Universidad Nacional de Rosario

Facultad de Ciencias Económicas y Estadística

Índices Socio-económicos desde el Enfoque de Reducción Suficiente de Dimensiones

Rodrigo García Arancibia

Tesis presentada para obtener el grado de Doctor en Estadística

Directora: Liliana Forzani, Ph.D. (*FIQ, Universidad Nacional del Litoral*)

Co-Director: Diego Tomassi, Dr. (*FICH-FIQ, Universidad Nacional del Litoral*)

Agradecimientos

En primer lugar quisiera agradecer a la Dra. Liliana Forzani, quién realizó una labor más que sobresaliente como directora, compartiendo su conocimiento y experiencia, contagiando su pasión, entusiasmo y energía, estando presente en todo momento para ayudarme. No sólo me enriquecí en lo académico sino también con la amistad que pudimos forjar. Sin su confianza, apoyo y atención, este trabajo no hubiera sido posible.

A mi co-director de tesis, el Dr. Diego Tomassi por su constante predisposición para responder mis consultas, brindándome un apoyo fundamental para la implementación computacional de los métodos de la tesis y aportándome valiosas recomendaciones para la escritura del manuscrito.

Quiero agradecer a la Dra. Pamela Llop por ser una excelente compañera de trabajo, acompañándome en todo el transcurso de la tesis, desde las primeras cuentas hasta la últimas correcciones del manuscrito. Su colaboración fue crucial para que pudiera emprender y terminar este trabajo.

A la Dra. Marta Quaglino por haber atendido siempre mis consultas, realizando una tarea estupenda en la dirección del Doctorado en Estadística de la FCEyE-UNR, permanentemente preocupada en cubrir las necesidades particulares de cada doctorando y procurando que los mismos puedan culminar con éxito la carrera.

Quiero agradecer a la Dra. Edith Depetris Guiguet, mi actual directora de beca Postdoctoral, quién me incentivó a que realice y culmine este doctorado.

Agradezco a mis padres, hermanos y demás familiares que permanentemente me alientan en mis estudios y en mi carrera.

Por último, pero no menos importante, quiero darles las gracias y dedicarles este trabajo a mis amores, Natalia y Cosme, quiénes constante e incondicionalmente me apoyan para seguir haciendo lo que me gusta y apasiona.

Resumen

Los métodos de reducción de dimensiones son utilizados en una gran variedad de aplicaciones en ciencias sociales, biológicas y de la salud. En particular, para la construcción de índices de estatus socio-económico con el fin de clasificar a individuos u hogares y predecir algún fenómeno social de interés resumido en una variable respuesta. En la práctica, los datos contienen una mezcla de variables de diferente naturaleza, como ser continuas, categóricas ordinales y dicotómicas. Por ello, algunos métodos usuales de reducción, sea componentes principales o selección de variables en modelos de regresión, han sido extendidos para contemplar otro tipo de variables además de las continuas. En esta tesis nos proponemos extender el enfoque de Reducción Suficiente de Dimensiones basado en modelos, a problemas de regresión en los que coexisten predictores continuos, ordinales y binarios. Adoptando el enfoque de regresión inversa, en primer lugar abordamos el problema de reducción suficiente de dimensiones para una regresión que involucra sólo predictores categóricos ordinales y una variable respuesta de cualquier naturaleza. Suponiendo la existencia de variables latentes subyacentes a las ordinales, distribuidas normalmente, identificamos una reducción suficiente para la regresión, sin imponer ningún supuesto sobre la distribución condicional de la variable respuesta. Para esta reducción suficiente, proponemos un estimador de máxima verosimilitud, utilizando un algoritmo iterativo tipo EM para hacer factible la estimación en términos computacionales y prácticos. Luego extendemos la metodología para problemas de regresión con predictores continuos, ordinales y dicotómicos. Para ello proponemos una determinada factorización de la densidad conjunta condicionada de los predictores. A partir de dicha factorización, usando un enfoque de variables latentes para los predictores ordinales, un modelo normal para el subconjunto de variables continuas, y un modelo Bernoulli tipo Ising para el subconjunto de variables dicotómicas, identificamos una reducción suficiente. Asimismo obtenemos estimadores de máxima verosimilitud con un método iterativo que combina al procedimiento EM con modelos tipo logit condicionales. Para los métodos propuestos se presentan las correspondiente versiones regularizadas, para realizar conjuntamente selección de variables y reducción de dimensiones. El desempeño de los métodos propuestos se muestran por medio de una serie de simulaciones y de aplicaciones con datos reales para la construcción de índices de estatus socio-económico con fines predictivos. Los resultados son comparados con los arrojados por otros métodos alternativos de reducción de dimensiones, obteniendo conclusiones a favor del uso del enfoque propuesto de reducción suficiente para variables mixtas.

Palabras clave: Subespacio de Reducción Suficiente; Variables Latentes; Familias Exponenciales; Componentes Principales Ajustadas; Algoritmo EM; Modelo Ising; Selección de Variables; índices SES; Encuesta de Hogares.

Índice general

Índice de Figuras

Índice de Tablas

Introducción

En ciencias sociales y de la salud, diferentes medidas del nivel o estatus socio-económico de los agentes son ampliamente utilizadas como predictoras de distintos comportamientos individuales o de resultados sociales (e.g. Roy and Chaudhuri 2009, Murasko 2007, Kamakura and Mazzon 2013, Mazzonna 2014, Feeny et al. 2014). Para los gobiernos y organizaciones no gubernamentales que llevan adelante diferentes políticas o programas de protección social, resulta crucial la predicción del nivel socio-económico de un hogar a través de la observación de variables simples que lo caracterizan (Mokomane 2012, Richardson and Bradshaw 2012). Desde esta manera, los índices socio-económicos (índices *SES*) tienen como objetivo describir fenómenos sociales tales como el ingreso, la pobreza o cuestiones relacionadas con la salud (sintetizados en una variable respuesta Y), utilizando para su construcción variables indirectas (i.e. predictoras $\mathbf{X}$), generalmente de naturaleza categórica, que son más fáciles de obtener y de medir que la variable de interés a predecir Y.

La noción de estatus socio-económico contempla diversas categorías, incluyendo el ingreso monetario, la riqueza (e.g. posesión de activos), el nivel educativo alcanzado, el tipo de empleo u ocupación y otras aspectos que muestran una cierta estratificación acerca de la posición económica o social de individuos, hogares u otro agregado social (Bollen et al. 2001). El ingreso del hogar constituye una variable clave para representar el nivel socio-económico de los hogares, y por ello es usada en los enfoques más tradicionales de análisis de la pobreza. Un claro ejemplo lo constituye el enfoque de la *línea de la pobreza* en base a ingresos relevados de una encuesta de hogares, usado por muchos países para inferir la situación socio-económica de la población (Mokomane 2012, Richardson and Bradshaw 2012).

Sin embargo, el ingreso monetario presenta varios problemas de captación en términos de disponibilidad y credibilidad (Vyas and Kumaranayake 2006, Doocy and Burnham 2006). Por

esta razón, comúnmente se elabora un índice de estatus socio-económico que busca ser una *proxy* o predictor del ingreso, y está basado en variables que resultan más fáciles de observar y captar, tales como los activos del hogar (e.g. TV, radio, transporte), las condiciones habitacionales (e.g. materiales de los techos, pisos y paredes de la vivienda) y otras variables que caracterizan socialmente a los miembros, como su escolaridad y ocupación. La técnica más utilizada para la elaboración de este tipo de índice (comúnmente denominado índice SES: *Socio-Economic Status index*) es la de Componentes Principales (PCA) (Merola and Baulch 2014, Hoque 2014), proponiéndose recientemente algunas extensiones para variables categóricas ordinales (Kolenikov and Angeles 2009).

A pesar del uso extensivo de PCA en la aplicaciones mencionadas, tal método no explota toda la información contenida en las datos de entrenamiento, sea para predecir una variable de interés o bien para estimar el efecto marginal del nivel socio-económico sobre una cierta respuesta que se busca explicar. Más precisamente, dicho método no contempla la existencia de alguna variable respuesta de interés Y (por ejemplo ingreso, pobreza monetaria, fertilidad, consumo, etcétera), perdiendo así información relevante para los fines predictivos. Es decir que si se desea predecir el *ingreso* de un hogar (i.e. ingreso como variable respuesta) utilizando como predictoras un conjunto de variables *proxy* observables y se cuenta con una muestra o sub-muestra de hogares con sus respectivos ingresos monetarios, el índice SES derivado de la reducción vía PCA puede utilizarse para predecir el ingreso de un hogar fuera de dicha muestra pero no utilizará la información del ingreso relevado por los hogares contenidos en la misma. El enfoque de Reducción Suficiente de Dimensiones (SDR), al igual que PCA, busca reducir el espacio de covariables o predictoras $\mathbf{X}$ pero a diferencia de PCA utiliza información de la variable respuesta que se está modelando. Específicamente, para un vector de p covariables $\mathbf{X} \in \mathbb{R}^p$ y una variable respuesta Y, SDR busca una reducción $\mathbf{R}(\mathbf{X}) \in \mathbb{R}^d$, $d \leq p$, de forma tal que $Y|\mathbf{X} =_d Y|\mathbf{R}(\mathbf{X})$, donde $Y|\mathbf{X}$ denota la distribución condicional de Y dado $\mathbf{X}$, y '$=_d$' denota la igualdad en distribución. Gran parte de la literatura metodológica asociada a SDR está basada en el enfoque de *regresión inversa* (i.e. $\mathbf{X}|Y$) desarrollado por Cook y varios coautores (e.g. Cook 1998*b*, Cook and Weisberg 1991*b*, Cook 1994, 1998*a*, 2007, Cook and Lee 1999, Bura and Cook 2001*a*, Cook and Yin 2001, Chiaromonte et al. 2002, Cook and Ni 2005*b*, Cook and Forzani 2008, 2009).

Los métodos de estimación de la reducción $\mathbf{R}(\mathbf{X})$ en el marco de SDR pueden clasificarse

básicamente en aquellos basados en los momentos de la distribución condicional de $\mathbf{X}|Y$ (e.g. Li 1991*b*, Cook and Weisberg 1991*b*, Li 1992*a*, Bura and Cook 2001*a*, Xia et al. 2002*b*, Li et al. 2005*a*, Cook and Ni 2005*b*, Zhu and Zeng 2006*a*, Cook and Li 2002*b*, Li and Wang 2007) y aquellos basados en modelos para la regresión inversa $\mathbf{X}|Y$ (Cook 2007, Cook and Forzani 2008, 2009). Estas metodologías han sido desarrolladas originalmente en problemas de regresión que involucran predictores continuos. Sin embargo, dadas las características de las variables comúnmente disponibles en las bases de datos sociales, resulta necesario extender la metodología de SDR a otros tipos de variables, en particular a variables categóricas.

Una extensión que puede aplicarse para variables dicotómicas o multinomiales (tipo Bernoulli) ha sido recientemente desarrollada en el marco de *modelos lineales generalizados* por Bura et al. (2015). Sin embargo para la aplicación de construcción de índices, gran parte de las variables usadas tienen una naturaleza ordinal, no pudiendo encuadrarse su distribución dentro de la familia exponencial para la aplicación de tal extensión.

La presente tesis busca contribuir a la literatura existente extendiendo los métodos de estimación de SDR para el caso de variables categóricas ordinales. A su vez, dado que en general puede presentarse una mezcla de variables continuas, ordinales y dicotómicas, se propondrá un estimador de la reducción que admita una combinación de estos tipos de variables, usando el método que se desarrollará en primer lugar para las ordinales y la metodología ya propuesta en Bura et al. (2015).

Por lo tanto, el objetivo general de la presente investigación doctoral es extender los métodos de reducción suficiente de dimensiones para su aplicación a la construcción de índices de estatus socio-económico con el objeto de predecir una variable respuesta de interés cuando tenemos predictores de naturaleza heterogénea.

Los objetivos específicos son:

1. Desarrollar métodos de reducción suficiente para modelos de regresión que involucran covariables categóricas ordinales, tomando como respuesta tanto variables continuas como categóricas.

2. Identificar la reducción suficiente de dimensiones en problemas de regresión con predictores de naturaleza mixta (continuas, ordinales y binarias).

3. Extender métodos de estimación de la reducción suficiente que permitan combinar predictores continuos, ordinales y dicotómicos.

4. Aplicar estas metodologías a la construcción de índices de estatus socio-económico para modelos que involucran diferentes tipos de respuestas, como ser ingreso per cápita para respuesta continua y condición de pobreza, como respuesta discreta.

5. Evaluar el poder explicativo y predictivo de los indicadores de estatus socio-económico construidos con las metodología desarrollada, comparando los resultados con métodos clásicos de reducción utilizados actualmente para la construcción de indicadores.

La tesis se organiza de la siguiente manera: En el Capítulo 1 se expone el problema de construcción de índices de estatus socio-económico, la definición de índice utilizada y una revisión de los métodos actuales usados en la literatura y práctica reciente. Posteriormente, en el Capítulo 2 se realiza una revisión de antecedentes sobre Reducción Suficiente de Dimensiones, exponiendo algunas definiciones y resultados que constituyen la base para las extensiones propuestas en la presente tesis. En el Capítulo 3 se presenta un modelo para estudiar reducción suficiente para el caso de predictores ordinales, proponiendo un estimador de máxima verosimilitud, y mostrando algunas simulaciones para evaluar el método, y su aplicación con datos reales. El Capítulo 4 presenta la extensión para estudiar reducción suficiente en problemas de regresión en los que coexisten predictores continuos, ordinales y binarios, evaluando el método con simulaciones y con una aplicación para la construcción de índices. Se finaliza la tesis con trabajos a futuro en Capítulo 5.

Capítulo 1

Índices de Estatus Socio-económico

1.1. Introducción

En ciencias sociales existe un trasfondo tanto teórico como metodológico, en donde una serie de conceptos que surgen de un amplio abanico de teorías económicas, sociológicas, éticas y políticas, se traducen en cuantificar un cierto fenómeno social de interés a través diferentes metodologías y técnicas estadísticas. Es por ello que en este primer capítulo de la tesis, abordamos el problema de medición, a partir de detallar algunos conceptos, enfoques y definiciones que hacen a la motivación central de la tesis; esto es, la construcción de índices socio-económicos a partir de una nueva propuesta estadística.

Además de introducir algunos conceptos generales, vamos a dar la definición precisa de índices que adoptaremos a lo largo de la tesis, y que será objeto de las aplicaciones posteriores. También expondremos el alcance y las principales limitaciones de las metodologías estándares para la construcción de índices que constituyen las metodologías más usuales adoptadas por muchos gobiernos del mundo y organismos internacionales, para medir pobreza, y diseñar políticas y programas para diagnosticarla.

La construcción de índices de estatus socio-económico pueden tener varias motivaciones en la aplicación. Entre las más usuales podemos nombrar:

(i) Necesidad de obtener simples mediciones para describir resumidamente la realidad social y dar lineamientos para políticas de mediano y largo plazo.

(ii) Conocer la asociación o el impacto que tiene el estatus socio-económico sobre otros resultados sociales, tales como la salud o el desempeño escolar de los niños (e.g. índice como covariable en modelos de regresión).

(iii) Construcción de índices con fines predictivos de diagnóstico para su uso en política focalizadas de detección de pobreza. Ésta es la motivación para la construcción de índices de estatus socio-económico en esta tesis. Un ejemplo de ello lo constituyen aquellos programas focalizados de reducción de la pobreza llevados a cabo por gobiernos y ONGs, en los que se busca clasificar a hogares o individuos entre diferentes grupos socio-económicos a fin de facilitar una ayuda focalizada en aquellos hogares más vulnerables. Este tipo de ayuda suele definirse vía un índice focalizado (comúnmente denominado *Índice de Focalización de Pobreza*) y ha sido utilizado para implementar varios programas de reducción de pobreza (e.g., el CAS en Chile, Sisben en Colombia, SISFOH en Perú, Tekoporá en Paraguay, SIERP en Honduras, y PANES en Uruguay, entre otros). En general, el interés está en la pobreza medida a partir del ingreso (medición directa), por ello dichos programas suelen instrumentarse vía transferencias monetarias o de bienes materiales no asequibles por algunos hogares debido a limitaciones que se derivan de no poder percibir un cierto nivel de ingreso. Por lo tanto es necesario conocer el monto del ingreso percibido por el hogar al momento de evaluar al potencial beneficiario. Sin embargo, al basarse en un auto-reporte, el ingreso declarado por el potencial beneficiario muy probablemente esté sesgado debido a los incentivos que existen en torno a la ayuda económica y su relación con lo declarado (Doocy and Burnham 2006). Por ende, el objetivo aquí es predecir el ingreso o la pobreza (monetaria) de los hogares utilizando un conjunto de variables más fáciles y rápido de recolectar que el propio ingreso, a fin de agilizar el programa. Sin embargo, si para una muestra de entrenamiento se recolectan los ingresos, luego la información contenida en el ingreso de esta muestra puede utilizarse para construir el índice de estatus socio-económico (SES) que prediga la pobreza a partir de predictores confiables de obtener, esperando con ello obtener un mejor poder predictivo, y con ello una política social más eficiente.

Cabe destacar que tanto para los puntos (i) como (ii), con fines más descriptivos que predictivos, la metodología de reducción suficiente presentada en esta tesis puede aportar un nuevo enfoque para índices socio-económicos.

Lo que resta del presente capítulo se organiza de la siguiente manera. En la Sección 1.2 se esbozan los conceptos de estatus socio-económicos y pobreza. Luego en la Sección 1.3 se presenta una definición formal de índices que adoptaremos en la tesis, exponiendo asimismo la metodología estadística que corrientemente se usa para estimar los parámetros (ponderaciones) que involucra dicha definición de índices. Por último, en la Sección 1.4 se presenta la forma en que vamos a concebir la construcción de índices bajo el enfoque de reducción suficiente de dimensiones.

1.2. Marco Conceptual

El estudio de los estándares de vida de la población constituye un objetivo en común de la Economía Política desde Adam Smith hasta nuestros días (Sen 1984). Sin embargo, a lo largo de la historia han existido diferentes concepciones respecto a lo que representan las nociones de estándar de vida, bienestar y pobreza. Dichas concepciones resultan cruciales a la hora de pensar en cómo medir el bienestar de una sociedad, a lo efectos de contar con un *termómetro* social que constituya una guía para diseñar e implementar medidas de política, directas e indirectas, que tengan como fin el desarrollo económico y social.

Por lo tanto, antes de tratar el problema de la medición, conviene realizar un repaso sobre un par de conceptos estrechamente relacionados, y enmarcados en la noción general de estándares de vida y bienestar. En particular, vamos a focalizarnos en dos conceptos ligados entre sí: el de Estatus Socio-económico y el de Pobreza. Al mismo tiempo, al momento en que tratamos de precisar qué entendemos por estatus socio-económico o pobreza, van apareciendo las variables relativas a dichos conceptos, de forma tal que de una manera progresiva se va arribando al problema de la medición.

Básicamente, la forma en que estructuramos esta presentación, y la idea de presentar los significados de estatus socio-económico y pobreza de acuerdo a la literatura predominante, tiene que ver con la meta de la tesis y con las aplicaciones que se realizarán con la metodología estadística propuesta. Puntualmente, el objetivo de la aplicación es construir un índice de estatus socio-económico (de aquí en adelante, índice SES: *Socio-Economic Status*) para poder predecir ingreso y pobreza de los hogares utilizando una muestra de entrenamiento. Por ende, debemos tener en claro qué conceptos sociológicos y económicos existen para fundamentarlos,

a los efectos de saber qué definición operativa se adopta y qué es lo que precisamente se está midiendo al momento en que se construye un índice.

1.2.1. Estatus Socio-económico

No existe una definición cerrada, única y predominante de estatus socio-económico, sino que el mismo queda conceptualizado de una forma unitaria (en el sentido de la unidad del concepto), por una serie de categorías sociales y económicas englobadas. Siguiendo a Bollen et al. (2001), el estatus socio-económico se refiere a la posición de los individuos, familias, hogares u otro agregado, respecto a una o varias categorías conceptuales de estratificación social. Tales categorías incluyen el ingreso, el nivel educativo, el prestigio, el nivel cultural, el tipo de ocupación, u otros aspectos que los miembros de una sociedad determinada consideran relevantes.

A pesar de que esta perspectiva teórica general incluye varias categorías, existen concepciones sobre estatus socio-económico que son esencialmente unicategóricas (conceptualmente) y más aún, univariadas (i.e. factibles de representar con una sola variable), cuya ventaja principal es la de brindar una interpretación más parsimoniosa del concepto. En esta línea, se encuentra el poder adquisitivo para representar el estándar de vida material (O'Donnell et al. 2007), cuya variable de asociación directa es el ingreso, y sin dudas representa la primera categoría (y variable) que es tenida en cuenta en la definición de estatus socio-económico para el análisis de pobreza y bienestar. Desde el punto de vista teórico, la elección del poder adquisitivo como categoría conceptual crucial, encuentra su fundamento en la influyente hipótesis del ingreso permanente de Friedman (1957), donde se pone el acento en la categoría más económica. Pero la noción de ingreso de Friedman es más compleja (y completa) que el ingreso corriente que usualmente se reporta las encuestas de hogares, al incluir en el mismo aspectos más relacionados con la riqueza.

En general, ninguna de las categorías conceptuales quedan directamente representadas por determinadas variables, sino que en su mayoría deben ser operacionalizadas vía variables *proxy* para su representación cuantificable. Aún para captar la categoría más económica, muchos trabajos recurren a la medición del SES a través de los activos del hogar o de variables relacionadas al nivel educativo y a la situación ocupacional de los miembros del hogar (Bollen et al. 2001,

Fernald 2007). Mientras que el primer tipo de variables se asocia a la noción de ingreso permanente (y riqueza), las últimas dos también pueden representar otras categorías del estatus socio-económico asociadas a la calidad de vida, el prestigio y el nivel cultural, entre otras (Sen 1999). Por su parte, Angeles and You (2007) analizando una de las principales encuestas de hogares con una gran cobertura mundial (*The DHS Program: Demographic and Health Surveys*), muestran que en la mayoría de los países el estatus socio-económico queda representado por variables que hacen referencia a la posesión de activos y bienes durables, y a la calidad de la vivienda. Por otro lado, Mazzonna (2014) incluye el número de libros que existen en la vivienda además de algunas características demográficas y de la ocupación del jefe/a de hogar, a los efectos de captar el nivel cultural en el SES.

Por ende, podemos afirmar que conceptualmente el estatus socio-económico comprende varias categorías, mas allá de la puramente económica que es representada por el ingreso como poder adquisitivo para alcanzar cierto bienestar material. A su vez, tanto el poder adquisitivo como las otras categorías que caracterizan al estatus socio-económico, terminan siendo operativas por medio de variables *proxy* que las representan; y la elección de las mismas para la tarea de medición empírica, muchas veces queda supeditada a la disponibilidad que exista en la base de datos, más que a la adopción de una postura teórica específica.

1.2.2. Pobreza

El concepto de pobreza está muy ligado al de estatus socio-económico (SES), en el sentido de que un SES bajo está asociado a privaciones de bienes materiales, culturales y de derechos sociales (Sen 1999). Desde una perspectiva más estándar, la pobreza en general se define por las privaciones en el bienestar que resultan de la incapacidad de cubrir las necesidad básicas individuales o familiares (World Bank 2000). Claramente, la categoría a la que refiere tal concepción está asociada al ingreso o consumo necesario para cubrir ciertas necesidades básicas (alimentarias y no), y es lo que precisamente define la comúnmente llamada *línea de pobreza*, computada a partir de estimar el costo de una canasta básica alimentaria (según necesidades calóricas) y de vestimenta, vivienda, entre otros bienes de consumo considerados necesarios. Este concepto de pobreza apunta específicamente a la denominada *pobreza monetaria o de ingresos* (Wang et al. 2016).

Amarya Sen es quien introduce el concepto de pobreza multidimensional, postulando que la pobreza, y con ella el bienestar general, no puede ser representada por una sola categoría, derivada de la falta de ingresos o de consumo, sino también a partir de privaciones de las capacidades humanas (e.g. Sen 1984, 1993, 1999). Con esta concepción, parte del bienestar tiene que ver con lo monetario por lo que puede representarse con el ingreso para cubrir las condiciones básicas de vida, pero también existe un dominio no monetario expresado en la exclusión social, entendida a partir de privaciones en educación, salud y vivienda que surgen de un difícil contexto social (Wang et al. 2016). Es por ello que en la última década se ha difundido el uso de índices de pobreza multidimensionales basados en este enfoque de las capacidades, buscando en este proceso, las categorías relevantes que caracterizan el nivel de pobreza y bienestar (e.g. Alkire and Foster 2011, Alkire and Santos 2014, Caruso et al. 2015, Wang et al. 2016, entre otros).

Asumiendo el carácter multidimensional de la pobreza, resulta razonable esperar una relación directa entre la pobreza de ingresos y la pobreza multidimensional. Wang et al. (2016) muestran cómo un incremento en el ingreso reduce significativamente la pobreza multidimencional en cada categoría incorporada, aunque dicho impacto es limitado, por lo que hay información adicional en las medidas de pobreza multidimensional que no logra ser captada por la medida de ingresos.

Por lo tanto, al igual que con el estatus socio-económico, en el concepto de pobreza, el ingreso resulta ser una categoría muy relevante, aunque no llega a abarcar una concepción más general, como ser la del enfoque de capacidades.

En la presente tesis, sólo nos concentraremos en la pobreza medida a partir del ingreso monetario corriente (i.e. considerando la línea de pobreza) a la que tomaremos como variable respuesta, mientras que para el SES (tomada como predictor) contemplaremos varias categorías, desde las privaciones habitacionales, hasta la educación y situación laboral de los miembros del hogar. Esta elección, lejos de ser arbitraria, tienen que ver con el fin de la aplicación planteada sobre el uso de índices SES, como detallamos la siguiente sección.

1.3. Índices

En términos generales, los índices socio-económicos tienen como objetivo sintetizar o resumir cuantitativamente algún fenómeno social o económico de interés (e.g. inflación, desempleo, crecimiento, desarrollo, distribución o pobreza, entre otros). Por lo tanto, para el caso del estatus socio-económico (SES) se busca un índice que capte sus categorías conceptuales a partir de un conjunto de variables.

En el siguiente apartado se presentan unas definiciones formales de índice SES, que son las adoptadas comúnmente en este contexto, como así también en las aplicaciones de la presente tesis.

1.3.1. Definición General

Supongamos que tenemos un conjunto de p variables tal que operacionalizan las categorías consideradas dentro del concepto de SES que deseamos abordar. La siguiente definición, formaliza la idea de un índice unitario (en relación al concepto y a la medida), pero multivariado y multicategórico (en términos conceptuales).

Definición 1.1 *Sean $X_1, \ldots, X_p$ diferentes variables que caracterizan al hogar en términos económicos y sociales, se define por índice de estatus socio-económico (SES) a*

$$I_{SES} = \varphi(X_1, \ldots, X_p)$$

donde $\varphi : \mathbb{R}^p \longrightarrow \mathbb{R}$ es una función conocida de las variables socio-económicas.

Esta definición supone que la función φ es conocida. La metodología más estándar para el cómputo de índices SES, es asumir que φ es lineal, de forma tal que el índice se construye simplemente como una suma ponderada de las p variables. De esta manera, tenemos la siguiente definición:

Definición 1.2 *Sean $X_1, \ldots, X_p$ un conjunto de variables que caracterizan al hogar en términos económicos y sociales, el índice de estatus socio-económico (SES)* **lineal** *está dado por la*

combinación lineal de dichas variables para un determinado vector de pesos fijos $(\omega_1, \ldots, \omega_p)$, *i.e.*

$$I_{SES} = \omega_1 X_1 + \ldots + \omega_p X_p$$

Con esta definición, el problema de medición se traduce en encontrar los valores de ω_j. Estos coeficientes *a priori* no se conocen, por lo que deben ser estimados. Por ejemplo, en el marco de políticas focalizadas de pobreza, con una muestra (denominada de entrenamiento) debería estimar $(\omega_1, \ldots, \omega_p)$ y que me de un buen índice I_{SES} para implementar dicha política de manera coherente y eficiente.

Existen enfoques, comúnmente denominados *naive*, donde la elección de los ω_j es arbitraria, o bien el indicador es construido como la suma de variables dummy que indican si el individuo u hogar posee cierto activo (O'Donnell et al. 2007). Para superar la arbitrariedad de estos enfoques *naive*, los metodologías estadísticas de análisis factorial y componentes principales (PCA) se han utilizado profusamente en la literatura empírica, prevaleciendo aún en la actualidad (Merola and Baulch 2014, Hoque 2014).

1.3.2. Índice SES como Componente Principal

Dada la Definición 1.2 de índice SES, Filmer and Pritchett (1998, 2001) proponen el uso de Componentes Principales (PCA) para estimar las ponderaciones ω_j. De esta manera, el índice SES queda definido como la combinación lineal de las variables originales utilizando la primer componente del PCA.

Específicamente, para las variables originales $\mathbf{X} = (X_1, \ldots, X_p)^T$, PCA define un nuevo conjunto de k variables $(k \leq p)$, $(P_1, \ldots, P_k)$ no correlacionadas y con varianza decreciente, dadas por $P_j = \mathbf{a}_j^T \mathbf{X}$ siendo $\mathbf{a}_j$ el j-ésimo autovector de la matriz de covarianza de $\mathbf{X}$, ordenados de acuerdo al j-ésimo autovalor, en forma decreciente. Para el caso de índices SES se toma el autovalor corresponiente al mayor autovalor, y se construye

$$I_{SES} \equiv P_1 = a_{11} X_1 + a_{12} X_2 + \cdots + a_{1p} X_p,$$

donde $\mathbf{a}_1 = (a_{11}, a_{12}, \ldots, a_{1p})^T$ es el autovector correspondiente al mayor autovalor $\lambda_1 = Var(P_1)$. De esta manera queda definido el índice SES vía componentes principales.

Comparado con la elección arbitraria de la metodología *naive*, PCA resulta ser un criterio más razonable para la obtención de los pesos ω_j del I_{SES}: logramos reducir la información de las $\mathbf{X}$ en una sola dimensión maximizando la varianza (información contenida en ellas), y la componente seleccionada como ponderador no está correlacionada con las componentes restantes. A su vez, posee la ventajas de ser fácil de aplicar y comprender. Por ello, muchos organismos internacionales como el Banco Mundial y el programa DHS han adoptado esta técnica para la construcción de índices de estatus socio-económico. También es lo que ha prevalecido para el análisis empírico en ciencias sociales y de la salud, cuando se estudia el rol del SES sobre algun fenómeno en particular cpara la construcción de índices SES, la metodología del PCA es lo que ha prevalecido en los últimos veinte años.

A pesar de ser una metodología parsimoniosa para la construcción del índice SES, la metodología del PCA descansa sobre el supuesto de que $\mathbf{X}$ es normal multivariada, por lo que su aplicación resulta más acorde para datos continuos. Sin embargo, una de las características de las bases de datos sociales de individuos y/u hogares, es que la mayoría de las variables recolectadas son categóricas. El enfoque de Filmer and Pritchett (1998, 2001) tiene una motivación más practica, a los efectos de tener un indicador del poder adquisitivo o de la riqueza, en ausencia de datos de ingreso o gastos de consumo, aplicando PCA estándar independientemente de cómo sea la naturaleza de las variables incluidas. Para remediar esta propuesta *ad hoc* a este problema de medición, Kolenikov and Angeles (2009) proponen una extensión de PCA para datos ordinales, en vista de que la mayoría de las variables socio-económicas son categóricas y, por lo general (aunque no siempre), tienen un orden natural respecto al estatus socio-económico. Por ejemplo, la posesión de activos del hogar (TV, radio, internet, lavarropas, etcétera), en general se imputan con variables binarias (i.e. si posee o no posee el bien) y estas tienen un orden natural respecto al bienestar material del hogar (i.e. si tiene el bien, mayor SES). Otro ejemplo lo constituyen las variables relacionadas con los materiales de la vivienda (pisos, techo o paredes), que generalmente son capturadas con variables politómicas, con un orden que va desde los materiales más precarios a aquellos de mejor calidad.

Básicamente, la solución de Kolenikov and Angeles (2009) consiste en utilizar correlaciones policóricas entre las variables contenidas en $\mathbf{X}$ en lugar de la matriz de covarianza estándar correspondiente. Linting and van der Kooij (2009) proponen extensión más flexible del PCA que contempla la existencia conjunta de variables continuas y categóricas, como así también la

posibilidad de relaciones no lineales entre las variables. Básicamente este método consiste en re-escalar las variables originales, de forma que las variables transformadas poseen un dominio numérico continuo. Tal re-escalamiento busca ser óptimo, en el sentido de que la varianza de las variables transformadas, para un número de dado de componentes, sea maximizada.

Estas metodologías basadas en componentes principales, comparten una propiedad común, y es que no contemplan la existencia de alguna variable respuesta, perdiendo así información relevante respecto al fenómeno social particular que el índice busca explicar o predecir. El enfoque de reducción suficiente de dimensiones se propone como una metodología superadora en este sentido.

1.4. Índices bajo el enfoque de Reducción Suficiente de Dimensiones

En esta sección vamos ilustrar el interés práctico o aplicado del enfoque de Reducción Suficiente de Dimensiones (SDR: *Sufficient Dimension Reduction*) en el contexto de construcción de índices SES. No entraremos en el detalle mas formal sobre este enfoque, puesto que a ello se aboca el Capítulo 2. Aquí queremos dejar sentado cómo el enfoque de SDR constituye una forma novedosa de concebir la tarea de medición social, donde la existencia de una variable considerada como respuesta juega un rol crucial respecto a la información que se extraiga del conjunto de variables utilizadas para la elaboración de un índice.

Como antes, supongamos que tenemos un conjunto de p variables $\mathbf{X}$ relacionadas a las características socio-económicas de hogares que queremos reducir a un índice con el fin de predecir una cierta variable respuesta Y, para la cual tenemos datos para una cierto sub-conjunto de hogares que conforman la denominada muestra de entrenamiento. Usando el enfoque de SDR es posible construir un índice SES que utilice la información de $\mathbf{X}$ que es relevante para Y. De esta manera, se obtendría un índice de estatus socio-económico funcional a lo que se desea explicar o predecir. Ergo, podemos clasificar a los individuos u hogares no sólo por su similaridad respecto las variables incluidas para medir el SES sino también por la contribución o influencia del SES sobre la variable que deseamos conocer.

Suponer que existe una variable respuesta de interés, constituye más una regla que una excepción en las aplicaciones de índices SES. En general el mismo es incorporado como predictor para explicar o predecir un cierto fenómeno social, tal como el índice de masa corporal y la obesidad infantil (Murasko 2007, Fernald 2007), la fertilidad (Bollen et al. 2001, Kolenikov and Angeles 2009), la estructura familiar (Mokomane 2012), el desempeño cognitivo y económico en edad adulta (Mazzonna 2014), el estado de salud en general (O'Donnell et al. 2007), el nivel de consumo (Kamakura and Mazzon 2013), y hasta el nivel de felicidad (Feeny et al. 2014), entre otros.

Para el caso de los programas focalizados de pobreza, la variable respuesta es el ingreso o la condición de pobreza basada en ingresos. De esta manera, si existe una muestra de entrenamiento donde se revelaron los ingresos, además de las otras variables predictoras consideradas para el SES, el enfoque de SDR usará esta información del ingreso para la construcción del índice. Es decir que SDR ponderará cada variable en función de la relevancia que tiene sobre la respuesta Y. Dada esta mayor información incorporada en la construcción del índice, se esperaría obtener mejores resultados predictivos, y con ello una ejecución mas eficiente del programa. Sobre el escenario planteado por este ejemplo, se mostrarán las aplicaciones del método de SDR propuesto, y se comparará con las otras técnicas usuales de reducción de dimensiones, a fin de mostrar el desempeño comparativo de esta metodología. No obstante debe quedar claro que, dado el abanico de posibilidades para la variable respuesta, la aplicación de índices SES basados en SDR no se agota en este ejemplo, sino que consideramos que el enfoque de SDR tiene un gran potencial para aplicaciones futuras sobre índices y medición de la pobreza.

Capítulo 2

Reducción Suficiente de Dimensiones: Antecedentes

2.1. Introducción

El presente capítulo constituye la base sobre el cual se desarrollan las extensiones propuestas en la presente tesis. Específicamente, se brindarán algunas definiciones y resultados básicos dentro del campo conocido como Reducción Suficiente de Dimensiones (SDR: *Sufficient Dimension Reduction*), realizando una breve revisión de antecedentes sobre enfoques y metodologías. Lejos de ser exhaustiva, la exposición está pensada para que los desarrollos presentados en los dos próximos capítulos tengan un punto de apoyo en común, y que el aporte al conocimiento de las extensiones de la tesis resulte más directo de captar. Por ende, los antecedentes presentados están acotados a lo necesario para desarrollar los modelos y las metodologías propuestas en la tesis. Por lo tanto, muchos tópicos dentro del área de SDR no son tratados; no por su relevancia, sino por su menor conexión directa con los temas de la tesis. Para un análisis más exhaustivo y detallado sobre SDR puede consultarse Cook (2007, 1998*b*), y para una introducción de reducción de dimensiones en general, puede verse Burges (2009).

2.2. Reducción Suficiente de Dimensiones

2.2.1. Definiciones Básicas

La idea básica de reducción suficiente de dimensiones (SDR) en un problema de regresión de una variable respuesta $Y \in \mathbb{R}$ sobre un conjunto de p predictores $\mathbf{X} \in \mathbb{R}^p$, es encontrar una función $\mathbf{R}(\mathbf{X}) \in \mathbb{R}^d$, con $d \leq p$, que contenga toda la información de $\mathbf{X}$ que es relevante para la respuesta Y. De Cook (2007), podemos dar la siguiente definición formal para SDR.

Definición 2.1 *Una reducción* $\mathbf{R} : \mathbb{R}^p \longrightarrow \mathbb{R}^d$ *con* $d \leq p$ *es* **suficiente** *para la regresión de* Y *sobre* $\mathbf{X}$ *si satisface*

$$Y|\mathbf{X} \stackrel{d}{=} Y|\mathbf{R}(\mathbf{X}).$$

En la literatura de SDR suele presentarse el concepto de reducción suficiente a partir de definiciones equivalentes, que así mismo permiten enfocar el problema de reducción suficiente desde diferentes ángulos. Tales equivalencias se presentan en el siguiente lema.

Lema 2.1 *Para el vector de predictores* $\mathbf{X} \in \mathbb{R}^p$ *y la variable respuesta* $Y \in \mathbb{R}$*, si* $\mathbf{R}(\mathbf{X})$ *es una función medible en el conjunto de predictores, entonces tenemos que las siguientes afirmaciones son equivalentes:*

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

(ii) $Y|\mathbf{X} \stackrel{d}{=} Y|\mathbf{R}(\mathbf{X})$.

(iii) $\mathbf{X}|(\mathbf{R}(\mathbf{X}), Y) \stackrel{d}{=} \mathbf{X}|\mathbf{R}(\mathbf{X})$.

Mientras que la definición (ii) enfoca el problema de SDR a partir de la regresión original de interés, la definición (iii) enfoca el problema a partir de la regresión inversa, i.e. $\mathbf{X}|Y$; y la definición (i) refleja de mejor manera el sentido coloquial de SDR, en el sentido de que muestra que la información contenida en $\mathbf{X}$ que no está contenida en $\mathbf{R}(\mathbf{X})$, no aporta información a la variable respuesta. Mas allá de este sentido intuitivo, metodológicamente el enfoque con mayor difusión es el de regresión inversa. La principal motivación de ello radica en las ventajas de visualización de la dependencia de los datos. En lugar de analizar una variable respuesta sobre

p covariables (con $p > 2$, ya el análisis gráfico queda descartado), el enfoque de regresión inversa conlleva a analizar p regresiones con una sola covariable, lo cual facilita la modelización a partir de una primera inspección gráfica. Además permite pensar a la reducción $\mathbf{R}(\mathbf{X})$ sin necesidad de imponer un modelo *a priori* para $Y|\mathbf{X}$.

Otra definición que necesitamos introducir es la de minimalidad.

Definición 2.2 *Un reducción suficiente* $\mathbf{R}(\mathbf{X})$ *para la regresión de* Y *sobre* $\mathbf{X}$ *es* **minimal**, *si para toda otra reducción suficiente* $\mathbf{U}(\mathbf{X})$ *para la regresión* $Y|\mathbf{X}$, $\mathbf{R}(\mathbf{X})$ *es función de* $\mathbf{U}(\mathbf{X})$.

