\right|\left[S_\tau^2\left(H_{P^*,\boldsymbol{\theta}^*}\right)\right]^p = \left|\Delta^*\right|\left[S_\tau^2\left(H_{P,W^{-1}(\boldsymbol{\theta}^*)}\right)\right]^p && (129)\\
&= |A|^2\left|A^{-1}\Delta^*\left(A^{-1}\right)^T\right|\left[S_\tau^2\left(H_{P,W^{-1}(\boldsymbol{\theta}^*)}\right)\right]^p = |A|^2\Phi_{B,P}\left(W^{-1}\left(\boldsymbol{\theta}^*\right)\right) && (130)\\
&\geq |A|^2\Phi_{B,P}\left(\boldsymbol{\theta}_B\left(P\right)\right) = |A|^2\Phi_{B,P}\left(W^{-1}\left(W\left(\boldsymbol{\theta}_B\left(P\right)\right)\right)\right)\\
&= \Phi_{B,P^*}\left(\left(W\left(\boldsymbol{\theta}_B\left(P\right)\right)\right)\right)
\end{aligned}$$

donde la desigualdad vale por la definición de $\boldsymbol{\theta}_B\left(P\right)$, y la última igualdad es consecuencia de (129) y (130). Además,

$$S_\tau^2\left(H_{P^*,W(\boldsymbol{\theta}_B(P))}\right) = S_\tau^2\left(H_{P,\boldsymbol{\theta}_B(P)}\right) = \kappa_2,$$

lo cual prueba que $\boldsymbol{\theta}_B\left(P^*\right) = W\left(\boldsymbol{\theta}_B\left(P\right)\right) = \left(A\boldsymbol{\mu}\left(P\right) + \mathbf{b}, AB\left(P\right), A\Delta\left(P\right)A^T\right)$. ■

F. Apéndice del Capítulo 6

F.1. Demostraciones de existencia

Lema F.1 *(S_ε) es equivalente a la siguiente condición*

(C_ε) P y $\mathbf{f}$ son tales que

$$\delta_\varepsilon = \text{ínf}\left\{\delta : P\left(H\left(\boldsymbol{\alpha}, l, \delta\right)\right) \geq \varepsilon, \|\boldsymbol{\alpha}\| = 1, \boldsymbol{\alpha} \in \mathbb{R}^{p+r}, l \in \mathbb{R}, \delta \geq 0\right\}$$

es positivo.

Demostración.

C_ε **implica** S_ε : Sea P que satisface C_ε, y elijamos $\mathbf{a} \in \mathbb{R}^p, \mathbf{b} \in \mathbb{R}^r, c \in \mathbb{R}$. Sea $H = H_{(\mathbf{a},\mathbf{b},c)}$ un hiperplano. Consideremos la faja $H\left(\left(\mathbf{a}, \mathbf{b}\right), c, 0\right)$ con $\delta = 0 < \delta_\varepsilon$. Entonces $P\left(H\left(\boldsymbol{\alpha}, l, 0\right)\right) < \varepsilon$, lo cual prueba la implicación.

S_ε **implica** C_ε : Fijemos $\varepsilon > 0$.

Observemos que

$$\begin{aligned}\delta_\varepsilon &= \text{ínf}\left\{\delta : P\left(H\left(\boldsymbol{\alpha}, l, \delta\right)\right) \geq \varepsilon, \|\boldsymbol{\alpha}\| = 1, \boldsymbol{\alpha} \in \mathbb{R}^{p+r}, l \in \mathbb{R}, \delta \geq 0\right\} \\ &= \frac{1}{2}\,\text{ínf}\left\{\eta : P\left(H^0\left(\boldsymbol{\alpha}, l, \eta\right)\right) \geq \varepsilon, \|\boldsymbol{\alpha}\| = 1, \boldsymbol{\alpha} \in \mathbb{R}^{p+r}, l \in \mathbb{R}, \eta \geq 0\right\}\end{aligned}$$

donde

$$H^0\left(\boldsymbol{\alpha}, l, \eta\right) = \left\{\left(\mathbf{x}, y\right) : \boldsymbol{\alpha}^T\left(\mathbf{x}, \mathbf{f}\left(y\right)\right) \in \left[l - \eta, l + \eta\right]\right\}$$

es una **faja centrada** de tamaño 2η alrededor del hiperplano $H_{(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,l)}$. Sea

$$A\left(\boldsymbol{\alpha}, l\right) := \left\{\eta : P\left(H^0\left(\boldsymbol{\alpha}, l, \eta\right)\right) \geq \varepsilon\right\}$$

Entonces, (C_ε) es equivalente a que

$$\eta_\varepsilon = \inf_{\substack{\boldsymbol{\alpha} \in \mathbb{R}^{p+r}, \|\boldsymbol{\alpha}\| = 1 \\ l \in \mathbb{R}}} \left\{\eta : P\left(H^0\left(\boldsymbol{\alpha}, l, \eta\right)\right) \geq \varepsilon\right\} > 0.$$

Primer paso. Fijemos $\boldsymbol{\alpha}$ y l que cumplen $\in \mathbb{R}^{p+r}, \|\boldsymbol{\alpha}\| = 1$ y $l \in \mathbb{R}$. Sea

$$\eta\left(\boldsymbol{\alpha}, l\right) = \text{ínf}\left\{\eta : P\left(H^0\left(\boldsymbol{\alpha}, l, \eta\right)\right) \geq \varepsilon\right\}$$

Veamos que si P satisface S_ε, entonces $\eta\left(\boldsymbol{\alpha}, l\right) > 0$, para todo par $(\boldsymbol{\alpha},l)$ que cumpla las condiciones. Si P satisface S_ε, entonces

$$P\left(H^0\left(\boldsymbol{\alpha}, l, 0\right)\right) < \varepsilon,$$

por ser $H^0\left(\boldsymbol{\alpha}, l, 0\right)$ un hiperplano. También,

$$H^0\left(\boldsymbol{\alpha}, l, 0\right) = \bigcap_{n \in \mathbb{N}} H^0\left(\boldsymbol{\alpha}, l, \frac{1}{n}\right),$$

y como $\left\{H^0\left(\boldsymbol{\alpha}, l, \frac{1}{n}\right)\right\}_n$ es una sucesión decreciente de eventos, por la continuidad de la probabilidad, tenemos que

$$P\left(H^0\left(\boldsymbol{\alpha}, l, 0\right)\right) = \lim_{n\to+\infty} P\left(H^0\left(\boldsymbol{\alpha}, l, \frac{1}{n}\right)\right) < \varepsilon.$$

Como la sucesión $\left\{P\left(H^0\left(\boldsymbol{\alpha}, l, \frac{1}{n}\right)\right)\right\}_n$ es decreciente, existe un entero n_0 tal que

$$P\left(H^0\left(\boldsymbol{\alpha}, l, \frac{1}{n}\right)\right) < \varepsilon, \quad \forall n \geq n_0 = n_0\left(\boldsymbol{\alpha}, l\right).$$

Luego, $A = \left\{\eta : P\left(H^0\left(\boldsymbol{\alpha}, l, \eta\right)\right) \geq \varepsilon\right\} \subset \left(\frac{1}{n_0}, +\infty\right)$. Por otro lado, como

$$1 = P\left(\bigcup_{n\in\mathbb{N}} H^0\left(\boldsymbol{\alpha}, l, n\right)\right) = \lim_{n\to+\infty} P\left(H^0\left(\boldsymbol{\alpha}, l, n\right)\right),$$

existe un entero n_1 tal que

$$P\left(H^0\left(\boldsymbol{\alpha}, l, n\right)\right) \geq \varepsilon, \quad \forall n \geq n_1 = n_1\left(\boldsymbol{\alpha}, l\right).$$

Luego, el conjunto $A \neq \emptyset$, y $\eta\left(\boldsymbol{\alpha}, l\right) > \frac{1}{n_0} > 0$.

Observemos que,

$$P\left(H^0\left(\boldsymbol{\alpha}, l, \eta\left(\boldsymbol{\alpha}, l\right)\right)\right) \geq \varepsilon. \tag{131}$$

Esto es consecuencia de que, por definición,

$$\eta\left(\boldsymbol{\alpha}, l\right) = \text{ínf}\left\{\eta : P\left(H^0\left(\boldsymbol{\alpha}, l, \eta\right)\right) \geq \varepsilon\right\}.$$

Luego, existe una sucesión real $\left(\eta_n\right)_n$ tal que $\eta_n \downarrow \eta\left(\boldsymbol{\alpha}, l\right)$, y para la cual se verifica $P\left(H^0\left(\boldsymbol{\alpha}, l, \eta_n\right)\right) \geq \varepsilon$. Entonces

$$\begin{aligned} P\left(H^0\left(\boldsymbol{\alpha}, l, \eta\left(\boldsymbol{\alpha}, l\right)\right)\right) &= P\left(\bigcap_{n\in\mathbb{N}} H^0\left(\boldsymbol{\alpha}, l, \eta_n\right)\right) \\ &= \lim_{n\to+\infty} P\left(H^0\left(\boldsymbol{\alpha}, l, \eta_n\right)\right) \geq \varepsilon, \end{aligned}$$

lo cual prueba la observación.

Segundo paso. Por definición de η_ε, sabemos que existe una sucesión $\left(\boldsymbol{\alpha}_n, l_n\right)$ tal que

$$\eta\left(\boldsymbol{\alpha}_n, l_n\right) \searrow \eta_\varepsilon.$$

Como $\left\{\boldsymbol{\alpha}_n\right\}_n$ está en un compacto, existe un $\boldsymbol{\alpha}_0$ de norma uno y una subsucesión que converge a $\boldsymbol{\alpha}_0$. Sin pérdida de generalidad, suponemos que $\boldsymbol{\alpha}_n \to \boldsymbol{\alpha}_0$. Por el primer paso de la demostración, sabemos que

$$P\left(H^0\left(\boldsymbol{\alpha}_n, l_n, \eta\left(\boldsymbol{\alpha}_n, l_n\right)\right)\right) \geq \varepsilon, \quad \forall n.$$

Como P es acotada en probabilidad (*tight)*, para todo $0 < h < 1$ existe un subconjunto compacto $B_h \subset \mathbb{R}^{p+1}$:

$$P\left(\{(\mathbf{x}, y) \in B_h\}\right) \geq 1 - h.$$

Podemos asumir que dichos conjuntos son bolas y que la sucesión $(B_h)_h$ es creciente a $\mathbb{R}^{p+1}$. Luego, llamando H_n^0 a $H^0\left(\boldsymbol{\alpha}_n, l_n, \eta\left(\boldsymbol{\alpha}_n, l_n\right)\right)$ por (131) tenemos

$$P\left(H_n^0 \cap B_h\right) \geq P\left(H_n^0\right) + P\left(B_h\right) - 1 \geq \varepsilon - h.$$

Veamos que la sucesión real $(l_n)_n$ converge. En particular, tomando $h = \frac{\varepsilon}{2}$, tenemos que, para todo n,

$$H_n^0 \cap B_{\frac{\varepsilon}{2}} \neq \emptyset.$$

Más aún, como $\mathrm{P}(D_{\mathbf{f}}) = 0$, tenemos que $H_n^0 \cap B_{\frac{\varepsilon}{2}} \cap D_{\mathbf{f}} \neq \emptyset$, para todo n. Sea $(\mathbf{x}_n, y_n) \in H_n^0 \cap B_{\frac{\varepsilon}{2}} \cap D_{\mathbf{f}}$. Luego, llamando $\eta_n = \eta\left(\boldsymbol{\alpha}_n, l_n\right)$,

$$\begin{aligned} l_n - \eta_n \leq \boldsymbol{\alpha}_n^T\left(\mathbf{x}_n, \mathbf{f}\left(y_n\right)\right) \leq l_n + \eta_n \\ \left\{(\mathbf{x}_n, y_n)\right\}_n \in B_{\varepsilon/2} \end{aligned}$$

Como $B_{\varepsilon/2}$ es compacto, existe una subsucesión de $\left\{(\mathbf{x}_n, y_n)\right\}_n$ que converge a un punto $(\mathbf{x}_0, y_0) \in B_{\varepsilon/2}$. Sin pérdida de generalidad podemos suponer que $(\mathbf{x}_n, y_n) \to (\mathbf{x}_0, y_0)$. Como $(\mathbf{x}_n, y_n) \in D_{\mathbf{f}}$, tenemos que $\boldsymbol{\alpha}_n^T\left(\mathbf{x}_n, \mathbf{f}\left(y_n\right)\right)$ y $\eta\left(\boldsymbol{\alpha}_n, l_n\right)$ son sucesiones convergentes. Luego, existen reales m_1 y M_1 tales que, para todo n,

$$\begin{aligned} m_1 \leq \boldsymbol{\alpha}_n^T\left(\mathbf{x}_n, \mathbf{f}\left(y_n\right)\right) - \eta_n \leq l_n \\ l_n \leq \boldsymbol{\alpha}_n^T\left(\mathbf{x}_n, \mathbf{f}\left(y_n\right)\right) + \eta_n \leq M_1. \end{aligned}$$

Por lo tanto, la sucesión real $(l_n)_n$ está acotada, y entonces admite una subsucesión convergente a un valor l_0. Sin pérdida de generalidad, podemos asumir que toda la sucesión converge a l_0.

Tercer paso. Probemos que, sabiendo que $P\left(H_n^0\right) \geq \varepsilon$ para todo n, se deduce que $P\left(H^0\left(\boldsymbol{\alpha}_0, l_0, \eta_\varepsilon\right)\right) \geq \varepsilon$. Sea $(\mathbf{x}, y) \in S_n \cap B_h$ y $\delta > 0$. Esto implica que

$$l_n - \eta_n \leq \boldsymbol{\alpha}_n^T\left(\mathbf{x}, \mathbf{f}\left(y\right)\right) \leq l_n + \eta_n,$$

o, equivalentemente,

$$l_n - \eta_n - \left(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_0\right)^T\left(\mathbf{x}, \mathbf{f}\left(y\right)\right) \leq \boldsymbol{\alpha}_0^T\left(\mathbf{x}, \mathbf{f}\left(y\right)\right) \leq l_n + \eta_n - \left(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_0\right)^T\left(\mathbf{x}, \mathbf{f}\left(y\right)\right). \quad (132)$$

Como $\mathbf{f}$ es acotada, y B_h es compacto, existe una constante M_h tal que $\|(\mathbf{x}, \mathbf{f}\left(y\right))\| \leq M_h$. Como $\boldsymbol{\alpha}_n \to \boldsymbol{\alpha}_0$, existe un n_0 tal que para todo $n \geq n_0$ se tiene que $\|(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_0)\| < M_h \frac{\delta}{2}$. El valor de n_0 dependerá de $M_h \delta$, o sea, $n_0 = n_0\left(h, \delta\right)$. Para todo $n \geq n_0$ se tiene que

$$\left|\left(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_0\right)^T\left(\mathbf{x}, \mathbf{f}\left(y\right)\right)\right| \leq \|(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_0)\| \, \|(\mathbf{x}, \mathbf{f}\left(y\right))\| < \frac{\delta}{2}.$$

Como $l_n + \eta_n - (\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_0)^T (\mathbf{x}, \mathbf{f}(y))$ converge a $l_0 + \eta_\varepsilon$, de la desigualdad anterior y de (132) tenemos que,

$$\begin{aligned} l_0 - \eta_\varepsilon - \delta &\leq l_n - \eta_n - (\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_0)^T (\mathbf{x}, \mathbf{f}(y)) \leq \boldsymbol{\alpha}_0^T (\mathbf{x}, \mathbf{f}(y)) \\ &\leq l_n + \eta_n - (\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_0)^T (\mathbf{x}, \mathbf{f}(y)) \leq l_0 + \eta_\varepsilon + \delta. \end{aligned}$$

Luego, para todo $n \geq n_0 = n_0(h, \delta)$ se tiene que

$$S_n \cap B_h \subset H^0(\boldsymbol{\alpha}_0, l_0, \eta_\varepsilon + \delta) \cap B_h.$$

Luego,

$$\varepsilon - h \leq P(S_n \cap B_h) \leq P\left(H^0(\boldsymbol{\alpha}_0, l_0, \eta_\varepsilon + \delta) \cap B_h\right).$$

Entonces, para todo $0 < h < \varepsilon$, tenemos

$$\varepsilon - h \leq P\left(H^0(\boldsymbol{\alpha}_0, l_0, \eta_\varepsilon + \delta) \cap B_h\right).$$

Como $\left(B_{\frac{1}{n}}\right)_n$ es una sucesión creciente a todo el espacio, resulta que

$$\begin{aligned} \lim_{n \to +\infty} \left(\varepsilon - \frac{1}{n}\right) &\leq \lim_{n \to +\infty} P\left(H^0(\boldsymbol{\alpha}_0, l_0, \eta_\varepsilon + \delta) \cap B_{\frac{1}{n}}\right) \\ &= \lim_{n \to +\infty} P\left(\bigcup_{n=1}^{\infty} \left(H^0(\boldsymbol{\alpha}_0, l_0, \eta_\varepsilon + \delta) \cap B_{\frac{1}{n}}\right)\right) \\ &= \lim_{n \to +\infty} P\left(H^0(\boldsymbol{\alpha}_0, l_0, \eta_\varepsilon + \delta)\right) \\ \varepsilon &\leq P\left(H^0(\boldsymbol{\alpha}_0, l_0, \eta_\varepsilon + \delta)\right) \end{aligned}$$

Como esto vale para todo $\delta > 0$, y, los conjuntos $\left(H^0\left(\boldsymbol{\alpha}_0, l_0, \eta_\varepsilon + \frac{1}{n}\right)\right)_n$ decrecen a $H^0(\boldsymbol{\alpha}_0, l_0, \eta_\varepsilon)$, por la continuidad de la probabilidad resulta que $\varepsilon \leq P\left(H^0(\boldsymbol{\alpha}_0, l_0, \eta_\varepsilon)\right)$. Esto implica que $\eta_\varepsilon \geq \eta(\boldsymbol{\alpha}_0, l_0) > 0$. ■

Análogamente a la Observación 2.1 de Lopuhaä [1991], el siguiente lema será útil para probar uno de los resultados que siguen.

Lema F.2 *Si P satisface (S_ε), cualquier valor de $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ que minimiza a $\Phi_{C,P}(\boldsymbol{\theta})$ sujeta a*

$$E_{H_{P,\boldsymbol{\theta}}}(\rho_1(v)) = \kappa_1 \tag{133}$$

es también una solución al problema de minimización con la siguiente restricción, en vez de (133)

$$E_{H_{P,\boldsymbol{\theta}}}(\rho_1(v)) \leq \kappa_1. \tag{134}$$

Demostración. Consideremos la función $h : (0, +\infty) \to \mathbb{R}$ dada por

$$\begin{aligned}
h(s) &= \Phi_{C,P}(\boldsymbol{\mu}, B, s\Delta) \\
&= |s\Delta| \left(E_{H_{P,\boldsymbol{\theta}}} \rho_2(v)\right)^p \\
&= |s\Delta| \left(E_P \left(\rho_2 \left(\left\{(\mathbf{x} - \boldsymbol{\mu} - B\mathbf{f}(y))^T \frac{1}{s}\Delta^{-1}(\mathbf{x} - \boldsymbol{\mu} - B\mathbf{f}(y))\right\}^{1/2}\right)\right)\right)^p \\
&= |s\Delta| \left(E_P \left(\rho_2 \left(\left\{\left(\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right)^T \frac{1}{s}\Delta^{-1}\left(\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right)\right\}^{1/2}\right)\right)\right)^p \\
&= |\Delta| s^p \left(\int \rho_2 \left(\frac{\left\{\left(\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right)^T \Delta^{-1}\left(\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right)\right\}^{1/2}}{\sqrt{s}}\right) dP(\mathbf{x}, y)\right)^p \\
&= |\Delta| \left(\int s\rho_2 \left(\frac{d}{\sqrt{s}}\right) dH_{P,\boldsymbol{\theta}}(d)\right)^p,
\end{aligned}$$

Observemos que esta función es creciente para cualquier valor de $\boldsymbol{\theta} = (D, \Delta)$ en el espacio de parámetros. La derivada de $s\rho_2\left(\frac{d}{\sqrt{s}}\right)$ es

$$\begin{aligned}
\frac{\partial}{\partial s}\left[s\rho_2\left(\frac{d}{\sqrt{s}}\right)\right] &= \rho_2\left(\frac{d}{\sqrt{s}}\right) - \frac{1}{2}s\psi_2\left(\frac{d}{\sqrt{s}}\right) ds^{-3/2} \\
&= \rho_2\left(\frac{d}{\sqrt{s}}\right) - \frac{d}{2\sqrt{s}}\psi_2\left(\frac{d}{\sqrt{s}}\right).
\end{aligned}$$

Como P satisface (S_ε), es decir, no puede tener toda su masa concentrada en un hiperplano, por la condición adicional A3 impuesta sobre ρ_2 tenemos que $h'(s) > 0$, de modo que h es estrictamente creciente en $s > 0$.

Sea $\widetilde{\boldsymbol{\theta}}_C(P) = \left(\widetilde{D}_C(P), \widetilde{\Delta}_C(P)\right)$, y sea $\boldsymbol{\theta}_2(P) = (D_2(P), \Delta_2(P))$ que minimiza a $\Phi_{C,P}$ sujeto a la restricción (134). Queremos probar que

$$\Phi_{C,P}\left(\widetilde{\boldsymbol{\theta}}_C(P)\right) \leq \Phi_{C,P}(\boldsymbol{\theta}_2(P)), \tag{135}$$

puesto que la desigualdad inversa vale por definición de $\boldsymbol{\theta}_2(P)$. Si

$$E_{H_{P,\boldsymbol{\theta}_2(P)}} \rho_1(v) = \kappa_1,$$

entonces la desigualdad (135) se verifica. Si, en cambio,

$$E_{H_{P,\boldsymbol{\theta}_2(P)}} \rho_1(v) < \kappa_1,$$

entonces, por la Observación 5.1(i), resulta que $(\boldsymbol{\theta}_2(P))_\lambda = (D_2(P), \Delta_2(P)\lambda)$ satisface la igualdad a κ_1 en (43), con $\lambda = S_M^2\left(H_{P,\boldsymbol{\theta}_2(P)}\right)$. Como ρ_1 es creciente, dicho $\lambda < 1$. Como h es creciente, resulta que tomando $\boldsymbol{\theta} = \boldsymbol{\theta}_2(P)$ tenemos

$$h(\lambda) < h(1),$$

es decir

$$\Phi_{C,P}\left(\left(\boldsymbol{\theta}_2\left(P\right)\right)_\lambda\right) < \Phi_{C,P}\left(\boldsymbol{\theta}_2\left(P\right)\right).$$

Como además, $\left(\boldsymbol{\theta}_2\left(P\right)\right)_\lambda$ satisface la igualdad (134) que define la restricción que debe cumplir $\widetilde{\boldsymbol{\theta}}_C\left(P\right)$, por definición de $\widetilde{\boldsymbol{\theta}}_C\left(P\right)$ tenemos

$$\Phi_{C,P}\left(\widetilde{\boldsymbol{\theta}}_C\left(P\right)\right) \leq \Phi_{C,P}\left(\left(\boldsymbol{\theta}_2\left(P\right)\right)_\lambda\right).$$

Juntando las dos últimas desigualdades, hemos probado (135), y por lo tanto el resultado buscado. ■

Lema F.3 *Supongamos que* $\boldsymbol{\theta} = (D, \Delta) \in \widetilde{\boldsymbol{\Theta}}$ *satisface la restricción (43) y que* $\kappa_1 < a_1 = \max \rho_1$. *Entonces, existe una constante* $q > 0$, *que sólo depende de las funciones* ρ_1 *y* ρ_2 *y de la constante* κ_1, *tal que*

$$\int \rho_2 \left(\left\{ \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right] \right\}^{1/2} \right) dP\left(\mathbf{x}, y\right) \geq q.$$

Demostración. Sea A el conjunto

$$\begin{aligned} A &= \left[E\left(D, \Delta, \rho_1^{-1}\left(\frac{\kappa_1}{2}\right)\right)\right]^c \\ &= \left\{ (\mathbf{x}, y) : \sqrt{\left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right]} > \rho_1^{-1}\left(\frac{\kappa_1}{2}\right) \right\}. \end{aligned}$$

Entonces

$$\begin{aligned} \kappa_1 &= E_{H_{P,\boldsymbol{\theta}}}\left(\rho_1\left(v\right)\right) \\ &= E_P\left(\rho_1 \left(\left\{ \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right] \right\}^{1/2} \right) \right) \\ &= \int \rho_1 \left(\left\{ \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right] \right\}^{1/2} \right) dP\left(\mathbf{x}, y\right) \\ &= \int_A \rho_1 \left(\left\{ \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right] \right\}^{1/2} \right) dP\left(\mathbf{x}, y\right) \\ &+ \int_{A^c} \rho_1 \left(\left\{ \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y)\right] \right\}^{1/2} \right) dP\left(\mathbf{x}, y\right) \\ &\leq a_1 \int_A dP\left(\mathbf{x}, y\right) + \rho_1\left(\rho_1^{-1}\left(\frac{\kappa_1}{2}\right)\right) \int_{A^c} dP\left(\mathbf{x}, y\right) \\ &= a_1 P\left(A\right) + \frac{\kappa_1}{2}\left(1 - P\left(A\right)\right). \end{aligned}$$

Como $\kappa_1 < a_1$, esto implica que $P(A) \geq \frac{\kappa_1}{2a_1 - \kappa_1} > 0$. Luego

$$\begin{aligned}
&\int \rho_2 \left(\left\{ \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right] \right\}^{1/2} \right) dP(\mathbf{x}, y) \\
&= \int_A \rho_2 \left(\left\{ \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right] \right\}^{1/2} \right) dP(\mathbf{x}, y) \\
&+ \int_{A^c} \rho_2 \left(\left\{ \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right] \right\}^{1/2} \right) dP(\mathbf{x}, y) \\
&\geq \int_A \rho_2 \left(\left\{ \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right] \right\}^{1/2} \right) dP(\mathbf{x}, y) \\
&\geq \rho_2 \left(\rho_1^{-1} \left(\frac{\kappa_1}{2} \right) \right) \int_A dP(\mathbf{x}, y) \\
&= \rho_2 \left(\rho_1^{-1} \left(\frac{\kappa_1}{2} \right) \right) P(A) \\
&\geq \rho_2 \left(\rho_1^{-1} \left(\frac{\kappa_1}{2} \right) \right) \frac{\kappa_1}{2a_1 - \kappa_1} = q > 0. \quad \blacksquare
\end{aligned}$$

Lema F.4 *Sea* $\boldsymbol{\theta} = (D, \Delta) \in \widetilde{\boldsymbol{\Theta}}$, $0 < m_0 < \infty$, $0 < c < \infty$, *y* $0 < \varepsilon < 1$. *Sean* $\lambda_1(\Delta) \geq \lambda_1(\Delta) \geq \cdots \geq \lambda_p(\Delta) > 0$ *los autovalores de* Δ *ordenados de forma decreciente.*

i. Si P *satisface* (S_ε) *y* $P(E(D, \Delta, c)) \geq \varepsilon$, *entonces existe una constante* $k_1 > 0$, *que depende sólo de* ε, P *y* c, *tal que* $\lambda_p(\Delta) \geq k_1$.

ii. Supongamos que $\int \rho_1 \left(\frac{\|\mathbf{x}\|}{m_0} \right) dP(\mathbf{x}, y) \leq \kappa_1$ *y que* $\lambda_p(\Delta) \geq k_1 > 0$. *Entonces, existe una constante* $k_2 \in \mathbb{R}$, *que depende sólo de* k_1, m_0, ρ_1, ρ_2 *y* κ_1, *tal que si* $\lambda_1(\Delta) > k_2$, *el punto* (D, Δ) *no puede ser* $\widetilde{\boldsymbol{\theta}}_C(P)$, *es decir, no puede minimizar* $\Phi_{C,P}$ *sujeto a la restricción (134).*

iii. Asumamos que P *satisface* (S_ε), *que* $P(E(D, \Delta, c)) \geq \varepsilon$, *que* $0 < k_1 \leq \lambda_p(\Delta) \leq \lambda_1(\Delta) \leq k_2$, *y que* $\mathbf{f}$ *es continua o acotada. Entonces, existe un conjunto compacto* $K \subset \widetilde{\boldsymbol{\Theta}}$, *que sólo depende de* $\varepsilon, k_1, k_2, a_1$ $(= \arg\max \rho_1)$ *y* P, *tal que* $\boldsymbol{\theta} = (D, \Delta)$ *pertenece a* K.