Al igual que la definición de suficiencia, la de minimalidad está íntimamente conectada con la de estadístico minimal suficiente, como también el lema que presentaremos a continuación para encontrar reducciones minimales suficientes. Basados en el enfoque de regresión inversa, si tenemos un modelo para la regresión $\mathbf{X}|Y$ podemos utilizar una extension del Lema de Lehmann-Sheffé (ver Casella and Berger 2002) para encontrar una reducción suficiente minimal.

Lema 2.2 *Sea la distribución condicional de* $\mathbf{X}|Y$ *tal que su función de densidad o de probabilidad puntual viene dada por* $f_{\mathbf{X}|Y,\theta}$, *siendo* $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$ *el parámetro de la distribución correspondiente, luego* $\mathbf{R}(\mathbf{X})$ *es una reducción suficiente minimal para* $Y|\mathbf{X}$ *si las siguientes afirmaciones son equivalentes:*

(i) Para dos puntos muestrales cualesquiera $\mathbf{x}$ *y* $\mathbf{z}$, $\dfrac{f_{\mathbf{X}|Y,\theta}(\mathbf{x}|y,\boldsymbol{\theta})}{f_{\mathbf{X}|Y,\theta}(\mathbf{z}|y,\boldsymbol{\theta})}$ *no depende de* y.

(ii) $\mathbf{R}(\mathbf{x}) = \mathbf{R}(\mathbf{z})$.

En su mayoría, los desarrollos en SDR han estado más enfocados al caso de **reducciones lineales**. Esto es, cuando $\mathbf{R}(\mathbf{X}) = \boldsymbol{\Gamma}^T\mathbf{X}$, con $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$ y $d \leq p$. En la presente tesis se encuentran reducciones no necesariamente lineales, aunque sí lo serán respecto al estadístico suficiente. Si bien para la aplicación a ciencias sociales que consideraremos la reducción lineal tiene sus ventajas en términos de la interpretación del índice SES, ya que cada elemento de $\boldsymbol{\Gamma}$ constituiría la ponderación de cada variable para construir el índice, pueden existir interacciones de diferentes órdenes entre los predictores que aportan información relevante para la respuesta, siendo por ello incluidas en la reducción suficiente, obteniendo con ello una SDR no lineal.

Cuando consideramos reducciones lineales, si $\mathbf{\Gamma}^T\mathbf{X}$ es una reducción suficiente entonces $\boldsymbol{\beta}\mathbf{\Gamma}^T\mathbf{X}$ también lo será, para cualquier matriz $\boldsymbol{\beta}$ no singular $d \times d$. Luego, en el caso de reducciones lineales, lo que nos interesa identificar es el subespacio generado por las columnas de $\mathbf{\Gamma}$, denominado en la literatura como subespacio de reducción suficiente.

Definición 2.3 *Sea* $R(\mathbf{X}) = \mathbf{\Gamma}^T\mathbf{X}$ *una reducción suficiente para la regresión* $Y|\mathbf{X}$*, luego*

$$\mathcal{S}_{\mathbf{\Gamma}} \equiv \text{span}(\mathbf{\Gamma})$$

se denomina **Subespacio de Reducción Suficiente**.

Cabe notar que si tomamos $\mathbf{\Gamma} = \mathbb{I}_p$, luego $\mathbf{R}(\mathbf{X}) = \mathbf{\Gamma}^T\mathbf{X} = \mathbf{X}$ será una reducción suficiente. Dicha reducción es comúnmente denominada *reducción trivial*. No obstante, estamos interesados en obtener una reducción de dimensión lo más chica posible. De la misma manera que definimos reducción suficiente minimal, tenemos el análogo para el subespacio de reducción en el contexto de reducciones lineales.

Definición 2.4 *Sea* $\mathcal{S}$ *un subespacio de reducción suficiente correspondiente a la regresión* $Y|\mathbf{X}$*, si*

$$dim(\mathcal{S}) \leq dim(\mathcal{S}_{SDR})$$

para todo $\mathcal{S}_{SDR}$ *subespacio de reducción suficiente, luego* $\mathcal{S}$ *se denomina* **Subespacio de Reducción Suficiente Minimal**.

En particular, si existe un $\mathcal{S}$ subespacio de reducción suficiente tal que $\mathcal{S} \subset \mathcal{S}_{SDR}$ para todo $\mathcal{S}_{SDR}$ subespacio de reducción suficiente, dicho subespacio, denotado comúnmente por $\mathcal{S}_{Y|\mathbf{X}}$, se denomina **Subespacio Central** y constituye el único minimal (para un mayor desarrollo ver Cook (1998*b*)).

En el próximo apartado, resumiremos los principales enfoques de SDR, presentando los enfoques clásicos y los nuevos desarrollados basados en modelos, que son los que usaremos en esta tesis.

2.2.2. Enfoques y métodos en SDR: Una breve revisión

Una de las características de la mayoría de los métodos de reducción propuestos es que constituyen proyecciones del vector de predictores a subespacios de dimensión inferior. Esto significa que inmediatamente las reducciones suficientes son lineales, y con ello las metodologías asociadas. Una vez identificada la reducción suficiente o una parte de la misma, está el problema de estimación. Las primeras metodologías se basaban en que, bajo ciertas condiciones, el momento de orden 1 y/o 2 de $\mathbf{X}|Y$ está contenido en $\mathbf{\Sigma}^{-1}\mathcal{S}_{Y|\mathbf{X}}$ con $\mathbf{\Sigma} = cov(\mathbf{X})$, por lo que las primeras propuestas para estimar la reducción suficiente de dimensiones de la regresión $Y|\mathbf{X}$ se basaron en la utilización de estimadores de momentos de la regresión inversa.

Dentro del enfoque basado en los momentos, existen varias metodologías para estimar la reducción suficiente de la regresión de Y sobre $\mathbf{X}$; todas ellas fundamentadas en el estudio de los momentos o funciones de los momentos, de la distribución condicional de $\mathbf{X}|Y$. Ejemplos destacados de metodologías bajo este enfoque son, SIR de Li (1991a), SAVE de Cook and Weisberg (1991a), pHd deLi (1992b), PIR de Bura and Cook (2001b); MAVE de Xia et al. (2002a) Li et al. (2005b), Cook and Ni (2005a), Zhu and Zeng (2006b), Cook and Li (2002a), y DR de Li and Wang (2007)). Todos estos métodos fueron desarrollados para predictores continuos.

Con el fin de encontrar reducciones mas exhaustivas, se estudiaron reducciones no necesariamente lineales. Para ello se desarrollaron una serie de trabajos, combinando SDR con núcleos o *kernels* (Akaho 2001, Fukumizu et al. 2009, Fukumizu and Leng 2014, Wu et al. 2008, Hsing and Ren 2009, Yeh et al. 2009, Zhu and Li 2011). En esta misma línea, Li et al. (2011) usa *support vector machine* para estimar la reducción suficiente (denominado *Principal Support Vector Machine (PSVM)*), y constituye en sí un método no lineal, para el cual deben seleccionarse algunas funciones o parámetros, tal como son las funciones de núcleos y las ventanas utilizadas. Por su parte, Lee et al. (2013) considera el caso en que $Y \perp\!\!\!\perp \mathbf{X}|\mathrm{E}(Y|\mathbf{X})$ y propone una generalización de las metodologías basadas en momentos SIR (GSIR) y SAVE (GSAVE), usando el enfoque basado en núcleos. En general, estos enfoques mencionados siguen siendo intrínsecamente lineales en el sentido de que, si bien para estimar la reducción se trabaja sobre un espacio transformado de las variables $\mathbf{X}$ e Y para así captar la mayor información posible de los predictores y la respuesta, y con ello obtener reducciones no necesariamente lineales de

$\mathbf{X}$, se terminan aplicando métodos lineales de SDR tradicionales sobre dicho espacio transformado (Duarte 2016). Otra característica de estos métodos de estimación es que también están basados en momentos de la regresión inversa.

Cook (2007) introduce el enfoque basado en modelos sobre la regresión inversa $\mathbf{X}|Y$, con el fin de identificar reducciones suficientes para la regresión Y sobre $\mathbf{X}$. Cook (2007) y Cook and Forzani (2008, 2009) estudian el caso en que $\mathbf{X}|Y$ se distribuye normalmente, y a partir de aquí proponen dos métodos: Componentes Principales Ajustados (*Principal Fitted Components, PFC*) y el método LAD (*Likelihood Acquired Directions*). La ventaja de este enfoque es que permite obtener una reducción suficiente conteniendo toda la información de $\mathbf{X}$ que es relevante para Y, con la posibilidad de obtener además, estimadores de máxima verosimilitud (en base a un modelo) que son óptimos en términos de eficiencia y $\sqrt{n}$-consistentes aún cuando el modelo $\mathbf{X}|Y$ no es totalmente correcto.

Cuando suponemos que $\mathbf{X}|Y$ se distribuye normalmente con matriz de covarianza constante (respecto a Y) es posible obtener, utilizando máxima verosimilitud, reducciones minimales suficientes (Cook and Forzani 2008). No obstante cuando $\text{cov}(\mathbf{X}|Y)$ depende de Y, la minimalidad de la reducción suficiente no es posible obtenerla con estas metodologías (Cook and Forzani 2008). Sin embargo, ellos encuentran el subespacio central para este modelo.

Por otra parte, Bura and Forzani (2015) identifican reducciones no lineales en base al enfoque basado en modelos. Asumiendo una distribución elíptica para la regresión inversa de la forma $\mathbf{X}|Y \sim \mathcal{E}(\boldsymbol{\mu}_y, \boldsymbol{\Delta}, g_y)$ donde $\boldsymbol{\mu}_Y$ es la media, $\boldsymbol{\Delta}$ la varianza y g_Y la densidad, demuestran que la reducción minimal suficiente ya no es lineal, excepto para el caso normal de varianza constante. Es decir, prueban que si $\mathbf{X}|Y \sim \mathcal{E}(\boldsymbol{\mu}_y, \boldsymbol{\Delta}, g_y)$, existe una reducción suficiente lineal *no trivial* si g_Y es la densidad normal con varianza constante. Mas específicamente, cuando los datos no son normales pero tienen una distribución elíptica, muestran que $\boldsymbol{\Gamma}^T\mathbf{X}$ es suficiente para la regresión de Y sobre $\mathbf{X}$ sí y sólo si $\boldsymbol{\Gamma} = \mathbb{I}_p$. En este contexto, una reducción suficiente de $\mathbf{X}$ para la regresión de Y sobre $\mathbf{X}$, contiene una componente lineal y otra no lineal. Más aún, demuestran que si $\mathbf{X}|Y$ tiene una distribución elíptica, con los métodos de SDR lineales a lo sumo sólo pueden obtenerse algunos elementos del subespacio central, y nunca se podrá estimar exhaustivamente alguna reducción suficiente no trivial, por el hecho de que se omite la componente no lineal. De esta manera, este enfoque es el primero en dar reducciones suficientes genuinamente no lineales.

En la siguiente sección se expone el caso más general de SDR para familias exponenciales, y luego, como caso particular, se identifica la reducción para el caso normal con varianza constante, lo que da origen al método de Componentes Principales Ajustadas (PFC).

2.3. Reducción para Familias Exponenciales

Los métodos de reducción suficiente han tenido un mayor desarrollo para el caso en que los predictores $\mathbf{X}$ son continuos. Sin embargo, existen algunas contribuciones tempranas que abordaron el problema de reducción suficiente en problemas de regresión con predictores tanto continuos como categóricos. Tal es el caso de Chiaromonte et al. (2002), quienes proponen un método para SDR sobre los predictores continuos para cada sub-población definida a partir de las categorías (finitas) de los predictores categóricos. Por ello, tal enfoque está destinado más a encontrar una reducción sobre las variables continuas considerando la presencia de categóricas, que en reducir conjuntamente todo el vector de predictores (tanto continuos como categóricos). Además, adoptan supuestos muy restrictivos sobre la estructura de correlación entre las variables. En particular, suponen que no hay correlación entre las continuas y las categóricas, como así también que los predictores continuos tienen una estructura de covarianza que es invariante entre las diferentes sub-poblaciones que definen las variables categóricas. Wen and Cook (2007) presentan una propuesta superadora en este sentido, aunque mantienen la idea de encontrar una reducción suficiente para el conjunto de continuas en cada sub-población definida por la categoría de forma separada. No obstante, no abordan la de variables reducción de categóricas y continuas de modo conjunto, y su enfoque se torna intratable cuando se tienen muchas categóricas.

El trabajo de Cook and Li (2002*b*) constituye un primer aporte en SDR que contempla la reducción conjunta de predictores continuos y categóricos, proponiendo la idea de trabajar reducción suficiente en el marco de familias exponenciales para el conjunto de predictores $\mathbf{X}$ dada la variable respuesta Y. Una limitación relevante de esta metodología es el supuesto de independencia condicional entre los predictores contenidos en $\mathbf{X}$ dada la variable respuesta. Tal limitación es superada con la propuesta de Bura et al. (2015), sobre la cual se basa una parte importante de la presente tesis; y de hecho busca ser una extensión de la misma.

Por ende, en esta sección se exponen algunas definiciones y resultados de RSD para familias exponenciales, siguiendo a Bura et al. (2015) y Duarte (2016). Se propone también una generalización ante la presencia de parámetros tipo *nuisance*, en el sentido que no dependen de la respuesta Y. En general, para un vector de variables aleatorias $\mathbf{X} = (X_1, \ldots, X_p)^T$ con distribución $\mathcal{P}_{\boldsymbol{\theta}}$ de parámetro $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$, decimos que dicha variable pertenece a una familia exponencial de $k-$parámetros si la función de densidad o probabilidad puntual puede escribirse de la forma

$$f(\mathbf{X}, \boldsymbol{\theta}) = \exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^T \mathbf{T}(\mathbf{X}) - \psi(\boldsymbol{\theta})\Big) h(\mathbf{X}), \tag{2.1}$$

donde $\boldsymbol{\eta}(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), \ldots, \eta_k(\boldsymbol{\theta}))^T$ es el vector de parámetros naturales de la familia exponencial exponencial, los que son dos veces continuamente diferenciables respecto a $\boldsymbol{\theta}$ y la respectiva matriz Jacobiana tiene rango completo, $\mathbf{T}(\mathbf{X}) = (T_1(\mathbf{X}), \ldots, T_k(\mathbf{X}))^T$ es el estadístico suficiente de la familia para $\boldsymbol{\eta}(\boldsymbol{\theta})$, $h(\mathbf{X})$ es una función positiva de $\mathbf{X}$, y $\psi(\boldsymbol{\theta})$ es tal que $f(\cdot)$ constituye un densidad (o función de probabilidad puntual).

En muchos casos (principalmente para la inferencia estadística), interesa un cierto subconjunto de los parámetros, o una determinada función de los mismos, denominando al resto parámetros *nuisance* (Jørgensen and Labouriau 2012). Si escribimos $\boldsymbol{\eta}(\boldsymbol{\theta}) = (\boldsymbol{\eta}_1(\boldsymbol{\theta}), \boldsymbol{\eta}_2(\boldsymbol{\theta}))$ y el interés está sobre $\boldsymbol{\eta}_1(\boldsymbol{\theta})$, mientras que $\boldsymbol{\eta}_2(\boldsymbol{\theta})$ juega un rol de una función *nuisance* de los parámetros, luego la familia exponencial (2.1) puede escribirse de la forma

$$f(\mathbf{X}, \boldsymbol{\theta}) = \exp\Big(\boldsymbol{\eta}_1^T(\boldsymbol{\theta}) \mathbf{T}_1(\mathbf{X}) + \boldsymbol{\eta}_2^T(\boldsymbol{\theta}) \mathbf{T}_2(\mathbf{X}) - \psi(\boldsymbol{\eta}_1(\boldsymbol{\theta}), \boldsymbol{\eta}_2(\boldsymbol{\theta}))\Big) h(\mathbf{X}). \tag{2.2}$$

En tal caso, $\mathbf{T}_1(\mathbf{X})$ será un estadístico suficiente para $\boldsymbol{\eta}_1(\boldsymbol{\theta})$ y $\mathbf{T}_2(\mathbf{X})$ será un estadístico suficiente para $\boldsymbol{\eta}_2(\boldsymbol{\theta})$ (Chang and Liang 1994).

En el marco de estudio de SDR para la regresión de Y sobre p predictores $\mathbf{X}$, bajo el enfoque de reducción inversa se supone que $\mathbf{X}$ condicionado a Y tiene una distribución perteneciente a la familia exponencial. Luego, de acuerdo a (2.2), podemos extender la formulación de Bura et al. (2015) y suponer que su función de densidad viene dada por

$$f(\mathbf{X}|\boldsymbol{\eta}_y, Y = y) = \exp\Big(\boldsymbol{\eta}_y^T \mathbf{T}(\mathbf{X}) + \tilde{\boldsymbol{\eta}}^T \mathbf{U}(\mathbf{X}) - \psi(\boldsymbol{\eta}_y, \tilde{\boldsymbol{\eta}})\Big) h(\mathbf{X}), \tag{2.3}$$

donde esta representación divide al parámetro natural $\boldsymbol{\eta}_y^T = (\eta_{y1}, \ldots, \eta_{yk_1})^T$ en k_1 funciones de

Y y en k_2 parámetros independientes de Y dado por $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_{k_2})$, con $k_1 + k_2 = k \geq p$. El estadístico $\mathbf{T}(\mathbf{X})$ es suficiente para $\boldsymbol{\eta}_y^T$, mientras que $\mathbf{U}(\mathbf{X})$ es auxiliar (*ancillary*) para $\boldsymbol{\eta}_y^T$, en el sentido de que $E_Y(\mathbf{U}(\mathbf{X}))$ no depende de Y Lehmann and Casella (2003). En el marco de SDR, el parámetro de interés es $\boldsymbol{\eta}_y^T$ y en el marco de Modelos Lineales Generalizados, dicho parámetro natural se especifica linealmente como función de Y. Siguiendo a Bura et al. (2015), podemos escribir

$$\boldsymbol{\eta}_Y = \bar{\boldsymbol{\eta}} + \boldsymbol{\Gamma}\boldsymbol{\nu}_Y \tag{2.4}$$

donde $\bar{\boldsymbol{\eta}} = E_Y(\boldsymbol{\eta}_Y)$, $\boldsymbol{\Gamma} \in \mathbb{R}^{k_1 \times d}$ es una matriz semiortogonal de rango completo tal que sus columnas forman una base para el $\text{span}\{\boldsymbol{\eta}_y - \bar{\boldsymbol{\eta}} : y \in \mathcal{Y}\}$, siendo $\mathcal{Y}$ el espacio muestral de Y, $E_Y(\boldsymbol{\nu}_Y) = 0$ y $E_Y(\boldsymbol{\nu}_Y \boldsymbol{\nu}_Y^T)$ definida positiva. Puntualmente, siguiendo a Cook (2007) y Cook and Forzani (2008), $\boldsymbol{\nu}_Y$ puede ser modelado linealmente respecto a una función conocida de Y, $\mathbf{f}_Y \in \mathbb{R}^r$, de la forma $\boldsymbol{\nu}_Y = \boldsymbol{\beta}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)$, donde $\bar{\mathbf{f}}_Y = E_Y(\mathbf{f}_Y)$ y $\boldsymbol{\beta}$ es una matriz de orden $d \times r$ de rango $d \leq \text{mín}(k_1, r)$. Luego (2.4) puede escribirse de la forma

$$\begin{aligned} \boldsymbol{\eta}_Y &= \bar{\boldsymbol{\eta}} + \boldsymbol{\Gamma}\boldsymbol{\beta}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) \\ &= \bar{\boldsymbol{\eta}} + \mathbf{D}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y), \end{aligned} \tag{2.5}$$

donde $\mathbf{D} = \boldsymbol{\Gamma}\boldsymbol{\beta}$. En (2.5) se admiten dos representaciones para modelar $\boldsymbol{\eta}_Y$, una mediante un modelo lineal de rango reducido, y la otra con un modelo lineal generalizado sin restricción. En el siguiente teorema extendemos el Teorema 1 de Bura et al. (2015) para identificar la reducción suficiente para la regresión original $Y|\mathbf{X}$, la cual será lineal en el estadístico suficiente $\mathbf{T}(\mathbf{X})$ aunque no necesariamente en $\mathbf{X}$.

Teorema 2.1 *Asumiendo que* $\mathbf{X}|Y$ *tiene la densidad (2.3), la reducción suficiente minimal para la regresión* $Y|(\mathbf{X})$ *estará dada por*

$$\mathbf{R}(\mathbf{X}) = \mathbf{a}^T(\mathbf{T}(\mathbf{X}) - \text{E}(\mathbf{T}(\mathbf{X})),$$

donde $\mathbf{T}(\mathbf{X})$ *es el estadístico suficiente para* $\boldsymbol{\eta}_Y$ *de la familia exponencial definida por (2.3) y* $\mathcal{S}_{\mathbf{a}} = \text{span}\{\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}} : Y \in \mathcal{Y}\}$, *donde* $\mathcal{S}_{\mathbf{a}} \equiv \text{span}(\mathbf{a})$.

Prueba. Siguiendo a Bura et al. (2015), por el teorema de Lehmann-Sheffe, tenemos que la reducción $\mathbf{a}^T(\mathbf{T}(\mathbf{X}) - \text{E}(\mathbf{T}(\mathbf{X}))$ será suficiente minimal si para dos puntos muestrales cualesquiera $\mathbf{x}$ y $\tilde{\mathbf{x}}$, tener $\log\left(f_{\mathbf{X}|Y}(\mathbf{x})/f_{\mathbf{X}|Y}(\tilde{\mathbf{x}})\right)$ independiente de Y es equivalente a $\mathbf{a}^T\mathbf{T}(\mathbf{x}) = \mathbf{a}^T\mathbf{T}(\tilde{\mathbf{x}})$.

Por hipótesis $f_{\mathbf{X}|Y}$ está dada por (2.3), luego si $\log\left(f_{\mathbf{X}|Y}(\mathbf{x})/f_{\mathbf{X}|Y}(\tilde{\mathbf{x}})\right)$ es independiente de Y, vamos a tener que

$$\log\frac{f_{\mathbf{X}|Y}(\mathbf{x})}{f_{\mathbf{X}|Y}(\tilde{\mathbf{x}})} = \log\frac{h(\mathbf{x})}{h(\tilde{\mathbf{x}})} + \left\{\boldsymbol{\eta}_Y^T[\mathbf{T}(\mathbf{x}) - \mathbf{T}(\tilde{\mathbf{x}})] + \tilde{\boldsymbol{\eta}}^T[\mathbf{U}(\mathbf{x}) - \mathbf{U}(\tilde{\mathbf{x}})]\right\} = c, \tag{2.6}$$

siendo c una contante independiente de Y. Luego, al tomar esperanza con respecto a Y y considerando que $\tilde{\boldsymbol{\eta}}$ es independiente de Y, la ecuación (2.6) es equivalente a

$$(\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}})^T[\mathbf{T}(\mathbf{x}) - \mathbf{T}(\tilde{\mathbf{x}})] = 0. \tag{2.7}$$

Sea $\mathbf{a} \in \mathbb{R}^{k_1 \times d}$ tal que sus columnas generan el span$\{\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}, Y \in \mathcal{Y}\}$, luego $(\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}) = \mathbf{a}\boldsymbol{\nu}_Y$ para algún $\boldsymbol{\nu}_Y \in \mathbb{R}^d$ y para todo $Y \in \mathcal{Y}$. Luego, (2.7) puede escribirse como $\boldsymbol{\nu}_Y^T\mathbf{a}^T[\mathbf{T}(\mathbf{x}) - \mathbf{T}(\tilde{\mathbf{x}})] = 0$ puesto que por hipótesis $E_Y(\boldsymbol{\nu}_Y\boldsymbol{\nu}_Y^T)$ es definida positiva. Al valer para todo $Y \in \mathcal{Y}$, se tiene que $\mathbf{a}^T\mathbf{T}(\mathbf{x}) = \mathbf{a}^T\mathbf{T}(\tilde{\mathbf{x}})$ y por ende, $\mathbf{a}^T(\mathbf{T}(\mathbf{x}) - \mathrm{E}(\mathbf{T}(\mathbf{x}))$ es una reducción suficiente minimal para la regresión $Y|\mathbf{X}$. □

Para la estimación, tomando la especificación general de (2.5), Bura et al. (2015) proponen un estimador de máxima verosimilitud para una base del span($\mathbf{D}$) o del span($\boldsymbol{\Gamma}$), para un modelo de rango completo y de rango reducido, respectivamente, adaptando un algoritmo tipo IRLS para tales fines. Como se verá a continuación, para el caso normal, el estimador de máxima verosimilitud de la reducción tiene una forma explícita.

2.4. Caso Normal: Componentes Principales Ajustadas

Dentro del enfoque de regresión inversa, si $\mathbf{X}|Y$ se distribuye normalmente, tenemos un caso particular de familia exponencial. En particular, siguiendo a Cook (2007) y Cook and Forzani (2008), supongamos que $\mathbf{X}|Y = y \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Delta})$ con

$$\boldsymbol{\mu}_y \ = \bar{\boldsymbol{\mu}} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y + \boldsymbol{\epsilon}, \tag{2.8}$$

donde $\bar{\boldsymbol{\mu}} = E(\mathbf{X})$, $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$ es una matriz semiortogonal de rango completo tal que sus columnas forman una base para $\mathcal{S}_\Gamma = \text{span}\{\boldsymbol{\mu}_y - \bar{\boldsymbol{\mu}} : y \in \mathcal{Y}\}$, $\boldsymbol{\nu}_y = \boldsymbol{\Gamma}^T(\boldsymbol{\mu}_y - \bar{\boldsymbol{\mu}}) \in \mathbb{R}^d$ y $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Delta})$, independiente de Y. Para el modelo (2.8), Cook (2007) demuestra que $\mathbf{R}(\mathbf{X}) = \boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\mathbf{X}$ es

una reducción suficiente para la regresión $Y|\mathbf{X}$. En particular, cuando parametrizamos $\boldsymbol{\nu}_Y = \boldsymbol{\beta}\{\mathbf{f}_Y - E(\mathbf{f}_Y)\}$ para alguna función conocida $\mathbf{f}_Y \in \mathbb{R}^r$ de Y y para alguna matriz $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$ de rango $d \leq \text{mín}(r, p)$, estamos en presencia de lo modelos de Componentes Principales Ajustadas (PFC) (Cook and Forzani 2008, pp.3). De esta manera, el modelo (2.8) se escribe de la forma

$$\mathbf{X}|(Y = y) = \bar{\boldsymbol{\mu}} + \boldsymbol{\Gamma}\boldsymbol{\beta}\{\mathbf{f}_y - E(\mathbf{f}_y)\} + \epsilon = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \epsilon, \tag{2.9}$$

con $\boldsymbol{\mu} = \bar{\boldsymbol{\mu}} - \boldsymbol{\Gamma}\boldsymbol{\beta}E(\mathbf{f}_y)$. Dado el modelo (2.9), Cook and Forzani (2008) identifican la reducción suficiente minimal para $Y|\mathbf{X}$, la que enunciamos en el siguiente Teorema (equivalente al Teorema 2.1 de Cook and Forzani (2008)), pero realizamos una demostración alternativa utilizamos el resultado del Teorema 2.1, dado que el caso normal es un caso especial de familia exponencial.

Teorema 2.2 *Asumiendo que para* $\mathbf{X}|Y$ *vale el modelo (2.9), luego* $\mathbf{R}(\mathbf{X}) = \boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}(\mathbf{X}-E(\mathbf{X}))$ *constituye una reducción suficiente minimal para la regresión* $Y|\mathbf{X}$.

Prueba. Para aplicar el resultado del Teorema 2.1, en primer lugar debemos expresar la densidad de $\mathbf{X}|Y$ en términos de la forma (2.3). Específicamente, asumiendo el modelo (2.9), la densidad de $\mathbf{X}$ condicionada a Y puede escribirse de la forma

$$\begin{aligned}
f_{\mathbf{X}|Y}(\mathbf{X}|Y) &= 2\pi^{-p/2}|\boldsymbol{\Delta}|^{-1/2}\exp\{-\frac{1}{2}(\mathbf{X} - \bar{\boldsymbol{\mu}} - \boldsymbol{\Gamma}\boldsymbol{\beta}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y))^T\boldsymbol{\Delta}^{-1}(\mathbf{X} - \bar{\boldsymbol{\mu}} - \boldsymbol{\Gamma}\boldsymbol{\beta}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)\} \\
&= 2\pi^{-p/2}|\boldsymbol{\Delta}|^{-1/2}\exp\{-\frac{1}{2}(\mathbf{X} - \bar{\boldsymbol{\mu}})^T\boldsymbol{\Delta}^{-1}(\mathbf{X} - \bar{\boldsymbol{\mu}}) + (\mathbf{X} - \bar{\boldsymbol{\mu}})^T\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}\boldsymbol{\beta}(\mathbf{f}_Y - \bar{\mathbf{f}}_y) \\
&\qquad - \frac{1}{2}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)^T\boldsymbol{\beta}^T\boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}\boldsymbol{\beta}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)\} \\
&= 2\pi^{-p/2}|\boldsymbol{\Delta}|^{-1/2}\exp\{-\frac{1}{2}\text{vec}^T((\mathbf{X} - \bar{\boldsymbol{\mu}})(\mathbf{X} - \bar{\boldsymbol{\mu}})^T)\text{vec}(\boldsymbol{\Delta}^{-1}) \\
&\qquad + (\mathbf{X} - \bar{\boldsymbol{\mu}})^T\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}\boldsymbol{\beta}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \frac{1}{2}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)^T\boldsymbol{\beta}^T\boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}\boldsymbol{\beta}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)\}
\end{aligned} \tag{2.10}$$

En términos de la forma general de la familia exponencial (2.3), para (2.10) tomamos

$$h(\mathbf{X}) = 2\pi^{-p/2}, \tag{2.11}$$

$$\tilde{\boldsymbol{\eta}} = -\frac{1}{2}\text{vec}(\boldsymbol{\Delta}^{-1}), \tag{2.12}$$

$$\mathbf{U}(\mathbf{X}) = \text{vec}((\mathbf{X} - \bar{\boldsymbol{\mu}})(\mathbf{X} - \bar{\boldsymbol{\mu}})^T), \tag{2.13}$$

$$\psi(\boldsymbol{\eta}_Y, \tilde{\boldsymbol{\eta}}) = \frac{1}{2}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)^T \boldsymbol{\beta}^T \boldsymbol{\Gamma}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\Gamma} \boldsymbol{\beta} (\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \frac{1}{2} \log |\boldsymbol{\Delta}| + \bar{\boldsymbol{\mu}} \boldsymbol{\Delta}^{-1} \boldsymbol{\Gamma} \boldsymbol{\beta} (\mathbf{f}_Y - \bar{\mathbf{f}}_Y), \tag{2.14}$$

$$\mathbf{T}(\mathbf{X}) = \mathbf{X} \tag{2.15}$$

y

$$\boldsymbol{\eta}_Y = \boldsymbol{\Delta}^{-1} \boldsymbol{\Gamma} \boldsymbol{\beta} (\mathbf{f}_Y - \bar{\mathbf{f}}_Y). \tag{2.16}$$

Dado que $\bar{\mathbf{f}}_Y = E_Y(\mathbf{f}_Y)$ y $\bar{\boldsymbol{\eta}} = E_Y(\boldsymbol{\eta}_Y)$, luego $\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}} = \boldsymbol{\Delta}^{-1} \boldsymbol{\Gamma} \boldsymbol{\beta} (\mathbf{f}_Y - \bar{\mathbf{f}}_Y)$ y por el Teorema 2.1, $\boldsymbol{\Delta}^{-1} \boldsymbol{\Gamma}$ constituye una base del span$\{\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}} : Y \in \mathcal{Y}\}$, o lo que es lo mismo $\mathcal{S}_{\boldsymbol{\Gamma}} = \boldsymbol{\Delta}^{-1}$span$\{\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}} : Y \in \mathcal{Y}\}$. Es decir, $\mathbf{a} = \boldsymbol{\Delta}^{-1} \boldsymbol{\Gamma}$, y $\mathbf{R}(\mathbf{X}) = \mathbf{a}^T(\mathbf{T}(\mathbf{X}) - E(\mathbf{T}(\mathbf{X}))) = \boldsymbol{\Gamma}^T \boldsymbol{\Delta}^{-1}(\mathbf{X} - E(\mathbf{X})) = \boldsymbol{\Gamma}^T \boldsymbol{\Delta}^{-1}(\mathbf{X} - \bar{\boldsymbol{\mu}})$ es una reducción suficiente minimal. □

Si bien para el caso general de familias exponenciales no hay una forma explícita para el estimador de máxima verosimilitud de la reducción suficiente, para el caso normal, Cook and Forzani (2008) muestran que el estimador de máxima verosimilitud de $\boldsymbol{\Gamma}$ en este caso es

$$\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{\Sigma}}_{\mathbf{res}}^{-1/2} \mathbf{V}_d, \tag{2.17}$$

donde $\hat{\boldsymbol{\Sigma}}_{\mathbf{res}}$ es la matriz de covarianza de los errores de la regresión de $\mathbf{X}|Y$, y $\mathbf{V}_d$ es una matriz cuyas columnas son los primeros d autovectores de $\hat{\boldsymbol{\Sigma}}_{\mathbf{res}}^{-1/2} \hat{\boldsymbol{\Sigma}}_{\mathbf{fit}} \hat{\boldsymbol{\Sigma}}_{\mathbf{res}}^{-1/2}$, donde $\hat{\boldsymbol{\Sigma}}_{\mathbf{fit}}$ es la matriz de covarianza muestral de los vectores ajustados de la regresión lineal multivariada de $\mathbf{X}$ sobre $\mathbf{f}_Y$ (i.e. $\hat{\boldsymbol{\Sigma}}_{\mathbf{res}} = \hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_{\mathbf{fit}}$, con $\hat{\boldsymbol{\Sigma}} = (\mathbf{X}_n^T \mathbf{X}_n)/n$ siendo $\mathbf{X}_n \in \mathbb{R}^{n \times p}$ la matriz de los predictores observados centrada). Además encuentran el estimador de máxima verosimilitud para $\boldsymbol{\Delta}$, que esta dado por

$$\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\Sigma}}_{\mathbf{res}} + \hat{\boldsymbol{\Sigma}}_{\mathbf{res}}^{-1/2} \mathbf{V}_d \mathbf{K} \mathbf{V}_d^T \hat{\boldsymbol{\Sigma}}_{\mathbf{res}}^{-1/2}, \tag{2.18}$$

donde $\mathbf{K} = diag(0, \ldots, 0, \hat{\lambda}_{d+1}, \ldots, \hat{\lambda}_p)$, y $\hat{\lambda}_j$ representan los autovalores, ordenados de forma decreciente, de $\hat{\boldsymbol{\Sigma}}_{\mathbf{res}}^{-1/2} \hat{\boldsymbol{\Sigma}}_{\mathbf{fit}} \hat{\boldsymbol{\Sigma}}_{\mathbf{res}}^{-1/2}$.

En la presente tesis adoptamos este paradigma de SDR basada en modelos. En primer lugar, realizamos una extensión de la metodología de Componentes Principales Ajustados (PFC) para variables ordinales (PFCord), obteniendo una reducción suficiente para el modelo $Y|\mathbf{X}$ por máxima verosimilitud. Luego, usamos los resultados de SDR para la familia exponencial con el objetivo de generalizar el enfoque de SDR para predictores de naturaleza mixta. Por ello, los

modelos que presentaremos usarán fuertemente los resultados de los Teoremas 2.1 y 2.2. Una vez identificadas las reducciones suficientes, los estimadores de máxima verosimilitud de los parámetros del modelo que definen la reducción, tendrán una forma análoga a (2.17) y (2.18), en particular para el modelo de predictores ordinales, en cada paso del cálculo iterativo del método propuesto.

Capítulo 3

Reducción Suficiente de Dimensiones para Predictores Ordinales

3.1. Introducción

Los modelos de regresión con variables ordinales como predictores son muy comunes en varias aplicaciones. En particular en economía y ciencias sociales, la predominancia de variables ordinales es un rasgo común que tienen las bases de datos asociadas, tanto en lo que hace a las características de los agentes como a los comportamientos y resultados de fenómenos sociales que interesa estudiar. También una aplicación muy difundida en la última década son los sistemas automatizados de recomendación, a través de las preferencias de los consumidores, como es el conocido caso de Netflix (e.g. Bobadilla et al. 2012, Roberts 2014), donde la puntuación (variable ordinal) de una cierta película no vista por un usuario es predicha a partir de las puntuaciones realizadas por dicho usuario de otras películas y de la información de otros usuarios sobre la valuación de la película que se busca predecir.

En este contexto, cuando el número de predictores es muy elevado, interesa reducir la dimensión del problema a través de la combinación de una pocas variables que contengan la mayor información de las variables originales con el fin de ganar en eficiencia, como en el entendimiento del fenómeno a modelar. Como dijimos en e lCapítulo 1, las técnicas comúnmente usadas para variables ordinales son adaptaciones al análisis estándar de componentes principales (Linting and van der Kooij 2009, Kolenikov and Angeles 2009). De esta manera, en el caso del

Índice de Focalización de Pobreza, la primera componente principal es utilizada para predecir si un hogar es pobre o no, aún cuando esta variable respuesta nunca es utilizada para construir los coeficientes que definen al indicador. Claramente, al no utilizar ninguna información sobre lo que se busca predecir (pobreza, en este caso), existe una perdida importante del poder predictivo del método de reducción utilizado para la construcción del índice.

Un enfoque alternativo al de reducción de dimensiones, es el de selección de variables sobre el conjunto original de predictores. Un método adaptado al caso de predictores ordinales ha sido propuesto por Gertheiss and Tutz (2010). Si bien este método utiliza la información de la variable respuesta para realizar selección de variables, asume un modelo paramétrico específico para la respuesta como función de los predictores. Tampoco sirve para elaborar un índice, en el sentido que le damos en el presente trabajo.

Con el enfoque de reducción suficiente de dimensiones (SDR) es factible reducir la dimensión de los predictores $\mathbf{X}$ usando información de la variable respuesta Y sin proponer un modelo para el problema de regresión original $Y|\mathbf{X}$. Como dijimos en el Capítulo 2, la mayoría de los métodos han sido diseñados para predictores continuos, con una extensión reciente a variables con distribución perteneciente a la familia exponencial (e.g. Bura et al. 2015), que admite SDR para predictores de diferente naturaleza dentro de la familia exponencial. Con el objeto de aplicar SDR para predictores ordinales, una primera opción podría ser considerarlos como variables categóricas politómicas ignorando su orden natural. Así, postulando una distribución multinomial para dichos predictores, la SDR podría pensarse en el marco de familias exponenciales. Sin embargo, los predictores ordinales usualmente no tienen una distribución multinomial, y si se adoptadara esta distribución para representar a las ordinales, debe tenerse en cuenta que la información sobre el orden natural que contienen dichas variables se perderá por completo con esta representación. Otra opción es usar variables *dummy* para representar a las ordinales, pero esto puede terminar introduciendo un problema de correlaciones espurias (Kolenikov and Angeles 2009). Otro enfoque es tratar a los predictores ordinales suponiendo que surgen de un proceso de discretización de variables latentes continuas. Esta metodología es la comúnmente utilizada en ciencias sociales, y se enmarca dentro de los denominados *modelos de variables latentes*. En este contexto, las variables latentes usualmente se modelan suponiendo una distribución normal o logística, obteniendo los así llamados modelos probit y logit, respectivamente (Greene and Hensher 2010, Long 1997). Dependiendo del fenómeno científico de interés, la va-

riable latente puede adoptar un significado particular (e.g., utilidad en problemas de elección, predisposición a contraer enfermedades en genética, o tolerancia a drogas en toxicología); o bien, una interpretación general de la variable latente puede realizarse a través de pensarla como una *propensión* a observar un cierto valor j de una variable categórica ordinal (Skrondal and Rabe-Hesketh 2004). Mas allá de su sentido filosófico y de las críticas realizadas en torno a su existencia real, las variables latentes resultan de gran utilidad para el modelado estadístico en general.

Dado el potencial de SDR, y la presencia frecuente de predictores ordinales en las aplicaciones, en este capítulo desarrollamos un método de SDR para predictores ordinales siguiendo un enfoque de variables latentes que se suponen normalmente distribuidas dada la respuesta.

El resto de este capítulo se organiza de la siguiente manera. En la Sección 3.2 describimos el modelo de regresión inversa para datos ordinales e identificamos su reducción suficiente. En la Sección 3.4 derivamos el estimador de Máxima Verosimilitud para la reducción suficiente y también presentamos un método de selección de variables. La Sección 3.5 presenta un test de permutación para seleccionar la dimensión de la reducción suficiente. Los resultados de las simulaciones son presentados en la Sección 3.6. La Sección 3.7 contiene dos aplicaciones del método propuesto. La primera corresponde a la aplicación que motiva la tesis, esto es, a la creación de un índice SES usando variables ordinales que caracterizan al hogar, con el fin de predecir ingreso per cápita y nivel de pobreza. En un segundo ejemplo, usamos la metodología propuesta para la recomendación de películas utilizando datos de Netflix. Por último, exponemos unas breves conclusiones en la Sección 3.8.

3.2. Modelo

Consideraremos un modelo de regresión en el cual tenemos una variable respuesta $Y \in \mathbb{R}$ sobre un conjunto de predictores $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T \in \mathbb{R}^p$, donde cada X_j, $j = 1, \ldots, p$ es una variable categórica ordenada, es decir $X_j \in \{1, 2, \ldots, G_j\}$, $j = 1, \ldots, p$. Para encontrar la SDR de $\mathbf{X}$, usaremos el enfoque de regresión inversa basado en modelos (Cook 2007). Para ello, dado el vector de p variables ordinales observadas $\mathbf{X}$, supondremos la existencia de un vector p-dimensional de variables latentes continuas no observadas subyacentes a cada variable

ordinal, $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)^T$, tal que satisfacen el siguiente modelo

$$\mathbf{Z}|Y = \boldsymbol{\mu}_Y + \boldsymbol{\epsilon}, \tag{3.1}$$

donde $\boldsymbol{\mu}_Y = E(\mathbf{Z}|Y)$ y el término de error $\boldsymbol{\epsilon}$ es independiente de Y, normalmente distribuido y con media $\mathbf{0}$ y matriz decovarianza $\boldsymbol{\Delta}$ definida positiva. Como es usual en los modelos de variables latentes para datos ordinales, debemos imponer algunas restricciones para identificar los parámetros del modelo, que en el presente caso serán: $[\boldsymbol{\Delta}]_{jj} \doteq \delta_j = 1$ y $E(\mathbf{Z}) = 0$ (ver Jackman 2009). En el presente contexto, cada variable observada X_j es una versión discretizada a partir de una variable latente Z_j de la siguiente manera: Para un conjunto de umbrales $\boldsymbol{\Theta}^{(j)} = \{\theta_0^{(j)}, \theta_1^{(j)}, \ldots, \theta_{G_j}^{(j)}\}$, dividimos la recta real en intervalos disjuntos $-\infty = \theta_0^{(j)} < \theta_1^{(j)} < \ldots < \theta_{G_{j-1}}^{(j)} < \theta_{G_j}^{(j)} = +\infty$ y luego tomamos

$$X_j = \sum_{g=1}^{G_j} g\mathbb{I}(\theta_{g-1}^{(j)} \leq Z_j < \theta_g^{(j)}),$$

donde $\mathbb{I}(A)$ es una función indicadora del conjunto A. Esto es, X_j es una variable escalonada derivada de dicretizar Z_j.

En el desarrollo que sigue, vamos a denotar $\boldsymbol{\Theta} \doteq \{\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(p)}\} = \{\theta_0^{(1)}, \ldots, \theta_{G_1}^{(1)}, \ldots\ldots, \theta_0^{(p)}, \ldots, \theta_{G_p}^{(p)}\}$ y $C(\mathbf{X}, \boldsymbol{\Theta}) = [\theta_{X_1-1}^{(1)}, \theta_{X_1}^{(1)}) \times \cdots \times [\theta_{X_p-1}^{(p)}, \theta_{X_p}^{(p)})$. Sea $\mathcal{Y}$ el espacio dominio de Y y $\mathcal{S}_{\boldsymbol{\Gamma}} = \text{span}\{\boldsymbol{\mu}_Y - E(\boldsymbol{\mu}_Y)|Y \in \mathcal{Y}\}$. Si $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$, con $d \leq p$, es una matriz semi-ortogonal, cuyas columnas forman una base para el subespacio d-dimensional $\mathcal{S}_{\boldsymbol{\Gamma}}$, siguiendo a Cook and Forzani (2008), podemos re-escribir el modelo (3.1) de la forma

$$\mathbf{Z}|Y = \boldsymbol{\Gamma}\boldsymbol{\nu}_Y + \boldsymbol{\epsilon}, \tag{3.2}$$

donde $\boldsymbol{\nu}_Y = \boldsymbol{\Gamma}^T\boldsymbol{\mu}_Y \in \mathbb{R}^d$ con $E(\boldsymbol{\nu}_Y) = 0$ y $\text{var}(\boldsymbol{\nu}_Y) > 0$. En este caso modelamos al vector de coordenadas de la forma $\boldsymbol{\nu}_Y = \boldsymbol{\xi}\{\mathbf{f}_Y - E(\mathbf{f}_Y)\}$ donde $\mathbf{f}_Y \in \mathbb{R}^r$ es un vector de r funciones conocidas de Y tales que $E((\mathbf{f}_Y - E(\mathbf{f}_Y)(\mathbf{f}_Y - E(\mathbf{f}_Y))^T)$ conforma una matriz definida positiva y $\boldsymbol{\xi} \in \mathbb{R}^{d\times r}$ es una matriz de rango completo d, con $d \leq r$ (ver Cook and Forzani 2008, Adragni and Cook 2009). Bajo este modelo, cada coordenada de $\mathbf{Z}|Y$ es modelada linealmente como función de un vector de predictores dado por $\mathbf{f}_Y$, y por lo tanto, cuando Y es cuantitativa, podemos usar gráficas inversas con el fin de obtener información para seleccionar la función $\mathbf{f}_Y$, lo que no es

posible en la regresión original de Y sobre $\mathbf{X}$ debido a sus dimensiones. Cuando Y es continua, $\mathbf{f}_Y$ usualmente quedará representada por un conjunto flexible de funciones básicas, como ser un polinomio en Y, lo cual resulta ser una opción parsimoniosa cuando el método gráfico se ve agotado o resulta poco práctico para todos los predictores. Cuando Y es categórica y toma valores tales como $\{C_1, \ldots, C_h\}$, podemos tomar $r = h - 1$ y especificar j-ésimo elemento de $\mathbf{f}_Y$ a través de $\mathbb{I}(y \in C_j)$, con $j = 1, \ldots, h$. Cuando Y es continua también podemos particionar sus valores en h categorías $\{C_1, \ldots, C_h\}$ y luego especificar para la coordenada j-ésima de $\mathbf{f}_Y$ de la misma manera que para el caso en que Y sea categórica. Para mayores detalles, ver Adragni and Cook (2009).

De esta manera, el modelo (3.2) puede expresarse como

$$\mathbf{Z}|Y = \boldsymbol{\Gamma}\boldsymbol{\xi}\{\mathbf{f}_Y - E(\mathbf{f}_Y)\} + \boldsymbol{\epsilon}, \tag{3.3}$$

donde $\boldsymbol{\epsilon}$ es independiente de Y, normalmente distribuido con media $\mathbf{0}$ y matriz de varianza-covarianza (definida positiva) $\boldsymbol{\Delta}$, con unos en la diagonal a los fines de la identificabilidad del modelo.

3.3. Reducción suficiente

Suponiendo la validez del modelo (3.3), el siguiente teorema identifica la reducción suficiente de dimensiones para la regresión $Y|\mathbf{X}$.

Teorema 3.1 *Sea* $\mathbf{X}$ *un vector p-dimensional de variables ordinales para las cuales existe un vector de variables latentes subyacentes* $\mathbf{Z}$ *que verifican el modelo (3.3). Luego* $\mathbf{R}(\mathbf{X}) = \boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\mathbf{X}$ *es una reducción suficiente para la regresión de* Y *sobre* $\mathbf{X}$*. Esto es,*

$$Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\mathbf{X}.$$

Prueba. Sea $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^d$ dado por $\mathbf{R}(\mathbf{W}) = \boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\mathbf{W}$. En base al Teorema 2.2 del Capítulo 2, $\mathbf{R}(\mathbf{Z})$ es la reducción minimal suficiente para la regresión de Y sobre $\mathbf{Z}$. Esto implica que $\mathbf{Z}|(\mathbf{R}(\mathbf{Z}), Y)$ no depende de Y. Sea $\boldsymbol{g} = (g_1, \ldots, g_p)$, $g_j = 1, \ldots, G_j$. Puesto que

$X_j = g_j \Leftrightarrow Z_j \in [\theta^{(j)}_{X_j-1}, \theta^{(j)}_{X_j})$ entonces

$$\begin{aligned}\Pr\{\mathbf{X} = \boldsymbol{g}|\mathbf{R}(\mathbf{X}) = \mathbf{R}(\boldsymbol{g}), Y\} &= \Pr\{\mathbf{Z} \in C(\mathbf{X}, \boldsymbol{\Theta})|\mathbf{R}(\mathbf{Z}) \in \mathbf{R}(C(\mathbf{X}, \boldsymbol{\Theta})), Y\} \\ &= \Pr\{\mathbf{Z} \in C(\mathbf{X}, \boldsymbol{\Theta})|\mathbf{R}(\mathbf{Z}) \in \mathbf{R}(C(\mathbf{X}, \boldsymbol{\Theta}))\} \\ &= \Pr\{\mathbf{X} = \boldsymbol{g}|\mathbf{R}(\mathbf{X}) = \mathbf{R}(\boldsymbol{g})\},\end{aligned}$$

y, como consecuencia tenemos que $\mathbf{X}|(\mathbf{R}(\mathbf{X}), Y)$ no depende de Y. La segunda igualdad sale del hecho de que $\mathbf{Z}|(\mathbf{R}(\mathbf{Z}), Y)$ es independiente de Y. Esto implica que $\mathbf{X}|(\mathbf{R}(\mathbf{X}), Y) =_d \mathbf{X}|\mathbf{R}(\mathbf{X})$. Usando la equivalencia del Lema 2.1 llegamos a que $\mathbf{R}(\mathbf{X})$ es una reducción suficiente para la regresión de Y sobre $\mathbf{X}$.□

Cabe notar que no es necesario ningún supuesto sobre la distribución de Y o de $Y|\mathbf{X}$ para la validez del Teorema 3.1. Esta observación tiene un implicación práctica relevante, pues nos dice que podemos utilizar las reducciones obtenidas con cualquier regla de predicción, sea en regresión o clasificación.

Ahora bien, si $R(\mathbf{X}) = \boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\mathbf{X}$ es una reducción suficiente, luego $R(\mathbf{X}) = \mathbf{A}\boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\mathbf{X}$ es una reducción suficiente para toda $\mathbf{A} \in \mathbb{R}^{d\times d}$ invertible (Cook 1998*b*). Por lo tanto el subespacio generado por las columnas de $\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$ es identificable, pero no lo es $\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$ en sí mismo. Como lo mencionamos en el Capítulo 2, en la literatura sobre SDR, el parámetro identificable del subespacio generado por las columnas de $\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$ es denominado *Subespacio de Reducción Suficiente*. Por otra parte, el Teorema 2.2 puede extenderse para predictores que satisfacen las condiciones de linealidad y cobertura (Cook and Forzani 2008), y por lo tanto, el Teorema 3.1 puede extenderse al caso de no normalidad siempre que los momentos de $\mathbf{Z}|Y$ satisfagan las hipótesis del Teorema 3.5 de Cook and Forzani (2008). Por lo tanto el el enfoque de variables latentes descrito aquí es útil aún cuando nos alejamos del supuesto de normalidad.

3.4. Estimación

Del Teorema 3.1, una matriz de coeficientes para la reducción suficiente es $\boldsymbol{\alpha} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$. Si $\mathbf{Z}$ fuera observada, el estimador de máxima versoimilitud de $\boldsymbol{\alpha}$ estaría dado por (2.17) y (2.18) del método de Componentes Principales Ajustadas (PFC). Esto es, $\widehat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\Sigma}}^{-1/2}_{\mathbf{res}}\mathbf{V}_d$, donde $\mathbf{V}_d$ son los primeros d autovectores de $\hat{\boldsymbol{\Sigma}}^{-1/2}_{\mathbf{res}}\hat{\boldsymbol{\Sigma}}_{\mathbf{fit}}\hat{\boldsymbol{\Sigma}}^{-1/2}_{\mathbf{res}}$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{fit}}$ es la matriz de covarianza muestral

de los vectores ajustados de la regresión lineal multivariada de $\mathbf{Z}$ sobre $\mathbf{f}_Y$ y $\hat{\mathbf{\Sigma}}_{\mathbf{res}} = \hat{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}_{\mathbf{fit}}$, siendo $\hat{\mathbf{\Sigma}}$ la matriz de covarianza muestral (marginal) de los predictores. Sin embargo, $\mathbf{Z}$ es una variable latente no observada, y por lo tanto, las matrices de covarianza (marginal y ajustada) no pueden observarse de forma directa. En vista de la robustez probada en Cook and Forzani (2008), podríamos considerar la aplicación de la metodología de PFC directamente sobre $\mathbf{X}$ de manera *naive*, y aún así obtendríamos un estimador $\sqrt{n}$-consistente. Este enfoque constituirá la base para comparar el método para ordinales. También será usado para obtener los valores iniciales del algoritmo propuesto para obtener estimadores de máxima verosimilitud bajo el modelo de variables latentes.

Para estimar los parámetros del modelo (3.3) y en particular los parámetros de la reducción suficiente, en base al Teorema 3.1 usaremos la parametrización $\boldsymbol{\alpha} = \mathbf{\Delta}^{-1}\mathbf{\Gamma}$ de forma tal que (3.3) puede ser re-expresado de la forma

$$\begin{aligned} \mathbf{Z}|Y &= \mathbf{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\{\mathbf{f}_Y - E(\mathbf{f}_Y)\} + \boldsymbol{\epsilon}, \\ \text{con } \ \boldsymbol{\alpha}^T\boldsymbol{\alpha} &= \mathbf{I} \ \text{ y } \ \boldsymbol{\epsilon} \sim N(0, \mathbf{\Delta}), \\ \text{y } \ X_j = g &\Leftrightarrow Z_j \in [\theta^{(j)}_{g-1}, \theta^{(j)}_g),\ j = 1, \ldots, p. \end{aligned} \tag{3.4}$$

Supongamos que tenemos una muestra aleatoria de n puntos $(y_i, \mathbf{x}_i)$ extraídos de la distribución conjunta de $(Y, \mathbf{X})$ tal que satisfacen el modelo (3.4). Necesitamos estimar el span$(\boldsymbol{\alpha})$. Supongamos por el momento que la dimensión d de la reducción es conocida (posteriormente en la Sección 3.5 se tratará la inferencia de d). Para obtener el estimador con los datos observados, necesitamos maximizar la función de log-verosimilitud

$$\sum_{i=1}^{n} \log f_{\mathbf{X}}(\mathbf{x}_i|y_i; \mathbf{\Theta}, \mathbf{\Delta}, \boldsymbol{\alpha}, \boldsymbol{\xi}). \tag{3.5}$$

Sea $C(\mathbf{X}, \mathbf{\Theta})$ el hiper-cubo $C(\mathbf{X}, \mathbf{\Theta}) = [\theta^{(1)}_{X_1-1}, \theta^{(1)}_{X_1}) \times \cdots \times [\theta^{(p)}_{X_p-1}, \theta^{(p)}_{X_p})$. Como para cada $g = 1, \ldots, G_j$, $X_j = g \Leftrightarrow Z_j \in [\theta^{(j)}_{X_j-1}, \theta^{(j)}_{X_j})$ y $\mathbf{Z}|Y$ está distribuida normalmente, para cada i, la

densidad truncada (no normalizada[1]) $f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i, \mathbf{z}_i|y_i; \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\alpha}, \boldsymbol{\xi})$ es

$$f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i, \mathbf{z}_i|y_i; \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\alpha}, \boldsymbol{\xi}) = (2\pi)^{-p/2}|\boldsymbol{\Delta}|^{-1/2} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Delta}^{-1}(\mathbf{z}_i - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i})(\mathbf{z}_i - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i})^T)} I_{\{\mathbf{z}_i \in C(\mathbf{x}_i, \boldsymbol{\Theta})\}},$$

donde $\bar{\mathbf{f}}_{y_i} \doteq \mathbf{f}_{y_i} - n^{-1}\sum_{i=1}^{n} \mathbf{f}_{y_i}$. Por lo tanto, para cada i, la densidad marginal no normalizada de $\mathbf{X}|Y$ será

$$f_{\mathbf{X}}(\mathbf{x}_i|y_i; \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\alpha}, \boldsymbol{\xi}) = \int f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i, \mathbf{z}_i|y_i; \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\alpha}, \boldsymbol{\xi})\, d\mathbf{z}_i.$$

Debido a que el cómputo exacto de la función de verosimilitud resulta muy dificultoso debido a las integrales múltiples contenidas en la misma, en estos casos, los estimadores de máxima verosimilitud comúnmente se obtienen usando un algoritmo iterativo EM (*Expectation-Maximization*). Este alternativa es usual en modelos con variables latentes, debido a que reduce la complejidad de cómputo de la verosimilitud conjunta de $(\mathbf{X}, \mathbf{Z})$. Por ello, adoptaremos esta metodología en el presente capítulo. El algoritmo se expone a continuación.

3.4.1. Algoritmo

En esta sección vamos a presentar el algoritmo EM, estrechamente ligado con el presentado por Guo et al. (2015) en el contexto de modelos gráficos, con el fin de estimar los parámetros del modelo (3.4). A lo largo de esta sección, vamos a usar superíndices de la forma $A^{(k)}$ para indicar el valor de una cierta cantidad A en la k-ésima iteración del algoritmo. El procedimiento comienza con el Paso 0, donde inicializamos los parámetros $\boldsymbol{\Delta}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\xi}^{(0)}$ utilizando los estimadores del método PFC aplicado directamente sobre los predictores ordinales observados $\mathbf{X}$. Luego, el algoritmo itera entre los dos siguientes pasos hasta alcanzar la convergencia: En el Paso 1 se estima $\boldsymbol{\Theta}^{(k)}$ dados $\boldsymbol{\Delta}^{(k-1)}, \boldsymbol{\alpha}^{(k-1)}, \boldsymbol{\xi}^{(k-1)}$; luego en el Paso 2 se obtienen $\boldsymbol{\Delta}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\xi}^{(k)}$ maximizando la esperanza condicional (dados $\boldsymbol{\Delta}^{(k-1)}, \boldsymbol{\alpha}^{(k-1)}, \boldsymbol{\xi}^{(k-1)}$ y $\boldsymbol{\Theta}^{(k)}$) de la función de log-verosimilitud conjunta (3.5). Este último paso se denomina EM.

Paso 1: Estimación de $\boldsymbol{\Theta}$: Dados $\boldsymbol{\Delta}^{(k-1)}, \boldsymbol{\alpha}^{(k-1)}, \boldsymbol{\xi}^{(k-1)}$ del Paso 0 o de un paso anterior,

[1]El término 'no normalizada' se utiliza, debido a que en un sentido estricto, no constituye una densidad al no integrar 1. No obstante, la literatura relacionada hace caso omiso de ello en el sentido de que sigue denominándola densidad.

tomamos $\boldsymbol{\Psi}^{(k-1)} \doteq \boldsymbol{\Delta}^{(k-1)}\boldsymbol{\alpha}^{(k-1)}\boldsymbol{\xi}^{(k-1)}$. Para cada $j = 1, \ldots, p$ y $g = 1, \ldots, G_j$ definimos

$$L_g(\theta) \doteq \#\{i : x_{ij} \leq g\} - \sum_{i=1}^{n} \Phi\left(\frac{\theta - \boldsymbol{\Psi}_j^{(k-1)}\bar{\mathbf{f}}_{y_i}}{\delta_j^{(k-1)}}\right),$$

donde Φ es la función de distribución acumulada de una normal estándar. Para cada j, $\delta_j^{(k-1)} = (\boldsymbol{\Delta}^{(k-1)})_{jj}$, $\boldsymbol{\Psi}_j^{(k-1)}$ indica la j-ésima fila de $\boldsymbol{\Psi}^{(k-1)}$, x_{ij} es la j-ésima coordenada de $\mathbf{x}_i$, y $\#S$ indica el cardinal del conjunto S. Luego tomamos $\hat{\theta}_0^{(j)} = -\infty =$, $\hat{\theta}_{G_j}^{(j)} = +\infty$. Para $g = 1, \ldots, G_j - 1$, asignamos a $\hat{\theta}_g^{(j)}$ la única solución de la ecuación $L_g(\theta) = 0$. De esta manera tomamos $\boldsymbol{\Theta}^{(k)} = \{\theta_0^{(1)}, \ldots, \theta_{G_1}^{(1)}, \ldots, \theta_0^{(p)}, \ldots, \theta_{G_p}^{(p)}\}$. Podemos notar que L_g está basada en el supuesto de normalidad de la distribución condicional de la latente. Precisamente, definimos a L_g como una *función de búsqueda* de los umbrales usando la función de distribución normal subyacente.

Paso 2: Estimación de $\boldsymbol{\Delta}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\xi}^{(k)}$: Dado $\boldsymbol{\Theta}^{(k)}$ calculado en el Paso 1 y $\boldsymbol{\Delta}^{(k-1)}, \boldsymbol{\alpha}^{(k-1)}, \boldsymbol{\xi}^{(k-1)}$ del Paso 0 o de un paso anterior, aplicamos el algoritmo EM para maximizar (3.5). El algoritmo EM consiste en la encontrar $\boldsymbol{\Omega}^{(k)} \doteq (\boldsymbol{\Delta}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\xi}^{(k)})$ tal que maximice

$$Q(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(k-1)}) = \sum_{i=1}^{n} E_{\mathbf{z}_i|\mathbf{x}_i, y_i; \boldsymbol{\Omega}^{(k-1)}}\left[\log f_{\mathbf{x}_i, \mathbf{z}_i}(\mathbf{x}_i, \mathbf{z}_i|y_i; \boldsymbol{\Omega})\big|y_i; \boldsymbol{\Omega}^{(k-1)}\right], \tag{3.6}$$

en $\boldsymbol{\Omega} \doteq (\mathbf{A}, \boldsymbol{\beta}, \boldsymbol{\Delta})$. Esto produce

$$\begin{aligned}
\boldsymbol{\alpha}^{(k)} &= \mathbf{S}^{-1/2}\hat{\boldsymbol{\zeta}}\mathbf{N}, \\
(\boldsymbol{\Delta}^{-1})^{(k)} &= \mathbf{S}^{-1} + \boldsymbol{\alpha}^{(k)}((\boldsymbol{\alpha}^{(k)})^T\mathbf{S}_{\text{res}}\boldsymbol{\alpha}^{(k)})^{-1}(\boldsymbol{\alpha}^{(k)})^T - \boldsymbol{\alpha}^{(k)}((\boldsymbol{\alpha}^{(k)})^T\mathbf{S}\boldsymbol{\alpha}^{(k)})^{-1}(\boldsymbol{\alpha}^{(k)})^T, \\
\boldsymbol{\xi}^{(k)} &= ((\boldsymbol{\alpha}^{(k)})^T\boldsymbol{\Delta}^{(k)}\boldsymbol{\alpha}^{(k)})^{-1}(\boldsymbol{\alpha}^{(k)})^T\mathbf{M}^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1},
\end{aligned}$$

donde $\mathbf{N}$ es una matriz tal que $(\boldsymbol{\alpha}^{(k)})^T\boldsymbol{\alpha}^{(k)} = \mathbf{I}_d$ y $\hat{\boldsymbol{\zeta}}$ son los primeros d autovectores de $\mathbf{S}^{-1/2}\mathbf{S}_{\text{fit}}\mathbf{S}^{-1/2}$, donde las matrices $\mathbf{S} \in \mathbb{R}^{p\times p}$ y $\mathbf{S}_{\text{fit}} \in \mathbb{R}^{p\times p}$ están dadas por

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} E_{\mathbf{z}_i|\mathbf{x}_i, y_i; \boldsymbol{\Omega}^{(k-1)}}(\mathbf{z}_i\mathbf{z}_i^T|\mathbf{x}_i, y_i; \boldsymbol{\Omega}^{(k-1)}) \quad \text{y} \quad \mathbf{S}_{\text{fit}} = n^{-1}\mathbf{M}^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{M}$$

con $\mathbf{F} \in \mathbb{R}^{n\times r}$ y $\mathbf{M} \in \mathbb{R}^{n\times p}$ matrices cuyas traspuestas están dadas por $\mathbf{F}^T = [\bar{\mathbf{f}}_{y_1}, \ldots, \bar{\mathbf{f}}_{y_n}]$ y

$\mathbf{M}^T = [E_{\mathbf{z}_1|\mathbf{x}_1,y_1;\boldsymbol{\Omega}^{(k-1)}}(\mathbf{z}_1|\mathbf{x}_1,y_1;\boldsymbol{\Omega}^{(k-1)}),\ldots,E_{\mathbf{z}_n|\mathbf{x}_n,y_n;\boldsymbol{\Omega}^{(k-1)}}(\mathbf{z}_n|\mathbf{x}_n,y_n;\boldsymbol{\Omega}^{(k-1)}))]$ y la matriz residual $\mathbf{S}_{\text{res}}$ está definida por $\mathbf{S}_{\text{res}} = \mathbf{S} - \mathbf{S}_{\text{fit}}$.

Los detalles del algoritmo EM son presentados en el Apéndice 3.9.1.

Paso 3: Chequeamos convergencia. Si la convergencia no se logra, se retorna al **Paso 1**. La convergencia se chequea simplemente observando si $Q(\boldsymbol{\Omega}^{(k)}|\boldsymbol{\Omega}^{(k-1)})$ deja de incrementarse de una iteración a la siguiente. Específicamente, chequeamos si $(Q(\boldsymbol{\Omega}^{(k)}|\boldsymbol{\Omega}^{(k-1)}) - Q(\boldsymbol{\Omega}^{(k-1)}|\boldsymbol{\Omega}^{(k-2)}))/Q(\boldsymbol{\Omega}^{(k-1)}|\boldsymbol{\Omega}^{(k-2)}) < \epsilon$, con ϵ usualmente fijado en 10^{-6}.

3.4.2. Estimación con selección de variables

Cuando realizamos reducción suficiente de dimensiones con el fin de obtener el menor conjunto de combinaciones lineales de los predictores originales, tal combinaciones lineales típicamente involucran a todas las variables originales. Esto significa que aún las variables no relevantes o redundantes son incluidas en el modelo final, dificultando más la interpretación. Para superar esta limitación, podemos realizar selección de variables y de este modo obtener combinaciones lineales que solo incluyan a las variables activas o relevantes. La maximización de (3.6) es equivalente a maximizar $-\frac{n}{2}\log|\boldsymbol{\alpha}^T\mathbf{S}_{\text{res}}\boldsymbol{\alpha}| - \frac{n}{2}\log|\mathbf{S}| + \frac{n}{2}\log|\boldsymbol{\alpha}^T\mathbf{S}\boldsymbol{\alpha}|$ (ver Apéndice 3.9.1), y Chen et al. (2010) probaron que la maximización de esta última expresión es equivalente a encontrar,

$$\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\alpha}} \big\{ -\ \text{tr}(\boldsymbol{\alpha}^T\mathbf{S}_{\text{fit}}\boldsymbol{\alpha})\big\}, \qquad \text{sujeto a } \boldsymbol{\alpha}^T\mathbf{S}\boldsymbol{\alpha} = \mathbf{I}_d. \tag{3.7}$$

Con esta equivalencia podemos inducir selección de variables en reducción de dimensiones, agregando una penalización a dicha función. Esta penalización es del tipo *group-lasso* ya que, con el fin de no elegir una variable X_j en particular, necesitamos hacer que toda la fila j-ésima de $\boldsymbol{\alpha}$ se iguale a 0. Por ello, siguiendo a Chen et al. (2010), utilizamos una norma mixta tipo ℓ_1/ℓ_2, donde la norma interna es la norma ℓ_2 de cada fila de $\boldsymbol{\alpha}$. Incorporando el término de penalización a (3.7), obtenemos

$$\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\alpha}} \Big\{ -\ \text{tr}(\boldsymbol{\alpha}^T\mathbf{S}^{-1/2}\mathbf{S}_{\text{fit}}\mathbf{S}^{-1/2}\boldsymbol{\alpha}) + \lambda\sum_{i=1}^{p}||\boldsymbol{\alpha}_i||_2\Big\}, \qquad \text{sujeto a } \boldsymbol{\alpha}^T\mathbf{S}\boldsymbol{\alpha} = \mathbf{I}_d.$$

El parámetro λ puede seleccionarse utilizando algún criterio de información, como ser el criterio de Akaike (AIC) o el de Bayes (BIC). Los detalles pueden encontrarse en Chen et al. (2010). Otra alternativa es encontrar el valor de λ que minimiza el error de predicción vía un experimento de validación cruzada, pero esto requiere la adopción de alguna regla de predicción.

Cabe destacar que este procedimiento realiza al mismo tiempo selección de variables y reducción suficiente de dimensiones sin necesidad de especificar un modelo para $Y|\mathbf{X}$. Entonces, la reducción obtenida puede utilizarse luego con cualquier regla de predicción. Este enfoque difiere de aquellos en los que el procedimiento de selección de variables hace uso de un modelo de regresión en particular, como por ejemplo en Gertheiss and Tutz (2010).

3.5. Elección de la dimensión d

En el desarrollo de la Sección 3.4 suponemos que la dimensión $d \leq \text{mín}(p, r)$ del subespacio de reducción suficiente era conocida. Si bien para la construcción de un índice unitario de Estatus Socioeconómico se fija $d = 1$, en otras aplicaciones prácticas, esta dimensión debería ser inferida a partir de los datos. Para el caso de reducción suficiente de dimensiones basada en modelos que admite el cálculo de la verosimilitud, para seleccionar d se ha propuesto el uso de cocientes de verosimilitud o de algún criterio de información tipo AIC o BIC (Cook and Forzani 2008). Sin embargo, estos métodos no son robustos ante desviaciones del modelo que se está asumiendo.

Cuando el principal objetivo es la predicción, un alternativa es evaluar diferentes valores de d en función de su desempeño predictivo fuera de la muestra, en un entorno de validación cruzada. Luego, el valor de d seleccionado es aquel que logra el mínimo error de predicción.

Otro camino para elegir la dimensión es a través de test de permutación, como lo hacen Cook and Yin (2001). Para ello, supongamos que $\boldsymbol{\alpha} \in \mathbb{R}^{p\times m}$ y $(\boldsymbol{\alpha}, \boldsymbol{\alpha}_0) \in \mathbb{R}^{p\times p}$ es unitario. El test de permutación se basa en el hecho de que $\mathcal{S}_{\boldsymbol{\alpha}}$ es un subespacio de reducción suficiente para la regresión de Y sobre $\mathbf{X}$ siempre que $(Y, \boldsymbol{\alpha}^T\mathbf{X}) \perp\!\!\!\perp \boldsymbol{\alpha}_0^T\mathbf{X}$. Notemos que esto implica que $m \geq d$. Para este test, consideraremos el estadístico $\hat{\Lambda}_m = 2(\mathcal{Q}_p(\mathbf{I}_p) - \mathcal{Q}_m(\widehat{\boldsymbol{\alpha}}))$, donde $\mathcal{Q}_r$ está dado por la función Q en (3.6), evaluada en el estimador obtenido vía el algoritmo EM previamente

descrito, para una dimensión fija de $\mathcal{S}_{\boldsymbol{\alpha}}$. Fijamos $m = 0$. El procedimiento para inferir d vía test de permutación involucra los siguientes pasos:

(i) Obtener $\widehat{\boldsymbol{\alpha}}$, el estimador de máxima verosimilitud (MV) de $\boldsymbol{\alpha}$, y computar $\hat{\Lambda}_m$. Obtener también $\widehat{\boldsymbol{\alpha}}_0$.

(ii) Para los datos $(Y_i, \widehat{\boldsymbol{\alpha}}^T\mathbf{X}_i, \widehat{\boldsymbol{\alpha}}_0^T\mathbf{X}_i)$, permutamos las columnas correspondientes a $\widehat{\boldsymbol{\alpha}}_0^T\mathbf{X}_i$ para obtener una nueva muestra $(Y_i, \mathbf{X}_i^*)$. Para cada nuevo conjunto de datos, obtenemos el estimador de MV y computamos $\hat{\Lambda}_m^*$.

(iii) Repetimos el paso (ii) B veces.

(iv) Calculamos la fracción de veces en que $\hat{\Lambda}_m^*$ excede a $\hat{\Lambda}_m$. Si este valor es mas pequeño que el nivel de significancia elegido y si $m < \text{mín}(r, p)$ tomamos $m = m + 1$ y volvemos al paso (i). En caso contrario, tenemos $d = m$.

De esta manera, el valor de d inferido es la menor m que falla en rechazar la hipótesis nula de independencia entre $(Y, \widehat{\boldsymbol{\alpha}}^T\mathbf{X})$ y $\widehat{\boldsymbol{\alpha}}_0^T\mathbf{X}$.

3.6. Simulaciones

En esta sección vamos a evaluar diferentes aspectos del método propuesto usando datos simulados. Un punto crítico de la implementación del método es el cómputo del paso E del algoritmo EM. El cálculo exacto de los momentos trucados involucrados en este paso hace que el método propuesto no sea factible en la práctica, aún cuando la dimensión de los predictores es de orden $5 < p < 10$, dependiendo del número de categorías de cada uno de ellos. Por ello, implementamos un método de estimación aproximado adaptado del propuesto por Guo et al. (2015). La idea principal de esta aproximación es el uso recursivo para computar iterativamente los momentos truncados de una distribución normal multivariada. La derivación y demás detalles están presentados en el Apéndice 3.9.2. Por lo tanto, el primer paso es validar este método de aproximación utilizado en el paso E, comparándolo con el cómputo exacto de los momentos truncados a fin de evaluar la desempeño del mismo. Posteriormente, vamos a comparar el desempeño del método de SDR propuesto respecto a la metodología estándar desarrollada para

datos continuos. Luego vamos a ilustrar el desempeño de la estrategia propuesta para inferir la dimensión del subespacio de reducción vía test de permutación y validación cruzada. Finalmente, evaluamos el funcionamiento del estimador regularizado propuesto en la Sección 3.4.2 en términos del poder predictivo.

3.6.1. Validación del paso E del algoritmo

La parte más demandante del método propuesto es el cómputo de los momentos truncados de la distribución normal multivariada del paso E. El método de aproximación iterativa sugerido para este cómputo, constituye el ingrediente principal para que la metodología de reducción presentada se aplicable en la práctica. En esta sección validaremos esta estrategia, comparando el cómputo aproximado *vis à vis* el exacto, utilizando para ello el algoritmo propuesto por Lee and Scott (2012). Dado que el cómputo exacto involucra un alto costo computacional, aún para una dimensión moderada de los predictores, tomamos $p = 5$ y $n = 100$. Adicionalmente, fijamos $G_j = 4$ para $j = 1, 2, \ldots, 5$. Los datos son generados de acuerdo al modelo (3.4), con $Y \sim N(0, 1)$. Para la matriz de base $\boldsymbol{\alpha}$, tomamos $\sqrt{p}\ \boldsymbol{\alpha} = (\mathbf{1}_p \quad \text{sign}(\mathbf{e}))$, con $\mathbf{1}_p$ un vector columna de unos, de tamaño $p = 5$, y $\mathbf{e} \sim N_p(\mathbf{0}, \mathbf{I})$. Para la matriz de covarianza $\boldsymbol{\Delta}$, tomamos $\boldsymbol{\Delta} = \mathbf{I} + \boldsymbol{\alpha}\mathbf{B}\boldsymbol{\alpha}^T$ siendo $\mathbf{B}$ una matriz aleatoria simétrica 2×2 fija desde el principio. También elegimos una base polinómica para $\mathbf{f}_Y$, con $r = 2$. La misma elección de $\mathbf{f}_Y$ fue usada para la estimación. El experimento es replicado 100 veces. En cada corrida, la misma muestra de entrenamiento es usada por ambos métodos (i.e. aproximado y exacto).

Para evaluar la exactitud de la estimación, medimos el ángulo entre los subespacios generados por el verdadero $\boldsymbol{\alpha}$ y el estimado $\widehat{\boldsymbol{\alpha}}$. Esta cantidad varía de 0 grados, si los dos subespacios son idénticos, a los 90 grados, si no comparten ninguna información. El ángulo medio obtenido con el método exacto fue de 10.21 grados con una desviación estándar de 6.37 grados, mientras que para la estimación obtenida con el método aproximado, el ángulo promedio fue de 12.65 grados con una desviación estándar de 5.70 grados. El intervalo del 95 % de confianza para la diferencia promedio entre los ángulos obtenidos con ambos métodos es igual a (2.07,2.81) grados. Estos valores sugieren que el precio a pagar por la introducción del cálculo aproximado es muy pequeña.

También resulta ilustrativo ver el impacto de la aproximación en predicción. Usando regresión lineal estándar para $Y|\widehat{\boldsymbol{\alpha}}^T\mathbf{X}$, el error cuadrático medio (MSE) de los residuos, promediados sobre las 100 corridas, es de 0.810 cuando estimamos la reducción usando la aproximación en el paso E, y es igual a 0.801 cuando se utiliza el método exacto. En ambos casos, la desviación estándar del MSE promedio es de 0.02. Por lo tanto, la diferencia en el MSE obtenido con el método aproximado representa menos del 1.5 % del MSE promedio obtenido con el método exacto.

La importancia del método aproximado para el paso E se resalta mejor cuando se toma nota de la gran diferencia en el tiempo de cálculo. Usando implementaciones en MATLAB para ambos métodos, el cálculo con el método exacto requiere un promedio de 1.13×10^2 segundos para cada ejecución, mientras que con el método aproximado este tiempo se redujo a 0.26 segundos, una diferencia de tres órdenes de magnitud.

En general, estos resultados muestran que el método aproximado para calcular los momentos truncados resulta ser una alternativa viable: se reduce el tiempo de cálculo para la aplicación práctica, sin una pérdida significativa de precisión.

3.6.2. Desempeño del método propuesto

En esta sección evaluamos el desempeño del método propuesto con datos simulados. Para este caso, fijamos $p = 20$, $d = 2$ y una base polinómica con $r = 2$ para $\mathbf{f}_Y$. Al igual que en la Sección 3.6.1, generamos los datos de acuerdo al modelo (3.4), con $Y \sim N(0,1)$, $\sqrt{p}\,\boldsymbol{\alpha} = (\mathbf{1}_p \;\; \text{sign}(\mathbf{e}))$, con $\mathbf{e} \sim N_p(\mathbf{0}, \mathbf{I})$, y matriz de varianza-covarianza $\boldsymbol{\Delta} = \mathbf{I} + \boldsymbol{\alpha}\mathbf{B}\boldsymbol{\alpha}^T$, con $\mathbf{B}$ una matriz aleatoria $d \times d$ simétrica, que dejamos fija al inicio. Los valores de G_j en este caso van desde 3 hasta 5. Para evaluar el desempeño del método, calculamos el ángulo entre los subespacios generados por el verdadero $\boldsymbol{\alpha}$ y el $\widehat{\boldsymbol{\alpha}}$ estimado.