Demostración. Antes de probar las tres afirmaciones, aplicaremos algo de álgebra lineal a nuestro problema. Como $\Delta \in PDS(p)$, existe una matriz ortogonal $U \in \mathbb{R}^{p \times p}$, $U = \left[\begin{array}{ccc} \mathbf{u}_1 & \cdots & \mathbf{u}_p \end{array} \right]$ con $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ una base ortonormal de $\mathbb{R}^p$, tal que

$$\Delta = U \cdot diag(\lambda_1(\Delta), \ldots, \lambda_p(\Delta)) \cdot U^T.$$

Luego

$$\Delta^{-1} = U \cdot diag\left(\lambda_1^{-1}(\Delta), \ldots, \lambda_p^{-1}(\Delta)\right) \cdot U^T.$$

Por lo tanto,

$$\begin{aligned} d^2\left((\mathbf{x},y),D,\Delta\right) &:= \left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]^T\Delta^{-1}\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right] \\ &= \left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]^T U\cdot diag\left(\lambda_1^{-1}(\Delta),\ldots,\lambda_p^{-1}(\Delta)\right)\cdot U^T\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right] \\ &= \sum_{i=1}^{p}\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]^T\mathbf{u}_i\frac{1}{\lambda_i(\Delta)}\mathbf{u}_i^T\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right] \\ &= \sum_{i=1}^{p}\frac{1}{\lambda_i(\Delta)}\left(\mathbf{u}_i^T\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]\right)^2 \\ &\geq \frac{1}{\lambda_j(\Delta)}\left(\mathbf{u}_j^T\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]\right)^2, \end{aligned}$$

para todo j entre 1 y p. Sea

$$\begin{aligned} V_j &:= \left\{(\mathbf{x},y):\left(\mathbf{u}_j^T\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]\right)^2\leq\lambda_j(\Delta)c^2\right\} \\ &= \left\{(\mathbf{x},y):\left|\mathbf{u}_j^T\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]\right|\leq\sqrt{\lambda_j(\Delta)}c\right\} \\ &= \left\{(\mathbf{x},y):-\sqrt{\lambda_j(\Delta)}c\leq\mathbf{u}_j^T\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]\leq\sqrt{\lambda_j(\Delta)}c\right\} \\ &= \left\{(\mathbf{x},y):\mathbf{u}_j^T\boldsymbol{\mu}-\sqrt{\lambda_j(\Delta)}c\leq\mathbf{u}_j^T\mathbf{x}-\mathbf{u}_j^TB\mathbf{f}(y)\leq\mathbf{u}_j^T\boldsymbol{\mu}+\sqrt{\lambda_j(\Delta)}c\right\}. \end{aligned}$$

V_j es una faja. La podemos escribir en la forma (48). El vector normal es $\mathbf{n}_j=\left(\mathbf{u}_j^T,-\left(B^T\mathbf{u}_j\right)^T\right)$, cuya norma es

$$\|\mathbf{n}_j\|=\|\mathbf{u}_j\|^2+\left\|B^T\mathbf{u}_j\right\|^2=1+\left\|B^T\mathbf{u}_j\right\|^2.$$

Entonces, si tomamos $\boldsymbol{\alpha}_j=\frac{\mathbf{n}_j}{\|\mathbf{n}_j\|}, l_j=\frac{\mathbf{u}_j^T\boldsymbol{\mu}-\sqrt{\lambda_j(\Delta)}c}{\|\mathbf{n}_j\|}, \delta_j=\frac{2\sqrt{\lambda_j(\Delta)}c}{\|\mathbf{n}_j\|}$, resulta que

$$E(D,\Delta,c)\subseteq V_j=H\left(\boldsymbol{\alpha}_j,l_j,\delta_j\right), \tag{136}$$

para todo $1\leq j\leq p$. Ahora nos abocamos a probar cada ítem.

i. En particular, tenemos que $E(D,\Delta,c)\subseteq H\left(\boldsymbol{\alpha}_p,l_p,\delta_p\right)$. De modo que

$$\varepsilon\leq P\left(E(D,\Delta,c)\right)\leq P\left(H\left(\boldsymbol{\alpha}_p,l_p,\delta_p\right)\right).$$

En virtud del Lema F.1, tenemos

$$\delta_\varepsilon\leq\delta_p=\frac{2\sqrt{\lambda_p(\Delta)}c}{\|\mathbf{n}_p\|}\leq 2\sqrt{\lambda_p(\Delta)}c$$

puesto que $\|\mathbf{n}_p\|\geq 1$. O, equivalentemente, $0<\frac{\delta_\varepsilon^2}{4c^2}\leq\lambda_p(\Delta)$, lo cual prueba (i).

ii. Buscamos $\boldsymbol{\theta} = (D, \Delta)$ que cumpla $\Phi_{C,P}(D, \Delta) \leq \Phi_{C,P}(D^*, \Delta^*)$ para todo $(D^*, \Delta^*) \in \widetilde{\boldsymbol{\Theta}}$, donde

$$\begin{aligned}\Phi_{C,P}(D, \Delta) &= |\Delta| \left(E_P \left(\rho_2 \left(\left\{ \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right] \right\}^{1/2} \right) \right) \right)^p \\ &= \prod_{i=1}^{p} \lambda_i(\Delta) \left(\int \rho_2 \left(\left\{ \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right] \right\}^{1/2} \right) dP(\mathbf{x}, y) \right)^p \\ &\leq \lambda_1(\Delta) \prod_{i=2}^{p} \lambda_i(\Delta) (a_2)^p\end{aligned}$$

Observemos que $\left(0, m_0^2 I\right)$ satisface la restricción (134). Entonces, de acuerdo al Lema F.2, si (D, Δ) es un candidato a ser $\widetilde{\boldsymbol{\theta}}_C(P)$, debe satisfacer

$$\begin{aligned}\Phi_{C,P}(D, \Delta) \leq \Phi_{C,P}\left(0, m_0^2 I\right) &= \left(m_0^2\right)^p \left(\int \rho_2 \left(\left\{ \frac{\mathbf{x}^T \mathbf{x}}{m_0^2} \right\}^{1/2} \right) dP(\mathbf{x}, y) \right)^p \\ &\leq \left(m_0^2\right)^p (a_2)^p\end{aligned}$$

lo cual implica que

$$\lambda_1(\Delta) \prod_{i=2}^{p} \lambda_i(\Delta) \leq \frac{\left(m_0^2\right)^p (a_2)^p}{\left(\int \rho_2 \left(\left\{ \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right] \right\}^{1/2} \right) dP(\mathbf{x}, y) \right)^p}.$$

Por el Lema F.3, sabemos que el denominador de la expresión de la derecha en la desigualdad previa está acotado por debajo por una constante q, de modo que

$$\begin{aligned}\lambda_1(\Delta) &\leq \frac{\left(m_0^2\right)^p (a_2)^p}{\left(\int \rho_2 \left(\left\{ \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right]^T \Delta^{-1} \left[\mathbf{x} - D\widetilde{\mathbf{f}}(y) \right] \right\}^{1/2} \right) dP(\mathbf{x}, y) \right)^p \prod_{i=2}^{p} \lambda_i(\Delta)} \\ &\leq \frac{\left(m_0^2\right)^p (a_2)^p}{q^p \left[\lambda_p(\Delta)\right]^{p-1}} \leq \frac{\left(m_0^2 a_2\right)^p}{q^p \left[k_1\right]^{p-1}},\end{aligned}$$

lo cual prueba (ii).

iii. Por hipótesis, y a partir de (136) tenemos que

$$\varepsilon \leq P(E(D, \Delta, c)) \leq P(H(\boldsymbol{\alpha}_j, l_j, \delta_j)). \tag{137}$$

Como se detalla en el Apéndice B.1.2, la 2-norma (o norma de Frobenius) de una matriz $p \times r$ es su norma Euclídea, tomando a la matriz como un vector de dimensión pr:

$$\|B\|_2^2 = \operatorname{traza}\left(B^T B\right) = \sum_{i=1}^{p} \sum_{j=1}^{r} B_{ij}^2$$

También

$$\left\|U^T B\right\|_2^2 = \text{traza}\left(\left(U^T B\right)^T U^T B\right) = \text{traza}\left(B^T B\right) = \|B\|_2^2 .$$

Asumimos que $\|B\|_2 > k_3$. Eligiendo apropiadamente el valor de k_3 contradiremos (137), como se verá a continuación. El cuadrado de la 2-norma es también la suma de los cuadrados de las normas de cada fila de $U^T B$, mirada como un vector de $\mathbb{R}^r$. Como

$$U^T B = \begin{bmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_p^T \end{bmatrix} B = \begin{bmatrix} \mathbf{u}_1^T B \\ \vdots \\ \mathbf{u}_p^T B \end{bmatrix}$$

entonces $\left\|U^T B\right\|_2^2 > k_3^2$ implica que existe al menos un j entre 1 y p tal que $\left\|\mathbf{u}_j^T B\right\|_2 > \frac{k_3}{\sqrt{p}}$. Entonces, si tomamos $k_3 > \frac{2\sqrt{rk_2}c}{\delta_\varepsilon}$, donde δ_ε es positivo, a través del Lema F.1, obtenemos que $\delta_j < \delta_\varepsilon$, de modo que, para dicho j tenemos que $P\left(H\left(\boldsymbol{\alpha}_j, l_j, \delta_j\right)\right) < \varepsilon$. Esto contradice (137). Luego, existe un valor positivo k_3, que depende solamente de ε, k_2 y P tal que $\|B\|_2 \leq k_3$.

Como la norma espectral de Δ satisface $\|\Delta\| = \lambda_1(\Delta)$ y tenemos que $0 < k_1 \leq \lambda_p(\Delta) \leq \lambda_1(\Delta) \leq k_2$, entonces Δ pertenece a un subconjunto acotado de $PDS(p)$.

Nos resta acotar $\boldsymbol{\mu}$. Como P es una probabilidad en $\mathbb{R}^{p+1}$, es tight (acotada en probabilidad). De modo que dado ε, existe B_ε un subconjunto compacto de $\mathbb{R}^{p+1}$ tal que $P(B_\varepsilon) \geq 1 - \varepsilon$. Como

$$\left\{(\mathbf{x}, y) : \|\mathbf{x} - B\mathbf{f}(y) - \boldsymbol{\mu}\|^2 \leq c^2 \lambda_p(\Delta)\right\} \subseteq E(D, \Delta, c)$$

y

$$\begin{aligned} E(D, \Delta, c) &\subseteq \left\{(\mathbf{x}, y) : \|\mathbf{x} - B\mathbf{f}(y) - \boldsymbol{\mu}\|^2 \leq c^2 \lambda_1(\Delta)\right\} \\ &\subseteq \left\{(\mathbf{x}, y) : \|\mathbf{x} - B\mathbf{f}(y) - \boldsymbol{\mu}\|^2 \leq c^2 k_2\right\} \end{aligned}$$

tenemos que $\|\mathbf{x} - B\mathbf{f}(y) - \boldsymbol{\mu}\| \leq c\sqrt{k_2}$ para algún $(\mathbf{x}, y) \in B_\varepsilon$. Sino,

$$\left\{(\mathbf{x}, y) : \|\mathbf{x} - B\mathbf{f}(y) - \boldsymbol{\mu}\|^2 \leq c^2 k_2\right\} \subset B_\varepsilon^c,$$

implicando que $E(D, \Delta, c) \subset B_\varepsilon^c$, lo cual contradiría el hecho de que $P(E(D, \Delta, c)) \geq \varepsilon$. Como B es acotado, $(\mathbf{x}, y) \in B_\varepsilon$ compacto,

$$\|\mathbf{x} - B\mathbf{f}(y)\| \leq \|\mathbf{x}\| + \|B\mathbf{f}(y)\| \leq \|\mathbf{x}\| + \|B\| \, \|\mathbf{f}(y)\| \leq k_4 + k_3 k_4,$$

si asumimos que $\mathbf{f}$ sea continua o acotada, k_4 depende solamente de $\mathbf{f}$, ε y P. Finalmente,

$$\|\boldsymbol{\mu}\| \leq \|\mathbf{x} - B\mathbf{f}(y) - \boldsymbol{\mu}\| + \|\mathbf{x} - B\mathbf{f}(y)\| \leq c\sqrt{k_2} + k_4 + k_3 k_4,$$

que es una constante que depende de $c, \varepsilon, k_2, \mathbf{f}$ y P, y la afirmación queda probada. ■

Demostración del Teorema 6.1. Sea $(D,\Delta)\in\Theta$ que satisface la restricción (43). Para que podamos utilizar el Lema F.4 (i), debemos probar que $P\left(E\left(D,\Delta,c\right)\right)\geq\varepsilon$ para alguna constante c positiva. Sea $E:=E\left(D,\Delta,c_1\right)$ un elipsoide generalizado, centrado en $D\widetilde{\mathbf{f}}(y)$ y de radio c_1. Entonces, sabemos que

$$\begin{aligned}
&\int\rho_1\left(\left\{\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]^T\Delta^{-1}\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right)dP\left(\mathbf{x},y\right)\\
&=\int_E\rho_1\left(\left\{\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]^T\Delta^{-1}\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right)dP\left(\mathbf{x},y\right) \qquad (138)\\
&+\int_{E^c}\rho_1\left(\left\{\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]^T\Delta^{-1}\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right)dP\left(\mathbf{x},y\right)\\
&=\int_E\rho_1\left(\left\{\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]^T\Delta^{-1}\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right)dP\left(\mathbf{x},y\right)\\
&+a_1\left(1-P\left(E\right)\right), \qquad (139)
\end{aligned}$$

puesto que ρ_1 es una función creciente en $[0,c_1]$ que alcanza su máximo en c_1 y es constante en $[c_1,+\infty)$, $a_1=\rho_1\left(c_1\right)$, y

$$E^c=\left\{(\mathbf{x},y):\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]^T\Delta^{-1}\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]\geq c_1^2\right\}.$$

Podemos reescribir (139) de la siguiente forma

$$\begin{aligned}
&\frac{1}{a_1}\int\rho_1\left(\left\{\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]^T\Delta^{-1}\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right)dP\left(\mathbf{x},y\right)\\
&=\frac{1}{a_1}\int_E\rho_1\left(\left\{\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]^T\Delta^{-1}\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right)dP\left(\mathbf{x},y\right)+\left(1-P\left(E\right)\right),
\end{aligned}$$

que, como $\frac{1}{a_1}\int_E\rho_1\left(\left\{\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]^T\Delta^{-1}\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right)dP\left(\mathbf{x},y\right)>0$, implica

$$\begin{aligned}
P\left(E\right)&\geq1-\frac{1}{a_1}\int\rho_1\left(\left\{\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]^T\Delta^{-1}\left[\mathbf{x}-D\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right)dP\left(\mathbf{x},y\right)\\
&=1-\frac{\kappa_1}{a_1}\geq\varepsilon,
\end{aligned}$$

puesto que (D,Δ) satisface la restricción (43). El Lema F.4 (i) implica que $\lambda_p\left(\Delta\right)\geq k_1>0$.

Como $\lim_{m\to+\infty}\int\rho_1\left(\frac{\|\mathbf{x}\|}{m}\right)dP\left(\mathbf{x},y\right)=0$, existe $m_0>0$ tal que

$$\int\rho_1\left(\frac{\|\mathbf{x}\|}{m_0}\right)dP\left(\mathbf{x},y\right)\leq\kappa_1.$$

Luego, el Lema F.4 (ii) asegura que $\lambda_1(\Delta) \leq k_2$. Finalmente, el Lema F.4 (iii) implica que para resolver el problema que permite encontrar a $\widetilde{\boldsymbol{\theta}}_C(P)$ uno puede restringirse a un subconjunto compacto $K \subset \widetilde{\boldsymbol{\Theta}}$. Como la función $\Phi_{C,P}$ es una función continua de D y Δ, debe alcanzar un mínimo en K. ■

F.2. Demostraciones de unicidad

A partir del Lema F.5, que es el Lema A.8 de García Ben et al. [2006], y el Lema F.6, que es el Lema 4.2 de Tatsuoka y Tyler [2000], puede obtenerse el lema clave para garantizar la unicidad (Lema F.7), que es similar al Lema A.10 de García Ben et al. [2006], pero adaptado a nuestro problema. Recordemos que las propiedades que definen a las $\rho-$funciones fueron definidas en la página 34.

Lema F.5 (Lema A.8, García Ben et al. [2006]) *Sea ρ que satisface A1-A3, y definamos $r : (0, +\infty) \to \mathbb{R}$ por*

$$r_H(t) = t^2 E_H\left(\rho\left(\frac{v}{t}\right)\right),$$

donde H es la distribución de la variable aleatoria positiva v. Entonces, r_H es una función no decreciente.

Demostración. Se deduce de

$$\begin{aligned} r'_H(t) &= 2tE_H\left(\rho\left(\frac{v}{t}\right)\right) + t^2 E_H\left(\psi\left(\frac{v}{t}\right)\frac{v}{t^2}(-1)\right) \\ &= tE_H\left[2\rho\left(\frac{v}{t}\right) - \psi\left(\frac{v}{t}\right)\frac{v}{t}\right] \end{aligned}$$

y la propiedad A3. ■

Lema F.6 (Lema 4.2, Tatsuoka y Tyler [2000]) *Sea $\rho : [0, +\infty) \to \mathbb{R}$ que satisface $\rho(0) = 0$, ρ es creciente (en sentido amplio) y ρ es continua a derecha en 0, y P, la distribución de $\mathbf{x}$ está en la clase $\mathcal{P}_p(W_p, M)$, sea $V \in PDS(p)$, entonces*

$$E_P\left[\rho\left((\mathbf{x}-\boldsymbol{\mu})^T V^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\right] \geq E_P\left[\rho\left(a\mathbf{x}^T\mathbf{x}\right)\right],$$

donde $a^p = \det(aI) = \det\left(V^{-1}\right)$. Más aún, si o bien ρ es estrictamente creciente en un intervalo donde la densidad de $\mathbf{x}$ es estrictamente positiva o bien, la densidad de $\mathbf{x}$ es estrictamente $M-$cóncava, entonces la desigualdad es estricta.

Demostración. Este es el Lema 4.2 de Tatsuoka y Tyler [2000]. ■

Lema F.7 *Sea* $\rho : [0, +\infty) \to \mathbb{R}$ *que satisface* $\rho(0) = 0$, ρ *es creciente (en sentido amplio) y* ρ *es continua a derecha en 0, y* P, *la distribución de* $\mathbf{u}$ *está en la clase* $\mathcal{P}_p(W_p, M)$. *Sea* $\mathbf{v}$ *un vector aleatorio en* $\mathbb{R}^p$ *independiente de* $\mathbf{u}$, $V \in PDS(p)$, $a^p = \det\left(V^{-1}\right)$. *Asumamos que* ρ *es estrictamente creciente en un intervalo donde la densidad de* $\mathbf{u}$ *es estrictamente positiva, o bien, que la densidad de* $\mathbf{u}$ *es estrictamente M-cóncava, y que vale o bien* (i) $P(\mathbf{v} \neq \mathbf{0}) > 0$, *o bien* (ii) $V \neq a^{-1}I$. *Entonces*

$$E\left[\rho\left(\left\{(\mathbf{u}-\mathbf{v})^T V^{-1}(\mathbf{u}-\mathbf{v})\right\}^{1/2}\right)\right] > E\left[\rho\left(\left\{a\mathbf{u}^T\mathbf{u}\right\}^{1/2}\right)\right].$$

Demostración. Supongamos que (i) sea cierta. Entonces,

$$E\left[\rho\left(\left\{(\mathbf{u}-\mathbf{v})^T V^{-1}(\mathbf{u}-\mathbf{v})\right\}^{1/2}\right) \mid \mathbf{v} = \boldsymbol{\mu}\right] = E\left[\rho\left(\left\{(\mathbf{u}-\boldsymbol{\mu})^T V^{-1}(\mathbf{u}-\boldsymbol{\mu})\right\}^{1/2}\right)\right]$$

donde la esperanza se toma respecto del vector aleatorio $(\mathbf{u}, \mathbf{v})$. Si ρ cumple las hipótesis del Lema F.6, entonces $\rho \circ \sqrt{\cdot}$ también las cumple, de modo que por el Lema F.6 tenemos

$$E\left[\rho\left(\left\{(\mathbf{u}-\boldsymbol{\mu})^T V^{-1}(\mathbf{u}-\boldsymbol{\mu})\right\}^{1/2}\right)\right] > E\left[\rho\left(\left\{a\mathbf{u}^T\mathbf{u}\right\}^{1/2}\right)\right]$$

y la desigualdad es estricta con probabilidad mayor a cero. Luego, el lema vale. La demostración es similar cuando vale (ii). ■

Lema F.8 *Supongamos que* ρ_1 *y* ρ_2 *satisfacen A1-A2, y también que* ρ_2 *satisface A3. Sea* $(\mathbf{x}, y)$ *que satisface el modelo PFC y asumamos que la distribución del error* $\mathbf{u}$ *está en la clase* $\mathcal{P}_p(W_p, M)$. *Sea* $\boldsymbol{\theta} = (\boldsymbol{\mu}, B, \Delta) \in \Theta$, *sea* $\lambda > 0$ *tal que* $\det(\lambda\Delta_0) = \det(\Delta)$ *y supongamos que* $(\boldsymbol{\mu}, B, \Delta) \neq (\boldsymbol{\mu}_0, B_0, \lambda\Delta_0) = \boldsymbol{\theta}_0$, *en el sentido que* $B_0 \neq B$, *o bien* $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}$, *o* $\lambda\Delta_0 \neq \Delta$. *Entonces*

$$\Phi_{B,P}(\boldsymbol{\mu}_0, B_0, \lambda\Delta_0) < \Phi_{B,P}(\boldsymbol{\mu}, B, \Delta),$$

donde

$$\Phi_{B,P}(\boldsymbol{\mu}, B, \Delta) = \det(\Delta)\left[S_M^2(H_{P,\boldsymbol{\theta}})\right]^p \left[E_{H_{P,\boldsymbol{\theta}}}\left(\rho_2\left(\frac{v}{S_M(H_{P,\boldsymbol{\theta}})}\right)\right)\right]^p$$

y P *es la medida de probabilidad de* $(\mathbf{x}, y)$.

Demostración. $(\mathbf{x}, y)$ satisface el modelo PFC, de modo que

$$\mathbf{x} = \boldsymbol{\mu}_0 + B_0\mathbf{f}(y) + \Delta_0^{1/2}\mathbf{u}$$

entonces, para cada $\Sigma \in PDS(p)$, tenemos

$$\begin{aligned}
&(\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y))^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y)) \\
&= \left(\boldsymbol{\mu}_0 + B_0\mathbf{f}(y) + \Delta_0^{1/2}\mathbf{u} - \boldsymbol{\mu} - B\mathbf{f}(y)\right)^T \Sigma^{-1} \\
&\left(\boldsymbol{\mu}_0 + B_0\mathbf{f}(y) + \Delta_0^{1/2}\mathbf{u} - \boldsymbol{\mu} - B\mathbf{f}(y)\right) \\
&= \left(\mathbf{u} - \left[\Delta_0^{-1/2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) + \Delta_0^{-1/2}(B-B_0)\mathbf{f}(y)\right]\right)^T \Delta_0^{1/2}\Sigma^{-1} \\
&\Delta_0^{1/2}\left(\mathbf{u} - \left[\Delta_0^{-1/2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) + \Delta_0^{-1/2}(B-B_0)\mathbf{f}(y)\right]\right) \\
&= (\mathbf{u}-\mathbf{v})^T \left(\Delta_0^{-1/2}\Sigma\Delta_0^{-1/2}\right)^{-1} (\mathbf{u}-\mathbf{v}) = (\mathbf{u}-\mathbf{v})^T V^{-1} (\mathbf{u}-\mathbf{v}) \qquad (140)
\end{aligned}$$

donde llamamos

$$\mathbf{v} = \Delta_0^{-1/2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) + \Delta_0^{-1/2}(B-B_0)\mathbf{f}(y) \qquad (141)$$

$$V = \Delta_0^{-1/2}\Sigma\Delta_0^{-1/2} \in PDS(p). \qquad (142)$$

Para poder comparar $\Phi_{B,P}((\boldsymbol{\theta}_0)_\lambda)$ con $\Phi_{B,P}(\boldsymbol{\theta})$ necesitamos comparar $S_M^2\left(H_{P,(\theta_0)_\lambda}\right)$ con $S_M^2(H_{P,\theta})$. Por definición de S_M^2,

$$\begin{aligned}
\kappa_1 &= E_{H_{P,\boldsymbol{\theta}}}\left(\rho_1\left(\frac{v}{S_M(H_{P,\boldsymbol{\theta}})}\right)\right) \\
&= E_P\left(\rho_1\left(\left\{\frac{(\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y))^T\Delta^{-1}(\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y))}{S_M^2(H_{P,\boldsymbol{\theta}})}\right\}^{1/2}\right)\right) \\
&= E_P\left(\rho_1\left(\left\{\frac{(\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y))^T(\Delta/\lambda)^{-1}(\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y))}{\lambda S_M^2(H_{P,\boldsymbol{\theta}})}\right\}^{1/2}\right)\right) \\
&= E_P\left(\rho_1\left(\left\{\frac{(\mathbf{u}-\mathbf{v})^T V^{-1}(\mathbf{u}-\mathbf{v})}{\lambda S_M^2(H_{P,\boldsymbol{\theta}})}\right\}^{1/2}\right)\right)
\end{aligned}$$

donde $\mathbf{v}$ y V fueron definidas en (141) y (142) con $\Sigma = \Delta/\lambda$. Entonces, como

$$\det\left(V^{-1}\right) = 1,$$

si tomamos $\rho(t) = \rho_1\left(\frac{t}{\sqrt{\lambda S_M^2(H_{P,\boldsymbol{\theta}})}}\right)$ entonces está en las hipótesis del Lema F.7 y a

partir de él tenemos

$$\begin{aligned}\kappa_1 &= E_P\left(\rho\left(\left\{(\mathbf{u}-\mathbf{v})^T V^{-1}(\mathbf{u}-\mathbf{v})\right\}^{1/2}\right)\right)\\ &> E_P\left(\rho\left(\left\{\mathbf{u}^T\mathbf{u}\right\}^{1/2}\right)\right)\\ &= E_P\left(\rho_1\left(\left\{\frac{1}{\left[\lambda S_M^2\left(H_{P,\boldsymbol{\theta}}\right)\right]}\mathbf{u}^T\mathbf{u}\right\}^{1/2}\right)\right).\end{aligned} \tag{143}$$

También,

$$\begin{aligned}\kappa_1 &= E_{H_{P,(\boldsymbol{\theta}_0)_\lambda}}\left(\rho_1\left(\frac{v}{S_M\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)}\right)\right)\\ &= E_P\left(\rho_1\left(\left\{\frac{(\mathbf{x}-\boldsymbol{\mu}_0-B_0\mathbf{f}(y))^T\Delta_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0-B_0\mathbf{f}(y))}{\lambda S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)}\right\}^{1/2}\right)\right)\\ &= E_P\left(\rho_1\left(\left\{\frac{1}{\lambda S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)}\mathbf{u}^T\mathbf{u}\right\}^{1/2}\right)\right).\end{aligned} \tag{144}$$

donde la última igualdad se deduce de (140) para $\boldsymbol{\mu}=\boldsymbol{\mu}_0, \Sigma=\Delta_0$, y $B=B_0$, donde obtenemos $\mathbf{v}=\mathbf{0}$ y $V=I$. De (143) y (144), como $\rho_1\circ\sqrt{\cdot}$ es no decreciente, tenemos

$$S_M^2\left(H_{P,\boldsymbol{\theta}}\right) > S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right). \tag{145}$$

Además, como ρ_2 satisface $A1-A3$, si tomamos $H=H_{P,\boldsymbol{\theta}}$ y la relación (145) obtenemos la siguiente primera desigualdad

$$\begin{aligned}&\Phi_{B,P}(\boldsymbol{\theta})\\ &= \det(\Delta)\left[S_M^2\left(H_{P,\boldsymbol{\theta}}\right)E_{H_{P,\boldsymbol{\theta}}}\left(\rho_2\left(\frac{v}{S_M\left(H_{P,\boldsymbol{\theta}}\right)}\right)\right)\right]^p\\ &\geq \det(\Delta)\left[S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)E_{H_{P,\boldsymbol{\theta}}}\left(\rho_2\left(\frac{v}{S_M\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)}\right)\right)\right]^p\\ &= \det(\lambda\Delta_0)\left[S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)\right]^p\\ &\cdot\left[E_P\left(\rho_2\left(\left\{\frac{(\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y))^T\Delta^{-1}(\mathbf{x}-\boldsymbol{\mu}-B\mathbf{f}(y))}{S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)}\right\}^{1/2}\right)\right)\right]^p\\ &= \det(\lambda\Delta_0)\left[S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)\right]^p\\ &\cdot\left[E_P\left(\rho_2\left(\left\{(\mathbf{u}-\mathbf{v})^T V^{-1}(\mathbf{u}-\mathbf{v})\right\}^{1/2}\right)\right)\right]^p\end{aligned}$$

donde estamos usando (140) para $\Sigma = \Delta S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)$. Entonces, como

$$\det\left(V^{-1}\right) = \left[\frac{1}{\lambda S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)}\right]^p$$

a partir del Lema F.7 obtenemos

$$\begin{aligned}
&\Phi_{B,P}(\boldsymbol{\theta}) \\
&\geq \det(\lambda\Delta_0)\left[S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)\right]^p\left[E_P\left(\rho_2\left(\left\{(\mathbf{u}-\mathbf{v})^T V^{-1}(\mathbf{u}-\mathbf{v})\right\}^{1/2}\right)\right)\right]^p \\
&> \det(\lambda\Delta_0)\left[S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)\right]^p\left[E_P\left(\rho_2\left(\left\{\frac{\mathbf{u}^T\mathbf{u}}{\lambda S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)}\right\}^{1/2}\right)\right)\right]^p \\
&= \det(\lambda\Delta_0)\left[S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)\right]^p\left[E_P\left(\rho_2\left(\left\{\frac{\mathbf{u}^T(I\lambda)^{-1}\mathbf{u}}{S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)}\right\}^{1/2}\right)\right)\right]^p \\
&= \det(\lambda\Delta_0)\left[S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)\right]^p \\
&\cdot\left[E_P\left(\rho_2\left(\left\{\frac{(\mathbf{x}-\boldsymbol{\mu}_0-B_0\mathbf{f}(y))^T(\lambda\Delta_0)^{-1}(\mathbf{x}-\boldsymbol{\mu}_0-B_0\mathbf{f}(y))}{S_M^2\left(H_{P,(\boldsymbol{\theta}_0)_\lambda}\right)}\right\}^{1/2}\right)\right)\right]^p \\
&= \Phi_{B,P}\left((\boldsymbol{\theta}_0)_\lambda\right)
\end{aligned}$$

donde la cuarta afirmación se debe a la identidad (140) para $\boldsymbol{\mu} = \boldsymbol{\mu}_0, \Sigma = \lambda\Delta_0$, y $B = B_0$, (de modo que obtenemos $\mathbf{v} = \mathbf{0}$ y $V = I\lambda$). Esto prueba el resultado. ■

Demostración del Teorema 6.2. (Existencia) La distribución del error $\mathbf{u}$, $F_{\mathbf{u}} \in \mathcal{P}_p(W_p, M)$, entonces $\mathbf{u}$ es absolutamente continua, luego, por la Observación 6.1, satisface la propiedad S_ε para todo $0 < \varepsilon \leq 1$, de modo que de acuerdo al Teorema 6.1 existe al menos una solución al problema planteado por el $\tau-$funcional de estimación para P_0.