La Figura 3.1-(a) muestra los diagramas de caja obtenidos del experimento replicado 100 veces con un tamaño de muestra $n = 500$. En el mismo se compara el ángulo obtenido cuando $\widehat{\boldsymbol{\alpha}}$ es computado utilizando el PFC estándar (para datos continuos) sobre las variables observadas $\mathbf{X}$, respecto de aquel obtenido cuando usamos el método propuesto para ordinales (que a partir de aquí lo denotaremos con PFCord). Los ángulos están medidos en grados. Podemos observar que el valor medio del ángulo es significativamente menor cuando utilizamos el método

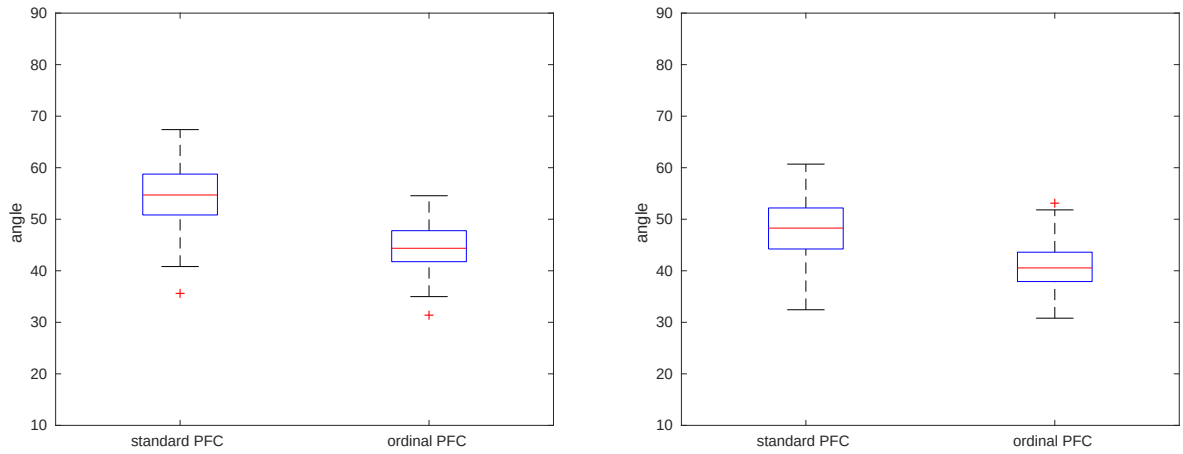
de reducción propuesto para ordinales; y ambos métodos muestran una varianza similar. El intervalo de confianza (normal) del 95 % para la diferencia entre los ángulos obtenidos por ambos estimadores es (9.59, 11.02), en grados. Estos resultados muestran que la estimación usando el PFCord claramente mejora respecto a la estimación obtenida con el PFC estándar usando los datos ordinales como si fueran continuos.

Algo que podríamos preguntarnos en el presente contexto, es si esta ganancia en términos del desempeño del estimador se mantiene luego de relajar el supuesto de normalidad de las variables latentes subyacentes. Para el PFC, Cook and Forzani (2008, Teorema 3.5) mostraron que el estimador es aún consistente cuando $\mathbf{X}|Y$ se desvía de la normalidad. En nuestro caso, para evaluar el método propuesto en tal sentido, generamos ahora los datos como lo hicimos anteriormente, pero tomando un $\boldsymbol{\epsilon}$ que no se distribuye normalmente. En particular, suponemos que $\boldsymbol{\epsilon}$ tiene una distribución chi-cuadrado de la forma $(\boldsymbol{\epsilon})_j \sim \chi^2(5)$. El resto de los parámetros de la simulación permanece como antes. Los resultados obtenidos se muestran en la Figura 3.1-(b). Puede observarse que los ángulos obtenidos con ambos métodos son muy similares a los obtenidos bajo normalidad condicional de los datos. En este caso, el intervalo de confianza del 95 % para la diferencia entre los ángulos obtenidos luego de la estimación está dado por (6.76, 8.28), mostrando así que las ventaja de usar el método diseñado para datos ordinales se mantiene, y la diferencia sigue siendo significativa.

3.6.3. Inferencia sobre la dimensión d

En esta sección llevamos a cabo un estudio de simulación para evaluar los métodos para inferir la dimensión del subespacio de reducción a partir de los datos. En particular, vamos a comparar la inferencia de d a partir de método de test de permutación con el de validación cruzada de 10 particiones (10-fold CV). Para esto, generamos los datos de la misma forma que en la Sección 3.6.2, con $p = 10$, $d = 2$ y usamos una base polinomial de grado $r = 4$ para $\mathbf{f}_Y$. Puesto que $d \leq \min(r, p)$, el valor verdadero de d lo debemos buscar dentro del conjunto $\{0, 1, 2, 3, 4\}$. Para el test de permutación, construimos la distribución de $\hat{\Lambda}_m$ extrayendo muestras a partir de los datos (re-muestreo) 500 veces, como fue discutido en la Sección 3.5, y usando un nivel de significancia de 0.01. Para validación cruzada usamos el promedio del error cuadrático medio de predicción sobre la partición de prueba como la medida de desempeño a seguir, utilizando

Figura 3.1: Desempeño del Estimador PFCord: $\measuredangle(\text{span}(\widehat{\boldsymbol{\alpha}}),\text{span}(\boldsymbol{\alpha}))$. (a) modelo normal para $\mathbf{Z}|(Y=y)$; (b) modelo no normal para $\mathbf{Z}|(Y=y)$.



(a) (b)

como regla de predicción el método de los k vecinos más cercanos (i.e. k-NN). El experimento fue repetido utilizando dos tamaños de muestra $n = 200$ y $n = 300$, y se corrió 500 veces para cada tamaño de muestra. La Tabla 3.1 muestra los resultados obtenidos. Para $n = 300$, el test de permutación encuentra la verdadera dimensión en el 81 % de los experimentos replicados, mientras que con validación cruzada se encuentra el 58 % de las veces. El intervalo de confianza del 95 % para la diferencia en la proporción de elección correcta de la dimensión es (0.177, 0.291), mostrando que dicha diferencia es estadísticamente significativa para el número utilizado de réplicas. Además, para este tamaño de muestra, ambos métodos casi nunca recogen menos de dos direcciones, y por lo tanto no se pierde ninguna información. La fracción de veces que al menos una dirección adicional es elegida, esto es, $\hat{d} = 2$ o $\hat{d} = 3$, es 0.898 con el test de permutación y 0.81 usando validación cruzada. El intervalo de confianza para esta diferencia es de (0.043, 0.133), confirmando la ventaja del test de permutación cuando el tamaño de la muestra es suficientemente grande. Asimismo, debemos notar que el procedimiento basado en permutación es más rápido que validación cruzada, ya que la reducción de la dimensión de los datos transformados permutados puede llevarse a cabo usando métodos estándares para datos continuos. De hecho, utilizando una PC estándar para los experimentos, el procedimiento del test de permutación es casi seis veces más rápido que el procedimiento requerido para validación cruzada.

Para un tamaño de muestra menor, específicamente tomando $n = 200$, el método de permutaciones es menos preciso. En este caso se encuentra la verdadera dimensión el 69 % de las veces, pero en el 22 % de las réplicas se recogió una sola dirección para la proyección en lugar de dos. Por su parte, con validación cruzada la verdadera dimensión se encontró el 57 % de las veces, pero tiende a sobrestimar la dimensión requerida d, generando con ello una potencial pérdida de eficiencia a favor de preservar la información. Realizando un contraste para la diferencia en la proporción de elecciones correctas de $\hat{d} \geq 2$, obtenemos un p-valor de $\approx 10^{-15}$, evidenciando así una ventaja estadísticamente significativa del método de validación cruzada, en este escenario con un tamaño de muestra pequeño, al ser más conservativo en términos de la pérdida de la información. Por lo tanto, podemos concluir que el test de permutaciones parece ser un mejor procedimiento para inferir la dimensión d de la reducción cuando el tamaño de la muestra es suficientemente grande, a la vez que la reducción del tiempo computacional es significativa. Sin embargo, con validación cruzada se obtiene una solución más segura, en términos de pérdida

Tabla 3.1: Fracción de veces en las que se elige un valor de d.

		Permutación	Validación Cruzada
$n = 200$	$d = 1$	0.220	0.000
	$d = 2$	0.690	0.572
	$d = 3$	0.078	0.220
	$d = 4$	0.012	0.208
$n = 300$	$d = 1$	0.082	0.000
	$d = 2$	0.810	0.576
	$d = 3$	0.088	0.223
	$d = 4$	0.020	0.190

de información, cuando la cantidad de datos es reducida.

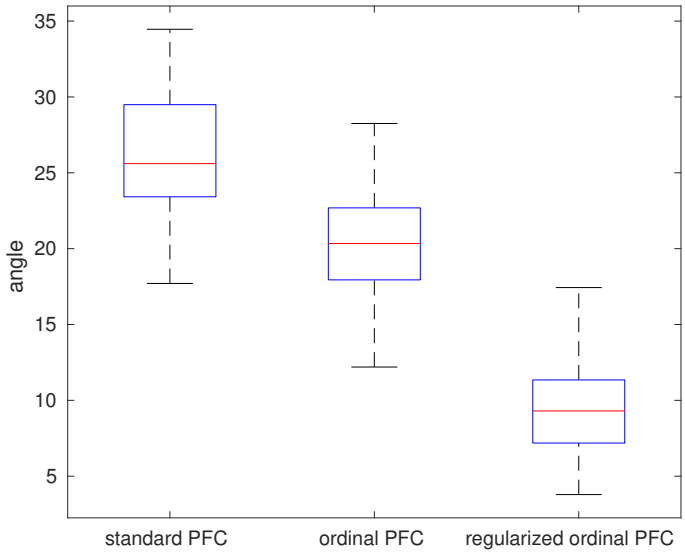
3.6.4. Desempeño del método incluyendo regularización

Finalmente, vamos a realizar un estudio de simulación para evaluar el desempeño de la versión regularizada (i.e. con selección de variables) del método propuesto para predictores ordinales. A diferencia de las configuraciones anteriores, ahora la reducción suficiente depende sólo de un subconjunto de predictores. En la presente simulación, solamente los primeros cuatro predictores aportan información acerca de la respuesta, esto es, $\boldsymbol{\alpha} = (\mathbf{A} \;\; \mathbf{0}_{2\times p-4})^T$, con

$$\mathbf{A} = \begin{pmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ -1/2 & 1/2 & 1/2 & -1/2 \end{pmatrix}.$$

Fijamos los valores para el resto de los parámetros de la misma forma que se hizo en la Sección 3.6.2. La reducción fue estimada usando PFC estándar para predictores continuos, la versión no regularizada del método de SDR, que denominamos PFCord, y la versión regularizada que propusimos en la Sección 2.4.2. En todos los casos hemos usado una base polinomial para $\mathbf{f}_Y$ de grado $r = 2$. Para cada estimador, calculamos el ángulo entre los subespacios $\mathcal{S}_{\boldsymbol{\alpha}}$ y $\mathcal{S}_{\hat{\boldsymbol{\alpha}}}$ en cada una de las 100 corridas del experimento. La Figura 3.2 muestra los resultados obtenidos para $p = 10$. Podemos ver que los ángulos obtenidos son menores que en el caso donde asumimos que todos los predictores son relevantes. Dado que todos los métodos se aplican a los mismos datos, del diagrama de caja queda claro que los estimadores diseñados exclusivamente para datos ordinales tienen un mejor desempeño que el PFC estándar para datos continuos. Por otra

Figura 3.2: Desempeño del estimador REG-PFCORD: $\measuredangle(\text{span}(\widehat{\boldsymbol{\alpha}}), \text{span}(\boldsymbol{\alpha}))$



parte, para esta situación, el estimador regularizado (que de ahora en adelante denominaremos REG-PFCORD) demuestra claramente ser superior a la versión no regularizada (PFCORD).

También realizamos otra serie de experimentos con el fin de evaluar la estabilidad del subconjunto de variables elegidas por el algoritmo del método regularizado propuesto. Denotamos por $S_0 \subset \{1, 2, \ldots, p\}$ al conjunto de índices para el subconjunto de variables que son realmente relevantes para la respuesta Y (variables activas), y por S_0^c a su complemento. De forma similar, denotamos por $\widehat{S}$ al subconjunto de variables elegidas por el estimador regularizado, en el sentido de que $\boldsymbol{\alpha}_j = \mathbf{0}$ para $j \notin \widehat{S}$, y por $\mathbf{Z}_{S_0}$ a un vector aleatorio con elementos $Z_j \neq 0$ para $j \in S_0$, realizando una definición similar para $\mathbf{Z}_{S_0^c}$. Estamos interesados en la evaluación de: (i) $\Pr(S_0 \subset \widehat{S})$, como un indicador de las variables relevantes que se conservan; (ii) la cardinalidad promedio del conjunto $\widehat{S}$ ($\#\widehat{S}$), como medida de la cantidad de variables no relevantes que son preservadas luego de la estimación. Por otra parte, estamos interesados en la evaluación de cómo varían estas medidas de performance para diferentes niveles de correlación entre $\mathbf{Z}_{S_0}$ y $\mathbf{Z}_{S_0^c}$. Para medir esta dependencia, usamos una medida de correlación basada en distancias como es definida en Székely et al. (2007) y Székely and Rizzo (2009), que aquí denotamos por $dCor(\mathbf{Z}_{S_0}, \mathbf{Z}_{S_0^c})$. Esta es una medida de correlación no paramétrica generalizada que resulta adecuada para vectores aleatorios de diferente tamaño y que no requiere ajuste de ningún

Tabla 3.2: Desempeño de los algoritmos de selección de variables cuando utilizamos REG-PFCORD o REG-PFC.

n	ρ	REG-PFCORD		REG-PFC	
		$\Pr(S_0 \subseteq \widehat{S})$	$\#\widehat{S}$	$\Pr(S_0 \subseteq \widehat{S})$	$\#\widehat{S}$
200	0.0	0.98	3.98 ($\pm$ 0.10)	0.92	3.88 ($\pm$ 0.63)
	0.2	0.91	3.91 ($\pm$ 0.27)	0.13	2.36 ($\pm$ 0.76)
	0.3	0.55	3.42 ($\pm$ 0.64)	0.04	2.23 ($\pm$ 0.46)
	0.5	0.16	2.91 ($\pm$ 0.67)	0.01	2.07 ($\pm$ 0.38)
500	0.0	1.00	4.00 ($\pm$ 0.00)	1.00	4.00 ($\pm$ 0.00)
	0.2	1.00	4.01 ($\pm$ 0.10)	0.96	4.52 ($\pm$ 0.73)
	0.3	0.96	4.02 ($\pm$ 0.38)	0.63	3.52 ($\pm$ 0.98)
	0.5	0.49	3.95 ($\pm$ 1.08)	0.17	2.53 ($\pm$ 0.83)

parámetro para su cálculo. Para controlar esta cantidad, ajustamos el valor de $\boldsymbol{\Delta}$ usado para generar los datos. En particular, fijamos $\boldsymbol{\Delta} = 4\mathbf{I}_p + \rho\boldsymbol{\alpha}\mathbf{B}\boldsymbol{\alpha}^T$ para diferentes valores de ρ. Con esto obtenemos $dCor_n(\mathbf{Z}_{S_0}, \mathbf{Z}_{S_0^c}) \approx 0.3$ para $\rho = 0.2$, 0.5 para $\rho = 0.3$ y aproximadamente 0.7 para $\rho = 0.5$. En todos los experimentos, el número de variables relevantes en S_0 se fijó en 4, con $p = 20$.

En la Tabla 3.2 se muestran los resultados obtenidos usando 100 réplicas del experimento, para cada condición evaluada. De la misma se observa que, para un tamaño de muestra suficientemente grande ($n = 500$), las versiones penalizadas de los métodos PFCORD y PFC logran una precisión perfecta en la selección del verdadero conjunto activo de variables, cuando los predictores no están correlacionados entre sí. Para un nivel moderado de correlación entre los predictores ($\rho = 0.3$), usando REG-PFCORD, el verdadero conjunto de variables activas S_0, está contenido en la solución el 96 % de las veces, con una fracción muy baja de asignaciones falsas. Por otro lado, si en el mismo escenario utilizamos la versión regularizada del PFC estándar (que denotamos por REG-PFC), observamos que dicho método permite recoger el verdadero conjunto de variables activas sólo el 63 % de las veces. Esta diferencia es estadísticamente significativa a un nivel del 0.001 (p-valor). Por otra parte, el número promedio de variables escogidas por el REG-PFC es de aproximadamente 3.50, lo que significa que algo de información se pierde con este procedimiento. Para niveles elevados de correlación entre los predictores ($\rho = 0.5$), el verdadero conjunto de variables activas es escogido sólo el 17 % de las veces cuando usamos REG-PFC, mientras REG-PFCORD aún lo escoge la mitad de las veces. La pérdida de precisión del REG-PFCORD está relacionada con el reemplazo de uno de los predictores del verdade-

ro conjunto de variables activas, por otro predictor altamente correlacionado, manteniendo el cardinal promedio del conjunto estimado $\widehat{S}$ con un valor cercano a 4.

Cuando el tamaño de la muestra es menor (i.e. $n = 200$), el desempeño del REG-PFC para seleccionar variables decae mucho más rápido respecto a su equivalente ordinal, a medida que aumenta la correlación entre los predictores. Cuando los predictores no están correlacionados, REG-PFC escoge el verdadero conjunto de variables activas S_0 un 92 % de las veces, mientras que el REG-PFCORD lo hace un 98 % de las veces. Pero con un nivel bajo de correlación entre los predictores ($\rho = 0.2$), el desempeño del REG-PFC decrece rápidamente a un nivel del 13 %, mientras que el método para variables ordinales sigue mostrando una precisión elevada, eligiendo el verdadero conjunto de variables activas en más del 90 % de las veces. A su vez, para niveles elevados de correlación, ambos procedimientos tienden a subestimar el número de variables relevantes; y esta tendencia es mayor para el REG-PFC.

Por lo tanto, estos resultados muestran que el uso del método propuesto especialmente adaptado para datos ordinales, proporciona una precisión significativamente mayor también en este caso, cuando se decide realizar reducción suficiente de dimensiones con selección de variables.

3.7. Análisis con datos reales

En esta sección vamos a aplicar el método propuesto en datos reales, analizando su desempeño y comparando con otros métodos alternativos supervisados y no supervisados.

En la primera sección se presenta la aplicación para la construcción de índices SES a lo fines de predecir ingreso per cápita y nivel de pobreza (monetaria) utilizando una muestra de entrenamiento. La segunda aplicación, analiza el poder predictivo de este método en el problema de puntuación y recomendación de películas de NETFLIX.

3.7.1. Construcción de índices de Estatus Socio-Económico (SES)

Como mencionamos en el Capítulo 1, existen varias aplicaciones por las que se busca un índice SES a partir de reducir un conjunto de variables económicas y sociales. En este capítulo,

como a lo largo de la tesis, la aplicación que se presenta de índices SES aborda el problema de predecir el ingreso per cápita del hogar y/o el nivel de pobreza de los hogares a partir de un índice SES construido aplicando reducción suficiente a un conjunto de variables observables que caracterizan el nivel social y económico del hogar. A pesar de sus limitaciones, como ya mencionamos en el Capítulo 1, el ingreso o el gasto de consumo constituyen las medidas tradicionales en los análisis de pobreza, lo que ha llevado a gran parte de los países del mundo a adoptar el enfoque de la línea de pobreza basada en el ingreso a partir de encuestas de hogares, con el objeto de inferir sobre la situación socioeconómica de la población (Mokomane 2012, Richardson and Bradshaw 2012).

Debido a la complejidad que encierra la captación correcta del nivel de ingreso, se busca una variable *proxy* del ingreso a partir de variables más fáciles de observar; esto es, un índice SES.

La principal metodología que se utilizada para obtener un índice SES, es el Análisis de Componentes Principales (PCA) (Merola and Baulch 2014, Hoque 2014). Dado que entre las variables observadas para construir el índice predominan las categóricas ordinales, es que Kolenikov and Angeles (2009) proponen una variante del PCA adaptada para datos ordinales utilizando correlaciones policóricas entre los predictores en lugar de la matriz de covarianza estándar.

En este ejemplo, vamos a mostrar un enfoque diferente para la construcción del índices SES, basado en la metodología propuesta de SDR para datos ordinales. La idea es obtener un índice para predecir una medida unidimensional de algún aspecto socio-económico, tal como el ingreso del hogar o la condición de pobreza en términos de ingresos. Por lo tanto, en este caso fijamos la dimensión de la reducción en $d = 1$ y derivamos el índice como una versión normalizada de la reducción suficiente $\widehat{\boldsymbol{\alpha}}^T\mathbf{X} \in \mathbb{R}$. Contrariamente a los índices tipo PCA, este nuevo enfoque usa la información acerca de la respuesta bajo análisis.

Los datos utilizados provienen de la base de microdatos de la *Encuesta Permanente de Hogares* (EPH) de Argentina, tomando, en particular, el cuarto trimestre de 2013. La EPH es la principal encuesta de hogares de Argentina y es realizada por el *Instituto Nacional de Estadísticas y Censos* (INDEC).

Consideramos nueve variables ordinales sobre las condiciones de vida de los hogares, y dos variables socio-económicas sobre el jefe/a de hogar (nivel de instrucción formal y situación laboral). Los detalles de estas variables pueden encontrarse en el Apéndice 3.9.3. A fin de tener en cuenta la heterogeneidad regional, estimamos índices SES de forma separada para cada una de las siguientes cinco regiones: región metropolitana del Gran Buenos Aires ($n = 2351$ hogares), región Pampeana ($n = 5003$), el Noroeste Argentino (NOA) ($n = 2852$), el Noreste Argentino (NEA) ($n = 1594$), y la Patagonia ($n = 2398$). Consideramos dos tipos de respuesta: una continua, dada por el ingreso per cápita familiar (`ipcf`), y una binaria basada en el ingreso (`pobreza`) que indica si el hogar está por encima o debajo de la linea de pobreza (i.e. si el hogar es pobre o no, en términos de ingresos). Estamos interesados en mostrar que el método REG-PFCORD constituye una alternativa superior a la metodologías tipo PCA para la construcción de índices SES, al mismo tiempo que logra tener un poder predictivo similar a considerar el conjunto entero de predictores (i.e. sin aplicar reducción). Para ello, el desempeño en predicción del índice propuesto (REG-PFCORD) se compara con las siguientes estrategias metodológicas:

- Consideración de todo el conjunto de predictores sin aplicar reducción, tratados como predictores continuos, en el sentido métrico. Vamos a denominar a este método FULL.
- Todo el conjunto de predictores incluidos sin aplicar reducción, incorporados por medio de variables *dummies*. Vamos a llamar a este enfoque FULL-I.
- Todo el conjunto de predictores es incluido pero usando un procedimiento tipo *group-lasso* para predictores ordinales (Gertheiss and Tutz 2010), realizando con ello selección de variables. A este método lo llamaremos LASSOORD.
- Una variante del PCA diseñada para predictores ordinales usando correlaciones policóricas (Kolenikov and Angeles 2009). Lo denominaremos PCAPOLY.
- Una variante no lineal del PCA que utiliza un escalamiento especial para aplicarse a categorías ordinales (Linting and van der Kooij 2009). A este método lo llamaremos NLPCA.

Las primeras tres estrategias son incluidas con el objetivo de tener una base de referencia sobre el desempeño que se obtiene cuando se usa el conjunto entero de predictores, pero debe quedar

claro que las mismas no brindan un índice. De hecho, solamente las dos últimas alternativas de la lista compiten en la construcción del índice SES con la metodología propuesta.

Para cada estrategia, ajustamos una regresión logística para la respuesta `pobreza` (discreta) y una regresión lineal para la variable respuesta `ipcf` (continua).

Cuando computamos la reducción suficiente, elegimos una $\mathbf{f}_Y$ diferente para cada respuesta. Para la respuesta continua, usamos una base polinómica de grado $r = 2$. Para la respuesta binaria, $\mathbf{f}_Y$ es simplemente una variable indicadora centrada.

Los datos son particionados en diez conjuntos disjuntos a fin de tener diez réplicas experimentales. En cada corrida, uno de los subconjuntos es usado como conjunto de prueba, mientras que el resto de ellos conforman la muestra de entrenamiento. Con cada método se obtiene el error cuadrático medio (MSE) de realizar validación cruzada con diez iteraciones (10-fold cross-validation). Los resultados se muestran en la Tabla 3.3, reportándose también los correspondientes desvíos estandar.

En primer lugar, de la tabla puede observarse que, para la respuesta continua, usar variables dummy para el conjunto entero de predictores (FULL-I) o realizar selección de variables con LASSOord, constituyen estrategias más efectivas que considerar el conjunto entero de predictores como variables continuas (FULL). Lo contrario ocurre cuando consideramos la respuesta binaria; esto es FULL en general brinda mejores resultados predictivos que FULL-I y LASSOord.

Para los índices SES, tanto para la respuesta continua como discreta, los resultados muestran que reg-PFCord es superior que PCApoly y NLPCA. A su vez, entre los índices tipo PCA, NLPCA muestra un mejor desempeño que PCApoly. Por otra parte, los errores de predicción obtenidos con el índice reg-PFCord son muy similares a aquellos que surgen de la predicción con FULL. Cuando consideramos la respuesta discreta, el reg-PFCord da mejores resultados predictivos que el LASSOord para tres de las regiones consideradas. También debemos remarcar que, contrariamente al LASSOord, los índices obtenidos usando la técnica de reducción suficiente de dimensiones, nos permite utilizarlos con cualquier método predictivo.

Para ilustrar mejor el ajuste obtenido, la Figura 3.3 muestra gráficamente el resultado del modelo de regresión de `ipcf` sobre el índice SES obtenido usando todos los datos. Un término

Tabla 3.3: MSE para el índice SES (10-fold cross-validation)

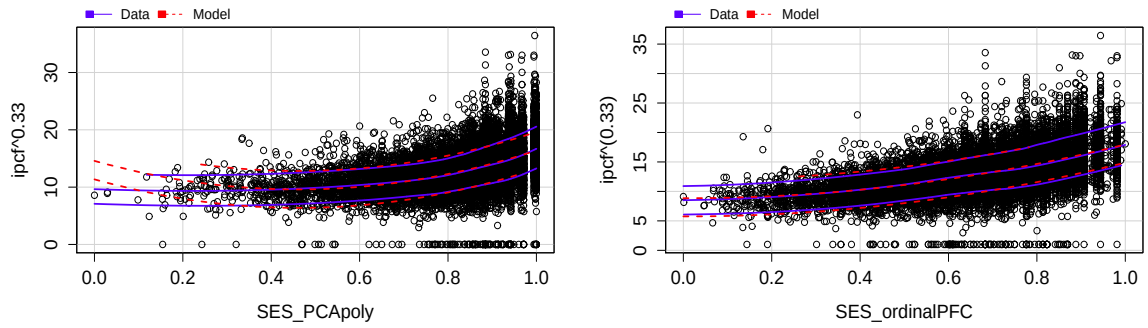
		Errores de Predicción -MSE				
Respuesta	Método	*GBA*	*Pampeana*	*NOA*	*NEA*	*Patagonia*
Ingreso per cápita	REG-PFCORD	7.26	4.73	4.71	3.30	13.20
(`ipcf`: continua)		(2.75)	(1.66)	(2.75)	(1.35)	(3.34)
	PCAPOLY	7.60	5.10	5.07	3.68	14.7
		(2.45)	(0.90)	(1.77)	(0.90)	(4.01)
	NLPCA	7.38	4.95	4.89	3.52	13.67
		(2.29)	(0.61)	(1.48)	(0.65)	(3.71)
	FULL	7.22	4.69	4.68	3.32	13.14
		(3.50)	(0.88)	(2.47)	(0.93)	(3.34)
	FULL-I	7.01	4.52	4.48	3.08	12.92
		(2.46)	(0.83)	(1.74)	(0.76)	(3.80)
	LASSOORD	7.00	4.51	4.42	3.05	12.88
		(2.46)	(0.83)	(1.77)	(0.76)	(3.84)
Línea de Pobreza	REG-PFCORD	0.208	0.167	0.279	0.287	0.129
(`pobreza`: discreta)		(0.015)	(0.011)	(0.028)	(0.028)	(0.022)
	PCAPOLY	0.229	0.204	0.366	0.390	0.132
		(0.028)	(0.018)	(0.023)	(0.063)	(0.025)
	NLPCA	0.228	0.186	0.302	0.290	0.161
		(0.021)	(0.019)	(0.019)	(0.018)	(0.016)
	FULL	0.202	0.162	0.274	0.287	0.129
		(0.021)	(0.008)	(0.026)	(0.036)	(0.020)
	FULL-I	0.206	0.171	0.286	0.298	0.126
		(0.023)	(0.021)	(0.024)	(0.065)	(0.028)
	LASSOORD	0.206	0.171	0.286	0.298	0.129
		(0.023)	(0.021)	(0.024)	(0.065)	(0.027)

Note: errores estándar entre paréntesis. Database: EPH (2013)

cuadrático SES^2 es incluido para corregir la curvatura en la función de regresión estimada y la variable respuesta fue transformada con ipcf $\leftarrow$ $\text{ipcf}^{1/3}$, a partir de realizar un análisis de regresión Box–Cox. Puede observarse que cuando usamos el índice SES construido con el PCAPOLY, los valores del índice están concentrados principalmente en el intervalo [0.5, 1.0] mientras que, para el SES modelado usando REG-PFCORD, los valores del indice se distribuyen más regularmente sobre todo el intervalo [0, 1.0]. Esto permite obtener un mejor ajuste con el modelo lineal, como se revela a partir un valor del R^2 igual a 0.302 comparado con un 0.231 obtenido con un índice SES basado en PCA.

Las Tablas 3.4 y 3.5 muestran los vectores de coeficientes estimados que definen el índice SES usando `ipcf` y `pobreza` como variable respuesta, respectivamente. Podemos notar que

Figura 3.3: Ajuste del modelo lineal del ingreso per cápita como función del índice SES. Lado izquierdo: usando índice SES con método tipo PCA. Lado derecho: resultado con el método propuesto.



para el método propuesto, varios coeficientes del $\widehat{\boldsymbol{\alpha}}$ han sido llevados a cero con la estimación regularizada, mientras que para PCApoly y NLPCA sólo *horas trabajadas* pareciera no ser muy relevante para la construcción del índice SES. Además, de los resultados reportados podemos apreciar varias diferencias entre el reg-PFCord y los enfoques basados en PCA (i.e., PCApoly y NLPCA). En primer lugar, la importancia de cada predictor para la construcción de índice SES difiere según el método. Por ejemplo, la variable *hacinamiento* obtiene la mayor ponderación con el reg-PFCord para todas las regiones y para ambas variables respuesta, mientras que la *instalación sanitaria* y la *ubicación del agua* muestran ser las más relevantes para la construcción del índice usando PCApoly, y la *instalación sanitaria* y el *desagüe* para el método NLPCA. En segundo lugar, observamos que los índices SES construidos usando ambos métodos basados en PCA dan similares ponderaciones a los predictores para todas las regiones. En cambio, el índice SES basado en reg-PFCord logra captura la divergencia económica regional explicada por diferencias en dotaciones de factores, productividad, niveles de actividad y patrones de crecimiento económico regional. Más aún, en las regiones urbanas más ricas (específicamente, Buenos Aires y la región Pampeana) la estimación regularizada del reg-PFCord tiende a fijar en cero las variables con mayor ponderación en los índices basados en PCA. Esta diferencia resulta atractiva, ya que estas regiones suelen tener una mejor infraestructura social básica general, por lo que las variables relacionadas con el drenaje, provisión de agua e la instalación sanitaria de la vivienda son menos importantes para medir el nivel socioeconómico, pues gran parte de la población estaría cubierta en este sentido. Con un fundamento similar encontramos que otras variables, tal como *hacinamiento* o *escolaridad*, son necesarias en orden de obtener un mejor índice SES para predecir el nivel de ingreso del hogar. En la misma línea, para regiones con mayores niveles de pobreza (NOA y NEA) el índice SES provisto por el método reg-PFCord muestra que otras variables, tal como *ubicación de la vivienda*, *provisión de agua* o *ubicación del agua*, pasan a ser relevantes en la determinación del nivel socio-económico.

Comparando ambas metodologías basadas en PCA, podemos notar que el NLPCA es más sensible a la heterogeneidad regional que el PCApoly, pero las diferencias en las ponderaciones del índice de estos métodos comparado con las del reg-PFCord permanecen sustanciales.

Adicionalmente, podemos apreciar que el índice SES obtenido vía reg-PFCord es sensible respecto a qué variable respuesta estamos usando para caracterizar el fenómeno social de interés.

Por ejemplo, en las regiones del GBA, Pampeana y Patagonia, la *escolaridad* tiene un peso considerable en el índice SES para explicar el ingreso per cápita, no así para predecir pobreza. Esto cobra sentido debido a que, en estas regiones relativamente más ricas, el acceso a niveles básicos de escolaridad está más garantizado en la población que habita en las mismas. Por otra parte, muchas veces el ingreso constituye un fuerte incentivo sobre la decisión individual de alcanzar mayores niveles de educación, reflejándose de hecho en la ponderación del índice para predecir ingreso. Además, para estas regiones más ricas, algunas variables, tal como *desagüe*, *instalación sanitaria* o *baño compartido* pasan a ser relevantes para predecir si el hogar es pobre o no (siguiendo el criterio de la línea de pobreza). Tales diferencias no pueden ser capturas por el índice SES basado en la metodología de PCA.

Tabla 3.4: Comparación de coeficientes del índice SES para las metodologías reg-PCFord, PCApoly y NLPCA para predecir ingreso per cápita del hogar.

Variables	GBA			Pampeana			NOA		
	reg-PCFord	PCApoly	NLPCA	reg-PCFord	PCApoly	NLPCA	reg-PCFord	PCApoly	NLPCA
ubicación de la vivienda	0	-0.1690	-0.0943	0	-0.1903	-0.0976	-0.1314	-0.1068	-0.0835
calidad de la vivienda	-0.1985	-0.3768	-0.2199	0.2591	-0.3557	-0.1985	0	-0.3278	-0.1849
combustible para cocinar	-0.4646	-0.3788	-0.2080	0.3627	-0.3609	-0.1678	-0.1070	-0.3287	-0.1582
hacinamiento	-0.7272	-0.2888	-0.1788	0.8300	-0.2351	-0.1329	-0.8798	-0.1991	-0.1194
escolaridad	-0.2676	-0.2275	-0.1474	0.2668	-0.2075	-0.1135	-0.3614	-0.2197	-0.1201
desagüe	-0.0873	-0.3381	-0.2047	0	-0.3519	-0.2333	-0.0901	-0.3623	-0.2462
instalación sanitaria	0	-0.4061	-0.2246	0	-0.4105	-0.2411	0	-0.4217	-0.2545
baño compartido	0	-0.2759	-0.1186	0	-0.3176	-0.1699	0	-0.2579	-0.1334
ubicación del agua	0.1700	-0.3918	-0.1790	0	-0.3933	-0.1941	-0.2054	-0.4202	-0.2309
provisión de agua	0	-0.2023	-0.1033	0	-0.2461	-0.1129	0	-0.3646	-0.1374
horas trabajadas	-0.3283	0	0	0.2029	0	0	-0.1277	0	0
	NEA			Patagonia					
	reg-PCFord	PCApoly	NLPCA	reg-PCFord	PCApoly	NLPCA			
ubicación de la vivienda	-0.1509	-0.1809	-0.0978	-0.1149	-0.1437	-0.0981			
calidad de la vivienda	0	-0.3727	-0.2130	-0.3046	-0.3258	-0.1844			
combustible para cocinar	-0.0742	-0.1648	-0.0646	-0.1797	-0.4026	-0.1810			
hacinamiento	-0.8496	-0.2052	-0.1040	-0.7263	-0.2207	-0.0984			
escolaridad	-0.3507	-0.1869	-0.1009	-0.3670	-0.1284	-0.0516			
desagüe	0	-0.3572	-0.2573	-0.1383	-0.4122	-0.2566			
instalación sanitaria	0	-0.4383	-0.2735	0	-0.4376	-0.2622			
baño compartido	-0.2284	-0.2921	-0.1344	-0.1204	-0.2937	-0.1734			
ubicación del agua	0	-0.4227	-0.2377	0.1877	-0.4169	-0.2196			
provisión de agua	-0.2574	-0.3733	-0.1384	0.2473	-0.1525	-0.0522			
horas trabajadas	-0.0922	0	0	-0.2637	0	0			

Tabla 3.5: Comparación de coeficientes del índice SES para las metodologías REG-PCFORD, PCAPOLY y NLPCA para predecir pobreza (respuesta discreta).

Variables	GBA			Pampeana			NOA		
	REG-PFCORD	PCAPOLY	NLPCA	REG-PFCORD	PCAPOLY	NLPCA	REG-PFCORD	PCAPOLY	NLPCA
ubicación de la vivienda	0	-0.1690	-0.0943	0	-0.1903	-0.0976	-0.2434	-0.1068	-0.0835
calidad de la vivienda	-0.4033	-0.3768	-0.2199	0.3347	-0.3557	-0.1985	0	-0.3278	-0.1849
combustible para cocinar	-0.5240	-0.3788	-0.2080	0.3579	-0.3609	-0.1678	0	-0.3287	-0.1582
hacinamiento	-0.7076	-0.2888	-0.1788	0.7216	-0.2351	-0.1329	-0.7939	-0.1991	-0.1194
escolaridad	0	-0.2275	-0.1474	0	-0.2075	-0.1135	-0.2094	-0.2197	-0.1201
desagüe	0	-0.3381	-0.2047	0	-0.3519	-0.2333	0	-0.3623	-0.2462
instalación sanitaria	0	-0.4061	-0.2246	0.3990	-0.4105	-0.2411	-0.1528	-0.4217	-0.2545
baño compartido	-0.1836	-0.2759	-0.1186	0	-0.3176	-0.1699	0	-0.2579	-0.1334
ubicación del agua	-0.1208	-0.3918	-0.1790	0.2647	-0.3933	-0.1941	-0.4933	-0.4202	-0.2309
provisión de agua	0	-0.2023	-0.1033	0	-0.2461	-0.1129	0	-0.3646	-0.1374
horas trabajadas	-0.1173	0	0	0.0990	0	0	0	0	0
	Northeast			Patagonia					
	REG-PFCORD	PCAPOLY	NLPCA	REG-PFCORD	PCAPOLY	NLPCA			
ubicación de la vivienda	-0.1982	-0.1809	-0.0978	-0.1187	-0.1437	-0.0981			
calidad de la vivienda	0	-0.3727	-0.2130	-0.3693	-0.3258	-0.1844			
combustible para cocinar	-0.2509	-0.1648	-0.0646	-0.2788	-0.4026	-0.1810			
hacinamiento	-0.7063	-0.2052	-0.1040	-0.3987	-0.2207	-0.0984			
escolaridad	-0.1442	-0.1869	-0.1009	-0.0766	-0.1284	-0.0516			
desagüe	0	-0.3572	-0.2573	-0.1887	-0.4122	-0.2566			
instalación sanitaria	0	-0.4383	-0.2735	-0.1313	-0.4376	-0.2622			
baño compartido	-0.3477	-0.2921	-0.1344	-0.1289	-0.2937	-0.1734			
ubicación del agua	0	-0.4227	-0.2377	0.2785	-0.4169	-0.2196			
provisión de agua	-0.5071	-0.3733	-0.1384	0.6585	-0.1525	-0.0522			
horas trabajadas	0	0	0	-0.1626	0	0			

3.7.2. El caso de NETFLIX

En esta sección utilizamos la base de Netflix[2] para ilustrar el desempeño del método con otra aplicación real diferente de la presentada para índices SES. Este conjunto de datos contiene la puntuación o calificación ("estrellas") sobre películas disponibles en Netflix, y dichas puntuaciones son realizadas por suscriptores de este servicio. Estos datos podemos pensarlos como unas matrices grandes, de n usuarios y de p películas. Como no todos los usuarios han valorado todas las peliculas disponibles, esta matriz es rala (*sparse*) y las películas no puntuadas, usualmente son tratadas como datos faltantes. Para el propósito de este ejemplo, consideramos sólo un subconjunto de los datos, omitiendo las puntuaciones faltantes. Para ello, buscamos un subconjunto de n usuarios que han evaluado las mismas p películas. Buscando en la base de datos para $p = 20$, se extrajo un subconjunto de $n = 8347$ usuarios.

Puntualmente tenemos una matriz $\mathbf{X} = (X_{ij})$ tal que $X_{ij} \in \{1, 2, 3, 4, 5\}$ representa la puntuación del usuario i de la película j y no hay puntuaciones faltante en $\mathbf{X}$. Denotamos por $\mathbf{X}_{-i,j}$ a la matriz que contiene las puntuaciones de la película j de todos excepto del usuario i-ésimo, y definimos $\mathbf{X}_{i,-j}$ de manera análoga. Nuestro objetivo es predecir la puntuación X_{ij} utilizando la información del resto de la base de datos; en particular, a partir de las puntuaciones realizadas por i de otras películas en $\mathbf{X}$.