(Unicidad) Primero chequeamos que $(\boldsymbol{\theta}_0)_c$ satisface la restricción necesaria sobre la $\tau-$escala. Por la Observación 5.1 (iii), tenemos

$$S_\tau^2\left(H_{P_0,(\boldsymbol{\theta}_0)_c}\right) = \frac{1}{c}S_\tau^2\left(H_{P_0,\boldsymbol{\theta}_0}\right) = \frac{\kappa_2}{S_\tau^2\left(H_{P_0,\boldsymbol{\theta}_0}\right)}S_\tau^2\left(H_{P_0,\boldsymbol{\theta}_0}\right) = \kappa_2.$$

Sea $\theta = (\boldsymbol{\mu}, B, \Delta) \in\Theta$ tal que $\Delta > 0$, $\operatorname{rank}(\Gamma) = d$ y $S_\tau^2\left(H_{P_0,\boldsymbol{\theta}}\right) = \kappa_2$.Queremos probar que $\det(c\Delta_0) < \det(\Delta)$. Sea $\lambda = \left(\frac{\det(\Delta)}{\det(\Delta_0)}\right)^{1/p}$, de modo que

$$\det(\lambda\Delta_0) = \lambda^p \det(\Delta_0) = \det(\Delta). \tag{146}$$

Si $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, ó $B \neq B_0$ ó $\Delta \neq \lambda\Delta_0$, por el Lema F.8 tenemos

$$\Phi_{B,P_0}\left(\boldsymbol{\mu}_0, B_0, \lambda\Delta_0\right) < \Phi_{B,P_0}\left(\boldsymbol{\mu}, B, \Delta\right),$$

o equivalentemente

$$\det\left(\lambda\Delta_0\right)\left[S_\tau^2\left(H_{P_0,(\boldsymbol{\theta}_0)_\lambda}\right)\right]^p < \det\left(\Delta\right)\left[S_\tau^2\left(H_{P_0,\boldsymbol{\theta}}\right)\right]^p$$

lo cual, por (146), es equivalente a

$$\left[S_\tau^2\left(H_{P_0,(\boldsymbol{\theta}_0)_\lambda}\right)\right]^p < \left[S_\tau^2\left(H_{P_0,\boldsymbol{\theta}}\right)\right]^p = \left[\kappa_2\right]^p. \tag{147}$$

Por la Observación 5.1 (iii), tenemos

$$S_\tau^2\left(H_{P_0,(\boldsymbol{\theta}_0)_\lambda}\right) = S_\tau^2\left(H_{P_0,(\boldsymbol{\theta}_0)_{\frac{\lambda}{c}c}}\right) = \frac{c}{\lambda}S_\tau^2\left(H_{P_0,(\boldsymbol{\theta}_0)_c}\right) = \frac{c}{\lambda}\kappa_2.$$

Entonces, (147) es equivalente a

$$c^p < \lambda^p = \left(\frac{\det\left(\Delta\right)}{\det\left(\Delta_0\right)}\right)$$

o

$$c^p \det\left(\Delta_0\right) = \det\left(c\Delta_0\right) < \det\left(\Delta\right),$$

lo cual completa la prueba. ■

G. Apéndice del Capítulo 7

G.1. Preliminares

Demostración del Lema 7.1. Sea $g_k(\mathbf{x}, y) = g(\mathbf{x}, y, \boldsymbol{\theta}_k)$ y $g_L(\mathbf{x}, y) = g(\mathbf{x}, y, \boldsymbol{\theta}_L)$. Entonces, el subconjunto

$$E = \left\{ \mathbf{z} = (\mathbf{x}, y) \in \mathbb{R}^{p+1} : \exists (\mathbf{z}_k)_k \text{ tales que } \mathbf{z}_k \to \mathbf{z} \text{ y } g_k(\mathbf{z}_k) \nrightarrow g(\mathbf{z}) \right\}$$

satisface $E \subset D_{\mathbf{f}}$ que tiene $P-$probabilidad cero. Luego, aplicamos el Teorema 7.2 (Teorema 5.5 Billingsley [1968]), por lo que resulta que $P_k g_k^{-1}$ converge débilmente a Pg^{-1}. En particular, para toda función acotada y (uniformemente) continua $w : [0, \infty) \to [0, \infty)$, tenemos que

$$\lim_{k \to \infty} E_{P_k g_k^{-1}}(w(t)) = E_{Pg^{-1}}(w(t))$$

Si tomamos $w(t) = tI_{[0,a_1]}(t) + a_1 I_{(a_1, +\infty)}(t)$, obtenemos, usando el teorema de cambio de variables (Teorema E.1) en la tercera igualdad que sigue,

$$\begin{aligned}
\lim_{k \to \infty} E_{P_k g_k^{-1}}(w(t)) &= E_{Pg^{-1}}(w(t)) \\
\lim_{k \to \infty} \int w(t)\, dP_k g_k^{-1}(t) &= \int w(t)\, dPg^{-1}(t) \\
\lim_{k \to \infty} \int w(g(\mathbf{x}, y, \boldsymbol{\theta}_k))\, dP_k(\mathbf{x}, y) &= \int w(g(\mathbf{x}, y, \boldsymbol{\theta}_L))\, dP(\mathbf{x}, y) \\
\lim_{k \to \infty} \int g(\mathbf{x}, y, \boldsymbol{\theta}_k)\, dP_k(\mathbf{x}, y) \overset{l}{-} &= \int g(\mathbf{x}, y, \boldsymbol{\theta}_L)\, dP(\mathbf{x}, y),
\end{aligned}$$

lo cual prueba el resultado. ■

G.2. Demostraciones de convergencia

Queremos usar el Teorema 3.4 de Rao [1962] para la clase de funciones $\mathcal{A} \subset C_2\left(\mathbb{R}^{p+1}\right)$ dada por

$$\mathcal{A} = \left\{ \begin{array}{c} g : \mathbb{R}^{p+1} \to \mathbb{R}^2 : g(\mathbf{x}, y) = (g_1(\mathbf{x}, y), g_2(\mathbf{x}, y)), g_1, g_2 \in \mathcal{G}, \\ \sup_{\|(\mathbf{x}, y)\| \leq 1} \{|g_1(\mathbf{x}, y)|, |g_2(\mathbf{x}, y)|\} = 1 \end{array} \right\}, \tag{148}$$

donde la clase $\mathcal{G}$ de funciones está definida en (53).

Lema G.1 *La clase $\mathcal{A}$ definida en (148) es (u.c.c.) compacta.*

Demostración. Claramente, $\mathcal{A} \subset C_2\left(\mathbb{R}^{p+1}\right)$, las funciones continuas de $\mathbb{R}^{p+1}$ a $\mathbb{R}^2$, puesto que la función $\mathbf{f}$ definida en el modelo PFC es continua. Debemos ver que

(i) $\mathcal{A}$ es cerrada.

(ii) Para cada $(\mathbf{x}, y) \in \mathbb{R}^{p+1}$,

$$\sup_{g \in \mathcal{A}} \|g(\mathbf{x}, y)\| = \sup_{g_1, g_2 \in \mathcal{G}: \|(g_1, g_2)\|_\infty = 1} \{|g_1(\mathbf{x}, y)|, |g_2(\mathbf{x}, y)|\} < \infty.$$

(iii) $\mathcal{A}$ es equicontinua en cada $x \in \mathcal{X}$.

La primera condición nos dirá que $\overline{\mathcal{A}} = \mathcal{A}$ y a partir de (ii) y (iii) tendremos que $\mathcal{A}$ es (u.c.c.) pre-compacta. Juntando ambas tenemos que $\mathcal{A}$ es (u.c.c.) compacta. Nos resta probar las tres propiedades recién enunciadas.

(i) Veamos que $\mathcal{A}$ es cerrada. Sea $(g_n)_{n \geq 1} = ((g_{1,n}, g_{2,n}))_{n \geq 1} \subset \mathcal{A}$ una sucesión tal que $g_n \overset{ucc}{\to} g = (g_1, g_2)$. Como converge uniformemente sobre compactos, y la bola unitaria en $\mathbb{R}^{p+1}$ es compacta, resulta que

$$\lim_{n \to \infty} \sup_{\|(\mathbf{x}, y)\| \leq 1} \max\{|g_{1,n}(\mathbf{x}, y) - g_1(\mathbf{x}, y)|, |g_{2,n}(\mathbf{x}, y) - g_2(\mathbf{x}, y)|\} = 0,$$

o equivalentemente

$$\lim_{n \to \infty} \|g_{1,n} - g_1\|_\infty = \lim_{n \to \infty} \|g_{2,n} - g_2\|_\infty = 0,$$

luego $g_n \overset{\|\cdot\|_\infty}{\to} g$, o equivalentemente, $g_{1,n} \overset{\|\cdot\|_\infty}{\to} g_1$ y $g_{2,n} \overset{\|\cdot\|_\infty}{\to} g_2$. A partir de esto se deducen dos cosas: como $g_{1,n}, g_{2,n} \in \mathcal{G}$ que es un subespacio vectorial de dimensión finita, es cerrado y por lo tanto $g_1, g_2 \in \mathcal{G}$. Y además, resulta $\|g_{i,n}\|_\infty \to \|g_i\|_\infty$ para $i = 1, 2$. Luego $g \in \mathcal{A}$, y hemos probado que $\mathcal{A}$ es cerrada.

(ii) Fijemos $(\mathbf{x}, y)$. Consideremos el siguiente operador

$$T_{(\mathbf{x},y)} : \mathcal{G} \times \mathcal{G} \subset C_2\left(\mathbb{R}^{p+1}\right) \to \left(\mathbb{R}^2, \|\cdot\|_\infty\right)$$

dado por

$$T_{(\mathbf{x},y)}(g) = T_{(\mathbf{x},y)}(g_1, g_2) = (g_1(\mathbf{x}, y), g_2(\mathbf{x}, y)),$$

es decir, $T_{(\mathbf{x},y)}$ es el operador *evaluar en* $(\mathbf{x}, y)$. Claramente, $T_{(\mathbf{x},y)}$ es un operador lineal. Como $\mathcal{G} \times \mathcal{G}$ es un espacio vectorial de dimensión finita, entonces todo operador lineal definido en él es continuo. Por definición, $\mathcal{A} \subset C_2\left(\mathbb{R}^{p+1}\right)$ es acotado en norma $\|\cdot\|_\infty$. Como $T_{(\mathbf{x},y)}$ es continuo, su imagen $T_{(\mathbf{x},y)}(\mathcal{A})$ es acotada en $\mathbb{R}^2$, es decir, existe c constante tal que,

$$\left\|T_{(\mathbf{x},y)}(g)\right\|_\infty = \max\{|g_1(\mathbf{x}, y)|, |g_2(\mathbf{x}, y)|\} \leq c,$$

para todo $g \in \mathcal{A}$, o, equivalentemente,

$$\sup_{g: \|g\|_\infty = 1} \left\|T_{(\mathbf{x},y)}(g)\right\|_\infty < \infty.$$

Luego, hemos probado (ii).

(iii) $\mathcal{G}$ es un espacio vectorial de dimensión finita, escribimos $\mathcal{G} = \operatorname{span}\{h_1, \dots h_M\}$. Sea $g \in \mathcal{A}$. Luego existen constantes $(a_{ji})_{j=1,2;1\leq i\leq M}$ tales que

$$g_j = \sum_{i=1}^{M} a_{ji} h_i, \quad j = 1, 2,$$

y

$$\|g(\mathbf{x}, y) - g(\mathbf{x}_0, y_0)\|_\infty = \sup_{j=1,2} \left| \sum_{i=1}^{M} a_{ji} \left[h_i(\mathbf{x}, y) - h_i(\mathbf{x}_0, y_0) \right] \right| \tag{149}$$

Como $g \in \mathcal{A}$ resulta $\|g_j\|_\infty \leq 1$. Como g_j es una expresión lineal en las constantes, la condición $|g_j(\mathbf{x}, y)| \leq 1$ $(j = 1, 2)$ para varios puntos $(\mathbf{x}, y)$ garantiza que existe una constante C que acota superiormente todas las constantes, es decir

$$|a_{ij}| \leq C$$

para $j = 1, 2; 1 \leq i \leq M$. De (149) tenemos

$$\begin{aligned} \|g(\mathbf{x}, y) - g(\mathbf{x}_0, y_0)\|_\infty &\leq \sup_{j=1,2} \sum_{i=1}^{M} |a_{ji}| \, |h_i(\mathbf{x}, y) - h_i(\mathbf{x}_0, y_0)| \\ &\leq C \sum_{i=1}^{M} |h_i(\mathbf{x}, y) - h_i(\mathbf{x}_0, y_0)| \end{aligned} \tag{150}$$

Sea $\varepsilon > 0$ y fijemos $(\mathbf{x}_0, y_0) \in \mathbb{R}^{p+1}$. Para cada i existe un entorno N_i de $(\mathbf{x}_0, y_0)$ para el cual

$$|h_i(\mathbf{x}, y) - h_i(\mathbf{x}_0, y_0)| < \frac{\varepsilon}{CM} \qquad \text{para todo } (\mathbf{x}, y) \in N_i.$$

Luego $N = \bigcap_{i=1}^{M} N_i$ es un entorno de $(\mathbf{x}_0, y_0)$. Sea $(\mathbf{x}, y) \in N$, entonces, de (150) resulta

$$\|g(\mathbf{x}, y) - g(\mathbf{x}_0, y_0)\|_\infty \leq C \sum_{i=1}^{M} \frac{\varepsilon}{CM} < \varepsilon,$$

probando que la familia $\mathcal{A}$ es equicontinua, y por lo tanto el resultado buscado. ■

El siguiente lema nos permite pasar de la convergencia débil de probabilidades a una convergencia uniforme en una familia de conjuntos, y será usado en la prueba del Teorema 7.4.

Lema G.2 *Sea $(P_k)_k$ una sucesión de medidas de probabilidad definidas en $\mathbb{R}^{p+1}$ que converge débilmente a una medida de probabilidad P cuando $k \to \infty$. Asumimos que P es absolutamente continua (con respecto a la medida de Lebesgue en $\mathbb{R}^{p+1}$). Entonces,*

$$\sup_{D \in \mathcal{D}_2} |P_k(D) - P(D)| \underset{k\to\infty}{\to} 0,$$

donde $\mathcal{D}_2$ es la clase de conjuntos definida en (54).

Demostración. Queremos usar el Teorema 3.4 de Rao [1962] para la clase de funciones $\mathcal{A} \subset C_2\left(\mathbb{R}^{p+1}\right)$ definida en (148). Por el lema anterior tenemos que $\mathcal{A}$ es (u.c.c.) compacta. Como P es absolutamente continua y cada función $g \in \mathcal{A}$ es continua, tenemos que $P \circ g^{-1}$ tiene distribuciones marginales continuas. Luego, por el Teorema 7.3 tenemos que (57) vale para los conjuntos A de la forma

$$A = \left\{(\mathbf{x}, y) \in \mathbb{R}^{p+1} : g_1(\mathbf{x}, y) \leq a_1,\ g_2(\mathbf{x}, y) \leq a_2\right\} \tag{151}$$

donde $g_1, g_2 \in \mathcal{G}$, tales que $\|g\|_\infty = \sup_{\|(\mathbf{x},y)\| \leq 1} \{|g_1(\mathbf{x}, y)|, |g_2(\mathbf{x}, y)|\} = 1$. Queremos probar que (57) es válido para los conjuntos de la clase $\mathcal{D}_2$. Sea A de la forma (151), entonces $g_1 \in \mathcal{G}$ si y sólo si $-g_1 + a_1 \in \mathcal{G}$, y podemos escribir

$$\begin{aligned} A &= \{: g_1(\mathbf{x}, y) \leq a_1\} \cap \{(\mathbf{x}, y) : g_2(\mathbf{x}, y) \leq a_2\} \\ &= \{0 \leq -g_1 + a_1\} \cap \{0 \leq a_2 - g_2\} \\ &= \{0 \leq \widetilde{g}_1\} \cap \{0 \leq \widetilde{g}_2\} \end{aligned}$$

con $\widetilde{g}_1, \widetilde{g}_2 \in \mathcal{G}$, de modo que $A \in \mathcal{D}_2$. Recíprocamente, si elegimos $A \in \mathcal{D}_2$, entonces existen conjuntos $C, D \in \mathcal{D}_1$ tales que $A = C \cap D$. Luego, existen $g_1, g_2 \in \mathcal{G}$:

$$C = \{g_1 \geq 0\}, \qquad D = \{g_2 \geq 0\}.$$

Entonces, llamando

$$a = \sup_{(\mathbf{x},y):\|(\mathbf{x},y)\| \leq 1} \{|g_1(\mathbf{x}, y)|, |g_2(\mathbf{x}, y)|\}$$

podemos definir

$$g = \frac{1}{a}(g_1, g_2).$$

Luego,

i. $g \in \mathcal{A}$ y podemos escribir

ii. $C = \left\{-\frac{g_1}{a} \leq 0\right\}, D = \left\{-\frac{g_1}{a} \leq 0\right\},$

de modo que A resulta de la forma (151). Y hemos probado el resultado. ■

Demostración del Teorema 7.4. Recordemos que

$$\widetilde{\boldsymbol{\theta}}_C(P) = \left(\widetilde{D}_C(P), \widetilde{\Delta}_C(P)\right) = \left(\widetilde{\boldsymbol{\mu}}_C(P), \widetilde{B}_C(P), \widetilde{\Delta}_C(P)\right),$$

entonces por el Lema 5.1 podemos asumir que $\widetilde{\boldsymbol{\mu}}_C(P) = \mathbf{0}$ y $\widetilde{\Delta}_C(P) = I$.

P satisface (S_ε) para algún $0 < \varepsilon < 1 - \frac{\kappa_1}{a_1}$, esto es

$$P\left(H_{(\mathbf{a},\mathbf{b},c)}\right) < \varepsilon$$

para todo $\mathbf{a} \in \mathbb{R}^p, \mathbf{b} \in \mathbb{R}^r, c \in \mathbb{R}$. Entonces, por (55) y el Lema G.2 tenemos

$$\left|P_k\left(H_{(\mathbf{a},\mathbf{b},c)}\right) - P\left(H_{(\mathbf{a},\mathbf{b},c)}\right)\right| \leq \sup_{D \in \mathcal{D}_2} |P_k(D) - P(D)| \underset{k \to \infty}{\to} 0,$$

luego para $\delta > 0$, existe $k_0 = k_0(\delta)$ tal que para todo $k \geq k_0$ se cumple

$$\sup_{D \in \mathcal{D}_2} |P_k(D) - P(D)| < \delta.$$

A partir de las últimas tres desigualdades, tenemos

$$P_k\left(H_{(\mathbf{a},\mathbf{b},c)}\right) < \delta + \varepsilon < 1 - \frac{\kappa_1}{a_1} \text{ para todo } k > k_0 \text{ y todo } (\mathbf{a},\mathbf{b},c),$$

si tomamos $\delta < \left(1 - \frac{\kappa_1}{a_1} + \varepsilon\right)\frac{1}{2}$. Entonces, para todo $k \geq k_0$, P_k satisface (S_ε) para algún $0 < \varepsilon < 1 - \frac{\kappa_1}{a_1}$ y, en virtud del Teorema 6.1, el problema de minimizar Φ_{B,P_k} sujeta a la restricción (43) tiene al menos una solución $\widetilde{\boldsymbol{\theta}}_C(P_k) = \left(\widetilde{D}_C(P_k), \widetilde{\Delta}_C(P_k)\right) = \boldsymbol{\theta}_k$.

Usando que $\boldsymbol{\theta}_k$ satisface la restricción (43) tenemos que

$$\begin{aligned}
\kappa_1 &= E_{H_{P_k,\boldsymbol{\theta}_k}}(\rho_1(v)) = \int \rho_1(v)\, dH_{P_k,\boldsymbol{\theta}_k} \\
&= \int \rho_1\left(\left\{\left[\mathbf{x} - \widetilde{D}_C(P_k)\,\widetilde{\mathbf{f}}(y)\right]^T \left[\widetilde{\Delta}_C(P_k)\right]^{-1} \left[\mathbf{x} - \widetilde{D}_C(P_k)\,\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right) dP_k(\mathbf{x}, y) \qquad (152) \\
&= \int_{E_k} + \int_{E_k^c}
\end{aligned}$$

tomando $E_k = E\left(\widetilde{D}_C(P_k), \widetilde{\Delta}_C(P_k), c_1\right)$ donde llamamos $c_1 = \text{mín}\{t > 0 : \rho_1(t) = a_1\}$ y E está definido en (51), y además, no escribimos la función a ser integrada pues es la misma que en (152). La segunda integral es igual a $a_1 P_k(E_k^c)$ y la primera es positiva, de modo que

$$\kappa_1 \geq a_1 P_k(E_k^c) = a_1(1 - P_k(E_k))$$

o equivalentemente,

$$P_k(E_k) \geq 1 - \frac{\kappa_1}{a_1} > \varepsilon \text{ para todo } k > k_0. \qquad (153)$$

Supongamos que $P(E_k) < \varepsilon$. Usamos nuevamente el Lema G.2 para mostrar que, para $\delta_1 > 0$ existe un $k_0' = k_0'(\delta_1)$ tal que

$$\begin{aligned}
|P(E_k) - P_k(E_k)| &\leq \sup_{D \in \mathcal{D}_1} |P_k(D) - P(D)| \\
&\leq \sup_{D \in \mathcal{D}_2} |P_k(D) - P(D)| < \delta_1 \text{ para todo } k > k_0'
\end{aligned}$$

de modo que,

$$P_k(E_k) \leq |P(E_k) - P_k(E_k)| + P(E_k) < \delta_1 + \varepsilon \text{ para todo } k > k_0'$$

lo cual contradice (153) si $\delta_1 + \varepsilon < 1 - \frac{\kappa_1}{a_1}$, o equivalentemente, $\delta_1 < 1 - \frac{\kappa_1}{a_1} - \varepsilon$ y $k > \text{máx}\,(k_0, k_0')$. Entonces, hemos probado que

$$P\left(E_k\right) = P\left(E\left(\widetilde{D}_C\left(P_k\right), \widetilde{\Delta}_C\left(P_k\right), c_1\right)\right) \geq \varepsilon. \tag{154}$$

para todo k suficientemente grande.

De acuerdo al Lema F.4(i), esto significa que existe una constante $k_1 > 0$ tal que $\lambda_p\left(\widetilde{\Delta}_C\left(P_k\right)\right) \geq k_1$ eventualmente.

Para $\eta > 0$, consideremos los siguientes subconjuntos

$$F_1 = \left\{(\mathbf{x}, y) : \left\{\left[\mathbf{x} - \widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right]^T I^{-1}\left[\mathbf{x} - \widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2} \leq c_1\right\}$$

$$F_2 = \left\{(\mathbf{x}, y) : c_1 < \left\{\left[\mathbf{x} - \widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right]^T \left[\mathbf{x} - \widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2} < (1+\eta)\,c_1\right\}$$

$$F_3 = \left\{(\mathbf{x}, y) : (1+\eta)\,c_1 \leq \left\{\left[\mathbf{x} - \widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right]^T \left[\mathbf{x} - \widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right\}.$$

Como ρ_1 es estrictamente creciente en $[0, c_1]$ para $(\mathbf{x}, y) \in F_1$ tenemos

$$\rho_1\left(\frac{\left\|\mathbf{x} - \widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right\|}{1+\eta}\right) < \rho_1\left(\left\|\mathbf{x} - \widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right\|\right)$$

luego

$$\int_{F_1} \rho_1\left(\frac{\left\|\mathbf{x} - \widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right\|}{1+\eta}\right) dP(\mathbf{x}, y) < \int_{F_1} \rho_1\left(\left\|\mathbf{x} - \widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right\|\right) dP(\mathbf{x}, y).$$

Para $(\mathbf{x}, y) \in F_2$ tenemos

$$\rho_1\left(\frac{\left\|\mathbf{x} - \widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right\|}{1+\eta}\right) < a_1 = \rho_1\left(\left\|\mathbf{x} - \widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right\|\right),$$

y para $(\mathbf{x}, y) \in F_3$

$$\rho_1\left(\frac{\left\|\mathbf{x} - \widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right\|}{1+\eta}\right) = a_1 = \rho_1\left(\left\|\mathbf{x} - \widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right\|\right).$$

De modo que

$$E_P\left(\rho_1\left(\frac{\left\|\mathbf{x}-\widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right\|}{1+\eta}\right)\right)=\int\rho_1\left(\frac{\left\|\mathbf{x}-\widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right\|}{1+\eta}\right)dP(\mathbf{x},y) \qquad (155)$$
$$<\int\rho_1\left(\left\|\mathbf{x}-\widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right\|\right)dP(\mathbf{x},y)$$
$$=E_P\left(\rho_1\left(\left\|\mathbf{x}-\widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right\|\right)\right)=\kappa_1$$

donde la última igualdad vale pues $\widetilde{\boldsymbol{\theta}}_C(P)=\left(\widetilde{D}_C(P),\widetilde{\Delta}_C(P)\right)=\left(\mathbf{0},\widetilde{B}_C(P),I\right)$ satisface (43). Como P_k converge débilmente a P, y $\rho_1\left(\frac{\left\|\mathbf{x}-\widetilde{D}_C(P)\widetilde{\mathbf{f}}(y)\right\|}{1+\eta}\right)$ es una función real medible, por el Teorema 7.1 tenemos que

$$E_{P_k}\left(\rho_1\left(\frac{\left\|\mathbf{x}-\widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right\|}{1+\eta}\right)\right)\to E_P\left(\rho_1\left(\frac{\left\|\mathbf{x}-\widetilde{D}_C(P)\,\widetilde{\mathbf{f}}(y)\right\|}{1+\eta}\right)\right)<\kappa_1,$$

por lo que, para k suficientemente grande resulta $E_{P_k}\left(\rho_1\left(\frac{\left\|\mathbf{x}-\widetilde{D}_C(P)\widetilde{\mathbf{f}}(y)\right\|}{1+\eta}\right)\right)\leq\kappa_1$.

Esto significa que el vector $\left(\mathbf{0},\widetilde{B}_C(P),(1+\eta)^2I\right)$ satisface (134) para P_k para k suficientemente grande. A partir del Lema F.2 y de la definición de $\widetilde{\boldsymbol{\theta}}_C(P_k)$ tenemos que para $\eta>0$, vale

$$\Phi_{C,P_k}\left(\widetilde{\boldsymbol{\theta}}_C(P_k)\right)\leq\Phi_{C,P_k}\left(\mathbf{0},\widetilde{B}_C(P),(1+\eta)^2I\right),$$

esto es

$$|\widetilde{\Delta}_C(P_k)|\left(E_{P_k}\left(\rho_2\left(\left\{\left[\mathbf{x}-\widetilde{D}_C(P_k)\,\widetilde{\mathbf{f}}(y)\right]^T\left[\widetilde{\Delta}_C(P_k)\right]^{-1}\left[\mathbf{x}-\widetilde{D}_C(P_k)\,\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right)\right)\right)^p$$
$$\leq|(1+\eta)^2I|$$
$$\left(E_{P_k}\left(\rho_2\left(\left\{\left[\mathbf{x}-\left(\mathbf{0},\widetilde{B}_C(P)\right)\widetilde{\mathbf{f}}(y)\right]^T\left[(1+\eta)^2I\right]^{-1}\left[\mathbf{x}-\left(\mathbf{0},\widetilde{B}_C(P)\right)\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right)\right)\right)^p.$$

Tomando $\lim\sup_k$ obtenemos, para $\eta>0$,

$$\lim\sup_{k\to\infty}\Phi_{C,P_k}\left(\widetilde{\boldsymbol{\theta}}_C(P_k)\right)\leq|(1+\eta)^2I|\left(E_P\left(\rho_2\left(\frac{1}{(1+\eta)}\left\|\mathbf{x}-\left(\mathbf{0},\widetilde{B}_C(P)\right)\widetilde{\mathbf{f}}(y)\right\|\right)\right)\right)^p$$
$$=\Phi_{C,P}\left(\mathbf{0},\widetilde{B}_C(P),(1+\eta)^2I\right).$$

Cuando tomamos $\lim_{\eta \searrow 0}$ obtenemos, en virtud del teorema de la convergencia dominada, usando (155) para ρ_2

$$\begin{aligned} \lim \sup_{k\to\infty} \Phi_{C,P_k}\left(\widetilde{\boldsymbol{\theta}}_C\left(P_k\right)\right) &\leq \left(E_P\left(\rho_2\left(\left\|\mathbf{x}-\left(\mathbf{0},\widetilde{B}_C\left(P\right)\right)\widetilde{\mathbf{f}}(y)\right\|\right)\right)\right)^p \\ &= \Phi_{C,P}\left(\mathbf{0},\widetilde{B}_C\left(P\right),I\right). \end{aligned} \tag{156}$$

Llamemos $s_k := \left(E_{P_k}\left(\rho_2\left(\left\{\left[\mathbf{x}-\widetilde{D}_C\left(P_k\right)\widetilde{\mathbf{f}}(y)\right]^T\left[\widetilde{\Delta}_C\left(P_k\right)\right]^{-1}\left[\mathbf{x}-\widetilde{D}_C\left(P_k\right)\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right)\right)\right)^p$.

Por el Lema F.3, como $\widetilde{\boldsymbol{\theta}}_C\left(P_k\right)$ y P_k satisfacen la restricción (43) existe una constante $q>0$, que dependen de ρ_1,ρ_2 y κ_1 (pero no de P_k) tal que

$$\begin{aligned} q &\leq \int \rho_2\left(\left\{\left[\mathbf{x}-\widetilde{D}_C\left(P_k\right)\widetilde{\mathbf{f}}(y)\right]^T\left[\widetilde{\Delta}_C\left(P_k\right)\right]^{-1}\left[\mathbf{x}-\widetilde{D}_C\left(P_k\right)\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right) dP_k\left(\mathbf{x},y\right) \\ &= E_{P_k}\left(\rho_2\left(\left\{\left[\mathbf{x}-\widetilde{D}_C\left(P_k\right)\widetilde{\mathbf{f}}(y)\right]^T\left[\widetilde{\Delta}_C\left(P_k\right)\right]^{-1}\left[\mathbf{x}-\widetilde{D}_C\left(P_k\right)\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right)\right). \end{aligned}$$

de modo que tenemos

$$\begin{aligned} q^p &\leq s_k, \text{ para todo } k \\ \lim \sup_{k\to\infty} |\widetilde{\Delta}_C\left(P_k\right)| s_k &\leq \Phi_{C,P}\left(\mathbf{0},\widetilde{B}_C\left(P\right),I\right) \end{aligned} \tag{157}$$

Entonces, $\left(\left|\widetilde{\Delta}_C\left(P_k\right)\right|\right)_k$ está acotado por encima, pues si no lo estuviera, existiría una subsucesión tal que $\left|\widetilde{\Delta}_C\left(P_{k_j}\right)\right| \to +\infty$, de modo que $\left|\widetilde{\Delta}_C\left(P_{k_j}\right)\right| s_{k_j} \to +\infty$ lo cual contradice (157). Entonces, existe un $A>0$ tal que

$$\left|\widetilde{\Delta}_C\left(P_k\right)\right| \leq A.$$

Como hemos probado que existe una constante k_1 tal que $k_1 \leq \left[\lambda_p\left[\widetilde{\Delta}_C\left(P_k\right)\right]\right]$, tenemos

$$k_1^{p-1}\lambda_1\left[\widetilde{\Delta}_C\left(P_k\right)\right] \leq \left[\lambda_p\left[\widetilde{\Delta}_C\left(P_k\right)\right]\right]^{p-1}\lambda_1\left[\widetilde{\Delta}_C\left(P_k\right)\right] \leq \prod_{i=1}^{p}\lambda_i\left[\widetilde{\Delta}_C\left(P_k\right)\right] = \left|\widetilde{\Delta}_C\left(P_k\right)\right| \leq A,$$

de modo que $\lambda_1\left[\widetilde{\Delta}_C\left(P_k\right)\right]$ está acotada superiormente, para todo k suficientemente grande. Luego, podemos aplicar el Lema F.4(iii) a P y $\widetilde{\boldsymbol{\theta}}_C\left(P_k\right)$ para probar que existe un subconjunto compacto K de $\boldsymbol{\Theta}$ tal que para k suficientemente grande, $\widetilde{\boldsymbol{\theta}}_C\left(P_k\right)$ estará contenida en K. Por lo tanto, basta mostrar que toda subsucesión convergente $\left(\widetilde{\boldsymbol{\theta}}_C\left(P_{k_j}\right)\right)_j$ tiene límite $\left(\mathbf{0},\widetilde{B}_C\left(P\right),I\right) = \widetilde{\boldsymbol{\theta}}_C\left(P\right)$.