Las siguiente reglas de predicción son consideradas para predecir X_{ij}:

- MEDIA: $\widehat{X}_{ij} = \sum_{k \neq i} X_{kj}$; esto es, la puntuación predicha es la puntuación media de la película realizada por el resto de los usuarios.
- PFC+LIN: $\widehat{X}_{ij} = \mathbf{B}^T \left(\widehat{\boldsymbol{\alpha}}_{\text{PFC}}^T \mathbf{X}_{i,-j} \right)$. Aquí obtenemos una reducción vía PFC sobre los predictores, usando $\mathbf{X}_{-i,j}$ como la variable respuesta y $\mathbf{X}_{-i,-j}$ como predictores, y luego usamos un modelo de predicción lineal sobre los datos reducidos.
- PFCORD+LIN: $\widehat{X}_{ij} = \mathbf{C}^T \left(\widehat{\boldsymbol{\alpha}}_{\text{ORD}}^T \mathbf{X}_{i,-j} \right)$. Aquí procedemos de la misma forma que antes, pero utilizamos REG-PFCORD para la reducción de dimensión.
- FULL+LIN: $\widehat{X}_{ij} = \mathbf{B}^T \mathbf{X}_{i,-j}$. Aquí ajustamos un modelo predictivo lineal usando todo el conjunto de predictores.

[2]Información respecto al *Netflix Prize* puede obtenerse en HTTP://WWW.NETFLIXPRIZE.COM//RULES.

- PFC+SVR: $\widehat{X}_{ij} = g_{\text{SVR}}(\widehat{\boldsymbol{\alpha}}_{\text{PFC}}^T \mathbf{X}_{i,-j})$. En este caso buscamos una reducción de los predictores con PFC, tomando $\mathbf{X}_{-i,j}$ como respuesta y $\mathbf{X}_{-i,-j}$ como predictores, y luego usamos *support vector regression* (SVR) como modelo predictivo, usando los datos reducidos.
- PFCORD+SVR: $\widehat{X}_{ij} = g_{\text{SVR}}(\widehat{\boldsymbol{\alpha}}_{\text{ORD}}^T \mathbf{X}_{i,-j})$. Similar a la estrategia anterior, pero usamos el método propuesto, REG-PFCORD, para obtener la SDR.
- FULL+SVR: $\widehat{X}_{ij} = g_{\text{SVR}}(\mathbf{X}_{i,-j})$. Aquí utilizamos SVR sobre todo el conjunto de predictores.

Usamos un kernel Gaussiano para el método SVR, con la varianza (bandwith) elegida utilizando validación cruzada de 5 iteraciones (5-fold cross-validation). La Tabla 3.6 muestra el error de predicción medio (MSE) obenido (*leave-one-out*) tanto para la dimensión $\hat{d}$ inferida usando el procedimiento basado en permutaciones, como también para diferentes valores de la dimensión d. Podemos observar que, como la dimensión de $\mathbf{f}_Y$ es $r = 4$ en este ejemplo, y sabemos que la dimensión de la reducción debe satisfacer $d \leq \text{mín}(r, p)$, las reducciones que tienen sentido para estos datos deben tener dimensión $d \leq 4$. Podemos observar que la predicción de la puntuación usando solo puntuaciones de la misma película realizada por otros usuarios, es claramente superada por todos los métodos que usan puntuaciones realizadas por el mismo usuario. Para todos los valores de d, el mejor resultado es aquel obtenido utilizando el método propuesto (REG-PFC) para realizar reducción de dimensiones junto con SVR como modelo predictivo. Más aún, para un d fijo, el método que hemos propuesto de reducción suficiente para variables ordinales supera al PFC, independientemente de la regla de predicción elegida. También podemos ver que con SVR, si reducimos los predictores con PFC se obtiene un peor resultado que usar el conjunto entero de predictores, mientras que si usamos REG-PFCORD obtenemos menores errores de predicción cuando $d = 4$. También podemos notar que los resultados para $\hat{d}$ son muy similares a aquellos arrojados con $d = 4$. Esto se debe a que la dimensión inferida de la reducción es $\hat{d} = 4$ en el 98 % de las veces.

Por lo tanto, los resultados mostrados con esta aplicación muestran que el método propuesto es superior al PFC estándar, resaltando los beneficios de reducir la dimensión cuando los predictores categóricos ordinales se reconocen como tales.

Tabla 3.6: MSE obtenido para la base de datos de NETFLIX

	Media	PFC+lin	PFCord+lin	FULL+lin	PFC+svr	PFCord+svr	FULL+svr
$d = \hat{d}$	...	0.645	0.568	...	0.583	0.551	...
$d = 4$	···	0.642	0.565	···	0.574	0.549	···
$d = 3$	···	0.889	0.663	···	0.691	0.633	···
$d = 2$	···	0.901	0.857	···	0.874	0.821	···
$d = 1$	···	0.909	0.871	···	0.882	0.836	···
$d = p$	0.914	···	···	0.642	···	···	0.566

3.8. Conclusiones

El método de estimación propuesto basado en un algoritmo tipo EM para la reducción de la dimensión de problemas de regresión con predictores ordinales, mostró ser superior a los métodos estándar derivados para predictores continuos y basados en la regresión inversa. Esto lo probamos, tanto con las simulaciones como con conjuntos de datos reales que involucran predictores categóricos ordenados. Los experimentos mostraron que esta ventaja se acentúa cuando usamos selección de variables, donde el método propuesto supera claramente a su contraparte para los datos continuos, cuando dicha contraparte se aplica ingenuamente a predictores ordinales. Esto no es un problema menor, ya que muchos análisis en las ciencias aplicadas suelen tratarlos como variables continuas, sin tener en cuenta su naturaleza ordinal. Además, con la propuesta de algoritmo aproximado EM se constituye un método factible (en términos computacionales) para un conjunto mucho mayor de problemas en comparación con el uso del cálculo exacto de los momentos truncados. Este ahorro también permite realizar test de permutación y procedimientos de validación cruzada para inferir la dimensión de la reducción, que resultó ser razonablemente preciso en las simulaciones. Por último, la aplicación de la metodología propuesta para la construcción de índices de estatus socio-económico (índices SES) mostró muchas ventajas con respecto a los índices basados en PCA. En particular, el método no sólo ayuda a obtener mejores predicciones sino que también permite obtener una comprensión mejor de las relaciones entre los predictores y la respuesta. De manera más precisa, para el índice de SES, el método propuesto brinda diferentes ponderaciones logrando capturar diferencias regionales, históricas y/o culturales, siendo al mismo tiempo sensible respecto a la medida usada como respuesta (como el ingreso familiar per cápita o la línea de pobreza), lo que no ocurre con las metodologías no supervisadas basadas en PCA. Esta propiedad del método de reducción suficiente de dimensiones basada en modelos tiene implicaciones relevantes para el análisis social

aplicado.

3.9. Apéndice del capítulo

3.9.1. El Algoritmo EM

Con el objetivo de simplificar la notación, siempre vamos a omitir el condicionante sobre $\mathbf{\Omega}^{(k-1)}$ cuando tomemos esperanza. También vamos a omitir los condicionantes sobre el resto de las variables en el sub-índice. Para obtener una forma explicita de Q, vamos a computar la esperanza condicional de la log-verosimilitud conjunta de la forma

$$Q(\mathbf{\Omega}|\mathbf{\Omega}^{(k-1)}) = \sum_{i=1}^{n} E_{\mathbf{z}_i|\mathbf{x}_i,y_i;\mathbf{\Omega}^{(k-1)}} \left[\log f_{\mathbf{x}_i,\mathbf{z}_i}(\mathbf{x}_i,\mathbf{z}_i|y_i;\mathbf{\Omega})\big|y_i;\mathbf{\Omega}^{(k-1)}\right] = \sum_{i=1}^{n} E_{\mathbf{z}_i}\left[\log f_{\mathbf{x}_i,\mathbf{z}_i}(\mathbf{x}_i,\mathbf{z}_i|y_i;\mathbf{\Omega})\big|y_i\right].$$

Por lo tanto,

$$\begin{aligned}
Q(\mathbf{\Omega}|\mathbf{\Omega}^{(k-1)}) &= \sum_{i=1}^{n} E_{\mathbf{z}_i}\left[\log f_{\mathbf{x}_i,\mathbf{z}_i}(\mathbf{x}_i,\mathbf{z}_i|y_i;\mathbf{\Omega})\big|y_i\right] \\
&= \sum_{i=1}^{n} E_{\mathbf{z}_i}\left[\log\left((2\pi)^{-p/2}|\mathbf{\Delta}|^{-1/2}e^{-\frac{1}{2}\operatorname{tr}(\mathbf{\Delta}^{-1}(\mathbf{z}_i-\mathbf{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i})(\mathbf{z}_i-\mathbf{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i})^T)}I_{\{\mathbf{z}_i\in C(\mathbf{x}_i,\mathbf{\Theta})\}}\right)\Big|y_i\right] \\
&= -\frac{pn}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{\Delta}| \\
&-\frac{n}{2}\operatorname{tr}\left[\mathbf{\Delta}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}E_{\mathbf{z}_i}(\mathbf{z}_i\mathbf{z}_i^T|\mathbf{x}_i,y_i) - \frac{2}{n}\mathbf{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\sum_{i=1}^{n}\bar{\mathbf{f}}_{y_i}E_{\mathbf{z}_i}(\mathbf{z}_i^T|\mathbf{x}_i,y_i) + \frac{1}{n}\mathbf{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\sum_{i=1}^{n}\bar{\mathbf{f}}_{y_i}f_{y_i}^T\boldsymbol{\xi}^T\boldsymbol{\alpha}^T\mathbf{\Delta}\right)\right] \\
&= -\frac{pn}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{\Delta}| - \frac{n}{2}\operatorname{tr}\left[\mathbf{\Delta}^{-1}(\mathbf{S} - \frac{2\mathbf{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{F}^T\mathbf{M}}{n} + \frac{\mathbf{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{F}^T\mathbf{F}\boldsymbol{\xi}^T\boldsymbol{\alpha}^T\mathbf{\Delta}}{n})\right], \qquad (3.8)
\end{aligned}$$

donde $\mathbf{S} \in \mathbb{R}^{p\times p}$, $\mathbf{F} \in \mathbb{R}^{n\times r}$ y $\mathbf{M} \in \mathbb{R}^{n\times p}$ están dados por $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}E_{\mathbf{z}_i}(\mathbf{z}_i\mathbf{z}_i^T|\mathbf{x}_i,y_i)$, $\mathbf{F}^T = [\bar{\mathbf{f}}_{y_1},\ldots,\bar{\mathbf{f}}_{y_n}]$ y $\mathbf{M}^T = [E_{\mathbf{z}_1}(\mathbf{z}_1|\mathbf{x}_1,y_1),\ldots,E_{\mathbf{z}_n}(\mathbf{z}_n|\mathbf{x}_n,y_n))]$, respectivamente.

Maximización de la función Q (3.8).

De (3.8), tenemos que

$$Q(\mathbf{A},\boldsymbol{\beta},\mathbf{\Delta}) = -\frac{pn}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{\Delta}| - \frac{n}{2}\operatorname{tr}(\mathbf{\Delta}^{-1}\mathbf{S}) + \operatorname{tr}(\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{F}^T\mathbf{M}) - \frac{1}{2}\operatorname{tr}(\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{F}^T\mathbf{F}\boldsymbol{\xi}^T\boldsymbol{\alpha}^T\mathbf{\Delta}).$$

Dado que Q es una forma cuadrática en $\boldsymbol{\xi}$, el máximo se obtiene en $\boldsymbol{\xi}^{(k)} = (\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T\mathbf{M}^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}$. Reemplazando $\boldsymbol{\xi}^{(k)}$ en la función Q, obtenemos la log-verosimilitud parcial

$$\begin{aligned} Q(\boldsymbol{\Delta}^{-1}, \boldsymbol{\alpha}) &= -\frac{pn}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Delta}| - \frac{n}{2}\operatorname{tr}(\boldsymbol{\Delta}^{-1}\mathbf{S}) + \frac{1}{2}\operatorname{tr}\left[(\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T\mathbf{M}^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{M}\boldsymbol{\alpha}\right] \\ &= -\frac{pn}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Delta}| - \frac{n}{2}\operatorname{tr}(\boldsymbol{\Delta}^{-1}\mathbf{S}) + \frac{n}{2}\operatorname{tr}\left[(\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T\mathbf{S}_{\text{fit}}\boldsymbol{\alpha}\right], \end{aligned} \tag{3.9}$$

donde $\mathbf{S}_{\text{fit}} = \frac{1}{n}\mathbf{M}^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{M}$. Para maximizar Q con respecto a to $\boldsymbol{\Delta}^{-1}$ observemos que, por la Proposición 5.14 de Eaton (1983), si $\boldsymbol{\alpha} \in \mathbb{R}^{p\times d}$ es fija y $\boldsymbol{\alpha}_0 \in \mathbb{R}^{p\times(p-d)}$ es el complemento semi-ortogonal de $\boldsymbol{\alpha}$, tenemos una correspondencia uno a uno entre $\boldsymbol{\Delta}^{-1}$ y $(\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3)$, y $\mathbf{H}_1 = \boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha}$; $\mathbf{H}_2 = (\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}_0)^{-1}$ y $\mathbf{H}_3 = (\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha}_0$. De Rao (1973) y Cook and Forzani (2009) obtenemos

$$\boldsymbol{\Delta}^{-1} = \boldsymbol{\alpha}(\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T + \boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}_0(\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}_0)^{-1}\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}^{-1}, \tag{3.10}$$

$$|\boldsymbol{\Delta}| = |\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}_0|^{-1}|\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha}|. \tag{3.11}$$

La identidad $(\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha}_0 = -\boldsymbol{\alpha}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}_0(\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}_0)^{-1}$ implica que

$$\boldsymbol{\alpha}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}_0 = -(\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}_0 = -\mathbf{H}_3\mathbf{H}_2^{-1},$$

la cual, junto con $\boldsymbol{\alpha}\boldsymbol{\alpha}^T + \boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^T = \mathbf{I}_d$,

$$\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}_0 = \boldsymbol{\alpha}\boldsymbol{\alpha}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\alpha}_0 = -\boldsymbol{\alpha}\mathbf{H}_3\mathbf{H}_2^{-1} + \boldsymbol{\alpha}_0\mathbf{H}_2^{-1} = (-\boldsymbol{\alpha}\mathbf{H}_3 + \boldsymbol{\alpha}_0)\mathbf{H}_2^{-1}.$$

Usando todo esto en (3.10) obtenemos

$$\boldsymbol{\Delta}^{-1} = \boldsymbol{\alpha}\mathbf{H}_1^{-1}\boldsymbol{\alpha}^T + (-\boldsymbol{\alpha}\mathbf{H}_3 + \boldsymbol{\alpha}_0)\mathbf{H}_2^{-1}(-\mathbf{H}_3^T\boldsymbol{\alpha}^T + \boldsymbol{\alpha}_0^T), \tag{3.12}$$

Por lo tanto, encontrar $\widehat{\mathbf{H}}_i$, $i = 1, 2, 3$ es equivalente a encontrar $(\boldsymbol{\Delta}^{-1})^{(k)}$. Para escribir la función Q en términos de $\mathbf{H}_1$, $\mathbf{H}_2$ y $\mathbf{H}_3$, vamos a escribir $\log|\boldsymbol{\Delta}|$ y $\operatorname{tr}(\boldsymbol{\Delta}^{-1}\mathbf{S})$ in términos de las misma. Por lo tanto, usando (3.12) y (3.11), la función Q luego puede ser escrita en términos

de $\mathbf{H}_1$, $\mathbf{H}_2$, y $\mathbf{H}_3$ de la forma

$$Q(\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3, \boldsymbol{\alpha}) = -\frac{pn}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{H}_1| - \frac{n}{2}\log|\mathbf{H}_2| - \frac{n}{2}\operatorname{tr}\left[\boldsymbol{\alpha}\mathbf{H}_1^{-1}\boldsymbol{\alpha}^T(\mathbf{S} - \mathbf{S}_{\text{fit}})\right] - \frac{n}{2}\operatorname{tr}\left[(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}\mathbf{H}_3)\mathbf{H}_2^{-1}(\boldsymbol{\alpha}_0^T - \mathbf{H}_3^T\boldsymbol{\alpha}^T)\mathbf{S}\right]. \quad (3.13)$$

Ahora, como Q es cuadrática en $\mathbf{H}_3$, el máximo de Q para $\mathbf{H}_3$ se obtiene en

$$\widehat{\mathbf{H}}_3 = (\boldsymbol{\alpha}^T\mathbf{S}\boldsymbol{\alpha})^{-1}(\boldsymbol{\alpha}^T\mathbf{S}\boldsymbol{\alpha}_0). \quad (3.14)$$

Reemplazando (3.14) en (3.13) y definiendo $\mathbf{S}_{\text{res}} = \mathbf{S} - \mathbf{S}_{\text{fit}}$ (la cual es definida positiva), obtenemos la función log-verosimilitud parcial

$$Q(\mathbf{H}_1, \mathbf{H}_2, \boldsymbol{\alpha}) = -\frac{pn}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{H}_1| - \frac{n}{2}\log|\mathbf{H}_2| - \frac{n}{2}\operatorname{tr}\left[\mathbf{H}_1^{-1}\boldsymbol{\alpha}^T\mathbf{S}_{\text{res}}\boldsymbol{\alpha}\right] - \frac{n}{2}\operatorname{tr}\left[\mathbf{H}_2^{-1}(\boldsymbol{\alpha}_0^T - \boldsymbol{\alpha}_0^T\mathbf{S}\boldsymbol{\alpha}(\boldsymbol{\alpha}^T\mathbf{S}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T)\mathbf{S}(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}(\boldsymbol{\alpha}^T\mathbf{S}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T\mathbf{S}\boldsymbol{\alpha}_0)\right].$$

El máximo de Q sobre $\mathbf{H}_1$ y $\mathbf{H}_2$ es alcanzado en

$$\begin{aligned}\widehat{\mathbf{H}}_1 &= \boldsymbol{\alpha}^T\mathbf{S}_{\text{res}}\boldsymbol{\alpha};\\ \widehat{\mathbf{H}}_2 &= \boldsymbol{\alpha}_0^T\mathbf{S}\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}_0^T\mathbf{S}\boldsymbol{\alpha}(\boldsymbol{\alpha}^T\mathbf{S}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T\mathbf{S}\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_0^T\mathbf{S}^{-1}\boldsymbol{\alpha}_0)^{-1}.\end{aligned}$$

Después de sustituir los puntos de máximo de $\mathbf{H}_1$, $\mathbf{H}_2$, y $\mathbf{H}_3$ en (3.12), obtenemos que el máximo para $\boldsymbol{\Delta}^{-1}$ es alcanzado en

$$(\boldsymbol{\Delta}^{-1})^{(k)} = \mathbf{S}^{-1} + \boldsymbol{\alpha}(\boldsymbol{\alpha}^T\mathbf{S}_{\text{res}}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T - \boldsymbol{\alpha}(\boldsymbol{\alpha}^T\mathbf{S}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T.$$

Puesto que esta matriz estimada no necesariamente tiene unos en su diagonal, la escalamos a fin de obtener una matriz con diagonal unitaria. Con este estimador de $\boldsymbol{\Delta}^{-1}$, la log-verosimilitud maximizada será

$$\begin{aligned}Q(\boldsymbol{\alpha}) &= -\frac{pn}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\alpha}^T\mathbf{S}_{\text{res}}\boldsymbol{\alpha}| - \frac{n}{2}\log|(\boldsymbol{\alpha}_0^T\mathbf{S}^{-1}\boldsymbol{\alpha}_0)^{-1}| - \frac{nd}{2} - \frac{n(p-d)}{2}\\ &= -\frac{pn}{2}[\log(2\pi) + 1] - \frac{n}{2}\log|\boldsymbol{\alpha}^T\mathbf{S}_{\text{res}}\boldsymbol{\alpha}| - \frac{n}{2}\log|\mathbf{S}| + \frac{n}{2}\log|\boldsymbol{\alpha}^T\mathbf{S}\boldsymbol{\alpha}|.\end{aligned}$$

donde para la última igualdad hemos usado (3.11). Finalmente, él máximo en $\boldsymbol{\alpha}$ es alcanzado en

$$\boldsymbol{\alpha}^{(k)} = \mathbf{S}^{-1/2}\hat{\boldsymbol{\zeta}}\mathbf{N},$$

donde $\hat{\boldsymbol{\zeta}}$ son los primeros d autovectores de $\mathbf{S}^{-1/2}\mathbf{S}_{\text{fit}}\mathbf{S}^{-1/2}$ y $\mathbf{N}$ es una matriz tal que $\widehat{\boldsymbol{\alpha}}^T\widehat{\boldsymbol{\alpha}} = \mathbf{I}_d$.

3.9.2. Aproximación de S y M

Dados $\boldsymbol{\Omega}^{(k-1)} = (\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Delta}^{(k-1)}, \boldsymbol{\alpha}^{(k-1)}, \boldsymbol{\xi}^{(k-1)})$ y y_i fija, necesitamos estimar $\mathbf{S}$ y $\mathbf{M}$ para poder computar la función Q. Cada entrada de la matriz $\mathbf{S}$ puede escribirse como $s_{jk} = \sum_{i=1}^{n} E_{\mathbf{z}_i}(z_{i,j}z_{i,k}|\mathbf{x}_i)$ with $j,k = 1,\ldots,p$. Por lo tanto, para $j = k$ tenemos el segundo momento condicional $E_{\mathbf{z}_i}(z_{i,j}^2|\mathbf{x}_i)$. Siguiendo a (Guo et al. 2015), cuando $j \neq k$ los términos $E_{\mathbf{z}_i}(z_{i,j}z_{i,k}|\mathbf{x}_i)$ pueden ser aproximados por $E_{\mathbf{z}_i}(z_{i,j}z_{i,k}|\mathbf{x}_i) \approx E_{\mathbf{z}_i}(z_{i,j}|\mathbf{x}_i)E_{\mathbf{z}_i}(z_{i,k}|\mathbf{x}_i)$. De esta manera, podemos obtener un estimador de $\mathbf{S}$ a través de la estimación de los primeros y segundos momentos. Lo que sigue, es una modificación del procedimiento para calcular estos momentos, desarrollado por Guo et al. (2015), adaptado para el caso de distribuciones condicionadas. En particular, podemos escribir $\mathbf{x}_i$ como $\mathbf{x}_i = (x_{i,j}, \mathbf{x}_{i,-j})$ y $\mathbf{z}_i$ como $\mathbf{z}_i = (z_{i,j}, \mathbf{z}_{i,-j})$ donde $\mathbf{x}_{i,-j} = (x_{i,1},\ldots,x_{i,j-1},x_{i,j+1},\ldots,x_{i,p})$ y $\mathbf{z}_{i,-j} = (z_{i,1},\ldots,z_{i,j-1},z_{i,j+1},\ldots,z_{i,p})$. Por lo tanto, el primer momento es

$$\begin{aligned}
E_{\mathbf{z}_i}(z_{i,j}|\mathbf{x}_i) &= \int_{\mathbb{R}^p} z_{i,j} f_{\mathbf{z}_i}(\mathbf{z}_i|\mathbf{x}_i)d\mathbf{z}_i \\
&= \int_{\mathbb{R}^p} z_{i,j} f_{z_{i,j}}(z_{i,j}|\mathbf{z}_{i,-j}, x_{i,j}, \mathbf{x}_{i,j}) f_{\mathbf{z}_{i,-j}}(\mathbf{z}_{i,-j}|\mathbf{x}_i)d\mathbf{z}_i \\
&= \int_{\mathbb{R}^{p-1}} \left[\int_{\mathbb{R}} z_{i,j} f_{z_{i,j}}(z_{i,j}|z_{i,-j}, x_{i,j})dz_{i,j}\right] f_{\mathbf{z}_{i,-j}}(\mathbf{z}_{i,-j}|\mathbf{x}_i)d\mathbf{z}_{i,-j} \\
&= E_{\mathbf{z}_{i,-j}}\left\{E_{z_{i,j}}(z_{i,j}|\mathbf{z}_{i,-j}, x_{i,j})|\mathbf{x}_i\right\}. \qquad (3.15)
\end{aligned}$$

De la misma mamnera, el segundo momento puede escribirse de la forma

$$E_{\mathbf{z}_i}(z_{i,j}^2|\mathbf{x}_i) = E_{\mathbf{z}_{i,-j}}\left\{E_{z_{i,j}}(z_{i,j}^2|\mathbf{z}_{i,-j}, x_{i,j})|\mathbf{x}_i\right\}. \qquad (3.16)$$

Dados y_i y $\boldsymbol{\Omega}^{(k-1)}$, $(z_{i,1}, z_{i,2}, \ldots, z_{i,p})$ tiene una distribución normal multivariada con media $\mu_i = \boldsymbol{\Psi}^{(k-1)}f_{y_i} = \boldsymbol{\Delta}^{(k-1)}\boldsymbol{\alpha}^{(k-1)}\boldsymbol{\xi}^{(k-1)}$, y matriz de varianza-covarianza $\boldsymbol{\Delta}^{(k-1)}$. Tomando $\boldsymbol{\Delta}_{j,j}^{(k-1)} = 1$, para cada $j = 1,\ldots,p$, podemos escribir

$$\boldsymbol{\Delta}^{(k-1)} = \begin{pmatrix} 1 & \boldsymbol{\Delta}_{j,-j}^{(k-1)} \\ \boldsymbol{\Delta}_{-j,j}^{(k-1)} & \boldsymbol{\Delta}_{-j,-j}^{(k-1)} \end{pmatrix} \qquad \text{y} \qquad \mu_i = \begin{pmatrix} (\boldsymbol{\Psi}^{(k-1)}\bar{\mathbf{f}}_{y_i})_j \\ (\boldsymbol{\Psi}^{(k-1)}\bar{\mathbf{f}}_{y_i})_{-j} \end{pmatrix},$$

y por lo tanto, la distribución condicionada de $z_{i,j}$ dado $\mathbf{z}_{i,-j}$ es

$$z_{i,j}|\mathbf{z}_{i,-j} \sim N(\widetilde{\mu}_{i,j}, \widetilde{\delta}_{i,j}) \ ,$$

donde la media es $\widetilde{\mu}_{i,j} = (\boldsymbol{\Psi}^{(k-1)}f_{y_i})_j + \boldsymbol{\Delta}_{j,-j}^{(k-1)}(\boldsymbol{\Delta}_{-j,-j}^{(k-1)})^{-1}\left(\mathbf{z}_{i,-j} - (\boldsymbol{\Psi}^{(k-1)}f_{y_i})_{-j}\right)^T$ y la varianza $\widetilde{\delta}_{i,j}^2 = 1 - \boldsymbol{\Delta}_{j,-j}^{(k-1)}(\boldsymbol{\Delta}_{-j,-j}^{(k-1)})^{-1}\boldsymbol{\Delta}_{-j,j}^{(k-1)}$. Adicionalmente, la distribución condicionada de $z_{i,j}$ sobre los datos observados $x_{i,j}$ es equivalente a condicionarla sobre $z_{i,j} \in C(x_{i,j}, \boldsymbol{\Theta}) = [\theta_{x_{i,j}-1}^{(j)}, \theta_{x_{i,j}}^{(j)})$, la cual se distribuye como una normal truncada con densidad

$$f(z_{i,j}|C(x_{i,j}, \boldsymbol{\Theta}), \mathbf{z}_{i,-j}) = \frac{\frac{1}{\widetilde{\delta}_{i,j}}\phi(\widetilde{z}_{i,j})}{\Phi(\widetilde{\theta}_{x_{i,j}}^{(j)}) - \Phi(\widetilde{\theta}_{x_{i,j}-1}^{(j)})} I_{\{z_{i,j} \in C(x_{i,j}, \boldsymbol{\Theta})\}}.$$

Aquí, $\widetilde{z}_{i,j} = (z_{i,j} - \widetilde{\mu}_{i,j})/\widetilde{\delta}_{i,j}$, $\widetilde{\theta}_{x_{i,j}}^{(j)} = (\theta_{x_{i,j}}^{(j)} - \widetilde{\mu}_{i,j})/\widetilde{\delta}_{i,j}$ y $\widetilde{\theta}_{x_{i,j}-1}^{(j)} = (\theta_{x_{i,j}-1}^{(j)} - \widetilde{\mu}_{i,j})/\widetilde{\delta}_{i,j}$. A partir de la función generatriz de momentos de una distribución normal truncada, el primer y segundo momento de $z_{i,j}$ estarán dados por

$$E\left(z_{i,j}|\mathbf{z}_{i,-j}, x_{i,j}\right) = \widetilde{\mu}_{i,j} + \widetilde{\delta}_{i,j}a_{i,j}, \tag{3.17}$$

$$E\left(z_{i,j}^2|\mathbf{z}_{i,-j}, x_{i,j}\right) = \widetilde{\mu}_{i,j}^2 + \widetilde{\delta}_{i,j}^2 + 2a_{i,j}\widetilde{\mu}_{i,j}\widetilde{\delta}_{i,j} + b_{i,j}\widetilde{\delta}_{i,j}^2, \tag{3.18}$$

donde

$$a_{i,j} = \frac{\phi(\widetilde{\theta}_{x_{i,j}-1}^{(j)}) - \phi(\widetilde{\theta}_{x_{i,j}}^{(j)})}{\Phi(\widetilde{\theta}_{x_{i,j}}^{(j)}) - \Phi(\widetilde{\theta}_{x_{i,j}-1}^{(j)})} \ , \qquad b_{i,j} = \frac{\widetilde{\theta}_{x_{i,j}-1}^{(j)}\phi(\widetilde{\theta}_{x_{i,j}-1}^{(j)}) - \widetilde{\theta}_{x_{i,j}}^{(j)}\phi(\widetilde{\theta}_{x_{i,j}}^{(j)})}{\Phi(\widetilde{\theta}_{x_{i,j}}^{(j)}) - \Phi(\widetilde{\theta}_{x_{i,j}-1}^{(j)})}.$$

Usando (3.17) y (3.18) en (3.15) y (3.16), se tiene

$$E_{\mathbf{z}_i}(z_{i,j}|\mathbf{x}_i) = E_{\mathbf{z}_{i,-j}}\left(\widetilde{\mu}_{i,j}|\mathbf{x}_i\right) + \widetilde{\delta}_{i,j}E_{\mathbf{z}_{i,-j}}\left(a_{i,j}|\mathbf{x}_i\right). \tag{3.19}$$

$$E_{\mathbf{z}_i}(z_{i,j}^2|\mathbf{x}_i) = E_{\mathbf{z}_{i,-j}}\left(\widetilde{\mu}_{i,j}^2|\mathbf{x}_i\right) + \widetilde{\delta}_{i,j}^2 + 2\widetilde{\delta}_{i,j}E_{\mathbf{z}_{i,-j}}\left(a_{i,j}\widetilde{\mu}_{i,j}|\mathbf{x}_i\right) + \widetilde{\delta}_{i,j}^2E_{\mathbf{z}_{i,-j}}\left(b_{i,j}|\mathbf{x}_i\right). \tag{3.20}$$

Aquí $\widetilde{\mu}_{i,j}$ es lineal en $\mathbf{z}_{i,-j}$ luego, para $\boldsymbol{\Delta}^{(k-1)}$, $\widehat{\boldsymbol{\alpha}}$ y y_i fijos, tenemos

$$\begin{aligned} E_{\mathbf{z}_{i,-j}}\left(\widetilde{\mu}_{i,j}|\mathbf{x}_i\right) = {} & (\boldsymbol{\Psi}^{(k-1)}f_{y_i})_j + \boldsymbol{\Delta}_{j,-j}^{(k-1)}(\boldsymbol{\Delta}_{-j,-j}^{(k-1)})^{-1}E_{\mathbf{z}_{i,-j}}\left(\mathbf{z}_{i,-j}^T|\mathbf{x}_i\right) \\ & - \boldsymbol{\Delta}_{j,-j}^{(k-1)}(\boldsymbol{\Delta}_{-j,-j}^{(k-1)})^{-1}(\boldsymbol{\Psi}^{(k-1)}f_{y_i})_{-j}^T \end{aligned} \tag{3.21}$$

y

$$E_{\mathbf{z}_{i,-j}}\left(\widetilde{\mu}_{i,j}^2|\mathbf{x}_i\right) = (\boldsymbol{\Psi}^{(k-1)}f_{y_i})_j^2 + 2\left(\boldsymbol{\Psi}^{(k-1)}f_{y_i}\right)_j \boldsymbol{\Delta}_{j,-j}^{(k-1)}(\boldsymbol{\Delta}_{-j,-j}^{(k-1)})^{-1}\left[E_{\mathbf{z}_{i,-j}}\left(\mathbf{z}_{i,-j}|\mathbf{x}_i\right) - \left(\boldsymbol{\Psi}^{(k-1)}f_{y_i}\right)_{-j}\right]^T$$
$$+ \boldsymbol{\Delta}_{j,-j}^{(k-1)}(\boldsymbol{\Delta}_{-j,-j}^{(k-1)})^{-1}\Bigg[E_{\mathbf{z}_{i,-j}}\left(\mathbf{z}_{i,-j}^T\mathbf{z}_{i,-j}|\mathbf{x}_i\right) - E_{\mathbf{z}_{i,-j}}\left(\mathbf{z}_{i,-j}^T|\mathbf{x}_i\right)\left(\boldsymbol{\Psi}^{(k-1)}f_{y_i}\right)_{-j}$$
$$- \left(\boldsymbol{\Psi}^{(k-1)}f_{y_i}\right)_{-j}^T E_{\mathbf{z}_{i,-j}}\left(\mathbf{z}_{i,-j}|x_i\right) + \left(\boldsymbol{\Psi}^{(k-1)}f_{y_i}\right)_{-j}^T\left(\boldsymbol{\Psi}^{(k-1)}f_{y_i}\right)_{-j}\Bigg](\boldsymbol{\Delta}_{-j,-j}^{(k-1)})^{-1}(\boldsymbol{\Delta}_{-j,-j}^{(k-1)})^T \tag{3.22}$$

Por otro lado, tenemos que las funciones $a_{i,j}$ y $b_{i,j}$ son no lineales en $\widetilde{\theta}^j_{x_{i,j}}$ y $\widetilde{\theta}^j_{x_{i,j}-1}$, las que son funciones lineales de $\widetilde{\mu}_{i,j}$ y, por ende, de $\mathbf{z}_{i,-j}$. Entonces podemos escribir a $a_{i,j}$ y $b_{i,j}$ como $a_{i,j}(\mathbf{z}_{i,-j})$ y $b_{i,j}(\mathbf{z}_{i,-j})$. Condicionando sobre $\mathbf{x}_i$, $\mathbf{z}_{i,-j}$ tiene una distribución normal truncada con media $\widetilde{\mathbf{v}}_{i,-j} = E_{\mathbf{z}_{i,-j}}(\mathbf{z}_{i,-j}|\mathbf{x}_i)$ y matriz de varianza-covarianza $\widetilde{\mathbf{V}} = E_{\mathbf{z}_{i,-j}}\left((\mathbf{z}_{i,-j} - \widetilde{\mathbf{v}}_{i,-j})(\mathbf{z}_{i,-j} - \widetilde{\mathbf{v}}_{i,-j})^T|\mathbf{x}_i\right)$. Si suponemos que $a_{i,j}$ y $b_{i,j}$ tienen primeras derivadas parciales continuas, por el método delta de primer orden tenemos que

$$n^{1/2}\left\{a_{i,j}(\mathbf{z}_{i,-j}|\mathbf{x}_i) - a_{i,j}(\widetilde{\mathbf{v}}_{i,-j})\right\} \xrightarrow{\mathcal{D}} N\left(0, \nabla a_{i,j}(\widetilde{\mathbf{v}}_{i,-j})\widetilde{\mathbf{V}}\nabla^T a_{i,j}(\widetilde{\mathbf{v}}_{i,-j})\right) \text{ y}$$
$$n^{1/2}\left\{b_{i,j}(\mathbf{z}_{i,-j}|\mathbf{x}_i) - b_{i,j}(\widetilde{\mathbf{v}}_{i,-j})\right\} \xrightarrow{\mathcal{D}} N\left(0, \nabla b_{i,j}(\widetilde{\mathbf{v}}_{i,-j})\widetilde{\mathbf{V}}\nabla^T b_{i,j}(\widetilde{\mathbf{v}}_{i,-j})\right),$$

entonces podemos aproximar la esperanza $E_{\mathbf{z}_{i,-j}}\left(a_{i,j}|\mathbf{x}_i\right)$ por $a_{i,j}(\widetilde{\mathbf{v}}_{i,-j})$ y $E_{\mathbf{z}_{i,-j}}\left(b_{i,j}|\mathbf{x}_i\right)$ por $b_{i,j}(\widetilde{\mathbf{v}}_{i,-j})$, es decir,

$$E_{\mathbf{z}_{i,-j}}\left(a_{i,j}|\mathbf{x}_i\right) \approx a_{i,j}(\widetilde{\mathbf{v}}_{i,-j}) = \frac{\phi(\overset{\sim}{\widetilde{\theta}}{}^{(j)}_{x_{i,j}-1}) - \phi(\overset{\sim}{\widetilde{\theta}}{}^{(j)}_{x_{i,j}})}{\Phi(\overset{\sim}{\widetilde{\theta}}{}^{(j)}_{x_{i,j}}) - \Phi(\overset{\sim}{\widetilde{\theta}}{}^{(j)}_{x_{i,j}-1})} \tag{3.23}$$

$$E_{\mathbf{z}_{i,-j}}\left(b_{i,j}|\mathbf{x}_i\right) \approx b_{i,j}(\widetilde{\mathbf{v}}_{i,-j}) = \frac{\overset{\sim}{\widetilde{\theta}}{}^{(j)}_{x_{i,j}-1}\phi(\overset{\sim}{\widetilde{\theta}}{}^{(j)}_{x_{i,j}-1}) - \overset{\sim}{\widetilde{\theta}}{}^{(j)}_{x_{i,j}}\phi(\overset{\sim}{\widetilde{\theta}}{}^{(j)}_{x_{i,j}})}{\Phi(\overset{\sim}{\widetilde{\theta}}{}^{(j)}_{x_{i,j}}) - \Phi(\overset{\sim}{\widetilde{\theta}}{}^{(j)}_{x_{i,j}-1})} \tag{3.24}$$

con

$$\overset{\sim}{\widetilde{\theta}}{}^{(j)}_{x_{i,j}-k} = \frac{\widetilde{\theta}^{(j)}_{x_{i,j}-k} - E_{\mathbf{z}_{i,-j}}(\widetilde{\mu}_{i,j}|\mathbf{x}_i)}{\widetilde{\delta}_{i,j}} = \frac{\widetilde{\theta}^{(j)}_{x_{i,j}-k} - \left(\boldsymbol{\Psi}^{(k-1)}f_{y_i}\right)_j + \boldsymbol{\Delta}_{j,-j}^{(k-1)}(\boldsymbol{\Delta}_{-j,-j}^{(k-1)})^{-1}\left[\widetilde{\mathbf{v}}_{i,-j} - \left(\boldsymbol{\Psi}^{(k-1)}f_{y_i}\right)_{-j}\right]^T}{\widetilde{\delta}_{i,j}}$$

Usando (3.21), (3.22), (3.23), (3.24) y la aproximación $E_{\mathbf{z}_{i,-j}}\left(a_{i,j}\widetilde{\mu}_{i,j}|\mathbf{x}_i\right) \approx E_{\mathbf{z}_{i,-j}}\left(a_{i,j}|\mathbf{x}_i\right)E_{\mathbf{z}_{i,-j}}\left(\widetilde{\mu}_{i,j}|\mathbf{x}_i\right)$, la esperanza condicional en (3.19) puede aproximarse de la

manera

$$E_{\mathbf{z}_i}\left(z_{i,j}|\mathbf{x}_i\right) \approx (\mathbf{\Psi}^{(k-1)}f_{y_i})_j + \mathbf{\Delta}^{(k-1)}_{j,-j}(\mathbf{\Delta}^{(k-1)}_{-j,-j})^{-1}\Big[E_{\mathbf{z}_{i,-j}}\left(\mathbf{z}^T_{i,-j}|\mathbf{x}_i\right) - (\mathbf{\Psi}^{(k-1)}f_{y_i})_{-j}\Big]^T + \widetilde{\delta}_{i,j}\frac{\phi(\overset{\sim}{\theta}^{(j)}_{x_{i,j}-1}) - \phi(\overset{\sim}{\theta}^{(j)}_{x_{i,j}})}{\Phi(\overset{\sim}{\theta}^{(j)}_{x_{i,j}}) - \Phi(\overset{\sim}{\theta}^{(j)}_{x_{i,j}-1})}, \tag{3.25}$$

y el segundo momento en (3.20) por

$$\begin{aligned}
E_{\mathbf{z}_i}\left(z^2_{i,j}|\mathbf{x}_i\right) \approx{}& \left(\mathbf{\Psi}^{(k-1)}f_{y_i}\right)^2_j + 2\left(\mathbf{\Psi}^{(k-1)}f_{y_i}\right)_j \mathbf{\Delta}^{(k-1)}_{j,-j}(\mathbf{\Delta}^{(k-1)}_{-j,-j})^{-1}\left[E_{\mathbf{z}_{i,-j}}\left(\mathbf{z}_{i,-j}|\mathbf{x}_i\right) - \left(\mathbf{\Psi}^{(k-1)}f_{y_i}\right)_{-j}\right]^T \\
&+ \mathbf{\Delta}^{(k-1)}_{j,-j}(\mathbf{\Delta}^{(k-1)}_{-j,-j})^{-1}\Bigg[E_{\mathbf{z}_{i,-j}}\left(\mathbf{z}^T_{i,-j}\mathbf{z}_{i,-j}|\mathbf{x}_i\right) - E_{\mathbf{z}_{i,-j}}\left(\mathbf{z}^T_{i,-j}|\mathbf{x}_i\right)\left(\mathbf{\Psi}^{(k-1)}f_{y_i}\right)_{-j} \\
&- \left(\mathbf{\Psi}^{(k-1)}f_{y_i}\right)^T_{-j} E_{\mathbf{z}_{i,-j}}\left(\mathbf{z}_{i,-j}|x_i\right) + \left(\mathbf{\Psi}^{(k-1)}f_{y_i}\right)^T_{-j}\left(\mathbf{\Psi}^{(k-1)}f_{y_i}\right)_{-j}\Bigg](\mathbf{\Delta}^{(k-1)}_{-j,-j})^{-1}(\mathbf{\Delta}^{(k-1)}_{-j,-j})^T + \widetilde{\delta}_{i,j} \qquad (3.26)\\
&+ 2\widetilde{\delta}_{i,j}\frac{\phi(\overset{\sim}{\theta}^{(j)}_{x_{i,j}-1}) - \phi(\overset{\sim}{\theta}^{(j)}_{x_{i,j}})}{\Phi(\overset{\sim}{\theta}^{(j)}_{x_{i,j}}) - \Phi(\overset{\sim}{\theta}^{(j)}_{x_{i,j}-1})}\left[(\mathbf{\Psi}^{(k-1)}f_{y_i})_j + \mathbf{\Delta}^{(k-1)}_{j,-j}(\mathbf{\Delta}^{(k-1)}_{-j,-j})^{-1}\Big[E_{\mathbf{z}_{i,-j}}\left(\mathbf{z}^T_{i,-j}|\mathbf{x}_i\right) - (\mathbf{\Psi}^{(k-1)}f_{y_i})_{-j}\Big]^T\right] \\
&+ \widetilde{\delta}^2_{i,j}\frac{\overset{\sim}{\theta}^{(j)}_{x_{i,j}-1}\phi(\overset{\sim}{\theta}^{(j)}_{x_{i,j}-1}) - \overset{\sim}{\theta}^{(j)}_{x_{i,j}}\phi(\overset{\sim}{\theta}^{(j)}_{x_{i,j}})}{\Phi(\overset{\sim}{\theta}^{(j)}_{x_{i,j}}) - \Phi(\overset{\sim}{\theta}^{(j)}_{x_{i,j}-1})}.
\end{aligned}$$

La ecuaciones (3.25) y (3.26) dan expresiones recursivas para computar (iterativamente) $\mathbf{S}$ y $\mathbf{M}$, respectivamente.