Sea $\boldsymbol{\theta}_{k_j} = \widetilde{\boldsymbol{\theta}}_C\left(P_{k_j}\right)$, $j \in \mathbb{N}$, una subsucesión para la cual $\lim_{j\to\infty} \boldsymbol{\theta}_{k_j} = \boldsymbol{\theta}_L$. Entonces, en virtud del Lema 7.1 para g definida en (56) tenemos

$$\lim_{j\to\infty} \int g\left(\mathbf{x}, y, \boldsymbol{\theta}_{k_j}\right) dP_{k_j}\left(\mathbf{x}, y\right) = \int g\left(\mathbf{x}, y, \boldsymbol{\theta}_L\right) dP\left(\mathbf{x}, y\right),$$

esto es

$$\kappa_1 = \lim_{j\to\infty} P_{k_j}\left(\rho_1\left(d\left(\cdot, \boldsymbol{\theta}_{k_j}\right)\right)\right) = E_P\left(\rho_1\left(d\left(\cdot, \boldsymbol{\theta}_L\right)\right)\right).$$

Esto significa que $\boldsymbol{\theta}_L$ satisface la restricción (43) que debe cumplir $\widetilde{\boldsymbol{\theta}}_C\left(P\right)$. Como este problema tiene a $\widetilde{\boldsymbol{\theta}}_C\left(P\right) = \left(\mathbf{0}, \widetilde{B}\left(P\right), I\right)$ como única solución, debe cumplirse que

$$\left|\Delta_L\right| \left\{E_P\left(\rho_2\left(d\left(\cdot, \boldsymbol{\theta}_L\right)\right)\right)\right\}^p = \Phi_{C,P}\left(\boldsymbol{\theta}_L\right) \geq \Phi_{C,P}\left(\widetilde{\boldsymbol{\theta}}_C\left(P\right)\right).$$

Como $\Delta_{k_j} \to \Delta_L$, tenemos $\left|\Delta_{k_j}\right| \to \left|\Delta_L\right|$. Por el Lema 7.1 aplicado a ρ_2 tenemos que

$$E_P\left(\rho_2\left(d\left(\cdot, \boldsymbol{\theta}_L\right)\right)\right) = \lim_{j\to\infty} E_{P_{k_j}}\left(\rho_2\left(d\left(\cdot, \boldsymbol{\theta}_{k_j}\right)\right)\right),$$

de modo que

$$\begin{aligned}
\Phi_{C,P}\left(\boldsymbol{\theta}_L\right) = \left|\Delta_L\right| \left\{E_P\left(\rho_2\left(d\left(\cdot, \boldsymbol{\theta}_L\right)\right)\right)\right\}^p &= \lim_{j\to\infty} \left|\Delta_{k_j}\right| \left\{E_{P_{k_j}}\left(\rho_2\left(d\left(\cdot, \boldsymbol{\theta}_{k_j}\right)\right)\right)\right\}^p \\
&= \lim_{j\to\infty} \Phi_{C,P_{k_j}}\left(\widetilde{\boldsymbol{\theta}}_C\left(P_{k_j}\right)\right) \\
&\leq \lim\sup_{k\to\infty} \Phi_{C,P_{k_j}}\left(\widetilde{\boldsymbol{\theta}}_C\left(P_k\right)\right) \leq \Phi_{C,P}\left(\widetilde{\boldsymbol{\theta}}_C\left(P\right)\right)
\end{aligned}$$

donde la última desigualdad se debe a (156). Por hipótesis, hay un único valor para $\widetilde{\boldsymbol{\theta}}_C\left(P\right)$, de modo que concluimos que $\boldsymbol{\theta}_L = \widetilde{\boldsymbol{\theta}}_C\left(P\right)$, lo cual prueba el resultado. ■

G.3. Demostraciones de consistencia

Necesitamos demostrar que la distribución empírica de las observaciones que siguen el modelo PFC cumple los requisitos de medibilidad establecidos en el Teorema 7.5 por Pollard, para la clase de subconjuntos $\mathcal{D}_2$, y también que $\mathcal{D}_2$ tiene discriminación polinomial.

Lema G.3 (Lema 18 Pollard [1984]) *Sea $\mathcal{G}$ un espacio vectorial de dimensión finita de funciones reales definidas en S. La clase de los conjuntos de la forma $\{g \geq 0\}$, para g en $\mathcal{G}$, tiene discriminación polinomial de grado no mayor que la dimensión de $\mathcal{G}$.*

Lema G.4 (Lema 15 Pollard [1984]) *Sean $\mathcal{C}$ y $\mathcal{D}$ dos clases de subconjuntos con discriminación polinomial,entonces también tienen discriminación polinomial cada una de las siguientes:*

(i) $\{D^c : D \in \mathcal{D}\}$;

(ii) $\{C \cup D : C \in \mathcal{C} \text{ y } D \in \mathcal{D}\}$;

(iii) $\{C \cap D : C \in \mathcal{C} \text{ y } D \in \mathcal{D}\}$.

Lema G.5 *Las clases de subconjuntos $\mathcal{E}$ y $\mathcal{H}$ (definidas en (50) y (52) respectivamente) tienen discriminación polinomial. También la tiene la clase $\mathcal{D}_2$ definida por (54),*

$$\mathcal{D}_2 = \{C \cap D : C \in \mathcal{D}_1 \text{ y } D \in \mathcal{D}_1\},$$

donde $\mathcal{D}_1$ es la clase de subconjuntos de la forma $\{g \geq 0 : g \in \mathcal{G}\}$, $\mathcal{G}$ fue definida en (53).

Demostración. Por el Lema G.3 sabemos que $\mathcal{D}_1$ tiene discriminación polinomial. Como observáramos al principio de esta Sección, $\mathcal{E}$ está contenida en $\mathcal{D}_1$, luego $\mathcal{E}$ tiene discriminación polinomial. También mostramos que $\mathcal{H} \subset \mathcal{D}_2$ en (55) por lo que el Lema G.4, $\mathcal{D}_2$ tiene discriminación polinomial, y de ahí se deduce que también la tiene $\mathcal{H}$. ■

Todavía nos resta demostrar que $\mathcal{D}_2$ satisface la condición de medibilidad del Teorema 7.5. Para ello, vamos a utilizar una serie de lemas y definiciones del mismo libro Pollard [1984] (Sección II) .

Definición G.1 *Se dice que una clase de funciones $\mathcal{F}$ es **universalmente separable** si existe una subclase numerable $\mathcal{F}_0$ tal que cada f en $\mathcal{F}$ puede escribirse como un límite puntual de una sucesión en $\mathcal{F}_0$.*

Lema G.6 *Sea $\mathcal{F}$ una clase de funciones medibles que es universalmente separable. Si $\mathcal{F}$ tiene una envolvente F, es decir, si existe una función F medible tal que $|f| \leq F$ para cada f en $\mathcal{F}$, para la cual $\int F(x)\, dP(x) < \infty$, entonces*

$$\sup_{f \in \mathcal{F}} \left| \int f(x)\, dP_k(x) - \int f(x)\, dP(x) \right| = \sup_{f \in \mathcal{F}} \left| \frac{1}{k} \sum_{i=1}^{k} f(X_i) - \int f(x)\, dP(x) \right|$$

es medible.

Demostración. Es el ejercicio 3, Capítulo II, de Pollard [1984]. ■

Lema G.7 *Sea $\mathcal{G}$ un espacio vectorial de dimensión finita de funciones reales de S, la clase $\mathcal{D}$ de los conjuntos de la forma $\{g \geq 0\}$ para g en $\mathcal{G}$, es universalmente separable.*

Demostración. Cada función g en $\mathcal{G}$ se puede expresar como una combinación lineal de alguna colección fija y finita de funciones no negativas. Sea $\mathcal{G}_0$ la subclase numerable generada al tomar coeficientes racionales. Para cada g en $\mathcal{G}$ existe una sucesión $(g_n)_n$ en $\mathcal{G}_0$ para la que $g_n \downarrow g$. Entonces $\{g_n \geq 0\} \downarrow \{g \geq 0\}$ puntualmente. Es el ejercicio 4, Capítulo II, del libro de Pollard [1984]. ■

Lema G.8 *Las operaciones del Lema G.4 preservan la separabilidad universal.*

Demostración. Es el ejercicio 5, Capítulo II, de Pollard [1984]. ■

Lema G.9 *Sea $(\mathbf{x}_i, y_i)_{i\geq 1}$ una sucesión de vectores aleatorios iid con distribución P que satisface el modelo PFC, y sea $(P_k)_{k\geq 1}$ la sucesión de medidas empíricas basadas en estas observaciones. Entonces*

$$\sup_{A\in\mathcal{D}_2} |P_k(A) - P(A)|$$

es una función medible. La clase de subconjuntos $\mathcal{D}_2$ fue definida en (54).

Demostración. Es una consecuencia de los últimos tres lemas, si consideramos a $\mathcal{F}$ como la clase de funciones $\mathcal{F} = \{I_A : A \in \mathcal{D}_2\}$, y I_A como la función indicadora . De esta manera la función constante 1 es una envolvente $P-$integrable para $\mathcal{F}$, y para cada subconjunto A tenemos $E_{P_k}(I_A) = P_k(A)$. ■

Demostración del Teorema 7.6. Debido al Teorema de Glivenko-Cantelli en $\mathbb{R}^{p+1}$, (ver por ejemplo, Shorack y Wellner [2009] , pág. 833) también tenemos que

$$\sup_{\mathbf{t}=(t_1,\ldots,t_d)\in\mathbb{R}^d} \left| P_k\left(\bigcap_{i=1}^{d}(-\infty,t_i]\right) - P\left(\bigcap_{i=1}^{d}(-\infty,t_i]\right)\right| \underset{k\to\infty}{\to} 0$$

casi seguramente, de modo que $(P_k)_k$ converge débilmente a P.

Además, por el Teorema 7.5, y los Lemas G.5 y G.9 tenemos que $\lim_{k\to\infty}\sup_{D\in\mathcal{D}_2}|P_k(D) - P(D)| = 0$ casi seguramente. ■

Demostración del Corolario 7.1. Para probar este resultado, hay que corregir a $\widetilde{\boldsymbol{\theta}}_C(P_k)$ para convertirlo en $\boldsymbol{\theta}_\tau(P_k)$. Sabemos que

$$\begin{aligned}\boldsymbol{\theta}_\tau(P_k) &= \left[\widetilde{\boldsymbol{\theta}}_C(P_k)\right]_{\frac{1}{\kappa_2}S_\tau^2\left(H_{P,\widetilde{\boldsymbol{\theta}}_C(P)}\right)} \\ &= \left(\widetilde{D}_C(P), \frac{1}{\kappa_2}S_\tau^2\left(H_{P,\widetilde{\boldsymbol{\theta}}_C(P)}\right)\widetilde{\Delta}_C(P)\right),\end{aligned} \tag{158}$$

donde recordemos que

$$S_\tau^2\left(H_{P_k,\widetilde{\boldsymbol{\theta}}_C(P_k)}\right) = S_M^2\left(H_{P_k,\widetilde{\boldsymbol{\theta}}_C(P_k)}\right) E_{H_{P_k,\widetilde{\boldsymbol{\theta}}_C(P_k)}}\left(\rho_2\left(\frac{v}{S_M\left(H_{P_k,\widetilde{\boldsymbol{\theta}}_C(P_k)}\right)}\right)\right).$$

Por la Observación 5.1 *i.* se tiene que, por definición, $S_M^2\left(H_{P_k,\boldsymbol{\theta}_C(P_k)}\right)=1$ para toda probabilidad tanto P_k como P, con lo cual

$$\begin{aligned}S_\tau^2\left(H_{P_k,\widetilde{\boldsymbol{\theta}}_C(P_k)}\right) &= E_{H_{P_k,\widetilde{\boldsymbol{\theta}}_C(P_k)}}\left(\rho_2\left(v\right)\right)\\ &= \int \rho_2\left(\left(\left\{\left[\mathbf{x}-\widetilde{D}_C\left(P_k\right)\widetilde{\mathbf{f}}(y)\right]^T\left[\widetilde{\Delta}_C\left(P_k\right)\right]^{-1}\left[\mathbf{x}-\widetilde{D}_C\left(P_k\right)\widetilde{\mathbf{f}}(y)\right]\right\}^{1/2}\right)\right)dP_k\left(\mathbf{x},y\right).\end{aligned}$$

En virtud del Teorema de Glivenko-Cantelli, del hecho de que $\widetilde{\boldsymbol{\theta}}_C\left(P_k\right)$ converge casi seguramente a $\widetilde{\boldsymbol{\theta}}_C\left(P\right)$ por el corolario anterior y del Lema 7.1, resulta que

$$\lim_{k\to\infty} S_\tau^2\left(H_{P_k,\widetilde{\boldsymbol{\theta}}_C(P_k)}\right)=S_\tau^2\left(H_{P,\widetilde{\boldsymbol{\theta}}_C(P)}\right)\text{ casi seguramente.}$$

Por (158), y el Corolario 7.6 tenemos entonces que

$$\begin{aligned}\lim_{k\to\infty}\boldsymbol{\theta}_\tau\left(P_k\right) &= \lim_{k\to\infty}\left(\widetilde{D}_C\left(P_k\right),\frac{1}{\kappa_2}S_\tau^2\left(H_{P_k,\widetilde{\boldsymbol{\theta}}_C(P_k)}\right)\widetilde{\Delta}_C\left(P_k\right)\right)\\ &= \left(\widetilde{D}_C\left(P\right),\frac{1}{\kappa_2}S_\tau^2\left(H_{P,\widetilde{\boldsymbol{\theta}}_C(P)}\right)\widetilde{\Delta}_C\left(P\right)\right)=\boldsymbol{\theta}_\tau\left(P\right),\end{aligned}$$

casi seguramente, probando la consistencia fuerte de los $\tau-$estimadores del modelo PFC. ■

H. Apéndice del Capítulo 8

H.1. Cálculo del valor del $\tau-$funcional

Al final de la demostración, en la Sección H.1.1, exhibimos todas las ecuaciones del Petersen y Pedersen [2008] que se utilizan para probar el Teorema 8.1.

Demostración del Teorema 8.1.

i. El valor de $\boldsymbol{\theta}= (\boldsymbol{\mu}, \Gamma, \beta, \Delta)$ que minimiza a $\ln(\Phi_{B,P}(\boldsymbol{\theta}))$ será en particular un punto crítico para dicha función. Por ello, queremos encontrar los valores de $\boldsymbol{\theta}$ que satisfacen

$$\bigtriangledown \ln(\Phi_{B,P}(\boldsymbol{\theta})) = 0. \tag{159}$$

Para simplificar la notación, llamemos $d(\boldsymbol{\theta}) = d(\mathbf{x},y,\boldsymbol{\theta})$ y $S(\boldsymbol{\theta}) = S_M(H_{P,\boldsymbol{\theta}})$. Recordemos que llamamos $\psi_k(t) = \rho_k'(t)$, para $k = 1, 2$.

Calculamos $\dfrac{\partial}{\partial \boldsymbol{\mu}} \ln(\Phi_{B,P}(\boldsymbol{\theta}))$.

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}} \ln(\Phi_{B,P}(\boldsymbol{\theta})) = &\frac{2p}{S(\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\mu}} S(\boldsymbol{\theta}) \\ &+ \frac{p}{E_P\left(\rho_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right)} \int \psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{\partial}{\partial \boldsymbol{\mu}} \left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) dP(\mathbf{x},y)\end{aligned} \tag{160}$$

Derivamos a la igualdad

$$E_P\left(\rho_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right) = \kappa_1$$

que define a $S(\boldsymbol{\theta})$, y obtenemos,

$$\begin{aligned}0 &= \int \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{\partial}{\partial \boldsymbol{\mu}} \left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) dP(\mathbf{x},y) \\ &= \int \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{1}{S^2(\boldsymbol{\theta})} \left[\frac{\partial d(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} S(\boldsymbol{\theta}) - d(\boldsymbol{\theta}) \frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}}\right] dP(\mathbf{x},y) \\ &= \int \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{1}{S(\boldsymbol{\theta})} \frac{\partial d(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} dP(\mathbf{x},y) - \int \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{d(\boldsymbol{\theta})}{S^2(\boldsymbol{\theta})} \frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} dP(\mathbf{x},y) \\ &= \int \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{1}{2d(\boldsymbol{\theta}) S(\boldsymbol{\theta})} \frac{\partial d^2(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} dP(\mathbf{x},y) \\ &- \frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} \int \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{d(\boldsymbol{\theta})}{S^2(\boldsymbol{\theta})} dP(\mathbf{x},y)\end{aligned} \tag{161}$$

donde la última igualdad vale pues S no depende de $(\mathbf{x},y)$, y hemos usado la igualdad (163) escrita a continuación, que vincula la derivada de $d(\boldsymbol{\theta})$ con la derivada de $d^2(\boldsymbol{\theta})$: a partir de

$$\frac{\partial}{\partial \boldsymbol{\mu}} d^2(\boldsymbol{\theta}) = 2d(\boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\mu}} d(\boldsymbol{\theta}) \tag{162}$$

se deduce que

$$\frac{\partial}{\partial \boldsymbol{\mu}} d\left(\boldsymbol{\theta}\right) = \frac{1}{2d\left(\boldsymbol{\theta}\right)} \frac{\partial}{\partial \boldsymbol{\mu}} d^2\left(\boldsymbol{\theta}\right). \tag{163}$$

Despejando la derivada de S a partir de la igualdad (161), tenemos

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\mu}} \left[S\left(\boldsymbol{\theta}\right)\right] &= \frac{\int \psi_1 \left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{1}{2d(\boldsymbol{\theta})S(\boldsymbol{\theta})} \frac{\partial d^2(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} dP\left(\mathbf{x},y\right)}{\int \psi_1 \left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{d(\boldsymbol{\theta})}{S^2(\boldsymbol{\theta})} dP\left(\mathbf{x},y\right)} \\ &= \frac{\int \psi_1 \left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{1}{2d(\boldsymbol{\theta})} \frac{\partial d^2(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} dP\left(\mathbf{x},y\right)}{\int \psi_1 \left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})} dP\left(\mathbf{x},y\right)} \end{aligned} \tag{164}$$

Luego, podemos expresar la derivada del cociente entre la distancia y S de la siguiente forma

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\mu}} \left(\frac{d\left(\boldsymbol{\theta}\right)}{S\left(\boldsymbol{\theta}\right)}\right) &= \frac{1}{S^2\left(\boldsymbol{\theta}\right)} \left[\frac{\partial d\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\mu}} S\left(\boldsymbol{\theta}\right) - d\left(\boldsymbol{\theta}\right) \frac{\partial S\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\mu}}\right] \\ &= \frac{1}{S^2\left(\boldsymbol{\theta}\right)} \left[\frac{S\left(\boldsymbol{\theta}\right)}{2d\left(\boldsymbol{\theta}\right)} \frac{\partial d^2\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\mu}} - d\left(\boldsymbol{\theta}\right) \frac{\int \psi_1 \left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{1}{2d(\boldsymbol{\theta})} \frac{\partial d^2(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} dP\left(\mathbf{x},y\right)}{\int \psi_1 \left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})} dP\left(\mathbf{x},y\right)}\right] \end{aligned}$$

Llamando $u\left(\boldsymbol{\theta}\right) = E_P\left(\rho_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right)$, reemplazamos en la ecuación (160) para obtener

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{\mu}} \ln\left(\Phi_{B,P}\left(\boldsymbol{\theta}\right)\right) \\ &= \frac{2p}{S\left(\boldsymbol{\theta}\right)} \frac{\partial}{\partial \boldsymbol{\mu}} S\left(\boldsymbol{\theta}\right) + \frac{p}{u\left(\boldsymbol{\theta}\right)} \int \psi_2 \left(\frac{d\left(\boldsymbol{\theta}\right)}{S\left(\boldsymbol{\theta}\right)}\right) \frac{\partial}{\partial \boldsymbol{\mu}} \left(\frac{d\left(\boldsymbol{\theta}\right)}{S\left(\boldsymbol{\theta}\right)}\right) dP\left(\mathbf{x},y\right) \\ &= \frac{2p}{S\left(\boldsymbol{\theta}\right)} \frac{\int \psi_1 \left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{1}{2d(\boldsymbol{\theta})} \frac{\partial d^2(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} dP\left(\mathbf{x},y\right)}{\int \psi_1 \left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})} dP\left(\mathbf{x},y\right)} \\ &+ \frac{p}{u\left(\boldsymbol{\theta}\right)} \int \psi_2 \left(\frac{d\left(\boldsymbol{\theta}\right)}{S\left(\boldsymbol{\theta}\right)}\right) \frac{1}{S^2\left(\boldsymbol{\theta}\right)} \left[\frac{S\left(\boldsymbol{\theta}\right)}{2d\left(\boldsymbol{\theta}\right)} \frac{\partial d^2\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\mu}} \right. \\ &\left. - d\left(\boldsymbol{\theta}\right) \frac{\int \psi_1 \left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{1}{2d(\boldsymbol{\theta})} \frac{\partial d^2(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} dP\left(\mathbf{x}',y'\right)}{\int \psi_1 \left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})} dP\left(\mathbf{x}',y'\right)}\right] dP\left(\mathbf{x},y\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{2p}{S(\boldsymbol{\theta}) E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]} \int \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{1}{2d(\boldsymbol{\theta})} \frac{\partial d^2(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} dP(\mathbf{x},y) \\
&+ \frac{p}{2u(\boldsymbol{\theta}) S^2(\boldsymbol{\theta})} \int \psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})} \frac{\partial d^2(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} dP(\mathbf{x},y) \\
&- \frac{p}{u(\boldsymbol{\theta}) S(\boldsymbol{\theta})} \left[\int \psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})} dP(\mathbf{x},y)\right] \frac{\int \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{1}{2d(\boldsymbol{\theta})} \frac{\partial d^2(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} dP(\mathbf{x}',y')}{E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{p}{2u(\boldsymbol{\theta}) S^2(\boldsymbol{\theta})} \int \psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})} \frac{\partial d^2(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} dP(\mathbf{x},y) \\
&+ \int \left\{ \frac{p}{S^2(\boldsymbol{\theta}) E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]} - \frac{p}{2u(\boldsymbol{\theta}) S^2(\boldsymbol{\theta}) E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]} E_P\left[\psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right] \right\} \\
&\cdot \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})} \frac{\partial d^2(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} dP(\mathbf{x},y)
\end{aligned}
$$

Finalmente, la ecuación (160) que da la derivada de la función objetivo respecto de $\boldsymbol{\mu}$ es

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\mu}} \ln\left(\Phi_{B,P}(\boldsymbol{\theta})\right) \\
&= \int \left\{ w^{(1)}(\boldsymbol{\theta}) \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})} + w^{(2)}(\boldsymbol{\theta}) \psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})} \right\} \frac{\partial d^2(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} dP(\mathbf{x},y)
\end{aligned} \tag{165}
$$

donde estamos llamando

$$
\begin{aligned}
w^{(1)}(\boldsymbol{\theta}) &= \frac{p}{S^2(\boldsymbol{\theta}) E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]} \left(1 - \frac{1}{2u(\boldsymbol{\theta})} E_P\left[\psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right]\right) \\
&= \frac{p\left(2u(\boldsymbol{\theta}) - E_P\left[\psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right]\right)}{2u(\boldsymbol{\theta}) S^2(\boldsymbol{\theta}) E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]} \\
&= \frac{pE_P\left[2\rho_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) - \psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right]}{2u(\boldsymbol{\theta}) S^2(\boldsymbol{\theta}) E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]} = \frac{pE_P\left[a\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]}{2u(\boldsymbol{\theta}) S^2(\boldsymbol{\theta}) E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]}
\end{aligned} \tag{166}
$$

$$
w^{(2)}(\boldsymbol{\theta}) = \frac{p}{2u(\boldsymbol{\theta}) S^2(\boldsymbol{\theta})}. \tag{167}
$$

Recordemos la definición de los pesos

$$
w(d,\boldsymbol{\theta}) = \left\{ w^{(1)}(\boldsymbol{\theta}) \frac{\psi_1(d)}{d} + w^{(2)}(\boldsymbol{\theta}) \frac{\psi_2(d)}{d} \right\},
$$

por lo que

$$w\left(\frac{d\left(\mathbf{x},y,\boldsymbol{\theta}\right)}{S\left(\boldsymbol{\theta}\right)},\boldsymbol{\theta}\right)=w^{(1)}\left(\boldsymbol{\theta}\right)\psi_1\left(\frac{d\left(\boldsymbol{\theta}\right)}{S\left(\boldsymbol{\theta}\right)}\right)\frac{S\left(\boldsymbol{\theta}\right)}{d\left(\boldsymbol{\theta}\right)}+w^{(2)}\left(\boldsymbol{\theta}\right)\psi_2\left(\frac{d\left(\boldsymbol{\theta}\right)}{S\left(\boldsymbol{\theta}\right)}\right)\frac{S\left(\boldsymbol{\theta}\right)}{d\left(\boldsymbol{\theta}\right)}. \quad (168)$$

A lo largo de la demostración, y para abreviar, notaremos a los pesos por $w\left(\boldsymbol{\theta}\right)$, aunque en verdad dependen tanto del parámetro como del valor del vector aleatorio $\left(\mathbf{x},y\right)$, es decir, notamos

$$w\left(\boldsymbol{\theta}\right):=w\left(\frac{d\left(\mathbf{x},y,\boldsymbol{\theta}\right)}{S\left(\boldsymbol{\theta}\right)},\boldsymbol{\theta}\right). \quad (169)$$

Por las hipótesis hechas sobre las funciones ρ : tenemos que $2\rho_2\left(t\right)-\psi_2\left(t\right)t>0$ para $t>0$, de modo que $E_P\left[a\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]>0$. También sabemos que ρ_k son estrictamente crecientes en $[0,c_k]$ y constantes en $[c_k,+\infty)$, para $k=1,2$. Luego $b\left(d\right)=\psi_1\left(d\right)d\geq 0$ (es >0 cuando $d<c_k$ y cero en $[c_k,+\infty)$, de modo que su esperanza será positiva), resulta que los pesos $w\left(\boldsymbol{\theta}\right)$ son positivos. Luego, en notación compacta tenemos la siguiente expresión equivalente de (165)

$$\frac{\partial}{\partial\boldsymbol{\mu}}\ln\left(\Phi_{B,P}\left(\boldsymbol{\theta}\right)\right)=\int w\left(\boldsymbol{\theta}\right)\frac{\partial d^2\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\mu}}dP\left(\mathbf{x},y\right). \quad (170)$$

De la ecuación (76) de Petersen y Pedersen [2008], obtenemos que

$$\frac{\partial d^2\left[\left(\mathbf{x},y\right),\boldsymbol{\theta}\right]}{\partial\boldsymbol{\mu}}=-2\Delta^{-1}\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)$$

y finalmente, (170) queda

$$\frac{\partial}{\partial\boldsymbol{\mu}}\ln\left(\Phi_{B,P}\left(\boldsymbol{\theta}\right)\right)=-2\Delta^{-1}\int w\left(\boldsymbol{\theta}\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)dP\left(\mathbf{x},y\right)=0 \quad (171)$$

Calculamos $\dfrac{\partial}{\partial\beta}\ln\left(\Phi_{B,P}\left(\boldsymbol{\theta}\right)\right)$.

Operando del mismo modo que antes, la ecuación (170), en lo que refiere a β se convierte en

$$\frac{\partial}{\partial\beta}\ln\left(\Phi_{B,P}\left(\boldsymbol{\theta}\right)\right)=\int w\left(\boldsymbol{\theta}\right)\frac{\partial d^2\left[\left(\mathbf{x},y\right),\boldsymbol{\theta}\right]}{\partial\beta}dP\left(\mathbf{x},y\right) \quad (172)$$

Escribimos

$$\begin{aligned}d^2\left[\left(\mathbf{x},y\right),\boldsymbol{\theta}\right]&=\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)^T\Delta^{-1}\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)\\&=\left(\mathbf{x}-\boldsymbol{\mu}\right)^T\Delta^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)-2\left(\mathbf{x}-\boldsymbol{\mu}\right)^T\Delta^{-1}\Gamma\beta\mathbf{f}\left(y\right)\\&+\left(\beta\mathbf{f}\left(y\right)\right)^T\Gamma^T\Delta^{-1}\Gamma\left(\beta\mathbf{f}\left(y\right)\right).\end{aligned}$$

Para diferenciar el segundo término de la suma, utilizamos la ecuación (62) de Petersen y Pedersen [2008], y para el tercero, usamos la ecuación (80) de la misma fuente para obtener

$$\frac{\partial d^2\left[(\mathbf{x},y),\boldsymbol{\theta}\right]}{\partial\beta} = -2\Gamma^T\Delta^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\mathbf{f}\left(y\right)^T + 2\Gamma^T\Delta^{-1}\Gamma\beta\mathbf{f}\left(y\right)\mathbf{f}\left(y\right)^T.$$

Si juntamos todo, obtenemos

$$\begin{aligned}&\frac{\partial}{\partial\beta}\ln\left(\Phi_{B,P}\left(\boldsymbol{\theta}\right)\right) \\ &= 2\Gamma^T\Delta^{-1}\int w\left(\boldsymbol{\theta}\right)\left[-\left(\mathbf{x}-\boldsymbol{\mu}\right)\mathbf{f}\left(y\right)^T + \Gamma\beta\mathbf{f}\left(y\right)\mathbf{f}\left(y\right)^T\right]dP\left(\mathbf{x},y\right) \\ &= 0\end{aligned} \tag{173}$$

Calculamos $\dfrac{\partial}{\partial\Delta}\ln\left(\Phi_{B,P}\left(\boldsymbol{\theta}\right)\right).$

Por la ecuación (51) de Petersen y Pedersen [2008], tenemos

$$\frac{\partial\ln\left|\Delta\right|}{\partial\Delta} = \Delta^{-1},$$

y por la ecuación (57) de la misma fuente se obtiene

$$\begin{aligned}\frac{\partial d^2\left[(\mathbf{x},y),\boldsymbol{\theta}\right]}{\partial\Delta} &= \frac{\partial}{\partial\Delta}\text{traza}\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)^T\Delta^{-1}\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right) \\ &= -\Delta^{-1}\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)^T\Delta^{-1}.\end{aligned} \tag{174}$$

Con ellas podemos escribir

$$\begin{aligned}&\frac{\partial}{\partial\Delta}\ln\left(\Phi_{B,P}\left(\boldsymbol{\theta}\right)\right) \\ &= \Delta^{-1} + \int w\left(\boldsymbol{\theta}\right)\frac{\partial}{\partial\Delta}d^2\left(\boldsymbol{\theta}\right)dP\left(\mathbf{x},y\right) \\ &= \Delta^{-1} + \int w\left(\boldsymbol{\theta}\right)\left[-\Delta^{-1}\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)^T\Delta^{-1}\right]dP\left(\mathbf{x},y\right) \\ &= \Delta^{-1} - \int w\left(\boldsymbol{\theta}\right)\Delta^{-1}\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)^T\Delta^{-1}dP\left(\mathbf{x},y\right) \\ &= 0\end{aligned} \tag{175}$$

Calculamos $\dfrac{\partial}{\partial\Gamma}\ln\left(\Phi_{B,P}\left(\boldsymbol{\theta}\right)\right).$

Procediendo de la misma forma que antes, la ecuación (170) en lo que respecta a Γ se convierte en

$$\frac{\partial}{\partial\Gamma}\ln\left(\Phi_{B,P}\left(\boldsymbol{\theta}\right)\right) = \int w\left(d\left(\boldsymbol{\theta}\right)\right)\frac{\partial d^2\left[(\mathbf{x},y),\boldsymbol{\theta}\right]}{\partial\Gamma}dP\left(\mathbf{x},y\right) \tag{176}$$

Por lo tanto, sólo debemos calcular $\dfrac{\partial d^2\left[(\mathbf{x},y),\boldsymbol{\theta}\right]}{\partial\Gamma}$. Podemos usar la ecuación (80) de Petersen y Pedersen [2008], para obtener

$$\begin{aligned}\frac{\partial d^2\left[(\mathbf{x},y),\boldsymbol{\theta}\right]}{\partial\Gamma} &= \frac{\partial}{\partial\Gamma}\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)^T\Delta^{-1}\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)\\ &= -2\Delta^{-1}\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)\left(\beta\mathbf{f}\left(y\right)\right)^T\\ &= -2\Delta^{-1}\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)\mathbf{f}^T\left(y\right)\beta^T.\end{aligned}$$

de modo que

$$\begin{aligned}\frac{\partial}{\partial\Gamma}\ln\left(\Phi_{B,P}\left(\boldsymbol{\theta}\right)\right) &= -2\int w\left(\boldsymbol{\theta}\right)\Delta^{-1}\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)\mathbf{f}^T\left(y\right)\beta^T dP\left(\mathbf{x},y\right) \qquad (177)\\ &= -2\Delta^{-1}\int w\left(\boldsymbol{\theta}\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)\mathbf{f}^T\left(y\right)\beta^T dP\left(\mathbf{x},y\right)\\ &= 0\end{aligned}$$

Sabemos que los valores críticos deben satisfacer las cuatro ecuaciones (171), (173), (175) y (177), lo cual prueba la parte (i) de este teorema, ya que estas cuatro igualdades son las que conforman (68).

ii. Comenzamos resolviendo este sistema de ecuaciones por la ecuación (175). Es equivalente a

$$\begin{aligned}\Delta^{-1} &= \Delta^{-1}\int w\left(\boldsymbol{\theta}\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)^T dP\left(\mathbf{x},y\right)\Delta^{-1}\\ I_p &= \Delta^{-1}E_P\left(w\left(\boldsymbol{\theta}\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)^T\right)\\ \Delta &= E_P\left(w\left(\boldsymbol{\theta}\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)^T\right)\\ &= \int w\left(\boldsymbol{\theta}\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)^T dP\left(\mathbf{x},y\right). \qquad (178)\end{aligned}$$

Nos enfocamos en resolver (171). Como $\Delta > 0$, la ecuación (171) es equivalente a

$$\int w\left(\boldsymbol{\theta}\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)dP\left(\mathbf{x},y\right)=0$$

$$\int w\left(\boldsymbol{\theta}\right)\left[\mathbf{x}-\Gamma\beta\mathbf{f}\left(y\right)\right]dP\left(\mathbf{x},y\right)=\boldsymbol{\mu}\int w\left(\boldsymbol{\theta}\right)dP\left(\mathbf{x},y\right)$$

Calculamos la integral de los pesos. Llamamos $W(\boldsymbol{\theta}) = \int w(\boldsymbol{\theta})\, dP(\mathbf{x},y)$.

$$\begin{aligned} W(\boldsymbol{\theta}) &= \int w\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S(\boldsymbol{\theta})},\boldsymbol{\theta}\right) dP(\mathbf{x},y) \\ &= \int \left\{ w^{(1)}(\boldsymbol{\theta})\, \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})} + w^{(2)}(\boldsymbol{\theta})\, \psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})} \right\} dP(\mathbf{x},y) \\ &= w^{(1)}(\boldsymbol{\theta}) \int \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})} dP(\mathbf{x},y) \\ &+ w^{(2)}(\boldsymbol{\theta}) \int \psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})} dP(\mathbf{x},y) \end{aligned}$$

Como

$$\begin{aligned} w^{(1)}(\boldsymbol{\theta}) &= \frac{pE_P\left[a\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]}{2u(\boldsymbol{\theta})\, S^2(\boldsymbol{\theta})\, E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]} \\ w^{(2)}(\boldsymbol{\theta}) &= \frac{p}{2u(\boldsymbol{\theta})\, S^2(\boldsymbol{\theta})} \end{aligned}$$

tenemos

$$\begin{aligned} W(\boldsymbol{\theta}) &= \frac{pE_P\left[a\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right] E_P\left[\psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})}\right]}{2u(\boldsymbol{\theta})\, S^2(\boldsymbol{\theta})\, E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]} + \frac{pE_P\left[\psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})}\right]}{2u(\boldsymbol{\theta})\, S^2(\boldsymbol{\theta})} \\ &= \frac{p}{2u(\boldsymbol{\theta})\, S^2(\boldsymbol{\theta})} \left(\frac{E_P\left[a\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right] E_P\left[\psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})}\right]}{E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]} \right. \\ &+ \left. E_P\left[\psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})}\right] \right) \\ &= \frac{p}{2u(\boldsymbol{\theta})\, S^2(\boldsymbol{\theta})\, E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]} \left(E_P\left[2\rho_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) - \psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right] \right. \\ &\cdot \left. E_P\left[\psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})}\right] + E_P\left[\psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})}\right] E_P\left[\psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right] \right). \end{aligned} \tag{179}$$

Observemos que si bien hemos llamado pesos a los $w(\boldsymbol{\theta})$, no son éstos los que integran uno, sino $\frac{w(\boldsymbol{\theta})}{W(\boldsymbol{\theta})}$. Observemos también que $W(\boldsymbol{\theta}) = E_P\left[w\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S(\boldsymbol{\theta})},\boldsymbol{\theta}\right)\right]$, es decir, la esperanza de la variable aleatoria $w\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S(\boldsymbol{\theta})},\boldsymbol{\theta}\right)$ cuando el vector $(\mathbf{x},y)$

tiene distribución P. Definamos los pesos que sí integran uno,

$$\begin{aligned} w^{*}(\boldsymbol{\theta}) &= \frac{w(\boldsymbol{\theta})}{W(\boldsymbol{\theta})} \\ &= \frac{E_P\left[a\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]\psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})} + E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]\psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})}}{E_P\left[a\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]E_P\left[\psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})}\right] + E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]E_P\left[\psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})}\right]} \end{aligned}$$

Luego,

$$\begin{aligned} \boldsymbol{\mu} &= \frac{1}{\int w(\boldsymbol{\theta})\, dP(\mathbf{x},y)} \int w(\boldsymbol{\theta})\left(\mathbf{x} - \Gamma\beta\mathbf{f}(y)\right) dP(\mathbf{x},y) \\ &= \frac{1}{W(\boldsymbol{\theta})} E_P\left[w(\boldsymbol{\theta})\left(\mathbf{x} - \Gamma\beta\mathbf{f}(y)\right)\right] \end{aligned} \tag{180}$$

Utilizando la notación $dP^{*}_{\boldsymbol{\theta}}(\mathbf{x},y)$ introducida en (66), es decir,

$$dP^{*}_{\boldsymbol{\theta}}(\mathbf{x},y) := \frac{1}{W(\boldsymbol{\theta})} w(\boldsymbol{\theta})\, dP(\mathbf{x},y) = w^{*}(\boldsymbol{\theta})\, dP(\mathbf{x},y).$$

podemos escribir la identidad (180) de la siguiente forma

$$\begin{aligned} \boldsymbol{\mu} &= \frac{1}{W(\boldsymbol{\theta})} E_P\left[w(\boldsymbol{\theta})\left(\mathbf{x} - \Gamma\beta\mathbf{f}(y)\right)\right] \\ &= E_P\left[\frac{1}{W(\boldsymbol{\theta})} w\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S(\boldsymbol{\theta})}, \boldsymbol{\theta}\right)\left[\mathbf{x} - \Gamma\beta\mathbf{f}(y)\right]\right] \\ &= E_{P^{*}_{\boldsymbol{\theta}}}\left[\mathbf{x} - \Gamma\beta\mathbf{f}(y)\right]. \end{aligned} \tag{181}$$

O bien,

$$E_{P^{*}_{\boldsymbol{\theta}}}\left(\mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y)\right) = 0. \tag{182}$$

Si reescribimos la ecuación (178), tenemos

$$I_p = \int w(\boldsymbol{\theta})\, \Delta^{-1}\left(\mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y)\right)\left(\mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y)\right)^{T} dP(\mathbf{x},y),$$

entonces,

$$\begin{aligned} \text{traza}(I) &= \text{traza}\left(\int w(\boldsymbol{\theta})\, \Delta^{-1}\left(\mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y)\right)\left(\mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y)\right)^{T} dP(\mathbf{x},y)\right) \\ &= \int w(\boldsymbol{\theta})\, \text{traza}\left(\Delta^{-1}\left(\mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y)\right)\left(\mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y)\right)^{T}\right) dP(\mathbf{x},y) \\ &= \int w(\boldsymbol{\theta})\, \text{traza}\left(\left(\mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y)\right)^{T}\Delta^{-1}\left(\mathbf{x} - \boldsymbol{\mu} - \Gamma\beta\mathbf{f}(y)\right)\right) dP(\mathbf{x},y) \\ &= \int w(\boldsymbol{\theta})\, \text{traza}\left(d^2\left[(\mathbf{x},y), \boldsymbol{\theta}\right]\right) dP(\mathbf{x},y) \\ &= \int w(\boldsymbol{\theta})\, d^2\left[(\mathbf{x},y), \boldsymbol{\theta}\right] dP(\mathbf{x},y), \end{aligned}$$

o, equivalentemente

$$p = \int w(\boldsymbol{\theta})\, d^2\left[(\mathbf{x},y),\boldsymbol{\theta}\right] dP(\mathbf{x},y) \tag{183}$$

Por supuesto, esta cuenta también la podríamos haber hecho directamente, utilizando la expresión para los pesos hallada en (168), (166) y (167), como figura a continuación

$$\int \left\{ w^{(1)}(\boldsymbol{\theta})\, \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})} + w^{(2)}(\boldsymbol{\theta})\, \psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{S(\boldsymbol{\theta})}{d(\boldsymbol{\theta})} \right\} d^2(\boldsymbol{\theta})\, dP(\mathbf{x},y)$$
$$= S(\boldsymbol{\theta}) \int \left\{ w^{(1)}(\boldsymbol{\theta})\, \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) + w^{(2)}(\boldsymbol{\theta})\, \psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \right\} d(\boldsymbol{\theta})\, dP(\mathbf{x},y).$$

Como

$$w^{(1)}(\boldsymbol{\theta}) = \frac{pE_P\left[a\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]}{2u(\boldsymbol{\theta})\, S^2(\boldsymbol{\theta})\, E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]}$$
$$w^{(2)}(\boldsymbol{\theta}) = \frac{p}{2u(\boldsymbol{\theta})\, S^2(\boldsymbol{\theta})}$$

tenemos

$$S(\boldsymbol{\theta}) \int \left\{ w^{(1)}(\boldsymbol{\theta})\, \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) + w^{(2)}(\boldsymbol{\theta})\, \psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) \right\} d(\boldsymbol{\theta})\, dP(\mathbf{x},y)$$
$$= S^2(\boldsymbol{\theta}) \left(w^{(1)}(\boldsymbol{\theta}) \int \psi_1\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})} dP(\mathbf{x},y) + w^{(2)}(\boldsymbol{\theta}) \int \psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})} dP(\mathbf{x},y) \right)$$
$$= S^2(\boldsymbol{\theta}) \left(w^{(1)}(\boldsymbol{\theta})\, E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right] + w^{(2)}(\boldsymbol{\theta})\, E_P\left[\psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right] \right)$$

$$= S^2(\boldsymbol{\theta}) \left(\frac{pE_P\left[a\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]}{2u(\boldsymbol{\theta})\, S^2(\boldsymbol{\theta})\, E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right]} E_P\left[b\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right] \right.$$
$$\left. + \frac{p}{2u(\boldsymbol{\theta})\, S^2(\boldsymbol{\theta})} E_P\left[\psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right] \right)$$
$$= \frac{p}{2u(\boldsymbol{\theta})} \left(E_P\left[a\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right] + E_P\left[\psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right] \right)$$
$$= \frac{p}{2u(\boldsymbol{\theta})} \left(E_P\left[2\rho_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right) - \psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})} + \psi_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right] \right)$$
$$= \frac{p}{2u(\boldsymbol{\theta})} E_P\left[2\rho_2\left(\frac{d(\boldsymbol{\theta})}{S(\boldsymbol{\theta})}\right)\right] = p.$$

A partir de la ecuación (173), tenemos

$$\Gamma^T \Delta^{-1} \int w(\boldsymbol{\theta}) \left[-(\mathbf{x}-\boldsymbol{\mu})\, \mathbf{f}(y)^T + \Gamma\beta \mathbf{f}(y)\, \mathbf{f}(y)^T \right] dP(\mathbf{x},y) = 0$$

o, si dividimos por $W(\boldsymbol{\theta})$ en ambos términos, escribimos la identidad anterior como

$$\Gamma^T\Delta^{-1}\int\frac{w(\boldsymbol{\theta})}{W(\boldsymbol{\theta})}\left[-(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}(y))\,\mathbf{f}(y)^T\right]dP(\mathbf{x},y)=0.$$

o, por (182) la podemos reescribir de la siguiente forma

$$\Gamma^T\Delta^{-1}\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}(y),\mathbf{f}(y)\right]=0$$

equivalentemente,

$$\Gamma^T\Delta^{-1}\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}(y)\right]=\Gamma^T\Delta^{-1}\Gamma\beta\mathrm{var}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{f}(y)\right].$$

Despejamos β :

$$\left(\Gamma^T\Delta^{-1}\Gamma\right)^{-1}\Gamma^T\Delta^{-1}\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}(y)\right]\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{f}(y)\right]\right)^{-1}=\beta. \tag{184}$$

Como $\Delta>0$, la ecuación (177) es equivalente a

$$E_P\left[w(\boldsymbol{\theta})(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}(y))\,\mathbf{f}^T(y)\right]\beta^T=0,$$

o,

$$E_{P^*_{\boldsymbol{\theta}}}\left[\left[\mathbf{x}-\Gamma\beta\mathbf{f}(y)-E_{P^*_{\boldsymbol{\theta}}}(\mathbf{x}-\Gamma\beta\mathbf{f}(y))\right]\mathbf{f}^T(y)\right]\beta^T=0,$$

o, equivalentemente

$$\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[(\mathbf{x}-\Gamma\beta\mathbf{f}(y)),\mathbf{f}(y)\right]\beta^T=0 \tag{185}$$

o, equivalentemente

$$\left(\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}(y)\right]-\Gamma\beta\mathrm{var}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{f}(y)\right]\right)\beta^T=0.$$

Podemos reemplazar a β por el valor obtenido para él en la ecuación (184) para obtener

$$\left(\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}(y)\right]-\Gamma\left(\Gamma^T\Delta^{-1}\Gamma\right)^{-1}\Gamma^T\Delta^{-1}\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}(y)\right]\right.$$
$$\left.\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{f}(y)\right]\right)^{-1}\mathrm{var}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{f}(y)\right]\right)\beta^T=0$$

$$\left(\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}(y)\right]-\Gamma\left(\Gamma^T\Delta^{-1}\Gamma\right)^{-1}\Gamma^T\Delta^{-1}\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}(y)\right]\right)\beta^T=0$$
$$\left(I-\Gamma\left(\Gamma^T\Delta^{-1}\Gamma\right)^{-1}\Gamma^T\Delta^{-1}\right)\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}(y)\right]\beta^T=0.$$

Como $\Delta^{-1}>0$, podemos premultiplicar por ella para obtener

$$\Delta^{-1}\left(I-\Gamma\left(\Gamma^T\Delta^{-1}\Gamma\right)^{-1}\Gamma^T\Delta^{-1}\right)\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}(y)\right]\beta^T=0$$
$$\Delta^{-1/2}\left(I-\Delta^{-1/2}\Gamma\left(\Gamma^T\Delta^{-1}\Gamma\right)^{-1}\Gamma^T\Delta^{-1/2}\right)\Delta^{-1/2}\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}(y)\right]\beta^T=0$$

Sea $Q := P_{\Delta^{-1/2}\Gamma} = \Delta^{-1/2}\Gamma\left(\Gamma^T\Delta^{-1}\Gamma\right)^{-1}\Gamma^T\Delta^{-1/2}$ la matriz de proyección que proyecta al subespacio generado por las columnas de $\Delta^{-1/2}\Gamma$, de modo que la igualdad anterior puede escribirse

$$\begin{aligned}&\Delta^{-1/2}\left(I-Q\right)\Delta^{-1/2}\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}\left(y\right)\right]\cdot\\&\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{f}\left(y\right)\right]\right)^{-1}\left(\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}\left(y\right)\right]\right)^T\Delta^{-1}\Gamma\left(\Gamma^T\Delta^{-1}\Gamma\right)^{-1}\\&=0\end{aligned}$$

equivalentemente,

$$\begin{aligned}\left(I-Q\right)\Delta^{-1/2}\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}\left(y\right)\right]\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{f}\left(y\right)\right]\right)^{-1}\\\left(\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}\left(y\right)\right]\right)^T\Delta^{-1/2}\left[\Delta^{-1/2}\Gamma\right]=0.\end{aligned}\tag{186}$$

Llamemos

$$\begin{aligned}\Pi\left(\boldsymbol{\theta}\right)&=\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}\left(y\right)\right]\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{f}\left(y\right)\right]\right)^{-1}\left(\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}\left(y\right)\right]\right)^T\\&=\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x},\mathbf{f}\left(y\right)\right]\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{f}\left(y\right)\right]\right)^{-1}\left(\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{f}\left(y\right),\mathbf{x}\right]\right)\end{aligned}\tag{187}$$

La igualdad (186) es equivalente a

$$\left(I-Q\right)\Delta^{-1/2}\Pi\left(\boldsymbol{\theta}\right)\Delta^{-1/2}\left[\Delta^{-1/2}\Gamma\right]=0,$$

entonces $\Delta^{-1/2}\Pi\left(\boldsymbol{\theta}\right)\Delta^{-1/2}\left[\Delta^{-1/2}\Gamma\right]$ pertenece al subespacio $\mathrm{span}\{\Delta^{-1/2}\Gamma\}$, pues $(I-Q)$ proyecta a $\mathrm{span}\{\Delta^{-1/2}\Gamma\}^{\perp}$. Luego $\left[\Delta^{-1/2}\Gamma\right]$ tiene por columnas a los generadores de un autoespacio de $\Delta^{-1/2}\Pi\left(\boldsymbol{\theta}\right)\Delta^{-1/2}$ (las restricciones de rango garantizan que sean linealmente independientes).

Como las columnas de $\Delta^{-1/2}\Gamma$ son autovectores de la matriz simétrica $\Delta^{-1/2}\Pi\left(\boldsymbol{\theta}\right)\Delta^{-1/2}$, estas resultan ortogonales. Sin pérdida de generalidad, las podemos elegir ortonormales, es decir

$$\Delta^{-1/2}\Pi\left(\boldsymbol{\theta}\right)\Delta^{-1/2}\left(\Delta^{-1/2}\Gamma\right)=\Delta^{-1/2}\Gamma\Omega\left(\boldsymbol{\theta}\right),\ \text{donde}\ \Omega\left(\boldsymbol{\theta}\right)\in\mathbb{R}^{d\times d}\ \text{es diagonal, y}\tag{188}$$

$$\left(\Delta^{-1/2}\Gamma\right)^T\left(\Delta^{-1/2}\Gamma\right)=I,\ \text{o equivalentemente,}\ \Gamma^T\Delta^{-1}\Gamma=I.\tag{189}$$

Más aún, a los elementos de la diagonal de $\Omega\left(\boldsymbol{\theta}\right)$ los podemos ordenar en orden decreciente. Repasemos las expresiones que tenemos (hasta ahora) para los puntos críticos

$$\boldsymbol{\mu}=E_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x}-\Gamma\beta\mathbf{f}\left(y\right)\right]$$

$$\begin{aligned}\Delta &= E_P\left[w\left(\boldsymbol{\theta}\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)^T\right] \\ &= W\left(\boldsymbol{\theta}\right)\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{x}-\Gamma\beta\mathbf{f}\left(y\right),\mathbf{x}-\Gamma\beta\mathbf{f}\left(y\right)\right] \qquad (190)\end{aligned}$$

$$\begin{aligned}\beta &= \left(\Gamma^T\Delta^{-1}\Gamma\right)^{-1}\Gamma^T\Delta^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{x},\mathbf{f}\left(y\right)\right]\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{f}\left(y\right)\right]\right)^{-1} \\ &= \Gamma^T\Delta^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{x},\mathbf{f}\left(y\right)\right]\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{f}\left(y\right)\right]\right)^{-1} \qquad (191)\end{aligned}$$

donde la última igualdad es cierta por (189).

Finalmente, una manera alternativa de escribir las ecuación para Δ es

$$\begin{aligned}\Delta\frac{1}{W\left(\boldsymbol{\theta}\right)} &= \mathrm{cov}_{P_{\boldsymbol{\theta}}^*}\left(\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right),\left(\mathbf{x}-\boldsymbol{\mu}-\Gamma\beta\mathbf{f}\left(y\right)\right)\right) \qquad (192) \\ &= \mathrm{cov}_{P_{\boldsymbol{\theta}}^*}\left(\left(\mathbf{x}-\Gamma\beta\mathbf{f}\left(y\right)-E_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{x}-\Gamma\beta\mathbf{f}\left(y\right)\right]\right),\mathbf{x}-E_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{x}\right]\right) \\ &\quad -\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}\left[\left(\mathbf{x}-\Gamma\beta\mathbf{f}\left(y\right)\right),\mathbf{f}\left(y\right)\right]\left(\Gamma\beta\right)^T. \qquad (193)\end{aligned}$$

Por (185), sabemos que el segundo sumando es cero, entonces

$$\begin{aligned}\Delta\frac{1}{W\left(\boldsymbol{\theta}\right)} &= \mathrm{var}_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{x}\right]-\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}\left[\Gamma\beta\mathbf{f}\left(y\right),\mathbf{x}\right] \\ &= \mathrm{cov}_{P_{\boldsymbol{\theta}}^*}\left(\mathbf{x},\mathbf{x}\right)-\Gamma\beta\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{f}\left(y\right),\mathbf{x}\right].\end{aligned}$$

Ahora reemplazamos la expresión para β en el segundo término para obtener

$$\begin{aligned}\Gamma\beta\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}\left(\mathbf{f}\left(y\right),\mathbf{x}\right) &= \Gamma\Gamma^T\Delta^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{x},\mathbf{f}\left(y\right)\right]\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{f}\left(y\right)\right]\right)^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{f}\left(y\right),\mathbf{x}\right] \\ &= \Gamma\Gamma^T\Delta^{-1}\Pi\left(\boldsymbol{\theta}\right)\end{aligned}$$

donde $\Pi\left(\boldsymbol{\theta}\right)$ fue definida en (187). A partir de (188), tenemos

$$\begin{aligned}\Delta^{-1/2}\Pi\left(\boldsymbol{\theta}\right)\Delta^{-1/2}\left(\Delta^{-1/2}\Gamma\right) &= \Delta^{-1/2}\Gamma\Omega\left(\boldsymbol{\theta}\right) \\ \Delta^{-1/2}\Pi\left(\boldsymbol{\theta}\right)\Delta^{-1}\Gamma &= \Delta^{-1/2}\Gamma\Omega\left(\boldsymbol{\theta}\right) \\ \Pi\left(\boldsymbol{\theta}\right)\Delta^{-1}\Gamma &= \Gamma\Omega\left(\boldsymbol{\theta}\right)\end{aligned}$$

de modo que

$$\begin{aligned}\Gamma\beta\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}\left(\mathbf{f}\left(y\right),\mathbf{x}\right) &= \Gamma\Gamma^T\Delta^{-1}\Pi\left(\boldsymbol{\theta}\right) = \Gamma\left(\Pi\left(\boldsymbol{\theta}\right)\Delta^{-1}\Gamma\right)^T = \Gamma\left(\Gamma\Omega\left(\boldsymbol{\theta}\right)\right)^T \\ &= \Gamma\Omega\left(\boldsymbol{\theta}\right)\Gamma^T = \sum_{i=1}^{d}\Omega_i\left(\boldsymbol{\theta}\right)\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^T,\end{aligned}$$

$\Gamma = \begin{bmatrix}\boldsymbol{\gamma}_1 & \boldsymbol{\gamma}_2 & \cdots & \boldsymbol{\gamma}_d\end{bmatrix} \in \mathbb{R}^{p\times d}$ y $\Omega\left(\boldsymbol{\theta}\right) = \mathrm{diag}\left(\lambda_1,\lambda_2,...,\lambda_d\right) \in \mathbb{R}^{d\times d}$, (asumimos, sin pérdida de generalidad, que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$) por lo que

$$\begin{aligned}\Delta\frac{1}{W\left(\boldsymbol{\theta}\right)} &= \mathrm{var}_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{x}\right]-\Gamma\beta\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{f}\left(y\right),\mathbf{x}\right] \qquad (194) \\ &= \mathrm{var}_{P_{\boldsymbol{\theta}}^*}\left[\mathbf{x}\right]-\Gamma\Omega\left(\boldsymbol{\theta}\right)\Gamma^T. \qquad (195)\end{aligned}$$

iii. Por el Teorema 6.2, la función $\Phi_{B,P}$ alcanza un mínimo, que debe ser uno de los puntos críticos. Para probar que dicho mínimo se alcanza en $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Gamma, \beta, \Delta)$ cuando los autovalores seleccionados para construir a Γ son aquellos correspondientes a los mayores autovalores, la idea es descartar que ese mínimo se alcance cuando en Γ se ubica cualquier otra elección de los autovectores que no sea la correspondiente a los autovectores asociados a los mayores autovalores.

La estrategia para la demostración es la siguiente. Sean $\lambda_1 \geq \cdots \geq \lambda_p$ los autovalores de $\Delta^{-\frac{1}{2}}\Pi(\boldsymbol{\theta})\Delta^{-\frac{1}{2}}$ y $\mathbf{z}_i$, $1 \leq i \leq p$, los autovectores correspondientes. Sea $\mathcal{S}_{\Gamma_0}$ el subespacio generado por $\Delta^{1/2}\mathbf{z}_{i_j}$, $j = 1, \ldots, d$, es decir, Γ_0 tiene a los vectores $\Delta^{1/2}\mathbf{z}_{i_j}$ por columnas y, sin pérdida de generalidad, podemos asumir que $\lambda_{i_1} \geq \lambda_{i_2} \geq \cdots \geq \lambda_{i_d}$ (o, equivalentemente $i_1 \geq i_2 \geq \cdots \geq i_d$). Supongamos que $i_d > d$ y que $\lambda_{i_d} < \lambda_d$. Entonces existe un $k \leq d$ tal que $k \neq i_1, \ldots, i_d$. Llamemos $\mathcal{S}_{\Gamma_t}$ al subespacio que se obtiene reemplazando a $\mathbf{z}_{i_d}$ por $(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k$ y llamemos Γ_t a la matriz que tiene a los vectores $\Delta^{1/2}\mathbf{z}_{i_j}$ por columnas, donde reemplazamos a $\mathbf{z}_{i_d}$ por $(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k$. Denotemos por $\boldsymbol{\theta}_0 = (\boldsymbol{\mu}_0, \Gamma_0, \beta_0, \Delta)$ al punto crítico. Definamos la curva $\boldsymbol{\gamma}$: $(-1, 1) \to \Theta$, dada por

$$\boldsymbol{\gamma}(t) = (\boldsymbol{\mu}_t, \Gamma_t, \beta_t, \Delta_t)$$

donde

$$\begin{aligned}
\Gamma_t &= \left[\Delta^{1/2}\mathbf{z}_{i_1} \quad \ldots \quad \Delta^{1/2}\mathbf{z}_{i_{d-1}} \quad \Delta^{1/2}\left[(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k\right]\right], \\
\beta_t &= \Gamma_t^T \Delta^{-1} \mathrm{cov}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x}, \mathbf{f}(y)\right] \left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{f}(y)\right]\right)^{-1} \\
\Delta_t &= \Delta \\
\boldsymbol{\mu}_t &= E_{P^*_{\boldsymbol{\theta}}}\left[\mathbf{x} - \Gamma_t \beta_t \mathbf{f}(y)\right].
\end{aligned}$$

A continuación, probaremos que la función $\Phi_{B,P} \circ \boldsymbol{\gamma} : (-1, 1) \to \mathbb{R}$ alcanza un máximo local en $t = 0$. Esto implicará que la función $\Phi_{B,P}(\boldsymbol{\theta})$ no alcanza un mínimo local en $\boldsymbol{\theta}_0 = \boldsymbol{\gamma}(0) = (\boldsymbol{\mu}_0, \Gamma_0, \beta_0, \Delta)$, descartando que el valor del $\tau-$funcional sea $\boldsymbol{\theta}_0$. Como este argumento puede repetirse para todo $\boldsymbol{\theta}$ crítico en el que se seleccionen autovectores no asociados a los d mayores autovalores, este argumento inhabilita a todo otro punto crítico para ser el mínimo de $\Phi_{B,P}$. Como por el teorema mencionado este mínimo existe, entonces, se alcanzará en el punto crítico que consiste en poner en la matriz Γ los autovectores asociados a los mayores autovalores de $\Delta^{-\frac{1}{2}}\Pi(\boldsymbol{\theta})\Delta^{-\frac{1}{2}}$.