3.9.3. Descripción de las variables para la construcción del índice SES

En este apéndice detallamos las variables predictoras utilizadas para la construcción de los índices SES:

- *Ubicación de la vivienda*: Indica si la vivienda está ubicada en una zona desfavorable o en una área vulnerable. Más precisamente, esta variable indica si la vivienda: (i) está ubicada en una zona inundable, (ii) o/y cerca de un basural, (iii) o/y en una villa de emergencia.

Esta variable tiene 4 categorías: vale 1 si la vivienda presenta conjuntamente las características (i)-(iii), 2 para viviendas que presentan dos de las características (i)-(iii), 3 si la vivienda posee solo una de estas características, y 4 si la vivienda no tienen ninguna de tales características.

- *Calidad de la vivienda*: Contempla de forma conjunta la calidad (materiales) del techo, paredes y pisos de la vivienda, en base a la la metodología CALMAT (INDEC 2003) usada para el Censo Poblacional de Argentina. Tiene 4 categorías en orden creciente en términos de la calidad de la vivienda (i.e. asume un valor mayor, cuando la calidad de la vivienda es superior).

- *Combustible para cocinar*: Indica el tipo de combustible predominante en la vivienda para la preparación de los alimentos. Tiene 3 categorías: vale 1 si el principal combustible para cocinar es kerosene, madera o carbón, 2 si tiene gas envasado, y 3 si posee gas natural por tubería.

- *Hacinamiento*: Caracteriza el hacinamiento del hogar y se deriva del cómputo del ratio entre ambientes de la vivienda y la cantidad de miembros del hogar. Tiene 4 categorías: 1 si este ratio es menor o igual a 1, 2 si el ratio está en el intervalo (1,2], 3 si este está en (2,3], y 4 el ratio es mayor a 3.

- *Escolaridad*: Indica el nivel de instrucción formal alcanzado por el jefe/a del hogar. Tiene 7 categorías: 1 si el jefe/a de hogar no posee educación formal, 2 si tiene primaria incompleta, 3 si tiene primaria completa, 4 si realizó secundaria incompleta, 5 si realizó escuela secundaria completa, 6 si tiene estudios superiores incompletos y 7 si el jefe/a posee titulo terciario o universitario.

- *Horas trabajadas*: Describe las situación laboral del jefe/a de hogar. Tiene 4 categorías: 1 si está desempleado o inactivo, 2 cuando el jefe/a de hogar trabaja menos de 40 horas semanales, 3 para 40-45 horas semanales de trabajo, y 4 cuando el jefe/a de hogar está empleado con más de 45 horas semanales.

- *Desagüe*: Indica el tipo de drenaje o desagüe que posee el baño la vivienda. Tiene 4 categorías: 1 si el desagüe consisten en un agujero, 2 si el desagüe es en un pozo negro, 3 si es pozo negro con cámara séptica, y 4 para tuberías de desagüe en una red pública.

- *Instalación sanitaria*: Indica el tipo de instalación sanitaria que posee la vivienda. Tiene 3 categorias: 1 para letrinas, 2 para baño con inodoro sin descarga de agua, y 3 para baños con descarga de agua.

- *Baño compartido*: Indica si el baño es compartido o no. Tiene 3 categorías: 1 si el baño está fuera de la vivienda y es compartido con otras, 2 si el baño es compartido con otros hogares dentro de la vivienda, y 3 si el baño es de uso exclusivo del hogar.

- *Ubicación del agua*: Indica la ubicación más cercana para obtener agua potable. Tiene 3 categorías: 1 si el agua potable se consigue fuera del terreno de la vivienda, 2 si el agua se obtiene dentro del terreno, pero fuera de la vivienda, y 3 si el agua potable se obtiene en el interior vivienda por tubería.

- *Provisión de agua*: Indica la fuente de donde proviene el agua de la vivienda. Tiene 3 categorías: 1 si el agua potable proviene de una bomba de mano o de un grifo público compartido con los vecinos, 2 si el agua potable se obtiene mediante una bomba de perforación automatizada, y 3 si la vivienda tiene agua potable por tubería.

Capítulo 4

Reducción Suficiente de Dimensiones para Predictores Mixtos

4.1. Introducción

Los problemas de regresión con predictores de diversa naturaleza constituyen más una regla que una excepción en las aplicaciones con datos económicos y sociales, como así también en otras disciplinas científicas aplicadas. Por ende, la extensión de los métodos de Reducción Suficiente de Dimensiones (SDR) para variables mixtas resulta crucial para que los mismos alcancen mayor difusión en aplicaciones con datos reales.

En el capítulo anterior nos focalizamos en el caso de regresión con predictores ordinales, motivados por el hecho de que gran parte de los micro-datos disponibles para el estudio de los estándares de vida de la población y la elaboración de programas sociales, están conformados por variables categóricas que generalmente presentan un orden natural respecto al fenómeno socio-económico que se busca describir o predecir. Sin embargo, usualmente contienen también algunas variables continuas, como ser la edad, las horas trabajadas, la cantidad de personas por grupo etario, la tasa de hacinamiento o el índice de masa corporal, entre otras, que podrían realizar un gran aporte para la construcción de un índice de estatus socio-económico. Aunque es posible extender la metodología propuesta para predictores ordinales al caso en el que coexisten tanto predictores ordinales como continuos, dicha extensión no surge de manera totalmente directa.

En este tipo de bases de datos también es frecuente encontrar dicotómicas o binarias. Si bien una gran mayoría revela algún aspecto ordinal respecto al nivel socio-económico, como ser la posesión de activos físicos (e.g. vivienda, TV, CPU, internet, etc.), otras variables categóricas, como ser el sexo o la adherencia religiosa, que no tienen a priori una connotación ordinal. Aún teniendo un sentido ordinal, puede resultar algo *naive* o muy forzado suponer para una binaria la existencia de una variable continua latente subyacente a la misma y que se distribuye normalmente. Por ello, resultaría más adecuado pensar en un proceso más natural de generación de este tipo de variables, como ser uno Bernoulli.

Si bien existen metodologías desarrolladas para SDR que permiten combinar algunos tipos de variables, como ser la extensión de Cook and Li (2002*b*), y la de Bura et al. (2015) para familias exponenciales, que admiten la combinación de continuas con binarias, no existen contribuciones de este enfoque que permitan conjuntamente predictores ordinales, continuos y binarios. Además como expresamos en la Capítulo 2, la metodología de Cook and Li (2002*b*) es bastante restringida al suponer predictores condicionalmente independientes. Por su parte, la metodología de Bura et al. (2015) fue aplicada sólo para el caso de tener predictores dicotómicos únicamente, o bien con un mezcla de predictores dada por una binaria y varias continuas.

El objetivo de este capítulo es estudiar la reducción suficiente de dimensiones para el problema de regresión en el que co-existen predictores continuos, dicotómicos y ordinales. Es decir, supondremos que el modelo de regresión de interés viene dado por

$$Y|\mathbf{X}, \mathbf{W}, \mathbf{H} \tag{4.1}$$

donde $\mathbf{W}^T = (W_1, W_2, \ldots, W_s)$ es un vector de s variables continuas, $\mathbf{X}^T = (X_1, X_2, \ldots, X_p)$ es un vector de p variables ordinales, $\mathbf{H}^T = (H_1, H_2, \ldots, H_q)$ un vector de q variables binarias, e Y la variable respuesta que en principio puede ser de continua o discreta. El objetivo es encontrar una función $\mathbf{R} : \mathbb{R}^p \times \mathbb{R}^s \times \mathbb{R}^q \to \mathbb{R}^d$ del vector de predictores $\mathbf{R}(\mathbf{X}, \mathbf{W}, \mathbf{H})$ tal que $F(Y|\mathbf{X}, \mathbf{W}, \mathbf{H}) = F(Y|\mathbf{R}(\mathbf{X}, \mathbf{W}, \mathbf{H}))$, donde $F(\cdot|\cdot)$ representa la función de distribución condicionada.

Siguiendo el enfoque de reducción suficiente de dimensiones basado en la regresión inversa, debemos especificar una distribución para $\mathbf{X}, \mathbf{W}, \mathbf{H}|Y$ a fin de obtener el estadístico suficiente para Y y por ende la reducción suficiente para Y dado $(\mathbf{X}, \mathbf{W}, \mathbf{H})$. En este capítulo se pre-

sentará una cierta factorización para la distribución de $\mathbf{X}, \mathbf{W}, \mathbf{H}|Y$ que permitirá identificar y estimar la reducción suficiente en este contexto.

Específicamente, la distribución conjunta condicionada de $\mathbf{X}, \mathbf{W}, \mathbf{H}|Y$ puede escribirse de la forma

$$f(\mathbf{X}, \mathbf{W}, \mathbf{H}|Y) = f(\mathbf{X}, \mathbf{W}|Y, \mathbf{H}) f(\mathbf{H}|Y). \tag{4.2}$$

Esta factorización implica modelar por un lado las variables $(\mathbf{X}, \mathbf{W})$ como función de Y y $\mathbf{H}$, y por otro lado, modelar las variables dicotómicas $\mathbf{H}$ en función de Y.

En base a (4.2), proponemos un modelo paramétrico para $\mathbf{X}, \mathbf{W}|Y, \mathbf{H}$ y otro para $\mathbf{H}|Y$, bajo los cuales identificamos la reducción suficiente para la regresión $Y|\mathbf{X}, \mathbf{W}, \mathbf{H}$, pudiéndose estimar esta reducción suficiente vía máxima verosimilitud.

El modelo y los métodos propuestos resultan ser una extensión combinada de los resultados del capítulo previo de SDR para ordinales y de la generalización para familias exponenciales de Bura et al. (2015). No obstante, como se podrá apreciar a lo largo del capítulo, existen sutilezas para nada triviales en esta extensión, obteniendo un modelo y método particular que se diferencia sustancialmente de los aportes previos en los que se basa.

Lo que resta del presente capítulo se organiza de la siguiente manera: En la Sección 4.2 se especifica el modelo, para luego identificar la reducción suficiente en la Sección 4.3. Posteriormente se presenta la metodología de estimación (Sección 4.4). En la secciones 4.5 y 4.6 se muestran resultados de simulaciones y aplicaciones con datos reales para la construcción de índices de estatus socio-económico para predicción de ingreso y pobreza, enriqueciendo el ejemplo de índices presentado en el capítulo previo. Las conclusiones del capítulo se presentan al cierre en la Sección 4.7.

4.2. Modelo

Partiendo de (4.2) asumiremos dos modelos paramétricos para las regresiones inversas $\mathbf{X}, \mathbf{W}|Y, \mathbf{H}$ y $\mathbf{H}|Y$ respectivamente. Específicamente, para $\mathbf{X}, \mathbf{W}|Y$ utilizaremos, como en el capítulo previo, el enfoque de variables latentes para la parte ordinal $\mathbf{X}$. De esta manera supondremos la existencia de un vector p-dimensional $\mathbf{Z}$ de variables continuas latentes para el

vector $\mathbf{X}$ de variables ordinales. Por ende, para cada $j = 1, \ldots, p$, supondremos que existe un conjunto de umbrales $-\infty = \theta_0^{(j)} < \theta_1^{(j)} < \ldots < \theta_{G_j}^{(j)} = \infty$ tales que

$$X_j = k \quad \text{sí y sólo si} \quad Z_j \in (\theta_{k-1}^{(j)}, \theta_k^{(j)}], \quad k = 1, \ldots, G_j. \tag{4.4}$$

Considerando conjuntamente a las variables continuas latentes ($\mathbf{Z}$) y a las observadas ($\mathbf{W}$), asumiremos que se distribuyen normalmente, una vez que condicionamos en Y y $\mathbf{H}$. Denotando por $\mathbf{V}^T = (\mathbf{Z}^T, \mathbf{W}^T)$ al vector t-dimensional ($t = p + s$) de variables continuas, asumiremos el siguiente modelo

$$\begin{aligned} \mathbf{V}|\mathbf{H}, Y &\sim N(\boldsymbol{\mu} + \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \boldsymbol{\beta}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}), \boldsymbol{\Delta}), \\ \Pr(X_j = k|\mathbf{H}, Y) &= \Pr(\theta_{k-1}^{(j)} \leq Z_j < \theta_k^{(j)}|\mathbf{H}, Y), \quad k = 1, \ldots, G_j, \end{aligned} \tag{4.5}$$

donde, $\boldsymbol{\mu} = E(\mathbf{V}|Y, \mathbf{H})$, $\boldsymbol{\Delta}$ es la matriz de covarianza $t \times t$, $\boldsymbol{\alpha}$ es $t \times d$, $\boldsymbol{\xi}$ es $d \times r$, $\mathbf{f}_Y$ es una función conocida de Y de dimensión $r \times 1$, $\bar{\mathbf{f}}_Y = E_Y(\mathbf{f}_Y)$, $\boldsymbol{\beta}$ es de orden $t \times q$ y $(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})$ tiene dimensión $q \times 1$, donde $\boldsymbol{\mu}_{\mathbf{H}} = E_{\mathbf{H}}(\mathbf{H})$. Denotamos por $\boldsymbol{\Theta} \equiv \{\boldsymbol{\Theta}^1, \ldots, \boldsymbol{\Theta}^p) \equiv \{\theta_1^{(1)}, \ldots, \theta_{G_1}^{(1)}, \ldots, \theta_p^{(p)}, \ldots, \theta_{G_p}^{(p)}\}$ el conjunto de umbrales. Adicionalmente, podemos descomponer la media y la varianza de la forma $\boldsymbol{\mu}^T = (\boldsymbol{\mu}_{\mathbf{z}}^T, \boldsymbol{\mu}_{\mathbf{w}}^T)$ y $\boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{\Delta}_{\mathbf{zz}} & \boldsymbol{\Delta}_{\mathbf{zw}} \\ \boldsymbol{\Delta}_{\mathbf{wz}} & \boldsymbol{\Delta}_{\mathbf{ww}} \end{pmatrix}$, donde $\boldsymbol{\Delta}_{\mathbf{zz}}$ y $\boldsymbol{\Delta}_{\mathbf{ww}}$ son las matrices de varianza-covarianza de las variables latentes y de las continuas observadas, respectivamente, y $\boldsymbol{\Delta}_{\mathbf{zw}}$ es la matriz de covarianza entre $\mathbf{Z}$ y $\mathbf{W}$, con $\boldsymbol{\Delta}_{\mathbf{wz}} = \boldsymbol{\Delta}_{\mathbf{zw}}^T$.

Dado que $\mathbf{H}$ es un vector de variables dicotómicas, es natural suponer una distribución Bernoulli multivariada. Debido a que esta distribución involucra todas las interacciones de segundo orden y superiores entre las variables, para obtener una representación más parsimoniosa de la distribución conjunta de variables Bernoulli es usual tomar el caso especial del denominado modelo Ising (Dai et al. 2013, Dai 2013, Cheng et al. 2014), el que incorpora solo interacciones de a pares. Siguiendo a Cheng et al. (2014), condicionando en Y, el vector de las q variables binarias $\mathbf{H}^T = (H_1, H_2, \ldots, H_q)$ tienen la siguiente función de densidad conjunta

$$\begin{aligned} P(\mathbf{H}|Y = y) &= P(H_1, \ldots, H_q|Y = y) \\ &= \frac{1}{G(\boldsymbol{\Gamma}^y)} \exp\left(\sum_{j=1}^{q} \gamma_{jj}^y H_j + \sum_{1 \leq j < j' \leq q} \gamma_{jj'}^y H_j H_{j'}\right), \end{aligned} \tag{4.6}$$

donde, para $j, j' = 1, \ldots, p$,

$$\gamma_{jj}^{y} = \log\left(\frac{\Pr(H_j = 1|\mathbf{H}_{-j} = \mathbf{0}, y)}{1 - \Pr(H_j = 1|\mathbf{H}_{-j} = \mathbf{0}, y)}\right),$$

$$\gamma_{jj'}^{y} = \log\frac{\Pr(H_j = 1, H_{j'} = 1|\mathbf{H}_{-j,-j'} = \mathbf{0}, y)\Pr(H_j = 0, H_{j'} = 0|\mathbf{H}_{-j,-j'} = \mathbf{0}, y)}{\Pr(H_j = 1, H_{j'} = 0|\mathbf{H}_{-j,-j'} = \mathbf{0}, y)\Pr(H_j = 0, H_{j'} = 1|\mathbf{H}_{-j,-j'} = \mathbf{0}, y)},$$

$$G(\mathbf{\Gamma}^y) = \sum_{j=1}^{q}\sum_{\{H_j=0,1\}}\exp\left(\sum_{j=1}^{q}\gamma_{jj}^{y}H_j + \sum_{1\leq j<j'\leq q}\gamma_{jj'}^{y}H_jH_{j'}\right),$$

con $\mathbf{H}_{-j} = (H_1, \ldots, H_{j-1}, H_{j+1}, \ldots, H_q)$, $\mathbf{H}_{-j,-j'} = (H_1, \ldots, H_{j-1}, H_{j+1}, \ldots, H_{j'-1}, H_{j'+1}, \ldots, H_q)$ y $\mathbf{\Gamma}^y$ es una matriz simétrica cuyo elementos vienen dados por $[\mathbf{\Gamma}^y]_{ij} = \gamma_{ij}^{y}$. Por lo tanto, usando la simetría de $\mathbf{\Gamma}^y$ y el hecho de que para todo j, $H_j^2 = H_j$, podemos re-escribir el modelo (4.6) como

$$P(\mathbf{H}|Y = y) = \frac{1}{G(\mathbf{\Gamma}^y)}\exp\left\{\,\mathrm{vech}^T(\mathbf{H}\mathbf{H}^T)\,\mathrm{vech}(\mathbf{\Gamma}^y)\right\}, \tag{4.7}$$

donde vech es la vectorización en columna de la parte triangular inferior de una matriz simétrica y $G(\mathbf{\Gamma}^y) = \sum_{\mathbf{H}\in\mathcal{H}}\exp\left(\,\mathrm{vech}^T(\mathbf{H}\mathbf{H}^T)\,\mathrm{vech}(\mathbf{\Gamma}^y)\right)$, donde $\mathcal{H} = \{$todas las posibles combinaciones $\mathbf{H} \in \{0,1\}^q\}$.

Siguiendo a Cheng et al. (2016), cada γ_{ij}^{y} es modelado como función lineal de la función de la variable respuesta Y, $\mathbf{f}_Y \in \mathbb{R}^r$, y $\mathbf{V}$. Específicamente, supondremos que

$$\gamma_{ij}^{y} = \tau_{ij,0}^{*} + \boldsymbol{\tau}_{ij}^{T}\left(\mathbf{f}_Y - \bar{\mathbf{f}}_Y\right), \qquad i, j = 1, \ldots, q \tag{4.8}$$

donde $\boldsymbol{\tau}_{ij}^{T} = (\tau_{ij,1}, \ldots, \tau_{ij,r})$ es un vector de parámetros (independientes de Y) y $\tau_{ij,0}^{*}$ es el intercepto para cada (i, j).

Para poder escribir $\mathrm{vech}(\mathbf{\Gamma}^y)$ de forma matricial, definimos las matrices de orden $q \times q$, $\boldsymbol{\tau}_0$ y $\boldsymbol{\tau}_k$ $(k = 1, \ldots, r)$, de la siguiente manera: $[\boldsymbol{\tau}_0^{*}]_{ij} = \tau_{ij,0}^{*}$, y $[\boldsymbol{\tau}_k]_{ij} = \tau_{ij,k}$ con $i, j = 1, \ldots, q$ y $k = 1, \ldots, r$. A partir de estas matrices, definimos el vector $\boldsymbol{\tau}_0 \doteq \mathrm{vech}(\boldsymbol{\tau}_0^{*})$ de dimensión $q(q+1)/2$, y las matriz $\boldsymbol{\tau} = [\,\mathrm{vech}(\boldsymbol{\tau}_1), \ldots, \mathrm{vech}(\boldsymbol{\tau}_r)] \in \mathbb{R}^{q(q+1)/2\times r}$. Adicionalmente supondremos que $\boldsymbol{\tau}$ es de rango $c \leq \min(r, q(q+1)/2)$ de forma tal que puede escribirse de la forma $\boldsymbol{\tau} = \boldsymbol{\kappa}\boldsymbol{\iota}$, donde

$\boldsymbol{\kappa} \in \mathbb{R}^{q(q+1)/2 \times c}$ y $\boldsymbol{\iota} \in \mathbb{R}^{c \times r}$ son matrices de rango completo. De esta manera, $\text{vech}(\boldsymbol{\Gamma}^y)$ puede modelarse de la forma

$$\text{vech}(\boldsymbol{\Gamma}^y) = \boldsymbol{\tau}_0 + \boldsymbol{\kappa}\boldsymbol{\iota}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y). \tag{4.9}$$

Por lo tanto, (4.7) vendrá dado por

$$P(\mathbf{H}|Y = y) = \frac{1}{G(\boldsymbol{\Gamma}^y)} \exp\left\{ \text{vech}^T(\mathbf{H}\mathbf{H}^T)[\boldsymbol{\tau}_0 + \boldsymbol{\kappa}\boldsymbol{\iota}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)] \right\}, \tag{4.10}$$

De esta manera, la regresión inversa $\mathbf{X}, \mathbf{W}, \mathbf{H}|Y$ queda caracterizada por los modelos (4.5) y (4.10). Para que los parámetros del modelo estén identificados, debemos imponer ciertas condiciones o restricciones sobre los mismos. En particular, debemos imponer ciertas condiciones para el modelo $\mathbf{X}, \mathbf{W}|\mathbf{H}, Y$, debido a que el modelo Ising $\mathbf{H}|Y$ como caso especial de un modelo Bernoulli multivariado es identificable (Dai et al. 2013). Tales condiciones de identificabilidad están explicitas en el siguiente teorema.

Teorema 4.1 *Bajo las condiciones,*

1. $[\boldsymbol{\Delta}_{\mathbf{zz}}]_{jj} = 1$ *para todo* $j = 1, \ldots, p$*;*
2. $\boldsymbol{\mu}_{\mathbf{z}} = \mathbf{0}$*;*
3. *La matriz* $E\left(\begin{bmatrix} (\mathbf{f}_Y - \bar{\mathbf{f}}_Y)(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)^T & (\mathbf{f}_Y - \bar{\mathbf{f}}_Y)(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})^T \\ (\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)^T & (\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})^T \end{bmatrix} \right)$ *es invertible;*

el modelo especificado por (4.5) es identificable.

Prueba. En la Sección 4.8.2 del Apéndice.

Presentado el modelo para esta factorización de la distribución conjunta de $\mathbf{X}, \mathbf{W}, \mathbf{H}|Y$, en la próxima sección identificaremos la reducción suficiente.

4.3. Reducción Suficiente

Para identificar la reducción suficiente del problema $Y|\mathbf{X}, \mathbf{W}, \mathbf{H}$, vamos a estudiar en primer lugar la reducción suficiente $Y|\mathbf{Z}, \mathbf{W}, \mathbf{H}$, de forma análoga a lo realizado en Capítulo 3. Para

simplificar notación, definimos $\mathbf{A} \doteq \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}$. Luego, bajo los modelos (4.5) y (4.10), la distribución para $\mathbf{V}, \mathbf{H}|Y$ vendrá dada por

$$\begin{aligned} f(\mathbf{V}, \mathbf{H}|Y = y) \quad &= (2\pi)^{-\frac{t}{2}} |\boldsymbol{\Delta}|^{-\frac{1}{2}} \exp\Big\{ -\tfrac{1}{2} \big((\mathbf{V} - \boldsymbol{\mu}) - \mathbf{A}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})\big)^T \\ &\qquad \boldsymbol{\Delta}^{-1} \big((\mathbf{V} - \boldsymbol{\mu}) - \mathbf{A}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})\big) \\ &\qquad + \mathrm{vech}^T(\mathbf{H}\mathbf{H}^T)[\boldsymbol{\tau}_0 + \boldsymbol{\tau}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)] - \log(G(\boldsymbol{\Gamma}^y)) \Big\}. \end{aligned} \tag{4.11}$$

Re-ordenando convenientemente los términos de (4.11), podemos ver que la distribución condicional conjunta de $\mathbf{V}, \mathbf{H}|Y$ tiene la forma de la función de densidad general para familias exponenciales,

$$f(\mathbf{V}, \mathbf{H}|Y = y) = h(\mathbf{V}, \mathbf{H}) \exp\{\mathbf{T}^T(\mathbf{V}, \mathbf{H})\boldsymbol{\eta}_y + \mathbf{U}^T(\mathbf{V}, \mathbf{H})\tilde{\boldsymbol{\eta}} - \psi(\boldsymbol{\eta}_y, \tilde{\boldsymbol{\eta}})\}, \tag{4.12}$$

donde

$$h(\mathbf{V}, \mathbf{H}) = (2\pi)^{-\frac{t}{2}}$$

$$\begin{aligned} \psi(\boldsymbol{\eta}_y, \tilde{\boldsymbol{\eta}}) = \frac{1}{2}\log|\boldsymbol{\Delta}| + \frac{1}{2}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)^T\mathbf{A}^T\boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \boldsymbol{\mu}^T\boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) \\ - \boldsymbol{\mu}_{\mathbf{H}}^T\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \log G(\boldsymbol{\Gamma}^y) + \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} + \frac{1}{2}\boldsymbol{\mu}_{\mathbf{H}}^T\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} \end{aligned}$$

$$\begin{aligned} \mathbf{U}^T(\mathbf{V}, \mathbf{H})\tilde{\boldsymbol{\eta}} = -\frac{1}{2}\mathrm{vec}^T(\mathbf{V}\mathbf{V}^T)\mathrm{vec}(\boldsymbol{\Delta}^{-1}) - \frac{1}{2}\mathrm{vec}^T(\mathbf{H}\mathbf{H}^T)\mathrm{vec}(\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}) \\ + \mathbf{V}^T\boldsymbol{\Delta}^{-1}(\boldsymbol{\mu} - \boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}}) + \mathbf{H}^T\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}(\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} - \boldsymbol{\mu}) \\ + \mathrm{vec}^T(\mathbf{H}^T\mathbf{V})\mathrm{vec}(\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}) + \mathrm{vech}^T(\mathbf{H}\mathbf{H}^T)\boldsymbol{\tau}_0; \end{aligned}$$

y

$$\mathbf{T}^T(\mathbf{V}, \mathbf{H})\boldsymbol{\eta}_y = \mathbf{V}^T\boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \mathbf{H}^T\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \mathrm{vech}^T(\mathbf{H}\mathbf{H}^T)\boldsymbol{\tau}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y).$$

De esto se obtiene que

$$\mathbf{U}(\mathbf{V},\mathbf{H}) = \begin{pmatrix} \mathbf{V} \\ \mathbf{H} \\ \text{vec}(\mathbf{H}^T\mathbf{V}) \\ \text{vech}(\mathbf{H}\mathbf{H}^T) \\ \text{vec}(\mathbf{V}\mathbf{V}^T) \end{pmatrix},$$

$$\tilde{\boldsymbol{\eta}} = (\boldsymbol{\Delta}^{-1}(\boldsymbol{\mu} - \boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}}), \boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}(\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}} - \boldsymbol{\mu}), \text{vec}(\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}), -\frac{1}{2}\text{vec}(\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}), -\frac{1}{2}\text{vec}(\boldsymbol{\Delta}^{-1}))^T$$

y

$$\mathbf{T}(\mathbf{V},\mathbf{H}) = \begin{pmatrix} \mathbf{V} \\ \text{vech}(\mathbf{H}\mathbf{H}^T) \end{pmatrix}. \tag{4.13}$$

Definiendo $\mathbf{L} \in \mathbb{R}^{q(q+1)/2\times r}$ como $\mathbf{L} = [\,\text{vech}(\text{diag}([\boldsymbol{\beta}^T\boldsymbol{\alpha}\boldsymbol{\xi}]_{\bullet 1}), \ldots, \text{vech}(\text{diag}([\boldsymbol{\beta}^T\boldsymbol{\alpha}\boldsymbol{\xi}]_{\bullet r})]$, donde $[\boldsymbol{\beta}^T\boldsymbol{\alpha}\boldsymbol{\xi}]_{\bullet k}$ representa la k-ésima columna de la matriz $q \times r$ $\boldsymbol{\beta}^T\boldsymbol{\alpha}\boldsymbol{\xi} = \boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A}$, $k = 1, \ldots, r$, obtenemos el siguiente parámetro natural

$$\boldsymbol{\eta}_y - \bar{\boldsymbol{\eta}}_y = \begin{pmatrix} \boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) \\ (\boldsymbol{\tau} - \mathbf{L})(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}\boldsymbol{\xi} \\ \boldsymbol{\kappa}\iota - \mathbf{L} \end{pmatrix}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y), \tag{4.14}$$

siendo $\bar{\boldsymbol{\eta}}_y = E_Y(\boldsymbol{\eta}_y)$.

Por lo tanto, la densidad conjunta pertenece a una familia exponencial, por lo que podemos utilizar los resultados del Teorema 2.1, que en el presente modelo se resumen en la siguiente proposición:

Proposición 4.1 *Suponiendo que* $(\mathbf{V},\mathbf{H})|Y$ *tiene la densidad (4.11), la reducción suficiente minimal para la regresión* $Y|(\mathbf{V},\mathbf{H})$ *está dada por*

$$\mathbf{R}(\mathbf{V},\mathbf{H}) = \mathbf{a}^T(\mathbf{T}(\mathbf{V},\mathbf{H}) - \text{E}(\mathbf{T}(\mathbf{V},\mathbf{H}))),$$

donde $\mathbf{T}(\mathbf{V},\mathbf{H})$ *viene dado por (4.13) y* $\mathbf{a}$ *es tal que* $\mathcal{S}_{\mathbf{a}} = \text{span}\{\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}, Y \in \mathcal{Y}\}$ *con* $\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}$ *dado en (4.14).*

Prueba. El resultado se sigue del Teorema 2.1 del Capítulo 2.

Debe notarse que si en lugar de tomar como estadístico minimal suficiente (4.13) y (4.14) como parámetro natural, consideráramos es su lugar

$$\mathbf{T}(\mathbf{V},\mathbf{H}) = \begin{pmatrix} \mathbf{V} \\ \mathbf{H} \\ \text{vech}(\mathbf{H}\mathbf{H}^T) \end{pmatrix}, \tag{4.15}$$

y

$$\boldsymbol{\eta}_y - \bar{\boldsymbol{\eta}}_y = \begin{pmatrix} \boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) \\ -\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) \\ \boldsymbol{\tau}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}\boldsymbol{\xi} \\ -\boldsymbol{\beta}^T\boldsymbol{\alpha}\boldsymbol{\xi} \\ \boldsymbol{\kappa}\boldsymbol{\iota} \end{pmatrix} (\mathbf{f}_Y - \bar{\mathbf{f}}_Y), \tag{4.16}$$

luego la reducción suficiente $\mathbf{R}(\mathbf{V},\mathbf{H}) = \mathbf{a}^T(\mathbf{T}(\mathbf{V},\mathbf{H}) - \mathrm{E}(\mathbf{T}(\mathbf{V},\mathbf{H})))$ es idéntica a la de la Proposición 4.1 cuando el estadístico suficiente es (4.13) y el parámetro natural (4.14). Considerando (4.15) y (4.16), es sencillo ver que

$$\mathbf{a} = \begin{cases} (\boldsymbol{\alpha}^T, -\boldsymbol{\alpha}^T\boldsymbol{\beta}, (\boldsymbol{\xi}\boldsymbol{\xi}^T)^{-1}\boldsymbol{\xi}\boldsymbol{\iota}^T\boldsymbol{\kappa}^T)^T & \text{si} \quad d \geq c \\ ((\boldsymbol{\iota}\boldsymbol{\iota}^T)^{-1}\boldsymbol{\iota}\boldsymbol{\alpha}^T, -(\boldsymbol{\iota}\boldsymbol{\iota}^T)^{-1}\boldsymbol{\iota}\boldsymbol{\alpha}^T\boldsymbol{\beta}, \boldsymbol{\kappa}^T)^T & \text{si} \quad d < c. \end{cases} \tag{4.17}$$

La forma de expresar la reducción utilizando (4.15) y (4.16), y con ello (4.17), puede ser conveniente para la estimación regularizada. Esto es así porque la regularización puede realizarse de forma separada sobre los parámetros de componente normal por un lado, y sobre el modelo Ising por otro, como se detallará en la próxima sección de estimación.

En el siguiente teorema probaremos que la reducción suficiente de dimensiones para la regresión de Y sobre $(\mathbf{X},\mathbf{W},\mathbf{H})$ es una combinación lineal del estadístico suficiente (4.13) donde los coeficientes forman una base del subespacio generado por las columnas de (4.14). Por el razonamiento anterior, lo mismo vale para el estadístico suficiente (4.15) y el subespacio generado por las columnas de (4.16). Este Teorema (y su prueba) constituyen una extensión del Teorema 3.1 para el caso de regresión con todos los predictores ordinales. Sin embargo, vale la pena replicarlo para este caso de predictores mixtos, a los fines de tener una exposición de la tesis más completa y detallada.

Teorema 4.2 *Sea* $\mathbf{H}^T = (H_1, H_2, \ldots, H_q)$ *un vector de* q *variables binarias tal que verifica el modelo (4.6),* $\mathbf{W}^T = (W_1, W_2, \ldots, W_s)$ *un vector de* s *variables continuas y* $\mathbf{X}^T =$

$(X_1, X_2, \ldots, X_p)$ un vector de p variables ordinales. Suponiendo que existe un variable latente $\mathbf{Z}$ *subyacente de* $\mathbf{X}$ *que es continua y tal que* $(\mathbf{Z}, \mathbf{W})$ *verifica el modelo (4.5). Luego, una reducción suficiente para la regresión de* Y *sobre* $(\mathbf{X}, \mathbf{W}, \mathbf{H})$ *está dada por*

$$\mathbf{R}(\mathbf{X}, \mathbf{W}, \mathbf{H}) = \mathbf{a}^T(\mathbf{T}(\mathbf{X}, \mathbf{W}, \mathbf{H}) - \mathrm{E}(\mathbf{T}(\mathbf{X}, \mathbf{W}, \mathbf{H}))),$$

donde $\mathbf{a}$ *es tal que* $\mathcal{S}_{\mathbf{a}} = \mathrm{span}\{\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}, Y \in \mathcal{Y}\}$ *con* $\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}$ *dado por (4.14) y*

$$\mathbf{T}(\mathbf{X}, \mathbf{W}, \mathbf{H}) = \begin{pmatrix} \mathbf{X} \\ \mathbf{W} \\ \mathrm{vech}(\mathbf{H}\mathbf{H}^T) \end{pmatrix}$$

Prueba. Por la proposición 4.1, $\mathbf{R}(\mathbf{V}, \mathbf{H}) = \mathbf{a}^T(\mathbf{T}(\mathbf{V}, \mathbf{H}) - \mathrm{E}(\mathbf{T}(\mathbf{V}, \mathbf{H})))$ con $\mathbf{T}(\mathbf{V}, \mathbf{H})$ dado por (4.13) y $\mathcal{S}_{\mathbf{a}} \equiv \mathrm{span}(\mathbf{a}) = \mathrm{span}\{\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}, Y \in \mathcal{Y}\}$ con $\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}$ dado en (4.14) constituye una reducción suficiente para $Y|(\mathbf{V}, \mathbf{H})$. De ello se deduce que $(\mathbf{V}, \mathbf{H})|(\mathbf{R}(\mathbf{V}, \mathbf{H}), Y)$ no depende de Y. Por lo tanto, en base al mismo razonamiento del Teorema 3.1, $(\mathbf{X}, \mathbf{W}, \mathbf{H})|(\mathbf{R}(\mathbf{X}, \mathbf{W}, \mathbf{H}), Y)$ no depende de Y pues, tomando $\boldsymbol{g} = (g_1, \ldots, g_p)$, $g_j = 1, \ldots, G_j$, dado que $X_j = g_j \Leftrightarrow Z_j \in [\theta^{(j)}_{X_j - 1}, \theta^{(j)}_{X_j})$ se tiene que

$$\begin{aligned}
&\Pr\left(\mathbf{X} = \mathbf{g}, \mathbf{W}, \mathbf{H} = \mathbf{h} | \mathbf{R}(\mathbf{X}, \mathbf{W}, \mathbf{H}) = \mathbf{R}(\mathbf{g}, \mathbf{W}, \mathbf{h}), Y\right) \\
&\qquad = \Pr\left(\mathbf{Z} \in C(\mathbf{X}, \boldsymbol{\Theta}), \mathbf{W}, \mathbf{h} | \mathbf{R}(\mathbf{Z}, \mathbf{W}, \mathbf{H}) \in \mathbf{R}(C(\mathbf{X}, \boldsymbol{\Theta}), \mathbf{W}, \mathbf{h}), Y\right) \\
&\qquad = \Pr\left(\mathbf{Z} \in C(\mathbf{X}, \boldsymbol{\Theta}), \mathbf{W}, \mathbf{h} | \mathbf{R}(\mathbf{Z}, \mathbf{W}, \mathbf{H}) \in \mathbf{R}(C(\mathbf{X}, \boldsymbol{\Theta}), \mathbf{W}, \mathbf{h})\right) \\
&\qquad = \Pr\left(\mathbf{X} = \mathbf{g}, \mathbf{W}, \mathbf{H} = \mathbf{h} | \mathbf{R}(\mathbf{X}, \mathbf{W}, \mathbf{H}) = \mathbf{R}(\mathbf{g}, \mathbf{W}, \mathbf{h})\right),
\end{aligned}$$

donde la segunda igualdad surge del hecho que $(\mathbf{Z}, \mathbf{W}, \mathbf{H})|(\mathbf{R}(\mathbf{Z}, \mathbf{W}, \mathbf{H}), Y)$ es independiente de Y. Por lo tanto,

$$(\mathbf{X}, \mathbf{W}, \mathbf{H})|(\mathbf{R}(\mathbf{X}, \mathbf{W}, \mathbf{H}), Y) =_d (\mathbf{X}, \mathbf{W}, \mathbf{H})|\mathbf{R}(\mathbf{X}, \mathbf{W}, \mathbf{H}),$$

y por las equivalencias del Lema 2.1 obtenemos que $\mathbf{R}(\mathbf{X}, \mathbf{W}, \mathbf{H})$ es suficiente para la regresión de Y sobre $(\mathbf{X}, \mathbf{W}, \mathbf{H})$.□

Identificada la reducción suficiente, en la siguiente sección se presenta el procedimiento de estimación de los parámetros que involucra la misma. Básicamente, la estimación descansa en la

factorización presentada para la regresión inversa conjunta, simplificando el problema mediante la maximización de funciones de verosimilitud parciales condicionadas.