La función objetivo es

$$\Phi_{B,P}(\boldsymbol{\theta}) = |\Delta| \cdot \left(S_M^2\left(H_{P,\boldsymbol{\theta}}\right)\right)^p \cdot \left(\int \rho_2\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_M\left(H_{P,\boldsymbol{\theta}}\right)}\right) dP(\mathbf{x},y)\right)^p,$$

por lo que

$$\ln\left(\Phi_{B,P}(\boldsymbol{\gamma}(t))\right) = \ln|\Delta| + 2p\ln\left(S_M\left(H_{P,\boldsymbol{\gamma}(t)}\right)\right) + p\ln\left(\int \rho_2\left(\frac{d(\mathbf{x},y,\boldsymbol{\gamma}(t))}{S_M\left(H_{P,\boldsymbol{\gamma}(t)}\right)}\right) dP(\mathbf{x},y)\right).$$

Recordemos la notación introducida en la Sección 8.1,

$$\begin{aligned} a(d) &= 2\rho_2(d) - \psi_2(d)\, d \\ b(d) &= \psi_1(d)\, d \\ u(\boldsymbol{\theta}) &= E_P\left(\rho_2\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_M(H_{P,\boldsymbol{\theta}})}\right)\right) \end{aligned}$$

A partir de ellos, se definen

$$\begin{aligned} w^{(1)}(\boldsymbol{\theta}) &= \frac{pE_P\left[a\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_M(H_{P,\boldsymbol{\theta}})}\right)\right]}{2u(\boldsymbol{\theta})\, S_M^2(H_{P,\boldsymbol{\theta}})\, E_P\left[b\left(\frac{d(\mathbf{x},y,\boldsymbol{\theta})}{S_M(H_{P,\boldsymbol{\theta}})}\right)\right]} \\ w^{(2)}(\boldsymbol{\theta}) &= \frac{p}{2u(\boldsymbol{\theta})\, S_M^2(H_{P,\boldsymbol{\theta}})}, \end{aligned}$$

y los pesos se definen por

$$w(d,\boldsymbol{\theta}) = \left\{ w^{(1)}(\boldsymbol{\theta})\frac{\psi_1(d)}{d} + w^{(2)}(\boldsymbol{\theta})\frac{\psi_2(d)}{d} \right\}.$$

Si los componemos con la curva $\boldsymbol{\gamma}$ obtenemos:

$$\begin{aligned} \widetilde{w}(t) := w\left(\frac{d(\mathbf{x},y,\boldsymbol{\gamma}(t))}{S_M\left(H_{P,\boldsymbol{\gamma}(t)}\right)}, \boldsymbol{\gamma}(t)\right) &= w^{(1)}(\boldsymbol{\gamma}(t))\,\psi_1\left(\frac{d(\mathbf{x},y,\boldsymbol{\gamma}(t))}{S_M\left(H_{P,\boldsymbol{\gamma}(t)}\right)}\right)\frac{S_M\left(H_{P,\boldsymbol{\gamma}(t)}\right)}{d(\mathbf{x},y,\boldsymbol{\gamma}(t))} \\ &+ w^{(2)}(\boldsymbol{\gamma}(t))\,\psi_2\left(\frac{d(\mathbf{x},y,\boldsymbol{\gamma}(t))}{S_M\left(H_{P,\boldsymbol{\gamma}(t)}\right)}\right)\frac{S_M\left(H_{P,\boldsymbol{\gamma}(t)}\right)}{d(\mathbf{x},y,\boldsymbol{\gamma}(t))}. \end{aligned}$$

Para abreviar la notación, llamemos

$$\begin{aligned} S(t) &:= S_M\left(H_{P,\boldsymbol{\gamma}(t)}\right) \\ d(t) &:= d(\mathbf{x},y,\boldsymbol{\gamma}(t)). \end{aligned}$$

Observemos que, en verdad, $d(t)$ es una función tanto de t como del vector aleatorio $(\mathbf{x},y)$, aunque este hecho no se vea reflejado en la notación $d(t)$. Notemos por

$$A(t) := E_P\left[a\left(\frac{d(t)}{S(t)}\right)\right] = E_P\left[2\rho_2\left(\frac{d(t)}{S(t)}\right) - \psi_2\left(\frac{d(t)}{S(t)}\right)\frac{d(t)}{S(t)}\right] > 0 \qquad (196)$$

$$B(t) := E_P\left[b\left(\frac{d(t)}{S(t)}\right)\right] = E_P\left[\psi_1\left(\frac{d(t)}{S(t)}\right)\frac{d(t)}{S(t)}\right] > 0 \qquad (197)$$

$$v(t) := u(\boldsymbol{\gamma}(t)) = E_P\left(\rho_2\left(\frac{d((\cdot),\boldsymbol{\gamma}(t))}{S_M\left(H_{P,\boldsymbol{\gamma}(t)}\right)}\right)\right) = E_P\left(\rho_2\left(\frac{d(t)}{S(t)}\right)\right)$$

Observemos que con toda esta notación, resulta

$$\begin{aligned}\widetilde{w}(t) &:= w\left(\frac{d(t)}{S(t)}, \boldsymbol{\gamma}(t)\right) \\ &= w^{(1)}(\boldsymbol{\gamma}(t))\,\psi_1\left(\frac{d(t)}{S(t)}\right)\frac{S(t)}{d(t)} + w^{(2)}(\boldsymbol{\gamma}(t))\,\psi_2\left(\frac{d(t)}{S(t)}\right)\frac{S(t)}{d(t)} \\ &= \frac{pA(t)}{2v(t)S^2(t)B(t)}\psi_1\left(\frac{d(t)}{S(t)}\right)\frac{S(t)}{d(t)} + \frac{p}{2v(t)S^2(t)}\psi_2\left(\frac{d(t)}{S(t)}\right)\frac{S(t)}{d(t)} \\ &= \frac{p}{2v(t)S^2(t)}\left\{\psi_2\left(\frac{d(t)}{S(t)}\right)\frac{S(t)}{d(t)} + \frac{A(t)}{B(t)}\psi_1\left(\frac{d(t)}{S(t)}\right)\frac{S(t)}{d(t)}\right\}. \end{aligned} \tag{198}$$

Sea

$$\begin{aligned} m(t) &:= \ln \Phi_{B,P}(\boldsymbol{\gamma}(t)) \\ &= \ln|\Delta| + 2p\ln[S(t)] + p\ln\left(\int \rho_2\left(\frac{d(t)}{S(t)}\right) dP(\mathbf{x},y)\right) \end{aligned}$$

la función que queremos estudiar. Tenemos

$$m'(t) = \frac{2p}{S(t)}\frac{\partial S(t)}{\partial t} + \frac{p}{v(t)}\int \psi_2\left(\frac{d(t)}{S(t)}\right)\frac{\partial}{\partial t}\left(\frac{d(t)}{S(t)}\right) dP(\mathbf{x},y). \tag{199}$$

Entonces hay que calcular $\dfrac{\partial}{\partial t}\left(\frac{d(t)}{S(t)}\right)$ y $\dfrac{\partial S(t)}{\partial t}$. Empecemos estudiando $d(t)$, para luego derivarla.

$$\begin{aligned}(\mathbf{x} - \boldsymbol{\mu}_t - \Gamma_t\beta_t\mathbf{f}(y)) &= \left(\mathbf{x} - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x} - \Gamma_t\beta_t\mathbf{f}(y)] - \Gamma_t\beta_t\mathbf{f}(y)\right) \\ &= \left(\mathbf{x} - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}] - \Gamma_t\beta_t\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right)\end{aligned}$$

Entonces

$$\begin{aligned} d^2(t) &= (\mathbf{x} - \boldsymbol{\mu}_t - \Gamma_t\beta_t\mathbf{f}(y))^T\Delta^{-1}(\mathbf{x} - \boldsymbol{\mu}_t - \Gamma_t\beta_t\mathbf{f}(y)) \\ &= \left(\mathbf{x} - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}]\right)^T\Delta^{-1}\left(\mathbf{x} - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}]\right) \\ &+ \left(\Gamma_t\beta_t\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right)^T\Delta^{-1}\left(\Gamma_t\beta_t\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right) \\ &- 2\left(\mathbf{x} - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}]\right)^T\Delta^{-1}\Gamma_t\beta_t\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right) \\ &= uno + dos + tres \end{aligned} \tag{200}$$

El primer término

$$uno = \left(\mathbf{x} - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}]\right)^T\Delta^{-1}\left(\mathbf{x} - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}]\right),$$

no depende de t. Como $\Delta^{-\frac{1}{2}}\Pi(\boldsymbol{\theta})\Delta^{-\frac{1}{2}}$ es simétrica los autovectores $\mathbf{z}_i$ pueden elegirse ortonormales (son ortogonales, y pueden elegirse de norma uno). Esto significa que

$$\mathbf{z}_i^T\mathbf{z}_j = \begin{cases} 1 & \text{si } i = j \\ \\ 0 & \text{si } i \neq j. \end{cases} \tag{201}$$

Entonces,

$$\begin{aligned}
&\Gamma_t^T\Delta^{-1}\Gamma_t \\
&= \begin{bmatrix} \mathbf{z}_{i_1}^T\Delta^{1/2} \\ \vdots \\ \mathbf{z}_{i_{d-1}}^T\Delta^{1/2} \\ \left[(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k\right]^T\Delta^{1/2} \end{bmatrix} \Delta^{-1}\left[\Delta^{1/2}\mathbf{z}_{i_1} \quad \ldots \quad \Delta^{1/2}\mathbf{z}_{i_{d-1}} \quad \Delta^{1/2}\left[(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k\right]\right] \\
&= \begin{bmatrix} \mathbf{z}_{i_1}^T \\ \vdots \\ \mathbf{z}_{i_{d-1}}^T \\ \left[(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k\right]^T \end{bmatrix} \left[\mathbf{z}_{i_1} \quad \ldots \quad \mathbf{z}_{i_{d-1}} \quad \left[(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k\right]\right] \\
&= \begin{bmatrix} \mathbf{z}_{i_1}^T\mathbf{z}_{i_1} & \mathbf{z}_{i_1}^T\mathbf{z}_{i_2} & \ldots & \mathbf{z}_{i_1}^T\mathbf{z}_{i_{d-1}} & \mathbf{z}_{i_1}^T\left[(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k\right] \\ & & & & \\ & & & & \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} = I_d
\end{aligned}$$

De modo que $\Gamma_t^T\Delta^{-1}\Gamma_t = I_d$, y el segundo término en la suma dada en (200) es

$$\begin{aligned}
dos &= \left[\left(\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)\right]^T \beta_t^T\Gamma_t^T\Delta^{-1}\Gamma_t\beta_t\left(\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right) \\
&= \left[\left(\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)\right]^T \beta_t^T\beta_t\left[\left(\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)\right],
\end{aligned}$$

Como también

$$\begin{aligned}
\beta_t &= \Gamma_t^T\Delta^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x}, \mathbf{f}(y)]\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}, \\
\beta_t^T\beta_t &= \left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x}, \mathbf{f}(y)]\,\Delta^{-1}\Gamma_t\Gamma_t^T\Delta^{-1}\cdot \\
&\quad \mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x}, \mathbf{f}(y)]\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}
\end{aligned}$$

y

$$\Gamma_t\Gamma_t^T$$

$$= \left[\Delta^{1/2}\mathbf{z}_{i_1} \quad \ldots \quad \Delta^{1/2}\mathbf{z}_{i_{d-1}} \quad \Delta^{1/2}\left[(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k\right]\right] \begin{bmatrix} \mathbf{z}_{i_1}^T\Delta^{1/2} \\ \vdots \\ \mathbf{z}_{i_{d-1}}^T\Delta^{1/2} \\ \left[(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k\right]^T\Delta^{1/2} \end{bmatrix}$$

$$= \Delta^{1/2}\left(\sum_{j=1}^{d-1}\mathbf{z}_{i_j}\mathbf{z}_{i_j}^T + \left[(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k\right]\left[(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k\right]^T\right)\Delta^{1/2}.$$

Resulta que el segundo término es

$$dos = \left[\left(\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)\right]^T\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x},\mathbf{f}(y)]^T\Delta^{-1}$$
$$\Gamma_t\Gamma_t^T\Delta^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x},\mathbf{f}(y)]\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}\left[\left(\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)\right].$$

Como

$$\Gamma_t\beta_t = \Gamma_t\Gamma_t^T\Delta^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x},\mathbf{f}(y)]\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1},$$

el último término es

$$tres = -2\left(\mathbf{x} - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{x}]\right)^T\Delta^{-1}\Gamma_t\Gamma_t^T\Delta^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x},\mathbf{f}(y)]\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}\left(\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)$$

Sumando ambos obtenemos

$$\{\left[\left(\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)\right]^T\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y),\mathbf{x}] - 2\left(\mathbf{x} - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{x}]\right)^T\}$$
$$\Delta^{-1}\Gamma_t\Gamma_t^T\Delta^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y),\mathbf{x}]\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}\left[\left(\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)\right]$$

Esta es la expresión que debemos derivar con respecto a t. Tenemos

$$\frac{\partial}{\partial t}\Gamma_t\Gamma_t^T = \frac{\partial}{\partial t}\left[\Delta^{1/2}\left(\sum_{j=1}^{d-1}\mathbf{z}_{i_j}\mathbf{z}_{i_j}^T + \left[(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k\right]\left[(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k\right]^T\right)\Delta^{1/2}\right]$$
$$= \frac{\partial}{\partial t}\left[\Delta^{1/2}\left(\left[(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k\right]\left[(1-t^2)^{1/2}\mathbf{z}_{i_d} + t\mathbf{z}_k\right]^T\right)\Delta^{1/2}\right]$$
$$= \frac{\partial}{\partial t}\left[\Delta^{1/2}\left[(1-t^2)\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T + t^2\mathbf{z}_k\mathbf{z}_k^T + t(1-t^2)^{1/2}\mathbf{z}_{i_d}\mathbf{z}_k^T + t(1-t^2)^{1/2}\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\Delta^{1/2}\right]$$

Sabemos que

$$\frac{\partial}{\partial t}(1-t^2) = -2t$$
$$\frac{\partial}{\partial t}t(1-t^2)^{1/2} = \frac{1-2t^2}{\sqrt{1-t^2}}$$
$$\frac{\partial}{\partial t}(t^2) = 2t$$

entonces

$$\frac{\partial}{\partial t}\Gamma_t\Gamma_t^T = \Delta^{1/2}\left[(-2t)\,\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T + 2t\mathbf{z}_k\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_{i_d}\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\Delta^{1/2},$$

dando lugar a

$$\begin{aligned}\frac{\partial}{\partial t}d^2(t) = &\{\left[\left(\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)\right]^T\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y),\mathbf{x}] - 2\left(\mathbf{x} - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{x}]\right)^T\}\\ &\Delta^{-1}\frac{\partial}{\partial t}\Gamma_t\Gamma_t^T\Delta^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x},\mathbf{f}(y)]\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}\left[\left(\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)\right]\\ &= \{\left[\left(\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)\right]^T\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y),\mathbf{x}] - 2\left(\mathbf{x} - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{x}]\right)^T\}\\ &\Delta^{-1}\Delta^{1/2}\left[(-2t)\,\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T + 2t\mathbf{z}_k\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_{i_d}\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\Delta^{1/2}\Delta^{-1}\\ &\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x},\mathbf{f}(y)]\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}\left[\left(\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)\right]\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial t}d^2(t) = &\{\left[\left(\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)\right]^T\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y),\mathbf{x}] - 2\left(\mathbf{x} - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{x}]\right)^T\}\\ &\Delta^{-1/2}\left[(-2t)\,\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T + 2t\mathbf{z}_k\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_{i_d}\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\Delta^{-1/2}\\ &\mathrm{cov}_{P_{\boldsymbol{\theta}}^*}[\mathbf{x},\mathbf{f}(y)]\left(\mathrm{var}_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)^{-1}\left[\left(\mathbf{f}(y) - E_{P_{\boldsymbol{\theta}}^*}[\mathbf{f}(y)]\right)\right]\end{aligned} \tag{202}$$

Finalmente,

$$\frac{\partial}{\partial t}d^2(\mathbf{x},y,\boldsymbol{\gamma}(t)) = 2d(\mathbf{x},y,\boldsymbol{\gamma}(t))\frac{\partial}{\partial t}d(\mathbf{x},y,\boldsymbol{\gamma}(t))$$

o, escrito de forma sintética

$$\frac{\partial}{\partial t}d^2(t) = 2d(t)\frac{\partial}{\partial t}d(t)$$

de donde se obtiene

$$\frac{\partial}{\partial t}d(t) = \frac{1}{2d(t)}\frac{\partial}{\partial t}d^2(t) \tag{203}$$

Hallemos ahora la derivada de $S(t)$. Por definición de S_M^2 tenemos

$$\int \rho_1\left(\frac{d(\mathbf{x},y,\boldsymbol{\gamma}(t))}{S_M\left(H_{P,\boldsymbol{\gamma}(t)}\right)}\right)dP(\mathbf{x},y) = \kappa_1$$

Entonces, si derivamos respecto de t obtenemos

$$0 = \int \psi_1\left(\frac{d(\mathbf{x},y,\boldsymbol{\gamma}(t))}{S_M\left(H_{P,\boldsymbol{\gamma}(t)}\right)}\right)\frac{\partial}{\partial t}\left(\frac{d(\mathbf{x},y,\boldsymbol{\gamma}(t))}{S_M\left(H_{P,\boldsymbol{\gamma}(t)}\right)}\right)dP(\mathbf{x},y),$$

Llamábamos $d(t) = d(\mathbf{x},y,\boldsymbol{\gamma}(t))$ y $S(t) = S_M\left(H_{P,\boldsymbol{\gamma}(t)}\right)$ tenemos

$$\begin{aligned}
0 &= \int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{\partial}{\partial t}\left(\frac{d(t)}{S(t)}\right)dP(\mathbf{x},y)\\
&= \int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{1}{S^2(t)}\left[\frac{\partial d(t)}{\partial t}S(t) - d(t)\frac{\partial S(t)}{\partial t}\right]dP(\mathbf{x},y)\\
&= \int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{1}{S(t)}\frac{\partial d(t)}{\partial t}dP(\mathbf{x},y) - \int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{d(t)}{S^2(t)}\frac{\partial S(t)}{\partial t}dP(\mathbf{x},y)\\
&= \int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{1}{2d(t)S(t)}\frac{\partial d^2(t)}{\partial t}dP(\mathbf{x},y) - \frac{\partial S(t)}{\partial t}\int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{d(t)}{S^2(t)}dP(\mathbf{x},y)
\end{aligned}$$

o, equivalentemente

$$\begin{aligned}
\frac{\partial}{\partial t}[S(t)] &= \frac{\int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{1}{2d(t)S(t)}\frac{\partial d^2(t)}{\partial t}dP(\mathbf{x},y)}{\int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{d(t)}{S^2(t)}dP(\mathbf{x},y)}\\
&= \frac{\int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{1}{2d(t)}\frac{\partial d^2(t)}{\partial t}dP(\mathbf{x},y)}{\int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{d(t)}{S(t)}dP(\mathbf{x},y)}
\end{aligned} \tag{204}$$

donde la última igualdad vale pues S no depende de $(\mathbf{x},y)$. Definamos

$$\begin{aligned}
H(t) &:= \int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{1}{2d(t)}\frac{\partial d^2(t)}{\partial t}dP(\mathbf{x},y)\\
&= \int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{\partial d(t)}{\partial t}dP(\mathbf{x},y)\\
&= \frac{1}{2S(t)}\int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{S(t)}{d(t)}\frac{\partial d^2(t)}{\partial t}dP(\mathbf{x},y)
\end{aligned} \tag{205}$$

Y recordemos de (197) que

$$B(t) := E_P\left[b\left(\frac{d(t)}{S(t)}\right)\right] = E_P\left[\psi_1\left(\frac{d(t)}{S(t)}\right)\frac{d(t)}{S(t)}\right]$$

luego

$$\begin{aligned}
\frac{\partial}{\partial t}\left(\frac{d(t)}{S(t)}\right) &= \frac{1}{S^2(t)}\left[\frac{\partial d(t)}{\partial t}S(t) - d(t)\frac{\partial S(t)}{\partial t}\right]\\
&= \frac{1}{S^2(t)}\left[\frac{S(t)}{2d(t)}\frac{\partial d^2(t)}{\partial t} - d(t)\frac{\int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{1}{2d(t)}\frac{\partial d^2(t)}{\partial t}dP(\mathbf{x},y)}{\int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{d(t)}{S(t)}dP(\mathbf{x},y)}\right]\\
&= \frac{1}{S^2(t)}\left[\frac{S(t)}{2d(t)}\frac{\partial d^2(t)}{\partial t} - d(t)\frac{H(t)}{B(t)}\right]
\end{aligned}$$

y

$$\frac{\partial}{\partial t}\left[S\left(t\right)\right]=\frac{H\left(t\right)}{B\left(t\right)}. \tag{206}$$

Poniendo todo junto, tenemos

$$\begin{aligned}
m'\left(t\right) &= \frac{2p}{S\left(t\right)}\frac{\partial S\left(t\right)}{\partial t}+\frac{p}{v\left(t\right)}\int\psi_2\left(\frac{d\left(t\right)}{S\left(t\right)}\right)\frac{\partial}{\partial t}\left(\frac{d\left(t\right)}{S\left(t\right)}\right)dP\left(\mathbf{x},y\right)\\
&= \frac{2p}{S\left(t\right)}\frac{\partial}{\partial t}\left[S\left(t\right)\right]+\frac{p}{v\left(t\right)}\int\psi_2\left(\frac{d\left(t\right)}{S\left(t\right)}\right)\frac{1}{S^2\left(t\right)}\left[\frac{S\left(t\right)}{2d\left(t\right)}\frac{\partial d^2\left(t\right)}{\partial t}-d\left(t\right)\frac{H\left(t\right)}{B\left(t\right)}\right]dP\left(\mathbf{x},y\right)\\
&= \frac{2p}{S\left(t\right)}\frac{H\left(t\right)}{B\left(t\right)}+\frac{p}{v\left(t\right)}\frac{1}{S^2\left(t\right)}\int\psi_2\left(\frac{d\left(t\right)}{S\left(t\right)}\right)\frac{S\left(t\right)}{2d\left(t\right)}\frac{\partial d^2\left(t\right)}{\partial t}dP\left(\mathbf{x},y\right)\\
&- \frac{p}{v\left(t\right)S\left(t\right)}\left[\frac{H\left(t\right)}{B\left(t\right)}\right]\int\psi_2\left(\frac{d\left(t\right)}{S\left(t\right)}\right)\frac{d\left(t\right)}{S\left(t\right)}dP\left(\mathbf{x},y\right)\\
&= \frac{p}{v\left(t\right)}\frac{1}{S^2\left(t\right)}\int\psi_2\left(\frac{d\left(t\right)}{S\left(t\right)}\right)\frac{S\left(t\right)}{2d\left(t\right)}\frac{\partial d^2\left(t\right)}{\partial t}dP\left(\mathbf{x},y\right)\\
&+ \frac{p}{S\left(t\right)}\left[\frac{H\left(t\right)}{B\left(t\right)}\right]\left(2-\frac{1}{v\left(t\right)}\int\psi_2\left(\frac{d\left(t\right)}{S\left(t\right)}\right)\frac{d\left(t\right)}{S\left(t\right)}dP\left(\mathbf{x},y\right)\right)
\end{aligned}$$

$$\begin{aligned}
m'\left(t\right) &= \frac{p}{v\left(t\right)}\frac{1}{S^2\left(t\right)}\int\psi_2\left(\frac{d\left(t\right)}{S\left(t\right)}\right)\frac{S\left(t\right)}{2d\left(t\right)}\frac{\partial d^2\left(t\right)}{\partial t}dP\left(\mathbf{x},y\right)\\
&+ \frac{p}{S\left(t\right)}\left[\frac{H\left(t\right)}{B\left(t\right)}\right]\frac{1}{v\left(t\right)}\left(\int\left\{2\rho_2\left(\frac{d\left(t\right)}{S\left(t\right)}\right)-\psi_2\left(\frac{d\left(t\right)}{S\left(t\right)}\right)\frac{d\left(t\right)}{S\left(t\right)}\right\}dP\left(\mathbf{x},y\right)\right)\\
&= \frac{p}{2v\left(t\right)S^2\left(t\right)}\int\psi_2\left(\frac{d\left(t\right)}{S\left(t\right)}\right)\frac{S\left(t\right)}{d\left(t\right)}\frac{\partial d^2\left(t\right)}{\partial t}dP\left(\mathbf{x},y\right)\\
&+ \frac{p}{2v\left(t\right)S^2\left(t\right)}\left[\frac{A\left(t\right)}{B\left(t\right)}\right]2S\left(t\right)H\left(t\right)\\
&= \frac{p}{2v\left(t\right)S^2\left(t\right)}\left\{\int\psi_2\left(\frac{d\left(t\right)}{S\left(t\right)}\right)\frac{S\left(t\right)}{d\left(t\right)}\frac{\partial d^2\left(t\right)}{\partial t}dP\left(\mathbf{x},y\right)\right.\\
&+ \left.\left[\frac{A\left(t\right)}{B\left(t\right)}\right]2S\left(t\right)\frac{1}{2S\left(t\right)}\int\psi_1\left(\frac{d\left(t\right)}{S\left(t\right)}\right)\frac{S\left(t\right)}{d\left(t\right)}\frac{\partial d^2\left(t\right)}{\partial t}dP\left(\mathbf{x},y\right)\right\}
\end{aligned}$$

donde hemos usado (196) en la segunda igualdad y (205) en la última igualdad.

$$\begin{aligned}
m'\left(t\right) &= \frac{p}{2v\left(t\right)S^2\left(t\right)}\int\left\{\psi_2\left(\frac{d\left(t\right)}{S\left(t\right)}\right)\frac{S\left(t\right)}{d\left(t\right)}+\frac{A\left(t\right)}{B\left(t\right)}\psi_1\left(\frac{d\left(t\right)}{S\left(t\right)}\right)\frac{S\left(t\right)}{d\left(t\right)}\right\}\frac{\partial d^2\left(t\right)}{\partial t}dP\left(\mathbf{x},y\right)\\
&= \int\widetilde{w}\left(t\right)\frac{\partial d^2\left(t\right)}{\partial t}dP\left(\mathbf{x},y\right)
\end{aligned} \tag{207}$$

Reemplazamos la expresión de $\frac{\partial d^2(t)}{\partial t}$ que calculamos en (202),

$$
\begin{aligned}
m'(t) &= \int \widetilde{w}(t) \frac{\partial d^2(t)}{\partial t} dP(\mathbf{x},y) \\
&= \text{traza}\{\int \widetilde{w}(t) \{\left[\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right]^T \left(\text{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1} \cdot \\
&\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y), \mathbf{x}] - 2\left(\mathbf{x} - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}]\right)^T\} \\
&\Delta^{-1/2}\left[(-2t)\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T + 2t\mathbf{z}_k\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_{i_d}\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\Delta^{-1/2} \\
&\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}, \mathbf{f}(y)]\left(\text{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\left[\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right] dP(\mathbf{x},y)\} \\
&= \int \widetilde{w}(t)\,\text{traza}\{\left[\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right]\{\left[\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right]^T \left(\text{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1} \\
&\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y), \mathbf{x}] - 2\left(\mathbf{x} - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}]\right)^T\} \\
&\Delta^{-1/2}\left[(-2t)\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T + 2t\mathbf{z}_k\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_{i_d}\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\Delta^{-1/2} \\
&\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}, \mathbf{f}(y)]\left(\text{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1} dP(\mathbf{x},y)\}
\end{aligned}
$$

$$
\begin{aligned}
&= \text{traza}\{\int \widetilde{w}(t)\left[\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right]\{\left[\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right]^T \left(\text{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1} \cdot \\
&\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y), \mathbf{x}] - 2\left(\mathbf{x} - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}]\right)^T\} dP(\mathbf{x},y) \\
&\Delta^{-1/2}\left[(-2t)\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T + 2t\mathbf{z}_k\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_{i_d}\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\Delta^{-1/2} \\
&\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}, \mathbf{f}(y)]\left(\text{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\} \\
&= \text{traza}\{\{E_{P^*_{\boldsymbol{\theta}}}\left(\left[\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right]\left[\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right]^T\right)\left(\text{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1} \cdot \\
&\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y), \mathbf{x}] - 2E_{P^*_{\boldsymbol{\theta}}}\left(\left[\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right]\left(\mathbf{x} - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}]\right)^T\right)\} \\
&\Delta^{-1/2}\left[(-2t)\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T + 2t\mathbf{z}_k\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_{i_d}\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\Delta^{-1/2} \\
&\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}, \mathbf{f}(y)]\left(\text{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\}.
\end{aligned}
$$

En $t=0$, $S(0)=1$ así que

$$
\begin{aligned}
m'(0) = \operatorname{traza}\{\{ & E_{P^*_{\boldsymbol{\theta}}}\left(\left[\left(\mathbf{f}(y)-E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right]\left[\left(\mathbf{f}(y)-E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right]^T\right)\left(\operatorname{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\cdot \\
& \operatorname{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]^T - 2E_{P^*_{\boldsymbol{\theta}}}\left(\left[\left(\mathbf{f}(y)-E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right]\left(\mathbf{x}-E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}]\right)^T\right)\} \\
& \Delta^{-1/2}\left[\mathbf{z}_{i_d}\mathbf{z}_k^T+\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\Delta^{-1/2}\operatorname{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\operatorname{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\} \\
& = \operatorname{traza}\{\{\operatorname{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\left(\operatorname{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\operatorname{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]^T \\
& - 2\operatorname{cov}_{P^*_{\boldsymbol{\theta}}}(\mathbf{f}(y),\mathbf{x})\} \\
& \Delta^{-1/2}\left[\mathbf{z}_{i_d}\mathbf{z}_k^T+\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\Delta^{-1/2}\operatorname{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\operatorname{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\}
\end{aligned}
$$

$$
\begin{aligned}
m'(0) = \operatorname{traza}\{\{ & \operatorname{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y),\mathbf{x}] - 2\operatorname{cov}_{P^*_{\boldsymbol{\theta}}}(\mathbf{f}(y),\mathbf{x})\}\Delta^{-1/2}\left[\mathbf{z}_{i_d}\mathbf{z}_k^T+\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\Delta^{-1/2} \\
& \operatorname{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\operatorname{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\} \\
& = -\operatorname{traza}\{\operatorname{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y),\mathbf{x}]\Delta^{-1/2}\left[\mathbf{z}_{i_d}\mathbf{z}_k^T+\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\Delta^{-1/2} \\
& \operatorname{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\operatorname{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\}
\end{aligned}
$$

$$
\begin{aligned}
m'(0) = & -\operatorname{traza}\{\operatorname{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\operatorname{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\operatorname{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y),\mathbf{x}]\Delta^{-1/2} \\
& \cdot\left[\mathbf{z}_{i_d}\mathbf{z}_k^T+\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\Delta^{-1/2}\} \\
& = -\operatorname{traza}\{\Delta^{-1/2}\Pi(\boldsymbol{\theta})\Delta^{-1/2}\left[\mathbf{z}_{i_d}\mathbf{z}_k^T+\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\} = -\operatorname{traza}\{\left[\lambda_{i_d}\mathbf{z}_{i_d}\mathbf{z}_k^T+\lambda_k\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\} \\
& = -\operatorname{traza}\{\left[\lambda_{i_d}\mathbf{z}_k^T\mathbf{z}_{i_d}+\lambda_k\mathbf{z}_{i_d}^T\mathbf{z}_k\right]\} = 0,
\end{aligned}
$$

donde la última igualdad vale por (201). Restaría derivar de nuevo a m y evaluarla en cero. Para verificar que la función m alcanza un máximo local en $t=0$ nos bastaría con ver que $m''(0)<0$. Ya observamos en (207) que $m'(t)$ puede escribirse del siguiente modo

$$
m'(t) = \int \widetilde{w}(t)\frac{\partial d^2(t)}{\partial t}dP(\mathbf{x},y),
$$

entonces acabamos de probar que

$$
m'(0) = \int \widetilde{w}(0)\frac{\partial d^2(0)}{\partial t}dP(\mathbf{x},y) = 0 \tag{208}
$$

La segunda derivada de m tendrá dos sumandos, que surgen de derivar cada uno de los dos términos que se multiplican para definir a $m'(t)$,

$$
m''(t) = \int\left(\frac{\partial}{\partial t}\widetilde{w}(t)\right)\frac{\partial d^2(t)}{\partial t}dP(\mathbf{x},y) + \int \widetilde{w}(t)\frac{\partial^2 d^2(t)}{\partial t^2}dP(\mathbf{x},y). \tag{209}
$$

Comencemos calculando el segundo término.

$$\begin{aligned}\frac{\partial^2 d^2(t)}{\partial t^2} = \frac{\partial}{\partial t}\{&\left[\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right]^T \left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1} \mathrm{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y), \mathbf{x}] - 2\left(\mathbf{x} - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}]\right)^T\} \\ &\Delta^{-1/2}\left[(-2t)\,\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T + 2t\mathbf{z}_k\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_{i_d}\mathbf{z}_k^T + \left(\frac{1-2t^2}{\sqrt{1-t^2}}\right)\mathbf{z}_k\mathbf{z}_{i_d}^T\right]\Delta^{-1/2} \\ &\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}, \mathbf{f}(y)]\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\},\end{aligned}$$

como

$$\frac{\partial}{\partial t}\left(\frac{1-2t^2}{\sqrt{1-t^2}}\right) = \left(\frac{t}{(1-t^2)^{\frac{3}{2}}}\left(2t^2-3\right)\right)$$

resulta

$$\begin{aligned}\frac{\partial^2 d^2(t)}{\partial t^2} = &\left\{\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^T \left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1} \mathrm{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y), \mathbf{x}] - 2\left(\mathbf{x} - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}]\right)^T\right\} \\ &\Delta^{-1/2}\left[(-2)\,\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T + 2\mathbf{z}_k\mathbf{z}_k^T + \left(\frac{t}{(1-t^2)^{\frac{3}{2}}}\left(2t^2-3\right)\right)\left(\mathbf{z}_{i_d}\mathbf{z}_k^T + \mathbf{z}_k\mathbf{z}_{i_d}^T\right)\right]\Delta^{-1/2} \\ &\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}, \mathbf{f}(y)]\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\end{aligned}$$

Luego

$$\begin{aligned}\frac{\partial^2 d^2(0)}{\partial t^2} &= \left.\frac{\partial^2 d^2(t)}{\partial t^2}\right|_{t=0} \\ &= \left\{\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^T \left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1} \mathrm{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y), \mathbf{x}] - 2\left(\mathbf{x} - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}]\right)^T\right\} \\ &\Delta^{-1/2}\left[(-2)\,\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T + 2\mathbf{z}_k\mathbf{z}_k^T\right]\Delta^{-1/2} \\ &\mathrm{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}, \mathbf{f}(y)]\left(\mathrm{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\left(\mathbf{f}(y) - E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\end{aligned}$$

y el segundo sumando queda

$$
\begin{aligned}
&\int \widetilde{w}(0)\frac{\partial^2 d^2(0)}{\partial t^2}dP(\mathbf{x},y)\\
&=E_{P^*_{\boldsymbol{\theta}}}\left(\frac{\partial^2 d^2(t)}{\partial t^2}\right)\\
&=\text{traza}\{\{E_{P^*_{\boldsymbol{\theta}}}\left(\left[\left(\mathbf{f}(y)-E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right]\left[\left(\mathbf{f}(y)-E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right]^T\right)\left(\text{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\cdot\\
&\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y),\mathbf{x}]-2E_{P^*_{\boldsymbol{\theta}}}\left(\left[\left(\mathbf{f}(y)-E_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\right]\left(\mathbf{x}-E_{P^*_{\boldsymbol{\theta}}}[\mathbf{x}]\right)^T\right)\}\\
&\Delta^{-1/2}\left[(-2)\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T+2\mathbf{z}_k\mathbf{z}_k^T\right]\Delta^{-1/2}\\
&\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\text{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\}\\
&=\text{traza}\{\{\left(\text{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)\left(\text{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y),\mathbf{x}]\\
&-2\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y),\mathbf{x}]\}\Delta^{-1/2}\left[(-2)\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T+2\mathbf{z}_k\mathbf{z}_k^T\right]\Delta^{-1/2}\\
&\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\text{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\}\\
&=\text{traza}\{\{\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y),\mathbf{x}]-2\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y),\mathbf{x}]\}\Delta^{-1/2}\left[(-2)\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T+2\mathbf{z}_k\mathbf{z}_k^T\right]\Delta^{-1/2}\\
&\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\text{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\}\\
&=-\text{traza}\{\Delta^{-1/2}\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{x},\mathbf{f}(y)]\left(\text{var}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y)]\right)^{-1}\\
&\text{cov}_{P^*_{\boldsymbol{\theta}}}[\mathbf{f}(y),\mathbf{x}]\Delta^{-1/2}\left[(-2)\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T+2\mathbf{z}_k\mathbf{z}_k^T\right]\\
&=-\text{traza}\{\Delta^{-1/2}\Pi(\boldsymbol{\theta})\Delta^{-1/2}\left[(-2)\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T+2\mathbf{z}_k\mathbf{z}_k^T\right]\}\\
&=-\text{traza}\{\left[(-2)\lambda_{i_d}\mathbf{z}_{i_d}\mathbf{z}_{i_d}^T+2\lambda_k\mathbf{z}_k\mathbf{z}_k^T\right]\}\\
&=-\text{traza}\{\left[(-2)\lambda_{i_d}\mathbf{z}_{i_d}^T\mathbf{z}_{i_d}+2\lambda_k\mathbf{z}_k^T\mathbf{z}_k\right]\}=-\left((-2)\lambda_{i_d}+2\lambda_k\right)\\
&=2(\lambda_{i_d}-\lambda_k) \qquad (210)
\end{aligned}
$$

Como $\lambda_k > \lambda_{i_d}$, resulta que $[\lambda_{i_d}-\lambda_k]<0$, o sea que $m''(0)$ es la suma de dos términos, según (209), el segundo de los cuales es negativo. Veamos el signo del primero.

$$\widetilde{w}(t)=\frac{p}{2v(t)S^2(t)}\left\{\psi_2\left(\frac{d(t)}{S(t)}\right)\frac{S(t)}{d(t)}+\frac{A(t)}{B(t)}\psi_1\left(\frac{d(t)}{S(t)}\right)\frac{S(t)}{d(t)}\right\}$$

entonces

$$\begin{aligned}\widetilde{w}'(t) &= \left(\frac{p}{2v(t)S^2(t)}\right)'\left\{\psi_2\left(\frac{d(t)}{S(t)}\right)+\frac{A(t)}{B(t)}\psi_1\left(\frac{d(t)}{S(t)}\right)\right\}\frac{S(t)}{d(t)}\\ &+\frac{p}{2v(t)S^2(t)}\left\{\psi_2'\left(\frac{d(t)}{S(t)}\right)+\frac{A(t)}{B(t)}\psi_1'\left(\frac{d(t)}{S(t)}\right)\right\}\frac{S(t)}{d(t)}\frac{\partial}{\partial t}\left(\frac{d(t)}{S(t)}\right)\\ &+\frac{p}{2v(t)S^2(t)}\left\{\psi_2\left(\frac{d(t)}{S(t)}\right)+\frac{A(t)}{B(t)}\psi_1\left(\frac{d(t)}{S(t)}\right)\right\}\frac{\partial}{\partial t}\left(\frac{S(t)}{d(t)}\right)\\ &+\frac{p}{2v(t)S^2(t)}\frac{\partial}{\partial t}\left(\frac{A(t)}{B(t)}\right)\psi_1\left(\frac{d(t)}{S(t)}\right)\left(\frac{S(t)}{d(t)}\right).\end{aligned}$$

Llamemos

$$\begin{aligned}g_1(t) &= \left(\frac{p}{2v(t)S^2(t)}\right)'\left\{\psi_2\left(\frac{d(t)}{S(t)}\right)+\frac{A(t)}{B(t)}\psi_1\left(\frac{d(t)}{S(t)}\right)\right\}\frac{S(t)}{d(t)}\\ g_2(t) &= \frac{p}{2v(t)S^2(t)}\left\{\psi_2'\left(\frac{d(t)}{S(t)}\right)+\frac{A(t)}{B(t)}\psi_1'\left(\frac{d(t)}{S(t)}\right)\right\}\frac{S(t)}{d(t)}\frac{\partial}{\partial t}\left(\frac{d(t)}{S(t)}\right)\end{aligned}$$

$$\begin{aligned}g_3(t) &= \frac{p}{2v(t)S^2(t)}\left\{\psi_2\left(\frac{d(t)}{S(t)}\right)+\frac{A(t)}{B(t)}\psi_1\left(\frac{d(t)}{S(t)}\right)\right\}\frac{\partial}{\partial t}\left(\frac{S(t)}{d(t)}\right)\\ g_4(t) &= \frac{p}{2v(t)S^2(t)}\frac{\partial}{\partial t}\left(\frac{A(t)}{B(t)}\right)\psi_1\left(\frac{d(t)}{S(t)}\right)\left(\frac{S(t)}{d(t)}\right)\end{aligned}$$

Usando que

$$S(0)=1$$

y notando

$$\frac{\partial}{\partial t}d^2(0)=\left.\frac{\partial d^2(t)}{\partial t}\right|_{t=0},$$

como hemos descompuesto a la derivada en la suma

$$\widetilde{w}'(t)=g_1(t)+g_2(t)+g_3(t)+g_4(t),$$

resulta que

$$m''(t)=\sum_{i=1}^{4}\int g_i(t)\frac{\partial}{\partial t}d^2(t)\,dP(\mathbf{x},y)+\int\widetilde{w}(t)\frac{\partial^2 d^2(t)}{\partial t^2}dP(\mathbf{x},y).$$

Analicemos cada uno de los primeros cuatro sumandos de esta expresión,

$$\begin{aligned}
G_1(t) &:= \int g_1(t)\frac{\partial}{\partial t}d^2(t)\,dP(\mathbf{x},y) \\
&= \int \left(\frac{p}{2v(t)S^2(t)}\right)'\left\{\psi_2\left(\frac{d(t)}{S(t)}\right)\frac{S(t)}{d(t)}+\frac{A(t)}{B(t)}\psi_1\left(\frac{d(t)}{S(t)}\right)\frac{S(t)}{d(t)}\right\}\frac{\partial}{\partial t}d^2(t)\,dP(\mathbf{x},y) \\
&= \int \frac{-p\left[v'(t)S^2(t)+v(t)2S(t)S'(t)\right]}{2v^2(t)S^4(t)}\left\{\left(\frac{2v(t)S^2(t)}{p}\right)\widetilde{w}(t)\right\}\frac{\partial}{\partial t}d^2(t)\,dP(\mathbf{x},y) \\
&= \int \frac{-\left[v'(t)S^2(t)+v(t)2S(t)S'(t)\right]}{v(t)S^2(t)}\widetilde{w}(t)\frac{\partial}{\partial t}d^2(t)\,dP(\mathbf{x},y) \\
&= \frac{-\left[v'(t)S^2(t)+v(t)2S(t)S'(t)\right]}{v(t)S^2(t)}\int \widetilde{w}(t)\frac{\partial}{\partial t}d^2(t)\,dP(\mathbf{x},y)
\end{aligned}$$

Como por (208) tenemos,

$$\int \widetilde{w}(0)\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y)=0$$

entonces,

$$G_1(0)=0. \tag{211}$$

$$\begin{aligned}
G_2(t) &= \int \frac{p}{2v(t)S^2(t)}\left\{\psi_2'\left(\frac{d(t)}{S(t)}\right)+\frac{A(t)}{B(t)}\psi_1'\left(\frac{d(t)}{S(t)}\right)\right\}\frac{S(t)}{d(t)}\frac{\partial}{\partial t}\left(\frac{d(t)}{S(t)}\right)\frac{\partial}{\partial t}d^2(t)\,dP(\mathbf{x},y) \\
G_2(0) &= \int \frac{p}{2v(0)}\left\{\psi_2'(d(0))+\frac{A(0)}{B(0)}\psi_1'(d(0))\right\} \\
&\quad \frac{1}{d(0)}\left(\frac{S(0)\frac{\partial}{\partial t}d(0)-d(0)\frac{H(0)}{B(0)}}{S^2(0)}\right)\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y) \\
&= \int \frac{p}{2v(0)}\left\{\psi_2'(d(0))+\frac{A(0)}{B(0)}\psi_1'(d(0))\right\} \\
&\quad \frac{1}{d(0)}\left(\left(\frac{\partial}{\partial t}d(0)\right)-d(0)\frac{H(0)}{B(0)}\right)\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y) \\
&= \int \frac{p}{v(0)}\left\{\psi_2'(d(0))+\frac{A(0)}{B(0)}\psi_1'(d(0))\right\}\left(\frac{\partial}{\partial t}d(0)\right)^2 dP(\mathbf{x},y) \\
&\quad -\frac{H(0)}{B(0)}\int \frac{p}{2v(0)}\left\{\psi_2'(d(0))+\frac{A(0)}{B(0)}\psi_1'(d(0))\right\}\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y)
\end{aligned} \tag{212}$$

donde usamos (203) para la última igualdad.

$$\begin{aligned}
G_3(t) &= \int \frac{p}{2v(t)}\left\{\psi_2\left(\frac{d(t)}{S(t)}\right)+\frac{A(t)}{B(t)}\psi_1\left(\frac{d(t)}{S(t)}\right)\right\}\frac{\partial}{\partial t}\left(\frac{S(t)}{d(t)}\right)\frac{\partial}{\partial t}d^2(t)\,dP(\mathbf{x},y) \\
&= \int \frac{p}{2v(t)}\left\{\psi_2\left(\frac{d(t)}{S(t)}\right)+\frac{A(t)}{B(t)}\psi_1\left(\frac{d(t)}{S(t)}\right)\right\}\left(\frac{\frac{H(t)}{B(t)}d(t)-S(t)\frac{\partial}{\partial t}d(t)}{d^2(t)}\right)\frac{\partial}{\partial t}d^2(t)\,dP(\mathbf{x},y)
\end{aligned}$$

Luego

$$\begin{aligned}
G_3(0) &= \int \frac{p}{2v(0)}\left\{\frac{\psi_2(d(0))}{d(0)} + \frac{A(0)}{B(0)}\frac{\psi_1(d(0))}{d(0)}\right\}\left(\frac{H(0)}{B(0)} - \frac{\frac{\partial}{\partial t}d(0)}{d(0)}\right)\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y) \\
&= \int \widetilde{w}(0)\left(\frac{H(0)}{B(0)} - \frac{1}{d(0)}\frac{\partial}{\partial t}d(0)\right)\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y) \\
&= \frac{H(0)}{B(0)}\int \widetilde{w}(0)\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y) - 2\int \widetilde{w}(0)\left(\frac{\partial}{\partial t}d(0)\right)^2 dP(\mathbf{x},y) \\
&= 0 - 2\int \widetilde{w}(0)\left(\frac{\partial}{\partial t}d(0)\right)^2 dP(\mathbf{x},y) \qquad (213)
\end{aligned}$$

usamos nuevamente (203) para la penúltima igualdad y (208) en la última.

$$\begin{aligned}
G_4(t) &= \int \frac{p}{2v(t)S^2(t)}\frac{\partial}{\partial t}\left(\frac{A(t)}{B(t)}\right)\psi_1\left(\frac{d(t)}{S(t)}\right)\left(\frac{S(t)}{d(t)}\right)\frac{\partial}{\partial t}d^2(t)\,dP(\mathbf{x},y) \\
G_4(0) &= \frac{p}{2v(0)}\,\frac{\partial}{\partial t}\left(\frac{A(t)}{B(t)}\right)\Big|_{t=0}\int \frac{\psi_1(d(0))}{d(0)}\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y) \\
&= \frac{p}{2v(0)}\left(\frac{A'(0)B(0) - A(0)B'(0)}{B^2(0)}\right)H(0)\,2
\end{aligned}$$

Como

$$B(t) = \int \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{d(t)}{S(t)}dP(\mathbf{x},y),$$

tenemos

$$\begin{aligned}
B'(t) &= \int \psi_1'\left(\frac{d(t)}{S(t)}\right)\frac{\partial}{\partial t}\left(\frac{d(t)}{S(t)}\right)\frac{d(t)}{S(t)} + \psi_1\left(\frac{d(t)}{S(t)}\right)\frac{\partial}{\partial t}\left(\frac{d(t)}{S(t)}\right)dP(\mathbf{x},y) \\
&= \int \left\{\psi_1'\left(\frac{d(t)}{S(t)}\right)\frac{d(t)}{S(t)} + \psi_1\left(\frac{d(t)}{S(t)}\right)\right\}\frac{\partial}{\partial t}\left(\frac{d(t)}{S(t)}\right)dP(\mathbf{x},y) \\
&= \int \left\{\psi_1'\left(\frac{d(t)}{S(t)}\right)\frac{d(t)}{S(t)} + \psi_1\left(\frac{d(t)}{S(t)}\right)\right\}\left(S(t)\frac{\partial}{\partial t}d(t) - d(t)\frac{H(t)}{B(t)}\right)\frac{1}{S^2(t)}dP(\mathbf{x},y) \\
B'(0) &= \int \left\{\psi_1'(d(0))d(0) + \psi_1(d(0))\right\}\left(\frac{\partial}{\partial t}d(0) - d(0)\frac{H(0)}{B(0)}\right)dP(\mathbf{x},y) \\
&= \int \left\{\psi_1'(d(0))d(0) + \psi_1(d(0))\right\}\left(\frac{1}{2d(0)}\frac{\partial}{\partial t}d^2(0) - d(0)\frac{H(0)}{B(0)}\right)dP(\mathbf{x},y)
\end{aligned}$$

$$\begin{aligned}
A(t) &= E_P\left[a\left(\frac{d(t)}{S(t)}\right)\right] = E_P\left[2\rho_2\left(\frac{d(t)}{S(t)}\right) - \psi_2\left(\frac{d(t)}{S(t)}\right)\frac{d(t)}{S(t)}\right] \\
A(0) &= E_P\left[2\rho_2(d(0)) - \psi_2(d(0))d(0)\right] \\
&= \int \left[2\rho_2(d(0)) - \psi_2(d(0))d(0)\right]dP(\mathbf{x},y)
\end{aligned}$$

$$A'(t) = \int \left[2\psi_2\left(\frac{d(t)}{S(t)}\right) - \psi_2'\left(\frac{d(t)}{S(t)}\right)\frac{d(t)}{S(t)} - \psi_2\left(\frac{d(t)}{S(t)}\right)\right]\frac{\partial}{\partial t}\left(\frac{d(t)}{S(t)}\right)dP(\mathbf{x},y)$$
$$= \int \left[\psi_2\left(\frac{d(t)}{S(t)}\right) - \psi_2'\left(\frac{d(t)}{S(t)}\right)\frac{d(t)}{S(t)}\right]\frac{\partial}{\partial t}\left(\frac{d(t)}{S(t)}\right)dP(\mathbf{x},y)$$

$$A'(0) = \int \left[\psi_2(d(0)) - \psi_2'(d(0))d(0)\right]\left(\frac{\partial}{\partial t}d(0) - d(0)\frac{H(0)}{B(0)}\right)dP(\mathbf{x},y)$$
$$= \int \left[\psi_2(d(0)) - \psi_2'(d(0))d(0)\right]\left(\frac{1}{2d(0)}\frac{\partial}{\partial t}d^2(0) - d(0)\frac{H(0)}{B(0)}\right)dP(\mathbf{x},y)$$

Reemplazamos estas expresiones en G_4

$$G_4(0) = \frac{p}{2v(0)}\left(\frac{A'(0)B(0) - A(0)B'(0)}{B^2(0)}\right)2H(0)$$
$$= \frac{pH(0)}{v(0)}\frac{1}{B(0)}\left\{\int \left[\psi_2(d(0)) - \psi_2'(d(0))d(0)\right]\left(\frac{1}{2d(0)}\frac{\partial}{\partial t}d^2(0) - d(0)\frac{H(0)}{B(0)}\right)dP(\mathbf{x},y)\right.$$
$$\left.-\frac{A(0)}{B(0)}\int \left\{\psi_1'(d(0))d(0) + \psi_1(d(0))\right\}\left(\frac{1}{2d(0)}\frac{\partial}{\partial t}d^2(0) - d(0)\frac{H(0)}{B(0)}\right)dP(\mathbf{x},y)\right\}$$

$$= \frac{H(0)p}{v(0)}\frac{1}{B(0)}\left\{\int \left[\psi_2(d(0)) - \psi_2'(d(0))d(0) - \frac{A(0)}{B(0)}\left\{\psi_1'(d(0))d(0) + \psi_1(d(0))\right\}\right]\cdot\right.$$
$$\frac{1}{2d(0)}\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y)$$
$$\left.-\int \left[\psi_2(d(0)) - \psi_2'(d(0))d(0) - \frac{A(0)}{B(0)}\left\{\psi_1'(d(0))d(0) + \psi_1(d(0))\right\}\right]d(0)\frac{H(0)}{B(0)}dP(\mathbf{x},y)\right\}$$

$$= \frac{H(0)p}{v(0)}\frac{1}{B(0)}\left\{\int \psi_2(d(0))\frac{1}{2d(0)}\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y)\right. \tag{214}$$
$$-\int \left(\psi_2'(d(0))d(0) + \frac{A(0)}{B(0)}\psi_1'(d(0))d(0)\right)\frac{1}{2d(0)}\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y) \tag{215}$$
$$-\int \frac{A(0)}{B(0)}\psi_1(d(0))\frac{1}{2d(0)}\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y) \tag{216}$$
$$-\int [\psi_2(d(0))]d(0)\frac{H(0)}{B(0)}dP(\mathbf{x},y) \tag{217}$$
$$+\int \left[\psi_2'(d(0)) + \frac{A(0)}{B(0)}\psi_1'(d(0))\right]d^2(0)\frac{H(0)}{B(0)}dP(\mathbf{x},y) \tag{218}$$
$$\left.-\frac{A(0)}{B(0)}\frac{H(0)}{B(0)}\int \psi_1(d(0))d(0)\,dP(\mathbf{x},y)\right\} \tag{219}$$

Recordemos (208)

$$\int \widetilde{w}(0)\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y) = 0,$$

es decir,

$$\frac{p}{2v(0)}\int\left(\frac{\psi_2(d(0))}{d(0)}+\frac{A(0)}{B(0)}\frac{\psi_1(d(0))}{d(0)}\right)\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y)=0$$

Luego

$$\begin{aligned}\frac{1}{2}\int\frac{\psi_2(d(0))}{d(0)}\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y)&=-\int\frac{A(0)}{B(0)}\frac{\psi_1(d(0))}{2d(0)}\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y)\\&=-H(0)\frac{A(0)}{B(0)}\end{aligned}$$

es decir, (214) es igual que (216) e igual a $-H(0)\frac{A(0)}{B(0)}$, pues

$$\begin{aligned}H(0)&=\frac{1}{2}\int\frac{\psi_1(d(0))}{d(0)}\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y)\\&=\int\psi_1(d(0))\frac{\partial}{\partial t}d(0)\,dP(\mathbf{x},y).\end{aligned}$$

Por otro lado, si sumamos (217) y (219) obtenemos

$$\begin{aligned}&\frac{H(0)p}{B(0)v(0)}\left\{-\int[\psi_2(d(0))]d(0)\frac{H(0)}{B(0)}dP(\mathbf{x},y)-\frac{A(0)}{B(0)}\frac{H(0)}{B(0)}\int\psi_1(d(0))d(0)\,dP(\mathbf{x},y)\right\}\\&=-\frac{H^2(0)}{B^2(0)}\int\frac{p}{v(0)}\left[\psi_2(d(0))+\frac{A(0)}{B(0)}\psi_1(d(0))\right]d(0)\,dP(\mathbf{x},y)\\&=-\frac{H^2(0)}{B^2(0)}\int\frac{p}{v(0)}\left[\frac{\psi_2(d(0))}{d(0)}+\frac{A(0)}{B(0)}\frac{\psi_1(d(0))}{d(0)}\right]d^2(0)\,dP(\mathbf{x},y)\\&=-\frac{H^2(0)2}{B^2(0)}\int\widetilde{w}(0)\,d^2(0)\,dP(\mathbf{x},y)\end{aligned}$$

Luego

$$\begin{aligned}G_4(0)&=\frac{p}{2v(0)}\left(\frac{A'(0)B(0)-A(0)B'(0)}{B^2(0)}\right)2H(0)\\&=\frac{H(0)p}{B(0)v(0)}\left\{-2H(0)\frac{A(0)}{B(0)}\right.\\&-\int\left(\psi_2'(d(0))d(0)+\frac{A(0)}{B(0)}\psi_1'(d(0))d(0)\right)\frac{1}{2d(0)}\frac{\partial}{\partial t}d^2(0)\,dP(\mathbf{x},y)\\&\left.+\int\left[\psi_2'(d(0))+\frac{A(0)}{B(0)}\psi_1'(d(0))\right]d^2(0)\frac{H(0)}{B(0)}dP(\mathbf{x},y)\right\}\\&-\frac{H^2(0)2}{B^2(0)}\int\widetilde{w}(0)\,d^2(0)\,dP(\mathbf{x},y)\end{aligned}\tag{220}$$

Entonces, finalmente,

$$\begin{aligned} m''(0) &= \int \widetilde{w}'(0) \frac{\partial}{\partial t} d^2(0)\, dP(\mathbf{x},y) + 2(\lambda_{i_d} - \lambda_k) \\ &= G(0) + \widetilde{G}(0) \end{aligned}$$

y de (211), (212), (213) y (220) resulta

$$\begin{aligned} G(0) &= G_1(0) + G_2(0) + G_3(0) + G_4(0) \\ &= 0 + \int \frac{p}{v(0)} \left\{ \psi_2'(d(0)) + \frac{A(0)}{B(0)} \psi_1'(d(0)) \right\} \left(\frac{\partial}{\partial t} d(0) \right)^2 dP(\mathbf{x},y) \qquad (221) \\ &- \frac{H(0)}{B(0)} \int \frac{p}{2v(0)} \left\{ \psi_2'(d(0)) + \frac{A(0)}{B(0)} \psi_1'(d(0)) \right\} \frac{\partial}{\partial t} d^2(0)\, dP(\mathbf{x},y) \qquad (222) \\ &- 2 \int \widetilde{w}(0) \left(\frac{\partial}{\partial t} d(0) \right)^2 dP(\mathbf{x},y) - \frac{2H^2(0) A(0) p}{B^2(0) v(0)} \\ &+ \frac{H(0) p}{B(0) v(0)} \left\{ - \int \left(\psi_2'(d(0)) d(0) + \frac{A(0)}{B(0)} \psi_1'(d(0)) d(0) \right) \frac{1}{2d(0)} \frac{\partial}{\partial t} d^2(0)\, dP(\mathbf{x},y) \right. \qquad (223) \\ &+ \left. \int \left[\psi_2'(d(0)) + \frac{A(0)}{B(0)} \psi_1'(d(0)) \right] d^2(0) \frac{H(0)}{B(0)} dP(\mathbf{x},y) \right\} \\ &- \frac{H^2(0) 2}{B^2(0)} \int \widetilde{w}(0) d^2(0)\, dP(\mathbf{x},y) \end{aligned}$$