4.4. Estimación

Para obtener la reducción suficiente, necesitamos estimar los parámetros correspondientes de la función de verosimilitud conjunta. Para ello, supongamos que tenemos una muestra aleatoria de n puntos $(y_i, \mathbf{x}_i, \mathbf{w}_i, \mathbf{h}_i)$ extraídos de la distribución conjunta $(Y, \mathbf{X}, \mathbf{W}, \mathbf{H})$ y que para las distribuciones condicionales rigen los modelos (4.5) y (4.10). Más específicamente, los estimadores se obtienen de maximizar el logaritmo de la función de verosimilitud para los datos observados $(y_i, \mathbf{x}_i, \mathbf{w}_i, \mathbf{h}_i)$, la que viene dada por

$$\sum_{i=1}^{n} \log f_{\mathbf{X},\mathbf{W},\mathbf{H}}(\mathbf{x}_i, \mathbf{w}_i, \mathbf{h}_i | y_i; \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\tau}_0, \boldsymbol{\kappa}, \boldsymbol{\iota}). \tag{4.18}$$

En base al enfoque de variables latentes, estamos suponiendo que existe un vector p-dimensional $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)$, de variables continuas subyacentes a las variables ordinales observadas. Luego podemos escribir

$$\begin{aligned} f_{\mathbf{X},\mathbf{W},\mathbf{H}}(\mathbf{x}_i, \mathbf{w}_i, \mathbf{h}_i | y_i; \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\tau}_0, \boldsymbol{\kappa}, \boldsymbol{\iota}) &= \int f_{\mathbf{X},\mathbf{Z},\mathbf{W},\mathbf{H}}(\mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i, \mathbf{h}_i | y_i; \boldsymbol{\Omega}, \boldsymbol{\Upsilon}) d\mathbf{z}_i \\ &= \left\{ \int f_{\mathbf{X},\mathbf{Z},\mathbf{W}}(\mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i | y_i, \mathbf{h}_i; \boldsymbol{\Omega}) f_{\mathbf{H}}(\mathbf{h}_i | y_i; \boldsymbol{\Upsilon}) d\mathbf{z}_i \right\} \\ &= \left\{ \int f_{\mathbf{X},\mathbf{Z},\mathbf{W}}(\mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i | y_i, \mathbf{h}_i; \boldsymbol{\Omega}) d\mathbf{z}_i \right\} f_{\mathbf{H}}(\mathbf{h}_i | y_i; \boldsymbol{\Upsilon}), \end{aligned} \tag{4.19}$$

donde $\boldsymbol{\Omega} \doteq \{\boldsymbol{\Delta}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\beta}\}$ y $\boldsymbol{\Upsilon} \doteq (\boldsymbol{\tau}_0, \boldsymbol{\kappa}, \boldsymbol{\iota})$. Por lo tanto, para maximizar la función (4.18), separamos el problema en dos partes: En primer lugar, estimamos $\boldsymbol{\Omega}$ vía maximización de

$$\sum_{i=1}^{n} \log \left\{ \int f_{\mathbf{X},\mathbf{Z},\mathbf{W}}(\mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i | y_i, \mathbf{h}_i; \boldsymbol{\Omega}) d\mathbf{z}_i \right\},$$

y en segundo lugar, estimamos $\boldsymbol{\Upsilon}$ a partir de maximizar el logaritmo de la función de verosimilitud condicionada

$$\sum_{i=1}^{n} \log f_{\mathbf{H}}(\mathbf{h}_i|y_i;\boldsymbol{\Upsilon}).$$

.

4.4.1. Estimación de $\boldsymbol{\Omega}$

Vamos a considerar el vector t-dimensional, $\mathbf{V}^T = (\mathbf{Z}^T, \mathbf{W}^T)$ con $t = p + s$ para el cual asumimos la distribución condicional

$$\mathbf{V}|\mathbf{H}, Y \sim N(\boldsymbol{\mu} + \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \boldsymbol{\beta}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}), \boldsymbol{\Delta}), \tag{4.20}$$

con $\bar{\mathbf{f}}_Y = E_Y(\mathbf{f}_Y)$, $\boldsymbol{\mu}_{\mathbf{H}} = E_{\mathbf{H}}(\mathbf{H})$, $\boldsymbol{\mu}^T = (\mathbf{0}, \boldsymbol{\mu}_{\mathbf{W}})^{\mathbf{T}}$ y matriz de covarianza $\boldsymbol{\Delta}$ de orden $t \times t$, la cual cumple que $[\boldsymbol{\Delta}_{\mathbf{zz}}]_{jj} = 1$ para todo $j = 1, \ldots, p$; $\boldsymbol{\alpha}$ es una matriz $t \times d$, $\boldsymbol{\xi}$ is de orden $d \times r$, $\mathbf{f}_Y$ es de orden $r \times 1$ y $\boldsymbol{\beta}$ de $t \times q$. Vamos a llamar a las variables centradas de la siguiente manera: $\bar{\mathbf{f}}_{y_i} = \mathbf{f}_{y_i} - \bar{\mathbf{f}}_y$ con $\bar{\mathbf{f}}_y = \frac{1}{n}\sum_{i=1}^{n} \mathbf{f}_{y_i}$ y $\bar{\mathbf{h}}_i = \mathbf{h}_i - \bar{\mathbf{h}}$ con $\bar{\mathbf{h}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}_i$. Observemos que si $C(\mathbf{X}, \boldsymbol{\Theta})$ es el hiper-cubo $C(\mathbf{X}, \boldsymbol{\Theta}) = [\theta^{(1)}_{X_1-1}, \theta^{(1)}_{X_1}) \times \cdots \times [\theta^{(p)}_{X_p-1}, \theta^{(p)}_{X_p})$, dado que para cada $g = 1, \ldots, G_j$, $X_j = g \Leftrightarrow Z_j \in [\theta^{(j)}_{X_j-1}, \theta^{(j)}_{X_j})$, luego tenemos que

$$f_{\mathbf{X},\mathbf{V}}(\mathbf{x}_i, \mathbf{v}_i|y_i, \mathbf{h}_i; \boldsymbol{\Omega}) = \frac{1}{((2\pi)^t|\boldsymbol{\Delta}|)^{1/2}} \times \exp\left[-\frac{1}{2}(\mathbf{v}_i - \boldsymbol{\mu} - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i} - \boldsymbol{\beta}\bar{\mathbf{h}}_i)^T\boldsymbol{\Delta}^{-1}(\mathbf{v}_i - \boldsymbol{\mu} - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i} - \boldsymbol{\beta}\bar{\mathbf{h}}_i)\right] I_{\mathbf{z}_i \in C(\mathbf{x}_i,\Theta)}.$$

Por lo tanto, para cada observación i la densidad marginal condicionada $(\mathbf{X}, \mathbf{W})|(Y, \mathbf{H})$ será

$$f_{\mathbf{X},\mathbf{W}}(\mathbf{x}_i, \mathbf{w}_i|y_i, \mathbf{h}_i; \boldsymbol{\Omega}) = \left\{\int f_{\mathbf{X},\mathbf{V}}(\mathbf{x}_i, \mathbf{v}_i|y_i, \mathbf{h}_i; \boldsymbol{\Omega}) d\mathbf{z}_i\right\}. \tag{4.21}$$

Como en el Capítulo 3, nos encontramos con el problema de cómputo de las integrales múltiples, por lo que los estimadores máximos verosímiles también en este caso serán obtenidos por medio del algoritmo EM. De todos modos, al tener conjuntamente la latente con la continua, como también las discretas como predictores, los estimadores presentan algunas variantes, como se verá a continuación.

Algoritmo

En esta sección presentaremos el algoritmo para estimar $\mathbf{\Omega}$ siguiendo el desarrollo expuesto en el capítulo anterior para el caso de SDR que toma solo predictores ordinales. Para identificar la k-ésima iteración del algoritmo, se utilizará el supraíndice (k). El procedimiento comienza con el Paso 0, donde los parámetros $\boldsymbol{\mu}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\xi}^{(0)}\boldsymbol{\beta}^{(0)}$ y $\mathbf{\Delta}^{(0)}$ son inicializados usando estimadores del PFC estándar de Cook and Forzani (2008). Luego el algoritmo itera entre los dos pasos siguientes, hasta lograr convergencia: En el Paso 1 estimamos $\mathbf{\Theta}^{(k)}$ dados $\boldsymbol{\mu}^{(k-1)}, \boldsymbol{\alpha}^{(k-1)}, \boldsymbol{\xi}^{(k-1)}, \boldsymbol{\beta}^{(k-1)}, \mathbf{\Delta}^{(k-1)}$ y el en Paso 2 obtenemos $\boldsymbol{\mu}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\xi}^{(k)}\boldsymbol{\beta}^{(k)}, \mathbf{\Delta}^{(k)}$ a través de maximizar la esperanza condicional (dado $\boldsymbol{\mu}^{(k-1)}$, $\boldsymbol{\alpha}^{(k-1)}$, $\boldsymbol{\xi}^{(k-1)}$,$\boldsymbol{\beta}^{(k-1)}$, $\mathbf{\Delta}^{(k-1)}$ y $\mathbf{\Theta}^{(k)}$) del logaritmo de la verosimilitud conjunta (4.21). Este paso es denominado paso EM.

Paso 1: Estimación de $\mathbf{\Theta}$: Dado $\boldsymbol{\mu}^{(k-1)}$, $\boldsymbol{\alpha}^{(k-1)}, \boldsymbol{\xi}^{(k-1)}, \mathbf{\Delta}^{(k-1)}, \boldsymbol{\beta}^{(k-1)}$, del Paso 0 o de un paso previo, para estimar los umbrales de cada variable ordinal X_j (i.e. $\theta_0^{(j)}, \ldots, \theta_{G_j}^{(j)}$, para $j = 1, \ldots, p$), tomamos $\hat{\theta}_0^{(j)} = -\infty =$, $\hat{\theta}_{G_j}^{(j)} = +\infty$ y, para $g = 1, \ldots, G_j - 1$, asignamos a $\hat{\theta}_g^{(j)}$ la solución de la ecuación $L_g(\theta) = 0$, donde

$$L_g(\theta) \doteq \#\{i : x_{ij} \leq g\} - \sum_{i=1}^{n} \Phi\left(\frac{\theta - \mathbf{A}_j^{(k-1)}\bar{\mathbf{f}}_{y_i} - \boldsymbol{\beta}_j^{(k-1)}\bar{\mathbf{h}}_i}{\delta_j^{(k-1)}}\right), \tag{4.22}$$

donde Φ representa la función de distribución de una normal estándar, $\delta_j^{(k-1)} = (\mathbf{\Delta}^{(k-1)})_{jj}$, $\mathbf{A}_j^{(k-1)}$ y $\boldsymbol{\beta}_j^{(k-1)}$ son la fila j de $\mathbf{A}^{(k-1)} \equiv \Delta^{(k-1)}\boldsymbol{\alpha}^{(k-1)}\boldsymbol{\xi}^{(k-1)}$ y $\boldsymbol{\beta}^{(k-1)}$, respectivamente. A su vez, x_{ij} es la coordenada j de $\mathbf{x}_i$ y $\#S$ es el cardinal del conjunto S.

Observemos que, para la identificabilidad, en (4.22) $\boldsymbol{\mu}_j^{(k)} = 0$ para todo k and j. Por lo tanto, en cada iteración k obtenemos $\mathbf{\Theta}^{(k)} = \{\theta_0^{(1)}, \ldots, \theta_{G_1}^{(1)}, \ldots, \theta_0^{(p)}, \ldots, \theta_{G_p}^{(p)}\}$. De esta manera, como en el caso de SDR que toma solo variables ordinales, aquí también la definición de L_g descansa en el supuesto de normalidad condicional de $\mathbf{V}$, y por lo tanto, sobre la variable continua latente.

Paso 2: Estimación de $\boldsymbol{\mu}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\xi}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\Delta}^{(k)}$: Dado $\boldsymbol{\Theta}^{(k)}$ estimado en Paso 1 y $\boldsymbol{\mu}^{(k)}$, $\boldsymbol{\alpha}^{(k-1)}$, $\boldsymbol{\xi}^{(k-1)}$, $\Delta^{(k-1)}$, $\boldsymbol{\beta}^{(k-1)}$ y $\boldsymbol{\Delta}^{(k-1)}$ del Paso 0 o de un paso previo, aplicamos el algoritmo EM para maximizar el log de la verosimilitud condicionada. Más precisamente, el algoritmo EM busca el $\boldsymbol{\Omega}^{(k)} \doteq (\boldsymbol{\Delta}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\xi}^{(k)}, \boldsymbol{\beta}^{(k)})$ tal que maximiza

$$Q(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(k-1)}) = \sum_{i=1}^{n} E_{\mathbf{z}_i|\mathbf{y}_i,\mathbf{h}_i,\mathbf{x}_i\boldsymbol{\Omega}^{(k-1)}} \left[\log f_{\mathbf{x}_i,\mathbf{z}_i,\mathbf{w}_i}(\mathbf{x}_i,\mathbf{z}_i,\mathbf{w}_i|y_i,\mathbf{h}_i,\boldsymbol{\Omega})\big|y_i,\mathbf{h}_i\boldsymbol{\Omega}^{(k-1)}\right], \tag{4.23}$$

en $\boldsymbol{\Omega} \doteq (\boldsymbol{\Delta}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\beta})$, de donde se obtienen los siguientes estimadores

$$\begin{aligned}
\boldsymbol{\mu}^{(k)} &= (\mathbf{0}, \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_i)^T, \\
\boldsymbol{\alpha}^{(k)} &= (\widetilde{\mathbf{S}}_n^{\mathbf{w}})^{-1/2}\hat{\zeta}\mathbf{N} \\
(\boldsymbol{\Delta}^{-1})^{(k)} &= (\widetilde{\mathbf{S}}_n^{\mathbf{w}})^{-1} + \boldsymbol{\alpha}^{(k)}((\boldsymbol{\alpha}^{(k)})^T\mathbf{S}_{\mathbf{res}}\boldsymbol{\alpha}^{(k)})^{-1}(\boldsymbol{\alpha}^{(k)})^T \\
&\qquad -\boldsymbol{\alpha}^{(k)}\left((\boldsymbol{\alpha}^{(k)})^T\widetilde{\mathbf{S}}_n^{\mathbf{w}}\boldsymbol{\alpha}^{(k)}\right)^{-1}(\boldsymbol{\alpha}^{(k)})^T \\
\boldsymbol{\xi}^{(k)} &= ((\boldsymbol{\alpha}^{(k)})^T\boldsymbol{\Delta}^{(k)}\boldsymbol{\alpha}^{(k)})^{-1}(\boldsymbol{\alpha}^{(k)})^T(\mathbf{S}_{\mathbf{F,M}} - \mathbf{S}_{\mathbf{F,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,M}})^T(\mathbf{S}_{\mathbf{F}} - \mathbf{S}_{\mathbf{F,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,F}})^{-1} \\
\boldsymbol{\beta}^{(k)} &= (\mathbf{S}_{\mathbf{M,H}} - \boldsymbol{\Delta}^{(k)}\boldsymbol{\alpha}^{(k)}\boldsymbol{\xi}^{(k)}\mathbf{S}_{\mathbf{F,H}})\mathbf{S}_{\mathbf{H}}^{-1},
\end{aligned}$$

donde $\mathbf{w}_i$ es un punto muestral de $\mathbf{W}$; $\widetilde{\mathbf{S}}_n^{\mathbf{w}} = \mathbf{S}_n^{\mathbf{w}} - \mathbf{S}_{\mathbf{M,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,M}}$ con

$$\begin{aligned}
\mathbf{S}_n^{\mathbf{w}} = \frac{1}{n}\sum_{i=1}^{n} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}(\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^T|y_i,\mathbf{h}_i,\mathbf{x}_i) &= \frac{1}{n}\sum_{i} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}\left[\begin{pmatrix} \mathbf{z}_i\mathbf{z}_i^T & \mathbf{z}_i\bar{\mathbf{w}}_i^T \\ \mathbf{z}_i\bar{\mathbf{w}}_i^T & \bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^T \end{pmatrix}\middle|(y_i,\mathbf{h}_i,\mathbf{x}_i)\right] \\
&= \begin{pmatrix} \mathbf{S}_n & \frac{1}{n}\mathbf{M}_n^T\mathbf{W}_n \\ \frac{1}{n}\mathbf{W}_n^T\mathbf{M}_n & \frac{1}{n}\mathbf{W}_n^T\mathbf{W}_n \end{pmatrix},
\end{aligned}$$

$$\begin{aligned}
\mathbf{S}_{\mathbf{H}} &= \frac{1}{n}\mathbf{H}_n^T\mathbf{H}_n \\
\mathbf{S}_{\mathbf{M,H}} &= \frac{1}{n}(\mathbf{M}_n^{\mathbf{w}})^T\mathbf{H}_n \quad \text{y} \quad \mathbf{S}_{\mathbf{H,M}} = \mathbf{S}_{\mathbf{M,H}}^T,
\end{aligned}$$

con $\bar{\mathbf{v}}_i \doteq (\mathbf{z}_i^T, \mathbf{w}_i^T - \bar{\mathbf{w}}^T)^T \doteq (\mathbf{z}_i^T, \bar{\mathbf{w}}_i^T)^T$ para $\bar{\mathbf{w}} = n^{-1}\sum_{i=1}^{n}\mathbf{w}_i$; $\mathbf{M}_n^{\mathbf{w}} = [\mathbf{M}_n, \mathbf{W}_n] \in \mathbb{R}^{n\times t}$ con $\mathbf{M}_n^T = [E_{\mathbf{z}_1|\mathbf{x}_1,\mathbf{h}_1,y_1}(\mathbf{z}_1|\mathbf{x}_1,\mathbf{h}_1,y_1), \ldots, E_{\mathbf{z}_n|\mathbf{x}_n,\mathbf{h}_n,y_n}(\mathbf{z}_n|\mathbf{x}_n,\mathbf{h}_n,y_n)]$ y $\mathbf{W}_n^T = [\bar{\mathbf{w}}_1, \ldots, \bar{\mathbf{w}}_n]$; y $\mathbf{H}_n$ es

una matriz de orden $n \times q$ donde $\mathbf{H}_n^T = [\bar{\mathbf{h}}_1, \ldots, \bar{\mathbf{h}}_n]$. La matriz $\hat{\zeta}$ esta compuesta por los primero d autovectores de $(\widetilde{\mathbf{S}}_n^{\mathbf{w}})^{-1/2}\mathbf{S}_{\mathbf{fit}}(\widetilde{\mathbf{S}}_n^{\mathbf{w}})^{-1/2}$ y $\mathbf{N}$ es una matriz tal que $\hat{\boldsymbol{\alpha}}^T\hat{\boldsymbol{\alpha}} = \mathbf{I}_d$, siendo $\mathbf{S}_{\mathbf{fit}} \doteq (\mathbf{S}_{\mathbf{F,M}} - \mathbf{S}_{\mathbf{F,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,M}})^T(\mathbf{S}_{\mathbf{F}} - \mathbf{S}_{\mathbf{F,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,F}})^{-1}(\mathbf{S}_{\mathbf{F,M}} - \mathbf{S}_{\mathbf{F,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,M}})$ y $\mathbf{S}_{\mathbf{res}} = \widetilde{\mathbf{S}}_n^{\mathbf{w}} - \mathbf{S}_{\mathbf{fit}}$. Aquí

$$\begin{aligned}\mathbf{S}_{\mathbf{F,M}} &= \frac{1}{n}\mathbf{F}_n^T\mathbf{M}_n^{\mathbf{w}} \quad \text{and} \quad \mathbf{S}_{\mathbf{M,F}} = \mathbf{S}_{\mathbf{M,F}}^T, \\ \mathbf{S}_{\mathbf{F,H}} &= \frac{1}{n}\mathbf{F}_n^T\mathbf{H}_n \quad \text{and} \quad \mathbf{S}_{\mathbf{H,F}} = \mathbf{S}_{\mathbf{F,H}}^T, \mathbf{S}_{\mathbf{H}} = \frac{1}{n}\mathbf{H}_n^T\mathbf{H}_n\end{aligned}$$

con $\mathbf{F}_n \in \mathbb{R}^{n\times r}$, tal que $\mathbf{F}_n^T = [\bar{\mathbf{f}}_{y_1}, \ldots, \bar{\mathbf{f}}_{y_n}]$. Los pasos detallados sobre cómo se obtuvieron estos estimadores están explicados en el Apéndice 4.8.1.

Paso 3: Chequeamos convergencia a través de la inspección de $Q(\boldsymbol{\Omega}^{(k)}|\boldsymbol{\Omega}^{(k-1)})$ en cada paso. Si la misma sigue creciendo entre un paso y otro, se retorna al **Paso 1** del algoritmo.

4.4.2. Estimación de $\boldsymbol{\Upsilon}$

Usando la parametrización (4.8), la distribución conjunta para el modelo Ising (4.6) puede escribirse de la forma

$$\begin{aligned}P(\mathbf{H}|y) = \exp\Bigg(&\sum_{j=1}^{q}\tau_{jj0}^*H_j + \sum_{j=1}^{q}\boldsymbol{\tau}_{jj}^T(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)H_j \\ &+ \sum_{1\le j<j'\le q}\tau_{jj'0}^*H_jH_{j'} + \sum_{1\le j<j'\le q}\boldsymbol{\tau}_{jj'}^T(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)H_jH_{j'}\Bigg)\frac{1}{G(\boldsymbol{\Gamma}_Y)}\end{aligned} \tag{4.24}$$

Tomando como base la especificación de Cheng et al. (2014) para modelos gráficos con predictores, si consideramos una variable dicotómica particular j y condicionamos sobre las restantes $(\mathbf{H}_{-j} \equiv (H_1, \ldots, H_{j-1}, H_{j+1}, \ldots, H_q))$, obtenemos

$$\log\frac{P(H_j = 1|\mathbf{H}_{-j}, Y)}{P(H_j = 0|\mathbf{H}_{-j}, Y)} = \tau_{jj0}^* + \boldsymbol{\tau}_{jj}^T(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \sum_{j\neq j'}\tau_{jj'0}^*H_{j'} + \sum_{j<j'}\boldsymbol{\tau}_{jj'}^T(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)H_{j'} \tag{4.25}$$

De esta manera se obtiene que el log-odds condicional de una determinada variable dicotómica H_j es lineal en los parámetros, de forma tal que los estimadores máximos verosímiles (condicionados) de dichos parámetros pueden obtenerse mediante una regresión logística tomando H_j como respuesta con los predictores $(\mathbf{f}_Y, \mathbf{H}_{-j}, \mathbf{f}_Y\mathbf{H}_{-j})$. De esta manera, ajustando q modelos logísticos univariados se obtienen estimadores para $\boldsymbol{\tau}_0$ y $\boldsymbol{\tau}$. En particular, para los puntos muestrales $(\mathbf{h}_i, y_i) \equiv (h_{i1}, \ldots, h_{iq}, y_i)$, se tiene que para cada dicotómica j, con $j = 1, \ldots, q$, la función log-verosimilitud, vendrá dada por

$$\ell_j(\boldsymbol{\tau}_0, \boldsymbol{\tau}; \mathbf{h}_i, y_i) = \frac{1}{n}\sum_{i=1}^{n} \log P(h_{ij}|\mathbf{h}_{i,-j}, y_i) = \frac{1}{n}\sum_{i=1}^{n} \left(h_{ij}\epsilon_{ij} - \log(1 + \exp(\epsilon_{ij}))\right) \tag{4.26}$$

con

$$\epsilon_{ij} = \log \frac{P(h_{ij} = 1|h_{i,-j}, y)}{P(\mathbf{h}_{ij} = 0|\mathbf{h}_{i,-j}, y_i)} = \tau^*_{jj0} + \boldsymbol{\tau}^T_{jj}\bar{\mathbf{f}}_{y_i} + \sum_{j\neq j'} \tau^*_{jj'0} h_{ij'} + \sum_{j\neq j'} \boldsymbol{\tau}^T_{jj'}\bar{\mathbf{f}}_{y_i} h_{ij'}.$$

De esta manera puede obtenerse un estimador para $\boldsymbol{\Upsilon}$ de forma tal que que para cada j se maximicen (4.26). Por simetría, $\gamma^y_{jj'} = \gamma^y_{j'j}$, por lo que para tal parámetro tenemos dos estimadores, i.e. el de la regresión logística de H_j sobre $(\mathbf{f}_Y, \mathbf{H}_{-j}, \mathbf{f}_Y\mathbf{H}_{-j})$ y el de $H_{j'}$ sobre $(\mathbf{f}_Y, \mathbf{H}_{-j'}, \mathbf{f}_Y\mathbf{H}_{-j'})$, pues $H_{j'} \in \mathbf{H}_{-j}$ y $H_j \in \mathbf{H}_{-j'}$, y como nada nos garantiza que de la estimación obtengamos $\tau^*_{jj'0} = \tau^*_{j'j0}$ y $\boldsymbol{\tau}_{jj'} = \boldsymbol{\tau}_{j'j}$, debemos seleccionar algún criterio para que se cumpla la simetría. En la próxima sección describimos las posibilidades de simetrización en el proceso de regularización.

4.4.3. Estimación con selección de variables

Como en el Capítulo 3, podemos realizar selección de variables a fin de incorporar solo las variables relevantes en al estimar la reducción. Mas aún en este caso, donde la modelización para los predictores binarios conlleva a la estimación de una cantidad considerable de parámetros al considerar las interacciones.

De (4.17) puede observarse que la regularización del método puede llevarse a cabo de forma separada para el modelo normal (correspondiente a las continuas y ordinales) por un lado, y para el modelo Ising (el de las binarias) por otro lado, facilitándose así su implementación. Para el modelo normal estimado vía EM, aplicamos selección de variable del modo análogo al realizado para la reducción con predictores ordinales en base a Chen et al. (2010). Esto es,

como antes, la maximización de (4.23) es equivalente a encontrar en cada iteración,

$$\boldsymbol{\alpha}^{(k)} = \arg\min_{\boldsymbol{\alpha}} \big\{ - \operatorname{tr}(\boldsymbol{\alpha}^T \mathbf{S}_{\text{fit}} \boldsymbol{\alpha}) \big\}, \qquad \text{sujeto a } \boldsymbol{\alpha}^T \widetilde{\mathbf{S}}_n^{\mathbf{w}} \boldsymbol{\alpha} = \mathbf{I}_d. \tag{4.27}$$

De esta manera, la selección de variables puede realizarse incorporando un termino de penalización en (4.27) utilizando la norma mixta ℓ_1/ℓ_2 de la manera

$$\boldsymbol{\alpha}^{(k)} = \arg\min_{\boldsymbol{\alpha}} \Big\{ - \operatorname{tr}(\boldsymbol{\alpha}^T (\widetilde{\mathbf{S}}_n^{\mathbf{w}})^{-1/2} \mathbf{S}_{\text{fit}} (\widetilde{\mathbf{S}}_n^{\mathbf{w}})^{-1/2} \boldsymbol{\alpha}) + \lambda \sum_{i=1}^{p} ||\boldsymbol{\alpha}_i||_2 \Big\}, \qquad \text{sujeto a } \boldsymbol{\alpha}^T \widetilde{\mathbf{S}}_n^{\mathbf{w}} \boldsymbol{\alpha} = \mathbf{I}_d.$$

Como antes, el parámetro de regularización en este caso se elige utilizando un criterio de información (AIC o BIC).

Para la selección de las variables binarias, utilizamos la metodología propuesta por Cheng et al. (2014) para el modelo Ising con covariables. Específicamente, utilizando regularización con la norma ℓ_1, propone dos métodos para maximizar (4.26). En ambos solo se penaliza $\boldsymbol{\tau}$, pero uno lo hace de forma separada para cada j, mientras que el otro lo hace de manera conjunta. La regularizacián para regresiones logísticas de forma separada la realiza vía

$$\min_{\boldsymbol{\tau}_j \in \mathbb{R}^{(r+1)q}} -\ell_j(\boldsymbol{\tau}_0, \boldsymbol{\tau}; \mathbf{h}_i, y_i) + \lambda ||\boldsymbol{\tau}_{j \setminus 0}||_1, \tag{4.28}$$

donde $\boldsymbol{\tau}_j \doteq (\tau_{j10}, \boldsymbol{\tau}_{j1}^T, \ldots, \tau_{jq0}, \boldsymbol{\tau}_{jq}^T)$ y $\boldsymbol{\tau}_{j \setminus 0} \doteq \boldsymbol{\tau}_j \setminus \{\tau_{jj0}\}$. Para imponer la simetría, la propuesta de Cheng et al. (2014), en base a Meinshausen and Bühlmann (2006), sugiere la comparación de las magnitudes estimadas (pre-regularización) para τ_{jkl} y τ_{kjl} (con $j, k = 1, \ldots, q$ y $l = 1, \ldots, r$), eligiendo el mínimo o el máximo de ellos, dependiendo si se busca un método más o menos conservativo (mayor o menor cantidad de ceros estimados).

Por otra parte, la regularización conjunta termina estimando $(\boldsymbol{\tau}_0, \boldsymbol{\tau})$ de forma conjunta de la siguiente manera

$$\min_{(\boldsymbol{\tau}_0, \boldsymbol{\tau})} \sum_{j=1}^{q} -\ell_j(\boldsymbol{\tau}_0, \boldsymbol{\tau}; \mathbf{h}_i, y_i) + \lambda ||\boldsymbol{\tau}||_1. \tag{4.29}$$

Una ventaja de esta regularización conjunta es que la simetrización se impone automáticamente cuando se resuelve (4.29), sin embargo el costo computacional de este método es mayor. En las aplicaciones del presente capítulo, al no tener un $\mathbf{H}$ de dimensiones muy elevadas, se utiliza

este último método, usando el código para Matlab provisto por los autores[1].

4.5. Simulaciones

En esta sección estudiamos el comportamiento del método de reducción dimensional propuesto para datos que contienen variables de naturaleza mixta, al que llamaremos PFCmix. En una primera etapa evaluamos el desempeño en la estimación de la reducción para distintos tamaños muestrales, cuando los datos cumplen con el modelo propuesto. En una segunda etapa, evaluamos el desempeño en predicción cuando los datos cumplen con el modelo propuesto y también cuando no lo hacen. En este caso comparamos los resultados obtenidos contra los logrados por PFCord y los alcanzados sin utilizar reducción dimensional.

4.5.1. Desempeño en estimación

En este apartado queremos evaluar el desempeño del método propuesto, a partir de estudiar la calidad de la estimación obtenida en función de la cantidad n de observaciones disponibles para la estimación de parámetros. Para ello generamos dos conjuntos de datos independientes de acuerdo al modelo propuesto:

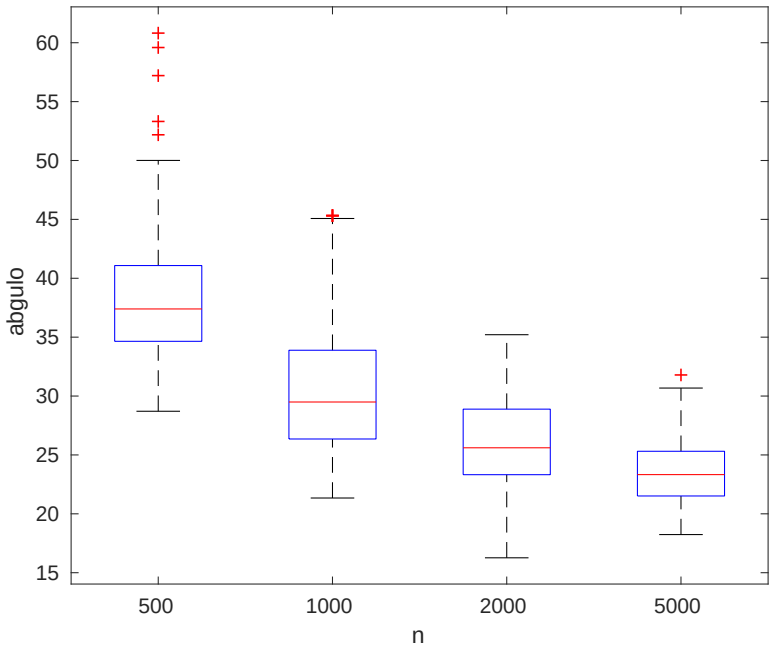
(i) Un conjunto de entrenamiento de n observaciones, con el que obtenemos la estimación $\hat{\mathbf{a}}$.

(ii) Un conjunto de m observaciones que se reducen de acuerdo a $\hat{\mathbf{a}}$ y sirven para evaluar la calidad de la estimación obtenida.

Los datos se generan siguiendo la factorización (4.2). En una primera etapa generamos $Y \sim$ Bernoulli(p_Y) con $p_Y = 0{,}4$. Luego generamos $\mathbf{H}|Y$ de acuerdo al modelo Ising. Para ello generamos los parámetros $\boldsymbol{\tau}_0$ y $\boldsymbol{\tau}$ de forma aleatoria de acuerdo a $\tau_{ij} = sJ$, donde s sigue una distribución normal estándar y J es una variable aleatoria Bernoulli de parámetro p_J. El valor de p_J sirve para controlar la proporción de interacciones en el modelo Ising y en esta simulación se usó el valor $p_J = 0{,}3$. Dados $\mathbf{H}$ e Y, a continuación generamos $\mathbf{Z}, \mathbf{W}|\mathbf{H}, Y \sim N_t(\boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{f}_Y + \boldsymbol{\beta}(\mathbf{H} - \boldsymbol{\mu}_\mathbf{H}), \boldsymbol{\Delta})$, donde $\mathbf{f}_Y$ en este caso es una variable indicadora de media cero. Los

[1] http://onlinelibrary.wiley.com/doi/10.1111/biom.12202/

Figura 4.1: Desempeño del Estimador: $\measuredangle(\text{span}(\hat{\mathbf{a}}), \text{span}(\mathbf{a}))$



parámetros del modelo normal fueron fijados de acuerdo al siguiente esquema: $\boldsymbol{\alpha} = \text{sign}(\mathbf{e})/\sqrt{t}$, con $\mathbf{e} \sim N_t(\mathbf{0}, \mathbf{I})$. La matriz de covarianza $\boldsymbol{\Delta}$ se fijó en $\boldsymbol{\Delta} = \mathbf{I} + \rho\boldsymbol{\alpha}\boldsymbol{\alpha}^T$, con $\rho = 0{,}25$, $\boldsymbol{\beta}$ se generó de acuerdo a una distribución normal estándar y se tomó $\boldsymbol{\xi} = 1$. Finalmente, las primeras $p < t$ componentes de $\mathbf{Z}$ fueron discretizadas de acuerdo a un conjunto de tres umbrales, $(-1, 0, 1)$, cada una, para así obtener $\mathbf{X}$ a partir de $\mathbf{Z}$. Para este ejemplo adoptamos $q = 6$, $p = 10$, $t = 14$.

Con los datos así generados usamos el algoritmo propuesto para estimar span($\hat{\mathbf{a}}$). En todos los casos usamos regularización. Medimos la calidad de la estimación calculando el ángulo entre span($\mathbf{a}$) y span($\hat{\mathbf{a}}$), El experimento se repitió para n tomando los valores 500, 1000, 2000 y 5000. Los resultados obtenidos se muestran la Figura 1, donde claramente se observa como mejora el estimador en precisión y eficiencia a medida que se incrementa el tamaño de la muestra. Al pasar de $n = 500$ a $n = 1000$ se observa principalmente un aumento en la precisión (menor ángulo) del estimador de la reducción. Luego con incrementos adicionales de n no solo se observa una reducción del ángulo promedio, sino también una reducción considerable de su varianza.

Tabla 4.1: Tasas de Error Promedio para el Escenario 1

	PFCmix	PFCord	NLPCA	Sin reducción	Oráculo
$n = 500$	0.223	0.207	0.281	0.195	0.171
$n = 1000$	0.186	0.198	0.258	0.183	0.171
$n = 2000$	0.183	0.192	0.254	0.179	0.171

Nota: Datos generados favorables para PFCmix.

4.5.2. Desempeño en predicción

En esta sección nos interesa evaluar el desempeño en predicción de las reducciones obtenidas con PFCmix, comparándolo con el obtenido por PFCord y con el uso de los datos crudos sin reducir. Usamos regresión logística como modelo predictivo. Para facilitar la comparación entre los resultados obtenidos con PFCmix y PFCord, las variables continuas $\mathbf{W}$ usadas en PFCmix fueron discretizadas en cuatro categorías ordenadas para ser incorporadas como variables ordinales en PFCord. Por su parte, las variables binarias $\mathbf{H}$ consideradas de acuerdo al modelo Ising en PFCmix fueron agregadas directamente como ordinales en PFCord.

Tuvimos en cuenta dos escenarios para la generación de los datos. Por un lado, generamos $\mathbf{X}, \mathbf{W}, \mathbf{H}, Y$ de forma idéntica a lo descrito en la sección anterior. Estas condiciones resultan óptimas para PFCmix.

En un segundo escenario, generamos $\mathbf{Z}|Y \sim N_{t+q}(\mathbf{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{f}_Y, \mathbf{\Delta})$ y luego discretizamos la primeras $p < t$ variables para generar $\mathbf{X}$, dejamos intactas las siguientes $t - p$ variables para generar $\mathbf{W}$ y tomamos las últimas q variables de $\mathbf{Z}$ y las discretizamos según $H_j = \text{sign}(Z_j)$ para generar $\mathbf{H}$. Este modelo de generación de datos es óptimo para PFCord. Los valores de los parámetros para el modelo normal de $\mathbf{Z}|Y$ se fijaron de forma análoga al escenario anterior.

Los resultados obtenidos bajo el primer escenario se resumen en la Tabla 4.1. Cuando $n = 500$ el método de reducción para variables ordinales (PFCord) es el que muestra el mejor desempeño en términos del error de predicción. Sin embargo, cuando incrementamos la cantidad de observaciones a $n = 1000$, la tasa de error del PFCmix se reduce considerablemente, mostrando en este caso mejor desempeño que el (PFCord). Para $n = 2000$, PFCmix sigue siendo el método de reducción que muestra mejor desempeño, acercándose al resultado de predecir con todas las variables (sin reducción) e incluso al error de predicción que surge de

Tabla 4.2: Tasa de Error Promedio para el Escenario 2

	PFCmix	PFCord	NLPCA	Sin reducción	Oráculo
$n = 500$	0.226	0.197	0.252	0.207	0.186
$n = 1000$	0.201	0.191	0.243	0.194	0.186
$n = 2000$	0.198	0.193	0.246	0.191	0.186

Nota: Datos generados favorables para PFCord.

utilizar el verdadero valor de los parámetros, que denominamos Oráculo. En todos los casos, componentes principales no lineales (NLPCA) muestra la mayor tasa de error.