Observemos que (222) y (223) son iguales, luego

$$\begin{aligned} G(0) &= \int \frac{p}{v(0)} \left\{ \psi_2'(d(0)) + \frac{A(0)}{B(0)} \psi_1'(d(0)) \right\} \left(\frac{\partial}{\partial t} d(0) \right)^2 dP(\mathbf{x},y) \\ &- \frac{H(0)}{B(0)} \int \frac{p}{v(0)} \left\{ \psi_2'(d(0)) + \frac{A(0)}{B(0)} \psi_1'(d(0)) \right\} \frac{\partial}{\partial t} d^2(0)\, dP(\mathbf{x},y) \\ &- 2 \int \widetilde{w}(0) \left(\frac{\partial}{\partial t} d(0) \right)^2 dP(\mathbf{x},y) - \frac{2H^2(0) A(0) p}{B^2(0) v(0)} \\ &+ \int \left[\psi_2'(d(0)) + \frac{A(0)}{B(0)} \psi_1'(d(0)) \right] d^2(0) \frac{H^2(0) p}{B^2(0) v(0)} dP(\mathbf{x},y) \\ &- \frac{H^2(0) p}{B^2(0)} \int \widetilde{w}(0) d^2(0)\, dP(\mathbf{x},y) \end{aligned}$$

$$G(0) = \int \frac{p}{v(0)} \left\{ \psi_2'(d(0)) + \frac{A(0)}{B(0)} \psi_1'(d(0)) \right\} \left(\frac{\partial}{\partial t} d(0) \right)^2 dP(\mathbf{x},y) \tag{224}$$

$$- \int \frac{p}{v(0)} \left\{ \psi_2'(d(0)) + \frac{A(0)}{B(0)} \psi_1'(d(0)) \right\} \frac{H(0)}{B(0)} 2d(0) \frac{\partial}{\partial t} d(0) \, dP(\mathbf{x},y) \tag{225}$$

$$- 2 \int \widetilde{w}(0) \left(\frac{\partial}{\partial t} d(0) \right)^2 dP(\mathbf{x},y) - \frac{2H^2(0) A(0) p}{B^2(0) v(0)}$$

$$+ \int \frac{p}{v(0)} \left[\psi_2'(d(0)) + \frac{A(0)}{B(0)} \psi_1'(d(0)) \right] d^2(0) \frac{H^2(0)}{B^2(0)} dP(\mathbf{x},y) \tag{226}$$

$$- \frac{H^2(0) 2}{B^2(0)} \int \widetilde{w}(0) d^2(0) \, dP(\mathbf{x},y)$$

Juntando (224) con (225) y (226)

$$G(0) = \int \frac{p}{v(0)} \left\{ \psi_2'(d(0)) + \frac{A(0)}{B(0)} \psi_1'(d(0)) \right\} \left[\left(\frac{\partial}{\partial t} d(0) \right) - \frac{H(0)}{B(0)} d(0) \right]^2 dP(\mathbf{x},y)$$

$$- 2 \int \widetilde{w}(0) \left(\frac{\partial}{\partial t} d(0) \right)^2 dP(\mathbf{x},y) - \frac{2H^2(0) A(0) p}{B^2(0) v(0)}$$

$$- \frac{H^2(0) 2}{B^2(0)} \int \widetilde{w}(0) d^2(0) \, dP(\mathbf{x},y)$$

Asumimos que $\frac{\psi_i(t)}{t}$ son no crecientes (para $i = 1, 2$), es decir, que, para todo $t > 0$

$$\frac{\partial}{\partial t} \frac{\psi_i(t)}{t} = \frac{\psi_i'(t) t - \psi_i(t)}{t^2} \leq 0,$$

o, equivalentemente, para todo $t > 0$ tenemos

$$\psi_i'(t) \leq \frac{\psi_i(t)}{t}.$$

Usando esta cota, como $\left[\left(\frac{\partial}{\partial t}d\left(0\right)\right)-\frac{H(0)}{B(0)}d\left(0\right)\right]^{2}\geq 0$, tenemos

$$\begin{aligned}
&\int\frac{p}{v\left(0\right)}\left\{\psi_{2}'\left(d\left(0\right)\right)+\frac{A\left(0\right)}{B\left(0\right)}\psi_{1}'\left(d\left(0\right)\right)\right\}\left[\frac{\partial}{\partial t}d\left(0\right)-\frac{H\left(0\right)}{B\left(0\right)}d\left(0\right)\right]^{2}dP\left(\mathbf{x},y\right)\\
&\leq\int\frac{p}{v\left(0\right)}\left\{\frac{\psi_{2}\left(d\left(0\right)\right)}{d\left(0\right)}+\frac{A\left(0\right)}{B\left(0\right)}\frac{\psi_{1}\left(d\left(0\right)\right)}{d\left(0\right)}\right\}\left[\frac{\partial}{\partial t}d\left(0\right)-\frac{H\left(0\right)}{B\left(0\right)}d\left(0\right)\right]^{2}dP\left(\mathbf{x},y\right)\\
&=\int2\widetilde{w}\left(0\right)\left[\frac{\partial}{\partial t}d\left(0\right)-\frac{H\left(0\right)}{B\left(0\right)}d\left(0\right)\right]^{2}dP\left(\mathbf{x},y\right)\\
&=\int2\widetilde{w}\left(0\right)\left[\frac{\partial}{\partial t}d\left(0\right)\right]^{2}dP\left(\mathbf{x},y\right)\\
&-4\int\widetilde{w}\left(0\right)\left[\frac{\partial}{\partial t}d\left(0\right)\right]\frac{H\left(0\right)}{B\left(0\right)}d\left(0\right)dP\left(\mathbf{x},y\right)\\
&+\int2\widetilde{w}\left(0\right)\left[\frac{H\left(0\right)}{B\left(0\right)}d\left(0\right)\right]^{2}dP\left(\mathbf{x},y\right)
\end{aligned}$$

Entonces

$$\begin{aligned}
G\left(0\right)&=\int\frac{p}{v\left(0\right)}\left\{\psi_{2}'\left(d\left(0\right)\right)+\frac{A\left(0\right)}{B\left(0\right)}\psi_{1}'\left(d\left(0\right)\right)\right\}\left[\left(\frac{\partial}{\partial t}d\left(0\right)\right)-\frac{H\left(0\right)}{B\left(0\right)}d\left(0\right)\right]^{2}dP\left(\mathbf{x},y\right)\\
&-2\int\widetilde{w}\left(0\right)\left[\frac{\partial}{\partial t}d\left(0\right)\right]^{2}dP\left(\mathbf{x},y\right)-\frac{2H^{2}\left(0\right)A\left(0\right)p}{B^{2}\left(0\right)v\left(0\right)}\\
&-\frac{H^{2}\left(0\right)2}{B^{2}\left(0\right)}\int\widetilde{w}\left(0\right)d^{2}\left(0\right)dP\left(\mathbf{x},y\right)\\
&\leq\int2\widetilde{w}\left(0\right)\left[\frac{\partial}{\partial t}d\left(0\right)\right]^{2}dP\left(\mathbf{x},y\right)-4\int\widetilde{w}\left(0\right)\left(\frac{\partial}{\partial t}d\left(0\right)\right)\frac{H\left(0\right)}{B\left(0\right)}d\left(0\right)dP\left(\mathbf{x},y\right)\\
&+\int2\widetilde{w}\left(0\right)\left[\frac{H\left(0\right)}{B\left(0\right)}d\left(0\right)\right]^{2}dP\left(\mathbf{x},y\right)-2\int\widetilde{w}\left(0\right)\left[\frac{\partial}{\partial t}d\left(0\right)\right]^{2}dP\left(\mathbf{x},y\right)\\
&-\frac{2H^{2}\left(0\right)A\left(0\right)p}{B^{2}\left(0\right)v\left(0\right)}-\frac{H^{2}\left(0\right)2}{B^{2}\left(0\right)}\int\widetilde{w}\left(0\right)d^{2}\left(0\right)dP\left(\mathbf{x},y\right)\\
&=-4\int\widetilde{w}\left(0\right)\left[\frac{\partial}{\partial t}d\left(0\right)\right]\frac{H\left(0\right)}{B\left(0\right)}d\left(0\right)dP\left(\mathbf{x},y\right)\\
&-\frac{2H^{2}\left(0\right)A\left(0\right)p}{B^{2}\left(0\right)v\left(0\right)}
\end{aligned}$$

ya que el primer y el cuarto sumando son iguales pero con signo distinto, y lo mismo

pasa con el tercero y el sexto. Además

$$\begin{aligned}
&-4\int \widetilde{w}\,(0)\left(\frac{\partial}{\partial t}d\,(0)\right)\frac{H\,(0)}{B\,(0)}d\,(0)\,dP\,(\mathbf{x},y)\\
&=(-2)\,\frac{H\,(0)}{B\,(0)}\int \widetilde{w}\,(0)\left(\frac{\partial}{\partial t}d\,(0)\right)2d\,(0)\,dP\,(\mathbf{x},y)\\
&=(-2)\,\frac{H\,(0)}{B\,(0)}\int \widetilde{w}\,(0)\left(\frac{\partial}{\partial t}d^2\,(0)\right)dP\,(\mathbf{x},y)=0
\end{aligned}$$

por (208). Finalmente

$$\begin{aligned}
m''\,(0)&=\int \widetilde{w}'\,(0)\,\frac{\partial}{\partial t}d^2\,(0)\,dP\,(\mathbf{x},y)+2(\lambda_{i_d}-\lambda_k)\\
&=G\,(0)+\widetilde{G}\,(0)\\
&\leq -\frac{2H^2\,(0)\,A\,(0)\,p}{B^2\,(0)\,v\,(0)}+2(\lambda_{i_d}-\lambda_k)<0.\ \blacksquare
\end{aligned}$$

H.1.1. Igualdades que involucran derivadas matriciales

A continuación exhibimos las ecuaciones del libro de Petersen y Pedersen [2008], citadas en esta tesis, en el orden en el que fueron citadas y con la numeración de la publicación original, para facilitar la lectura.

Sea $W\in\mathbb{R}^{n\times n}$ simétrica, $\mathbf{x},\mathbf{s}\in\mathbb{R}^n$, $A\in\mathbb{R}^{n\times n}$, entonces

$$\frac{\partial}{\partial \mathbf{s}}\left(\mathbf{x}-A\mathbf{s}\right)^T W\left(\mathbf{x}-A\mathbf{s}\right)=-2A^TW\left(\mathbf{x}-A\mathbf{s}\right). \qquad \text{(Petersen, ec 76)}$$

Sea $X\in\mathbb{R}^{n\times k}$ una matriz (o vector, si $k=1$), $\mathbf{a}\in\mathbb{R}^n$, $\mathbf{b}\in\mathbb{R}^k$, entonces

$$\frac{\partial}{\partial X}\left(\mathbf{a}^TX\mathbf{b}\right)=\mathbf{a}\mathbf{b}^T. \qquad \text{(Petersen, ec 62)}$$

Sea $W\in\mathbb{R}^{n\times n}$ simétrica, $\mathbf{x},\mathbf{s}\in\mathbb{R}^n$, $A\in\mathbb{R}^{n\times n}$

$$\frac{\partial}{\partial A}\left(\mathbf{x}-A\mathbf{s}\right)^T W\left(\mathbf{x}-A\mathbf{s}\right)=-2W\left(\mathbf{x}-A\mathbf{s}\right)\mathbf{s}^T. \qquad \text{(Petersen, ec 80)}$$

Sea X una matriz inversible,

$$\frac{\partial}{\partial X}\ln\left|\det\left(X\right)\right|=\left(X^{-1}\right)^T=\left(X^T\right)^{-1} \qquad \text{(Petersen, ec 51)}$$

Sea X una matriz inversible, y A, B matrices para las cuales los siguientes productos están bien definidos,

$$\frac{\partial}{\partial X}\text{traza}\left(AX^{-1}B\right)=-\left(X^{-1}BAX^{-1}\right)^T \qquad \text{(Petersen, ec 57)}$$

I. Apéndice del Capítulo 9

I.1. Procedimientos para calcular los ángulos principales

En esta sección presentamos varias formas (equivalentes) de calcular los ángulos ordenados entre dos subespacios $\mathcal{A}$ y $\mathcal{B}$, y también señalaremos algunos resultados y comentarios sobre el tema, que conciernen a los cálculos numéricos.

Procedimiento 1. Directamente derivado de la teoría anterior.

Sean $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ una base para $\mathcal{A}$, y $\{\mathbf{w}_1, \ldots, \mathbf{w}_l\}$ una base para $\mathcal{B}$. Supongamos, sin pérdida de generalidad, que $k \leq l$.

1. Hallar la matriz para $P_{\mathcal{A}}$ y $P_{\mathcal{B}}$ (en la base canónica). Esto se consigue encolumnando los vectores de la base de la siguiente forma: sean

$$V = \left[\begin{array}{ccc} \mathbf{v}_1 & \cdots & \mathbf{v}_k \end{array} \right] \in \mathbb{R}^{p \times k}$$
$$W = \left[\begin{array}{ccc} \mathbf{w}_1 & \cdots & \mathbf{w}_l \end{array} \right] \in \mathbb{R}^{p \times l}$$

   entonces

$$P_{\mathcal{A}} = V \left(V^T V\right)^{-1} V^T$$
$$P_{\mathcal{B}} = W \left(W^T W\right)^{-1} W^T.$$

2. Hallar los autovalores y autovectores de la matriz simétrica $P_{\mathcal{A}} P_{\mathcal{B}} P_{\mathcal{A}} \in \mathbb{R}^{p \times p}$, llamemos $\mu_1^2 \geq \cdots \geq \mu_k^2 > 0$ a los autovalores.

3. Entonces, como

$$(\cos \zeta_i)^2 = \mu_i^2, \qquad i = 1, \ldots, k,$$

   el $i-$ésimo ángulo principal es

$$\zeta_i = \arccos \sqrt{\mu_i^2}, \qquad i = 1, \ldots, k.$$

Como el $\arccos \sqrt{\mu^2}$ es una función decreciente de μ, los ángulos elegidos de esta forma resultan en el orden correcto: $\zeta_1 \leq \zeta_2 \leq \cdots \leq \zeta_k$.

Procedimiento 2. Supongamos que $\mathcal{A}$ y $\mathcal{B}$ tienen la misma dimensión k. Entonces $\mathcal{A} = \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ y $\mathcal{B} = \operatorname{span}\{\mathbf{b}_1, \ldots, \mathbf{b}_k\}$. Si aplicamos la transformación lineal $P_{\mathcal{B}} P_{\mathcal{A}}$ a los generadores de $\mathcal{B}$ obtenemos

$$P_{\mathcal{B}} P_{\mathcal{A}} \mathbf{b}_j = \sum_{i=1}^{k} \mu_i \mathbf{b}_i \mathbf{a}_i^T \mathbf{b}_j = \mu_j \mathbf{b}_j \mathbf{a}_j^T \mathbf{b}_j = \mu_j^2 \mathbf{b}_j,$$

donde la segunda igualdad es consecuencia del Teorema 9.1.(*iv*), de modo que $\mathbf{b}_j$ se encoje por $\mu_j^2 = \cos\boldsymbol{\zeta}_j$ $(0 < \mu_j \leq 1)$, y $\mathbf{b}_j$ es un autovector de $P_{\mathcal{B}}P_{\mathcal{A}}$ de autovalor μ_j^2.

En las aplicaciones, los subespacios $\mathcal{A}$ y $\mathcal{B}$ aparecen a través de una base de cada uno de ellos. Si las bases no son ortonormales, pueden ortonormalizarse usando las técnicas estándares para ello. Por lo tanto, asumamos que disponemos de bases ortonormales $\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$ y $\{\mathbf{r}_1, \ldots, \mathbf{r}_k\}$ de $\mathcal{A}$ y $\mathcal{B}$, respectivamente. Sea $Q_1 = \begin{bmatrix} \mathbf{q}_1 & \cdots & \mathbf{q}_k \end{bmatrix} \in \mathbb{R}^{p\times k}$ y $R_1 = \begin{bmatrix} \mathbf{r}_1 & \cdots & \mathbf{r}_k \end{bmatrix} \in \mathbb{R}^{p\times k}$. Queremos determinar los ángulos y los vectores principales entre los subespacios. Esto es equivalente a determinar las matrices $A_1 = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_k \end{bmatrix}$, $B_1 = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_k \end{bmatrix}$ y $D = \operatorname{diag}(\mu_1, \ldots, \mu_k)$ construidas en el Teorema 9.1.(*i*).

Lema I.1 *Sean $Q_1, A_1 \in \mathbb{R}^{p\times k}$ matrices con columnas linealmente independientes.*

Proposición I.1

i. Sus columnas generan el mismo subespacio, i.e. $\operatorname{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_k\} = \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ *si y sólo si existe una matriz $M_1 \in \mathbb{R}^{k\times k}$ no singular tal que $A_1 = Q_1M_1$.*

ii. Supongamos que $\operatorname{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_k\} = \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ *y que Q_1, A_1 tienen columnas ortonormales. Entonces, M_1 también tiene columnas ortonormales.*

A partir del resultado previo, tenemos $M_1, N_1 \in \mathbb{R}^{k\times k}$ matrices ortogonales (es decir, matrices que cumplen $M_1^{-1} = M_1^T$) tales que

$$A_1 = Q_1M_1 \text{ y } B_1 = R_1N_1.$$

Entonces

$$Q_1^TR_1 = \left(A_1M_1^T\right)^T\left(B_1N_1^T\right) = M_1A_1^TB_1N_1^T = M_1\operatorname{diag}(\mu_1, \ldots, \mu_k)\, N_1^T. \tag{227}$$

donde la última igualdad vale por el Teorema 9.1.(*iv*). Como $M_1, N_1 \in \mathbb{R}^{k\times k}$ son matrices ortogonales y $\operatorname{diag}(\mu_1, \ldots, \mu_k)$ es diagonal, $M_1\operatorname{diag}(\mu_1, \ldots, \mu_k)\, N_1^T$ es la descomposición SVD de $Q_1^TR_1$. Esto nos da una manera de calcular los ángulos y vectores principales.

Calcular

1. $Q_1^TR_1$.
2. La descomposición de ella: $Q_1^TR_1 = M_1\operatorname{diag}(\mu_1, \ldots, \mu_k)\, N_1^T$, donde $\mu_1 \geq \cdots \geq \mu_k$ son los valores singulares.
3. Tomar $\zeta_i = \arccos\mu_i$, $i = 1, \ldots, k$, como ángulos principales.
4. Tomar $A_1 = Q_1M_1$ y $B_1 = R_1N_1$ como vectores principales.

Procedimiento 3. Todavía resta resolver un problema para llevar a cabo estos cálculos, y es que la función de $\zeta = \arccos \mu$ está mal condicionada para μ cerca de uno (y ζ cerca de 0), como puede verse en la Figura 27a. Es decir, pequeñas perturbaciones en μ pueden causar perturbaciones relativamente grandes en ζ. Para valores de μ alejados de uno, $\arccos \mu$ está bien condicionada. Si queremos calcular ángulos principales pequeños con precisión, tenemos que encontrar otro método. Los detalles se pueden encontrar en Watkins [1991], Sección 7.5. Se basa en el siguiente resultado.

Teorema I.1 *Sean $\mathcal{A}$ y $\mathcal{B}$ subespacios $k-$dimensionales de $\mathbb{R}^p$ con ángulos principales $\zeta_1 \leq \zeta_2 \leq \cdots \leq \zeta_k$. Sean $\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$ y $\{\mathbf{q}_{k+1}, \ldots, \mathbf{q}_p\}$ bases ortonormales para $\mathcal{A}$ y $\mathcal{A}^\perp$, respectivamente, y sean $\{\mathbf{r}_1, \ldots, \mathbf{r}_k\}$ y $\{\mathbf{r}_{k+1}, \ldots, \mathbf{r}_p\}$ bases ortonormales para $\mathcal{B}$ y $\mathcal{B}^\perp$, respectivamente. Sean*

$$Q_1 = \begin{bmatrix} \mathbf{q}_1 & \cdots & \mathbf{q}_k \end{bmatrix} \in \mathbb{R}^{p\times k}, \qquad Q_2 = \begin{bmatrix} \mathbf{q}_{k+1} & \cdots & \mathbf{q}_p \end{bmatrix} \in \mathbb{R}^{p\times(p-k)}$$
$$R_1 = \begin{bmatrix} \mathbf{r}_1 & \cdots & \mathbf{r}_k \end{bmatrix} \in \mathbb{R}^{p\times k}, \qquad R_2 = \begin{bmatrix} \mathbf{r}_{k+1} & \cdots & \mathbf{r}_p \end{bmatrix} \in \mathbb{R}^{p\times(p-k)}.$$

Entonces, los valores singulares de $Q_2^T R_1$ y $Q_1^T R_2$ son

$$\operatorname{sen}\zeta_1 \leq \operatorname{sen}\zeta_2 \leq \cdots \leq \operatorname{sen}\zeta_k.$$

Análogamente a lo que comentamos antes para la inversa del coseno, resulta que el arc sen está bien condicionado cuando el ángulo es pequeño y mal condicionado cuando está cerca de $\pi/2$. En la Figura 27b vemos el gráfico de esta función. A partir de esta observación, podemos establecer otro procedimiento para calcular ángulos, que nos permitirá calcular a los pequeños con precisión.

Calcular

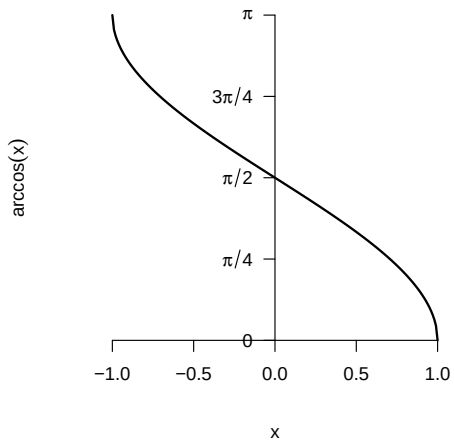
1. $\{\mathbf{q}_{k+1}, \ldots, \mathbf{q}_p\}$ una base ortonormal para $\mathcal{A}^\perp$. Sea $Q_2 = \begin{bmatrix} \mathbf{q}_{k+1} & \cdots & \mathbf{q}_p \end{bmatrix} \in \mathbb{R}^{p\times(p-k)}$.
2. Sea $\{\mathbf{r}_1, \ldots, \mathbf{r}_k\}$ una base ortonormal para $\mathcal{B}$, sea $R_1 = \begin{bmatrix} \mathbf{r}_1 & \cdots & \mathbf{r}_k \end{bmatrix} \in \mathbb{R}^{p\times k}$.
3. Calcular $Q_2^T R_1$.
4. Hallar los valores singulares $\sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_k$ de $Q_2^T R_1$.
5. Tomar $\zeta_i = \operatorname{arc\,sen}\sigma_i$, $i = 1, \ldots, k$, como los ángulos principales.

Observemos que la enumeración dada a los valores singulares es la inversa a la de los procedimientos anteriores porque $\sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_k$ están numerados en orden creciente, ya que la inversa del seno (arc sen) es una función creciente. De esta forma, los ángulos principales quedan ordenados en forma creciente, como se requiere, $\zeta_1 \leq \zeta_2 \leq \cdots \leq \zeta_k$.

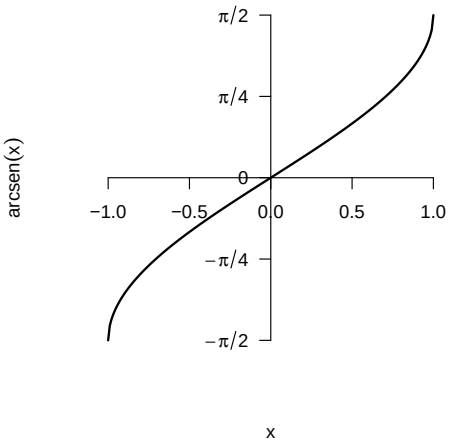
Este procedimiento permite hallar valores precisos de $\boldsymbol{\zeta}_i$ cuando corresponde a ángulos pequeños e imprecisos para los ζ_i que están cercanos a $\pi/2$.

Figura 27: Gráfico de la función $\arccos : [-1,1] \to [0,\pi]$ a la izquierda, y $\operatorname{arc\,sen} : [-1,1] \to \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, a la derecha. En ellos vemos que el arco coseno está mal condicionado para ángulos cercanos a 0 y a π, mientras que el arco seno está mal condicionado para ángulos cercanos a $\pm\frac{\pi}{2}$.

(a) Arco coseno



(b) Arco seno



Referencias

Absil, P.-A., Mahony, R., y Sepulchre, R. (2009). *Optimization algorithms on matrix manifolds*. Princeton University Press.

Anderson, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, *6*(2), 170–176.

Ash, R. B. (1972). *Real analysis and probability*. Probability and Mathematical Statistics. Academic Press, New York.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, *32*(2), 159–188.

Azzalini, A., y Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(2), 367–389.

Azzalini, A., y Capitanio, A. (2014). *The skew-normal and related families. institute of mathematical statistics monographs*. Cambridge University Press, Cambridge.

Azzalini, A., y Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, *83*(4), 715–726.

Basu, D., y Pereira, C. A. B. (1983). Conditional independence in statistics. *Sankhyā: The Indian Journal of Statistics, Series A*, *45*, 324–337.

Billingsley, P. (1968). *Convergence of probability measures*. John Wiley & Sons, Inc., New York.

Boente, G., y Fraiman, R. (1989). Robust nonparametric regression estimation for dependent observations. *The Annals of Statistics*, 1242–1256.

Boente, G., y Martínez, A. (2017). Marginal integration m-estimators for additive models. *TEST*, *26*(2), 231–260.

Bowman, A. W., y Azzalini, A. (2014). R package `sm`: nonparametric smoothing methods (version 2.2-5.4) [Manual de software informático]. University of Glasgow, UK and Università di Padova, Italia. Descargado de `http://www.stats.gla.ac.uk/~adrian/sm,http://azzalini.stat.unipd.it/Booksm`

Bura, E., y Cook, R. D. (2003). Rank estimation in reduced-rank regression. *Journal of Multivariate Analysis*, *87*(1), 159–176.

Bura, E., Duarte, S., y Forzani, L. (2016). Sufficient reductions in regressions with exponential family inverse predictors. *Journal of the American Statistical Association*, *111*(515), 1313–1329.

Bura, E., y Yang, J. (2011). Dimension estimation in sufficient dimension reduction: a unifying approach. *Journal of Multivariate Analysis*, *102*(1), 130–142.

Cook, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. En *Proceedings of the section on physical and engineering sciences* (pp. 18–25).

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, *91*(435), 983–992.

Cook, R. D. (1998). *Regression graphics: Ideas for studying regressions through graphics.* Wiley, New York.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science*, 1–26.

Cook, R. D., y Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statistical Science*, *23*(4), 485–501.

Cook, R. D., Forzani, L., y Tomassi, D. (2011). Ldr: A package for likelihood-based sufficient dimension reduction. *Journal of Statistical Software*, *39*(1), 1–20. Descargado de `https://www.jstatsoft.org/index.php/jss/article/view/v039i03`

Cook, R. D., Li, B., y Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, 927–960.

Cook, R. D., y Ni, L. (2005). Sufficient dimension reduction via inverse regression. *Journal of the American Statistical Association*, *100*(470), 410-428.

Cook, R. D., y Weisberg, S. (1991). Comment. *Journal of the American Statistical Association*, *86*(414), 328–332.

Davies, P. L. (1987). Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 1269–1292.

Durrett, R. (2010). *Probability: theory and examples.* Cambridge university press.

Eaton, M. L. (1983). *Multivariate statistics. a vector space approach.* John Wiley & Sons. Wiley Series in Probability and Mathematical Statistics. Probability and Mathematical Statistics. New York. John Wiley & Sons. XVI.

Edelman, A., Arias, T. A., y Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, *20*(2), 303–353.

Ferrer, J., García, M., y Puerta, F. (1994). Differentiable families of subspaces. *Linear algebra and its applications*, *199*, 229–252.

García Ben, M., Martínez, E., y Yohai, V. J. (2006). Robust estimation for the multivariate linear model based on a τ-scale. *Journal of Multivariate Analysis*, *97*(7), 1600–1622.

Gather, U., Hilker, T., y Becker, C. (2002). A note on outlier sensitivity of sliced inverse regression. *Statistics: A Journal of Theoretical and Applied Statistics*, *36*(4), 271–281.

Golub, G., y Van Loan, C. (2013). *Matrix computations 4th ed.* Johns Hopkins University Press, Baltimore, MD.

Hamm, J., y Lee, D. D. (2008). Grassmann discriminant analysis: a unifying view on subspace-based learning. En *Proceedings of the 25th international conference on machine learning* (pp. 376–383).

Hastie, T., Tibshirani, R., y Friedman, J. (2009). *The elements of statistical learning. Data mining, inference, and prediction. 2nd ed.* (2nd ed. ed.). New York, NY: Springer.

Huber, P. J. (1981). *Robust statistics*. Wiley, New York.

Li, B., y Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, *102*(479), 997–1008.

Li, B., Zha, H., y Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *Annals of statistics*, 1580–1616.

Li, K.-C. (1991a). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, *86*(414), 316–327.

Li, K.-C. (1991b). Sliced inverse regression for dimension reduction: Rejoinder. *Journal of the American Statistical Association*, *86*(414), 337–342.

Li, K.-C. (1992). On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma. *Journal of the American Statistical Association*, *87*(420), 1025–1039.

Lopuhaa, H. P. (1989). On the relation between s-estimators and m-estimators of multivariate location and covariance. *The Annals of Statistics*, 1662–1683.

Lopuhaä, H. P. (1991). Multivariate τ-estimators for location and scatter. *Canadian Journal of Statistics*, *19*(3), 307–321.

Maronna, R., Martin, D., y Yohai, V. (2006). *Robust statistics*. John Wiley & Sons, Chichester. ISBN.

Marshall, A. W., y Olkin, I. (1979). *Inequalities: Theory of majorization and its applications*. Academic Press.

Mosteller, F., y Tukey, J. W. (1968). Data analysis, including statistics. *Handbook of social psychology*, *2*, 80–203.

Muler, N., y Yohai, V. J. (2002). Robust estimates for arch processes. *Journal of Time Series Analysis*, *23*(3), 341–375.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, San Mateo, CA.

Pearl, J. (2009). *Causality*. Cambridge University Press, Cambridge. (Models, reasoning, and inference)

Petersen, K. B., y Pedersen, M. S. (2008). The matrix cookbook, 2008. , *3274*. Descargado de `http://www2.imm.dtu.dk/pubdb/p.php`

Pollard, D. (1984). *Convergence of stochastic processes*. Springer-Verlag, New York.

R Core Team. (2015). R: A language and environment for statistical computing [Manual de software informático]. Vienna, Austria. Descargado de `http://www.R-project.org/`

Rao, R. R. (1962). Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics*, 659–680.

Salibian-Barrera, M., y Yohai, V. J. (2006). A fast algorithm for s-regression estimates. *Journal of Computational and Graphical Statistics*, *15*(2), 414–427.

Scrucca, L. (2011). Model-based sir for dimension reduction. *Computational Statistics & Data Analysis*, *55*(11), 3010–3026.

Seber, G. A. F. (1984). *Multivariate observations*. John Wiley and Sons, New York, New York, USA.

Shorack, G. R., y Wellner, J. A. (2009). *Empirical processes with applications to statistics* (Vol. 59). Siam.

Szretter, M. E., y Yohai, V. J. (2009). The sliced inverse regression algorithm as a maximum likelihood procedure. *Journal of Statistical Planning and Inference*, *139*(10), 3570–3578.

Tatsuoka, K. S., y Tyler, D. E. (2000). On the uniqueness of s-functionals and m-functionals under nonelliptical distributions. *Annals of Statistics*, 1219–1243.

Todorov, V., y Filzmoser, P. (2009). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, *32*(3), 1–47.

Verma, T., y Pearl, J. (1988). Causal networks: Semantics and expressiveness. En *Proceedings of the fourth workshop on uncertainty in artificial intelligence, minneapolis, mn, mountain view, ca* (pp. 352–359).

Verma, T., y Pearl, J. (1990). Causal networks: semantics and expressiveness. En *Uncertainty in artificial intelligence, 4* (Vol. 9, pp. 69–76). North-Holland, Amsterdam.

Wang, H., y Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, *103*(482), 811–821.

Wang, J., Zamar, R., Marazzi, A., Yohai, V., Salibian-Barrera, M., Maronna, R., ... Konis, K. (2014). robust: Robust library [Manual de software informático]. Descargado de `https://CRAN.R-project.org/package=robust` (R package version 0.4-16)

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Watkins, D. S. (1991). *Fundamentals of matrix computations*. New York etc.: John Wiley &| Sons, Inc.

Weisberg, S. (2002). Dimension reduction regression in r. *Journal of Statistical Software*, *7*(1), 1–22.

Wen, X., y Cook, R. D. (2007). Optimal sufficient dimension reduction in regressions with categorical predictors. *Journal of statistical planning and inference*, *137*(6), 1961–1978.

Wong, Y.-C. (1967). Differential geometry of grassmann manifolds. *Proceedings of the National Academy of Sciences of the United States of America*, *57*(3), 589.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.*, *15*(2), 642–656.

Yohai, V. J., Stahel, W. A., y Zamar, R. H. (1991). A procedure for robust estimation and inference in linear regression. En *Directions in robust statistics and diagnostics: Part ii* (pp. 365–374). New York, NY: Springer New York.

Yohai, V. J., y Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American statistical association*, *83*(402), 406–413.

Yohai, V. J., y Zamar, R. H. (1997). Optimal locally robust m-estimates of regression. *Journal of Statistical Planning and Inference*, *64*(2), 309–323.