Para el segundo escenario, los resultados son reportados en la Tabla 4.2. En este caso, para todos los n el PFCord es el que muestra el mejor desempeño, lo que resultaba esperable dado que los datos se generaron favoreciendo dicho método. Sin embargo, a medida que aumenta n, el PFCmix va convergiendo al PFCord, como también al resultado sin reducción y al Oráculo. Por lo tanto, aún bajo este escenario que favorece al PFCord, los resultados muestran que el método propuesto para variables mixtas es al menos tan bueno como este; y si además tenemos en cuenta la ventaja que tiene el PFCmix de poder realizar reducción con otros tipos de variables además de las ordinales, la superioridad de este método se hace mas evidente.

Con la siguiente aplicación para datos reales, las conclusiones obtenidas respecto de las ventajas del PFCmix se verán reforzadas.

4.6. Aplicación con Datos Reales: Índice SES

Consideraremos, como en el Capítulo 3, la aplicación específica con datos reales para la construcción de índices de Estatus Socio-económico (índice SES) para su uso en predicción de ingreso y nivel de pobreza.

A diferencia de la base de datos usada para evaluar el PFCord, la utilizada en el presente capítulo contiene tanto variables ordinales como continuas y binarias. Por ello, *ex-ante*, el método extendido para predictores mixtos es preferido a los métodos acotados a un tipo de variables. En primer lugar, porque obviamente permite incorporar más variables que antes pudieran haberse omitido por su naturaleza. Y en segundo lugar, no menos importante, porque brinda la posibilidad de incorporar las variables tal cual fueron captadas sin necesidad de realizar

alguna transformación o re-construcción a partir de las mismas. Por ejemplo, para captar el hacinamiento del hogar en el índice SES, cuando se utilizó el PFCord se construyó una variable ordinal a partir del ratio entre número de personas en el hogar y la cantidad de ambientes de la vivienda que habitan. Dicho ratio constituye una variable continua, sin embargo a los fines de aplicar el método para ordinales se definieron algunos umbrales para construir a partir de la misma una variable ordinal. Esta definición de umbrales puede ser un tanto arbitraria, pudiendo responder o bien a alguna recomendación de un experto en el área o simplemente a partir de adoptar algún criterio del análisis de su distribución en la muestra (e.g. balanceo). Sin embargo esta interferencia puede terminar influyendo sobre el peso que se asigna a tal variable para la predicción de una respuesta de interés. Por ende, resulta favorable eliminar esta componente forzada de definición a priori de umbrales, utilizando la variable continua sin necesidad de transformación.

Otro ejemplo sobre esta conveniencia de tomar las variables directamente como vienen en la base, sin necesidad de definir nuevas variables que respondan a la naturaleza requerida por el método, lo constituye la construcción de una ordinal con más de dos categorías a partir de varias binarias. Para la aplicación del PFCord esto se hizo al incorporar aspectos de la localización de la vivienda; específicamente, se construyó una ordinal con cuatro categorías, a partir de 3 binarias. Esto respondió a motivos metodológicos, dado que el supuesto de existencia de una continua latente sobre una binaria es más fuerte y es esperable obtener una mejor estimación de la latente cuanto más categorías se observen. El método propuesto en este capítulo permite incorporar binarias asumiendo simplemente un proceso Bernoulli para las mismas sin necesidad de asumir una continua latente subyacente.

Para ser consistentes con la aplicación realizada en el capítulo anterior, los datos utilizados aquí también provienen de la base de microdatos de la Encuesta Permanente de Hogares (EPH) elaborada por el Instituto Nacional de Estadísticas y Censos (INDEC) correspondiente al cuarto trimestre de 2013, construyendo los índices para cada una de las cinco regiones consideradas (GBA, Pampeana, NOA, NEA y Patagonia).

Como predictores se tiene un total de 8 variables categóricas ordinales, 4 variables continuas y 7 variables binarias[2]. Las variables ordinales son las mismas que la presentadas en el capítulo

[2]Cabe volver a remarcar que hay variables binarias que son ordinales en el sentido del estatus socio-económico. Por ejemplo la tenencia de la vivienda (propietario o no) constituye una proxy de riqueza material, por lo tanto

anterior, solo que aquí hay un total de 8 variables en lugar de 11. Ello se debe a que 3 de dichas variables (*Hacinamiento*, *Horas trabajadas* y *Localización de la Vivienda*) fueron tomadas como se presentan directamente en la base; esto es, *Hacinamiento* y *Horas trabajadas* como continuas, y *Localización de la Vivienda* como tres binarias que indican si la vivienda está cercana a un basural, en una villa de emergencia o en una zona inundable. Además de estas últimas tres variables binarias de localización, se consideró la ocupación y el sexo del jefe/a de hogar, si tienen cobertura médica y si son propietarios de la vivienda. Para las variables continuas, además del ratio de hacinamiento y de las horas trabajadas en la ultima semana por el jefe/a de hogar, se tomó la edad de este último y la cantidad de menores en el hogar. En el Apéndice 2.9.3 se presenta una descripción detallada de las variables utilizadas.

También en esta aplicación consideramos dos tipos de respuesta: una continua dada por el ingreso per cápita del hogar (para mayor sencillez en la implementación del método tomamos $r = 1$) y una discreta que indica si el hogar es pobre o no de acuerdo con la linea de pobreza.

Para evaluar el desempeño predictivo del indice SES construido en base al PFCmix, lo comparamos con la predicción sin aplicar reducción utilizando todas las variables como son dadas (i.e. Full), con la que brinda la reducción no supervisada vía Componentes Principales No Lineales (NLPCA) y con el método propuesto para variables ordinales (PFCord). Específicamente para este último método, consideramos dos casos: utilizando solo las variables ordinales disponibles (i.e. PFCord($\mathbf{X}$)), y tomando también las variables binarias tratadas como ordinales (i.e. PFCord($\mathbf{X}, \mathbf{H}$)). Esto nos permite evaluar también el desempeño de esta metodología para el caso de contar con binarias que puedan ser consideradas como ordinales dentro del enfoque de variables latentes que presupone el PFCord. Del mismo modo, para el PFCmix evaluamos dos opciones PFCmix($\mathbf{X}, \mathbf{H}$) y PFCmix($\mathbf{X}, \mathbf{W}, \mathbf{H}$). Esto se realiza a los fines de poder comparar mejor con el PFCord en 'igualdad de condiciones', como así también para visualizar el efecto de tener variables continuas entre los predictores. Para todos los métodos supervisados de reducción suficiente, la estimación se realizan con selección de variables (i.e. métodos regularizados).

La predicción de los distintos índices como la realizada por el conjunto total de variables (Full) se realiza utilizando regresión logística para la respuesta discreta (i.e. pobreza) y re-

de no ser propietario a serlo, hay un orden. Sin embargo, solo consideramos como ordinales aquellas con más de una categoría por motivos ya expuestos.

gresión lineal para la respuesta continua (ingreso per cápita). Los datos se particionan en diez conjuntos disjuntos para replicar el experimento, tomando un conjunto cada vez como datos de prueba y el resto como muestra de entrenamiento. Los errores de predicción promedio se reportan en la Tabla 4.3. En primer lugar, podemos destacar que la predicción de pobreza y

Tabla 4.3: Validación Cruzada de 10 particiones (10-fold) para el índice SES.

		Error de Predicción - ECM				
Respuesta	**Método**	*GBA*	*Pampeana*	*NOA*	*NEA*	*Patagonia*
Pobreza (binaria)	PFCord($\mathbf{X}$)	0.2159	0.1933	0.3299	0.3892	0.1359
	PFCord($\mathbf{X}, \mathbf{H}$)	0.2008	0.1781	0.2751	0.3043	0.1277
	PFCmix($\mathbf{X}, \mathbf{W}, \mathbf{H}$)	0.1643	0.1312	0.2469	0.2419	0.0805
	PFCmix($\mathbf{X}, \mathbf{H}$)	0.1711	0.1435	0.2703	0.2997	0.0822
	NLPCA($\mathbf{X}, \mathbf{W}, \mathbf{H}$)	0.2335	0.2052	0.3727	0.3887	0.1318
	Full	0.2194	0.1919	0.3178	0.3736	0.1360
Ingreso (pc) (continua)	PFCord($\mathbf{X}$)	7.61	5.13	5.04	3.61	14.41
	PFCord($\mathbf{X}, \mathbf{H}$)	7.42	4.93	4.91	3.41	14.01
	PFCmix($\mathbf{X}, \mathbf{W}, \mathbf{H}$)	7.39	4.68	4.62	3.27	13.82
	PFCmix($\mathbf{X}, \mathbf{H}$)	7.45	4.71	4.93	3.46	13.88
	NLPCA($\mathbf{X}, \mathbf{W}, \mathbf{H}$)	8.99	6.03	5.72	4.02	16.28
	Full	7.58	5.12	5.04	3.61	14.39

del ingreso per cápita realizada vía el índice PFCord mejora cuando se incorporan variables binarias (i.e. PFCord($\mathbf{X}$) *vs.* PFCord($\mathbf{X}, \mathbf{H}$)). Por ende, el método de reducción suficiente para ordinales resulta funcional en términos predictivos para dicotómicas, tanto para el caso en el que son ordinales como no. Sin embargo, cuando utilizamos el método para mixtas para estos dos conjuntos de variables (i.e. PFCmix($\mathbf{X}, \mathbf{H}$) *vs.* PFCord($\mathbf{X}, \mathbf{H}$)), se obtiene un error de predicción considerablemente menor en todas las regiones cuando buscamos predecir pobre-

za. No obstante, para la predicción de ingreso per cápita del hogar, para algunas regiones (en particular, Pampeana, NOA y NEA), el error de predicción con el PFCord es menor, aunque los valores del ECM son muy similares a los arrojados por el PFCmix.

Evaluando los resultados del método propuesto para variables mixtas, puntualmente entre incorporar o no predictores continuos (PFCmix$(\mathbf{X}, \mathbf{W}, \mathbf{H})$ *vs.* PFCmix$(\mathbf{X}, \mathbf{H})$), se observa que siempre las variables continuas en la reducción permiten mejorar la predicción. Más aún, podemos concluir que con el índice SES elaborado con PFCmix$(\mathbf{X}, \mathbf{W}, \mathbf{H})$ se obtienen los mejores resultados para ambas variables respuesta y todas las regiones.

Por otro lado, se observa que el índice que no toma en cuenta la información de la variable respuesta (NLPCA) tiene los mayores errores de predicción, aun cuando incorpora todos los predictores. Incluso, si comparamos con el PFCord usando solo ordinales , este muestra mejor desempeño predictivo en tres de las cinco regiones cuando la respuesta es discreta y en todos los casos cuando la respuesta es continua.

Por último, al comparar con la predicción obtenida sin reducir (Full), en general los métodos de reducción supervisados dan mejores resultados, no así el no supervisado.

4.7. Conclusiones

En el presente capítulo hemos identificado una reducción suficiente para un modelo de regresión de una variable respuesta Y sobre un conjunto de predictores que pueden ser continuos, categóricos ordinales y binarios. Esta generalización tuvo como motivación central el hecho de que la mayor parte de los modelos aplicados con datos reales, y en particular en economía y ciencias sociales, se incluyen variables de naturaleza mixta. La aplicación de SDR a este tipo de modelos, demanda una nueva metodología que permita reducir de forma conjunta el vector de predictores, y no acotar las posibilidades de aplicación sólo en variables continuas o en aquellas no continuas pero que necesariamente su distribución pertenezca a la familia exponencial.

La metodología propuesta, que bautizamos como PFCmix al tener sus raices en el PFC de Cook and Forzani (2008), se nutre básicamente del enfoque de variables latentes para el subconjunto de variables ordinales y de una generalización del enfoque de SDR para familias exponenciales propuesto por Bura et al. (2015) que fue presentado en el Capítulo 2. Proponiendo

una factorización para la regresión inversa, identificamos una reducción suficiente presentando además estimadores de máxima verosimilitud para los parámetros involucrados. Adicionalmente, se propone selección de variables para este método, en base a lo desarrollado en el Capitulo 3 y a la propuesta de Cheng et al. (2014) en el marco de modelos Ising.

Tanto las simulaciones realizadas como la aplicación a índices SES, muestran resultados favorables para el método propuesto al compararlo con los PFC alternativos (i.e. el estándar diseñado para predictores continuos, y el PFCord diseñado para predictores ordinales) y con alternativas mejoradas del PCA, como ser su versión no lineal (NLPCA). Destacamos que además de obtener ventaja en términos predictivos, el método propuesto a priori resulta favorecido debido a que permite combinar variables de una naturaleza variada tal como están disponibles en la base de datos, sin necesidad de omitir algunas de ellas o aplicarle alguna transformación para que el método las admita.

Por último, debemos recalcar que hay un punto relevante no tratado en el presente capítulo, que es la elección de la dimension de la reducción, i.e $\tilde{d} = \text{máx}(d, c)$. Si bien para la aplicación de índices SES, que constituye el principal interés de aplicación de la tesis, fijamos de antemano $\tilde{d} = 1$ al concebir un concepto unitario en torno al estatus socioeconómico, en otras aplicaciones interesaría inferir de algún modo el valor de la dimensión del subespacio de reducción. Debido al modelo utilizado, la búsqueda de la dimensión implicaría inferir conjuntamente los valores de d y c. Tomando la idea de Cook et al. (2015), una alternativa sería utilizar algún criterio de información tipo AIC o BIC para la búsqueda simultánea de (d, c). Es decir, podemos realizar una búsqueda de (d, c) a partir de $(0, 0)$ a (r, r), eligiendo la combinación que da los menores valores del AIC o del BIC. También, como en la Capítulo 3, podría buscarse (c, d) vía validación cruzada, o bien diseñando pruebas de permutaciones, principalmente cuando el fin es predictivo. Por ende, queda como trabajo futuro implementar y comparar el desempeño de algunos de estos métodos para la búsqueda de $\tilde{d}$, para poder extender el método a otras aplicaciones donde un $\tilde{d} > 1$ sea relevante a los fines explicativos o predictivos.

4.8. Apéndice del Capítulo

4.8.1. EM para Datos Mixtos

Como en el caso de la explicación del EM para datos ordinales, para simplificar la notación, siempre omitimos $\mathbf{\Omega}^{(k-1)}$ cuando tomamos la esperanza condicionada. La forma explicita de la función Q utilizando viene dada por

$$\begin{aligned} Q(\mathbf{\Omega}|\mathbf{\Omega}^{(k-1)}) &= E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i;\mathbf{\Omega}^{(k-1)}} \log \left[\prod_{i=1}^{n} f_{\mathbf{x}_i,\mathbf{v}_i}(\mathbf{x}_i, \mathbf{v}_i|y_i, \mathbf{h}_i; \mathbf{\Omega}) \right] \\ &= \sum_{i=1}^{n} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i} \left[\log f_{\mathbf{x}_i,\mathbf{v}_i}(\mathbf{x}_i, \mathbf{v}_i|y_i, \mathbf{h}_i; \mathbf{\Omega}) \right]. \end{aligned} \tag{4.30}$$

Dados los supuestos sobre el modelo para las variables continuas (latentes y observadas), (4.30) toma la forma

$$\begin{aligned} \sum_{i=1}^{n} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i} \left[\log f_{\mathbf{x}_i,\mathbf{v}_i}(\mathbf{x}_i, \mathbf{v}_i|y_i, \mathbf{h}_i) \right] &= -\frac{tn}{2} \log 2\pi - \frac{n}{2} \log |\mathbf{\Delta}| \\ &- \frac{1}{2} \sum_{i=1}^{n} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i} \left[\operatorname{tr}[\mathbf{\Delta}^{-1}(\mathbf{v}_i - \boldsymbol{\mu} - \mathbf{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i} - \boldsymbol{\beta}\bar{\mathbf{h}}_i)(\mathbf{v}_i - \boldsymbol{\mu} - \mathbf{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i} - \boldsymbol{\beta}\bar{\mathbf{h}}_i)^T] | y_i, \mathbf{h}_i, \mathbf{x}_i \right]. \end{aligned}$$

Desarrollando convenientemente la igualdad anterior y usando que $\boldsymbol{\mu}^T = (\mathbf{0}, \boldsymbol{\mu}_{\mathbf{W}})^{\mathbf{T}}$ obtenemos,

$$\begin{aligned} Q(\mathbf{\Omega}|\mathbf{\Omega}^{(k-1)}) &\equiv \sum_{i=1}^{n} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i} \left[\log f_{\mathbf{x}_i,\mathbf{v}_i}(\mathbf{x}_i, \mathbf{v}_i|y_i, \mathbf{h}_i) \right] \\ &= -\frac{tn}{2} \log 2\pi - \frac{n}{2} \log |\mathbf{\Delta}| \\ &\quad - \frac{n}{2} \operatorname{tr} \Bigg\{ \mathbf{\Delta}^{-1} \Bigg[\frac{1}{n} \sum_{i=1}^{n} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}(\mathbf{v}_i \mathbf{v}_i^T | y_i, \mathbf{h}_i, \mathbf{x}_i) - \frac{2}{n} \sum_{i=1}^{n} \boldsymbol{\mu} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}(\mathbf{v}_i^T | y_i, \mathbf{h}_i, \mathbf{x}_i) \\ &\quad - \frac{2}{n} \sum_{i=1}^{n} \mathbf{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}(\mathbf{v}_i^T | y_i, \mathbf{h}_i, \mathbf{x}_i) - \frac{2}{n} \sum_{i=1}^{n} \boldsymbol{\beta}\bar{\mathbf{h}}_i E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}(\mathbf{v}_i^T | y_i, \mathbf{h}_i, \mathbf{x}_i) \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{2}{n} \sum_{i=1}^{n} \mathbf{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i} \boldsymbol{\mu}^T + \frac{2}{n} \sum_{i=1}^{n} \boldsymbol{\beta}\bar{\mathbf{h}}_i \boldsymbol{\mu}^T + \frac{2}{n} \sum_{i=1}^{n} \mathbf{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i} \bar{\mathbf{h}}_i^T \boldsymbol{\beta}^T \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} \mathbf{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i} \bar{\mathbf{f}}_{y_i}^T \boldsymbol{\xi}^T \boldsymbol{\alpha}^T \mathbf{\Delta} + \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\beta}\bar{\mathbf{h}}_i \bar{\mathbf{h}}_i^T \boldsymbol{\beta}^T \Bigg] \Bigg\}. \end{aligned} \tag{4.31}$$

Maximización de la función Q

Tomando derivada respecto a $\boldsymbol{\mu}$ y considerando la restricción de los parámetros (por identificabilidad), obtenemos que

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \mathbf{0} \\ \bar{\mathbf{w}} \end{pmatrix}, \quad \text{con} \quad \bar{\mathbf{w}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{w}_i.$$

Luego, definimos

$$\bar{\mathbf{v}}_i \doteq \mathbf{v}_i - \hat{\boldsymbol{\mu}} = \begin{pmatrix} \mathbf{z}_i \\ \mathbf{w}_i - \bar{\mathbf{w}} \end{pmatrix} \doteq \begin{pmatrix} \mathbf{z}_i \\ \bar{\mathbf{w}}_i \end{pmatrix}.$$

Reemplazando en (4.31) tenemos

$$\begin{aligned}
&\sum_{i=1}^{n} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}\left[\log f_{\mathbf{x}_i,\bar{\mathbf{v}}_i}(\mathbf{x}_i,\bar{\mathbf{v}}_i|y_i,\mathbf{h}_i)\right] \\
&\quad = -\frac{tn}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Delta}| \\
&\qquad - \frac{n}{2}\operatorname{tr}\Bigg\{\boldsymbol{\Delta}^{-1}\Bigg[\frac{1}{n}\sum_{i=1}^{n} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}(\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^T|y_i,\mathbf{h}_i,\mathbf{x}_i) - \frac{2}{n}\sum_{i=1}^{n}\boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}(\bar{\mathbf{v}}_i^T|y_i,\mathbf{h}_i,\mathbf{x}_i) \\
&\qquad - \frac{2}{n}\sum_{i=1}^{n}\boldsymbol{\beta}\bar{\mathbf{h}}_i E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}(\bar{\mathbf{v}}_i^T|y_i,\mathbf{h}_i,\mathbf{x}_i) + \frac{2}{n}\sum_{i=1}^{n}\boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i}\bar{\mathbf{h}}_i^T\boldsymbol{\beta}^T \qquad (4.32) \\
&\qquad + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i}\bar{\mathbf{f}}_{y_i}^T\boldsymbol{\xi}^T\boldsymbol{\alpha}^T\boldsymbol{\Delta} + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\beta}\bar{\mathbf{h}}_i\bar{\mathbf{h}}_i^T\boldsymbol{\beta}^T\Bigg]\Bigg\}.
\end{aligned}$$

donde

$$\bar{\mathbf{v}}_i = \begin{pmatrix} \mathbf{z}_i \\ \bar{\mathbf{w}}_i \end{pmatrix} \text{ y por lo tanto } \bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^T = \begin{pmatrix} \mathbf{z}_i\mathbf{z}_i^T & \mathbf{z}_i\bar{\mathbf{w}}_i^T \\ \mathbf{z}_i\bar{\mathbf{w}}_i^T & \bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^T \end{pmatrix}.$$

En (4.32) puede apreciarse que hay tres términos en los que aparece $\mathbf{z}_i$ para tomar esperanza. Primero,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}(\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^T|y_i,\mathbf{h}_i,\mathbf{x}_i) &= \frac{1}{n}\sum_{i} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}\left[\begin{pmatrix} \mathbf{z}_i\mathbf{z}_i^T & \mathbf{z}_i\bar{\mathbf{w}}_i^T \\ \mathbf{z}_i\bar{\mathbf{w}}_i^T & \bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^T \end{pmatrix}\middle|(y_i,\mathbf{h}_i,\mathbf{x}_i)\right] \\
&= \begin{pmatrix} \mathbf{S}_n & \frac{1}{n}\mathbf{M}_n^T\mathbf{W}_n \\ \frac{1}{n}\mathbf{W}_n^T\mathbf{M}_n & \frac{1}{n}\mathbf{W}_n^T\mathbf{W}_n \end{pmatrix} \equiv \mathbf{S}_n^{\mathbf{w}}, \qquad (4.33)
\end{aligned}$$

donde $\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{n} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}(\mathbf{z}_i\mathbf{z}_i^T|y_i,\mathbf{h}_i,\mathbf{x}_i) \in \mathbb{R}^{p\times p}$, la matriz $\mathbf{M}_n \in \mathbb{R}^{n\times p}$ está definida por $\mathbf{M}_n^T = [E_{\mathbf{z}_1|\mathbf{x}_1,\mathbf{h}_1,y_1}(\mathbf{z}_1|\mathbf{x}_1,\mathbf{h}_1,y_1),\ldots,E_{\mathbf{z}_n|\mathbf{x}_n,\mathbf{h}_n,y_n}(\mathbf{z}_n|\mathbf{x}_n,\mathbf{h}_n,y_n)]$ y $\mathbf{W}_n \in \mathbb{R}^{n\times s}$ con $\mathbf{W}_n^T = [\bar{\mathbf{w}}_1,\ldots,\bar{\mathbf{w}}_n]$. Luego $\mathbf{S}_n^{\mathbf{w}}$ es una matriz $t\times t$ (pues $\mathbf{S}_n$ es $p\times p$, $\mathbf{M}_n$ $n\times p$ y $\mathbf{W}_n$ $n\times s$, por lo que $\mathbf{M}_n^T\mathbf{W}_n$ es $p\times s$, $\mathbf{W}_n^T\mathbf{M}_n$ $s\times p$ y $\mathbf{W}_n^T\mathbf{W}_n$ $s\times s$).

Por su parte, el segundo término al incorporar la esperanza queda

$$\begin{aligned}-\frac{2}{n}\boldsymbol{\Delta\alpha\xi}\sum_i E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}\left(\bar{\mathbf{f}}_{y_i}\left[\mathbf{z}_i^T,\bar{\mathbf{w}}_i^T\right]|y_i,\mathbf{h}_i,\mathbf{x}_i\right) &= -\frac{2}{n}\boldsymbol{\Delta\alpha\xi}\sum_i E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}\left(\left[\bar{\mathbf{f}}_{y_i}\mathbf{z}_i^T,\bar{\mathbf{f}}_{y_i}\bar{\mathbf{w}}_i^T\right]|y_i,\mathbf{h}_i,\mathbf{x}_i\right)\\ &= -\frac{2}{n}\left[\boldsymbol{\Delta\alpha\xi}\mathbf{F}_n\mathbf{M}_n,\boldsymbol{\Delta\alpha\xi}\mathbf{F}_n\mathbf{W}_n\right]\\ &= -\frac{2}{n}\boldsymbol{\Delta\alpha\xi}\mathbf{F}_n\mathbf{M}_n^{\mathbf{w}},\end{aligned} \tag{4.34}$$

con $\mathbf{F}_n = [\bar{f}_{y_1},\ldots,\bar{f}_{y_n}] \in \mathbb{R}^{r\times n}$ y $\mathbf{M}_n^{\mathbf{w}} = [\mathbf{M}_n,\mathbf{W}_n] \in \mathbb{R}^{n\times t}$. Finalmente, para el tercer término tenemos que

$$\begin{aligned}-\frac{2}{n}\boldsymbol{\beta}\sum_i E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}\left(\bar{\mathbf{h}}_i\left[\mathbf{z}_i^T,\bar{\mathbf{w}}_i^T\right]|y_i,\mathbf{h}_i,\mathbf{x}_i\right) &= -\frac{2}{n}\boldsymbol{\beta}\left[\mathbf{H}_n^T\mathbf{M}_n,\mathbf{H}_n^T\mathbf{W}_n\right]\\ &= -\frac{2}{n}\boldsymbol{\beta}\mathbf{H}_n^T\mathbf{M}_n^{\mathbf{w}},\end{aligned} \tag{4.35}$$

donde $\mathbf{H}_n$ es de orden $n\times q$ con $[\mathbf{H}_n]_{ij} \in \{0,1\}$ y $\mathbf{H}_n^T = [\bar{\mathbf{h}}_1,\ldots,\bar{\mathbf{h}}_n]$. Usando (4.33), (4.34) y (4.35), (4.32) puede escribirse como

$$\begin{aligned}&Q(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(k-1)}) \equiv \sum_{i=1}^{n} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}\left[\log f_{\mathbf{x}_i,\mathbf{v}_i}(\mathbf{x}_i,\mathbf{v}_i|y_i,\mathbf{h}_i)\right]\\ &= -\frac{tn}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Delta}| - \frac{n}{2}\operatorname{tr}\left\{\boldsymbol{\Delta}^{-1}\left[\mathbf{S}_n^{\mathbf{w}} - \frac{2}{n}\boldsymbol{\Delta\alpha\xi}\mathbf{F}_n\mathbf{M}_n^{\mathbf{w}} - \frac{2}{n}\boldsymbol{\beta}\mathbf{H}_n^T\mathbf{M}_n^{\mathbf{w}}\right.\right.\\ &\quad\left.\left. + \frac{2}{n}\boldsymbol{\Delta\alpha\xi}\mathbf{F}_n\mathbf{H}_n\boldsymbol{\beta}^T + \frac{1}{n}\boldsymbol{\Delta\alpha\xi}\mathbf{F}_n\mathbf{F}_n^T\boldsymbol{\xi}^T\boldsymbol{\alpha}^T\boldsymbol{\Delta} + \frac{1}{n}\boldsymbol{\beta}\mathbf{H}_n^T\mathbf{H}_n\boldsymbol{\beta}^T\right]\right\}.\end{aligned} \tag{4.36}$$

Derivando (4.36) con respecto a $\boldsymbol{\beta}$ obtenemos

$$\widehat{\boldsymbol{\beta}} = (\mathbf{S}_{\mathbf{M},\mathbf{H}} - \boldsymbol{\Delta\alpha\xi}\mathbf{S}_{\mathbf{F},\mathbf{H}})\mathbf{S}_{\mathbf{H}}^{-1}, \tag{4.37}$$

donde,

$$\begin{aligned}
\mathbf{S}_{\mathbf{H}} &= \frac{1}{n}\mathbf{H}_n^T\mathbf{H}_n \\
\mathbf{S}_{\mathbf{M,H}} &= \frac{1}{n}(\mathbf{M}_n^{\mathbf{w}})^T\mathbf{H}_n \quad \text{and} \quad \mathbf{S}_{\mathbf{H,M}} = \mathbf{S}_{\mathbf{M,H}}^T \\
\mathbf{S}_{\mathbf{F,H}} &= \frac{1}{n}\mathbf{F}_n\mathbf{H}_n \quad \text{and} \quad \mathbf{S}_{\mathbf{H,F}} = \mathbf{S}_{\mathbf{F,H}}^T.
\end{aligned}$$

Además definimos

$$\begin{aligned}
\mathbf{S}_{\mathbf{F}} &= \frac{1}{n}\mathbf{F}_n\mathbf{F}_n^T \\
\mathbf{S}_{\mathbf{F,M}} &= \frac{1}{n}\mathbf{F}_n\mathbf{M}_n^{\mathbf{w}} \quad \text{and} \quad \mathbf{S}_{\mathbf{M,F}} = \mathbf{S}_{\mathbf{M,F}}^T
\end{aligned}$$

Por lo tanto,

$$\begin{aligned}
&\sum_{i=1}^{n} E_{\mathbf{z}_i|y_i,\mathbf{h}_i,\mathbf{x}_i}\left[\log f_{\mathbf{x}_i,\bar{\bar{\mathbf{v}}}_i}(\mathbf{x}_i,\bar{\mathbf{v}}_i|y_i,\mathbf{h}_i)\right] \\
&= -\frac{tn}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Delta}| \\
&\quad - \frac{n}{2}\operatorname{tr}\Bigg\{\boldsymbol{\Delta}^{-1}\Bigg[\mathbf{S}_n^{\mathbf{w}} - \frac{2}{n}\boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{F}_n\mathbf{M}_n^{\mathbf{w}} \\
&\quad - \frac{2}{n}(\mathbf{S}_{\mathbf{M,H}} - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{S}_{\mathbf{F,H}})\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{H}_n^T\mathbf{M}_n^{\mathbf{w}} + \frac{2}{n}\boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{F}_n\mathbf{H}_n\mathbf{S}_{\mathbf{H}}^{-1}(\mathbf{S}_{\mathbf{M,H}} - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{S}_{\mathbf{F,H}})^T \\
&\quad + \frac{1}{n}\boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{F}_n\mathbf{F}_n^T\boldsymbol{\xi}^T\boldsymbol{\alpha}^T\boldsymbol{\Delta} + \frac{1}{n}(\mathbf{S}_{\mathbf{M,H}} - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{S}_{\mathbf{F,H}})\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{H}_n^T\mathbf{H}_n\mathbf{S}_{\mathbf{H}}^{-1}(\mathbf{S}_{\mathbf{M,H}} - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{S}_{\mathbf{F,H}})^T\Bigg]\Bigg\} \\
&= -\frac{tn}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Delta}| \\
&\quad - \frac{n}{2}\operatorname{tr}\Bigg\{\boldsymbol{\Delta}^{-1}\Bigg[\mathbf{S}_n^{\mathbf{w}} - 2\boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{S}_{\mathbf{F,M}} \\
&\quad - 2(\mathbf{S}_{\mathbf{M,H}} - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{S}_{\mathbf{F,H}})\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,M}} + 2\boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{S}_{\mathbf{F,H}}\mathbf{S}_{\mathbf{H}}^{-1}(\mathbf{S}_{\mathbf{M,H}} - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{S}_{\mathbf{F,H}})^T \\
&\quad + \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{S}_{\mathbf{F}}\boldsymbol{\xi}^T\boldsymbol{\alpha}^T\boldsymbol{\Delta} + (\mathbf{S}_{\mathbf{M,H}} - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{S}_{\mathbf{F,H}})\mathbf{S}_{\mathbf{H}}^{-1}(\mathbf{S}_{\mathbf{M,H}} - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\mathbf{S}_{\mathbf{F,H}})^T\Bigg]\Bigg\} \\
&= -\frac{tn}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Delta}| \\
&\quad - \frac{n}{2}\operatorname{tr}\Bigg\{\boldsymbol{\Delta}^{-1}\Bigg[\mathbf{S}_n^{\mathbf{w}} - \mathbf{S}_{\mathbf{M,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,M}} - 2\boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}[\mathbf{S}_{\mathbf{F,M}} - \mathbf{S}_{\mathbf{F,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,M}}] \\
&\quad + \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}[\mathbf{S}_{\mathbf{F}} - \mathbf{S}_{\mathbf{F,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,F}}]\boldsymbol{\xi}^T\boldsymbol{\alpha}^T\boldsymbol{\Delta}\Bigg]\Bigg\}
\end{aligned}$$

Luego tenemos la función Q

$$\begin{aligned} Q(\boldsymbol{\Delta}, \boldsymbol{\alpha}, \boldsymbol{\xi}) = &-\frac{tn}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Delta}| \\ &- \frac{n}{2}\operatorname{tr}\left(\boldsymbol{\Delta}^{-1}\widetilde{\mathbf{S}}_n^{\mathbf{w}}\right) + n\operatorname{tr}\left[\boldsymbol{\alpha}\boldsymbol{\xi}(\mathbf{S}_{\mathbf{F,M}} - \mathbf{S}_{\mathbf{F,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,M}})\right] \\ &- \frac{n}{2}\operatorname{tr}\left[\boldsymbol{\alpha}\boldsymbol{\xi}(\mathbf{S}_{\mathbf{F}} - \mathbf{S}_{\mathbf{F,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,F}})\boldsymbol{\xi}^T\boldsymbol{\alpha}^T\boldsymbol{\Delta}\right], \end{aligned} \tag{4.38}$$

donde $\widetilde{\mathbf{S}}_n^{\mathbf{w}} \doteq \mathbf{S}_n^{\mathbf{w}} - \mathbf{S}_{\mathbf{M,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,M}}$. Esta función es una forma cuadrática en $\boldsymbol{\xi}$ análoga a la función Q para el caso en que tenemos todos los predictores ordinales. Luego, (4.38) es maximizada en

$$\hat{\boldsymbol{\xi}} = (\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T(\mathbf{S}_{\mathbf{F,M}} - \mathbf{S}_{\mathbf{F,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,M}})^T(\mathbf{S}_{\mathbf{F}} - \mathbf{S}_{\mathbf{F,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,F}})^{-1}. \tag{4.39}$$

Remplazando (4.39) en (4.38), obtenemos

$$Q(\boldsymbol{\Delta}, \boldsymbol{\alpha}) = -\frac{tn}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Delta}| - \frac{n}{2}\operatorname{tr}\left(\boldsymbol{\Delta}^{-1}\widetilde{\mathbf{S}}_n^{\mathbf{w}} - \boldsymbol{\alpha}(\boldsymbol{\alpha}^T\boldsymbol{\Delta}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T\mathbf{S}_{\mathbf{fit}}\right), \tag{4.40}$$

donde $\mathbf{S}_{\mathbf{fit}} \doteq (\mathbf{S}_{\mathbf{F,M}} - \mathbf{S}_{\mathbf{F,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,M}})^T(\mathbf{S}_{\mathbf{F}} - \mathbf{S}_{\mathbf{F,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,F}})^{-1}(\mathbf{S}_{\mathbf{F,M}} - \mathbf{S}_{\mathbf{F,H}}\mathbf{S}_{\mathbf{H}}^{-1}\mathbf{S}_{\mathbf{H,M}})$ y $\mathbf{S}_{\mathbf{res}} = \widetilde{\mathbf{S}}_n^{\mathbf{w}} - \mathbf{S}_{\mathbf{fit}}$, y tomando los resultados de la Sección 4.8.1, el máximo de (4.40) se obtiene en

$$\widehat{\boldsymbol{\Delta}}^{-1} = (\widetilde{\mathbf{S}}_n^{\mathbf{w}})^{-1} + \boldsymbol{\alpha}(\boldsymbol{\alpha}^T\mathbf{S}_{\mathbf{res}}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T - \boldsymbol{\alpha}(\boldsymbol{\alpha}^T\widetilde{\mathbf{S}}_n^{\mathbf{w}}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T, \tag{4.41}$$

y

$$\hat{\boldsymbol{\alpha}} = (\widetilde{\mathbf{S}}_n^{\mathbf{w}})^{-1/2}\hat{\zeta}\mathbf{N} \tag{4.42}$$

donde $\hat{\zeta}$ es una matriz conformada con los primeros d autovectores de $(\widetilde{\mathbf{S}}_n^{\mathbf{w}})^{-1/2}\mathbf{S}_{\mathbf{fit}}(\widetilde{\mathbf{S}}_n^{\mathbf{w}})^{-1/2}$ y $\mathbf{N}$ es una matriz tal que $\hat{\boldsymbol{\alpha}}^T\hat{\boldsymbol{\alpha}} = \mathbf{I}_d$.

4.8.2. Identificabilidad del Modelo: Prueba del Teorema 4.1

Para simplificar notación, en el modelo (4.5) vamos a tomar $\mathbf{A}$ de orden $t \times r$ definida por $\mathbf{A} = \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}$ con $\mathbf{A}^T = (\mathbf{A}_{\mathbf{z}}^T, \mathbf{A}_{\mathbf{w}}^T)$. Luego podemos re-escribir (4.5) en términos de los parámetros

$(\boldsymbol{\Delta}, \boldsymbol{\mu}, \mathbf{A}, \boldsymbol{\beta}, \boldsymbol{\Theta})$ de la siguiente manera

$$\begin{aligned} \mathbf{V} &= \boldsymbol{\mu} + \mathbf{A}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \boldsymbol{\beta}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) + \boldsymbol{\epsilon}, \\ \boldsymbol{\epsilon} &\sim N(0, \boldsymbol{\Delta}), \\ \Pr(X_j = g|\mathbf{H}, Y) &= \Pr(\theta_{g-1}^{(j)} \leq Z_j < \theta_g^{(j)}|\mathbf{H}, Y). \end{aligned} \tag{4.43}$$

Para la identificabilidad necesitamos probar que si el modelo (4.43) es verdadero para los parámetros $(\boldsymbol{\Delta}, \boldsymbol{\mu}, \mathbf{A}, \boldsymbol{\beta}, \boldsymbol{\Theta})$ y $(\boldsymbol{\Lambda}, \boldsymbol{\nu}, \mathbf{B}, \boldsymbol{\zeta}, \boldsymbol{\Omega})$, entonces $\boldsymbol{\Delta} = \boldsymbol{\Lambda}$, $\boldsymbol{\mu} = \boldsymbol{\nu}$, $\mathbf{A} = \mathbf{B}$, $\boldsymbol{\beta} = \boldsymbol{\zeta}$ y $\boldsymbol{\Theta} = \boldsymbol{\Omega}$. Vamos a suponer que el modelo (4.43) es válido para ambos conjuntos de parámetros. Observemos que en tal caso podemos escribir,

$$\begin{bmatrix} \mathbf{Z} \\ \mathbf{W} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{z}} \\ \boldsymbol{\mu}_{\mathbf{w}} \end{bmatrix} + \begin{bmatrix} \mathbf{A}_{\mathbf{z}} \\ \mathbf{A}_{\mathbf{w}} \end{bmatrix} (\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \begin{bmatrix} \boldsymbol{\beta}_{\mathbf{z}} \\ \boldsymbol{\beta}_{\mathbf{w}} \end{bmatrix} (\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) + \begin{bmatrix} \boldsymbol{\epsilon}_{\mathbf{z}} \\ \boldsymbol{\epsilon}_{\mathbf{w}} \end{bmatrix},$$

por lo tanto, si para $j = 1, \ldots, p$ llamamos $[\boldsymbol{\mu}_{\mathbf{z}}]_j \doteq \mu_{\mathbf{z}_j}$ y $[\boldsymbol{\epsilon}_{\mathbf{z}}]_j \doteq \epsilon_j$, el vector $1 \times r$ $[\mathbf{A}_{\mathbf{z}}]_j \doteq \mathbf{A}_{\mathbf{z}_j}$, y el vector $1 \times q$ $[\boldsymbol{\beta}_{\mathbf{z}}]_j \doteq \boldsymbol{\beta}_{\mathbf{z}_j}$, tenemos que

$$\begin{aligned} Z_j &= \mu_{\mathbf{z}_j} + \mathbf{A}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \boldsymbol{\beta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) + \epsilon_j, \quad j = 1, \ldots, p \\ \Pr(X_j = g|\mathbf{H}, Y) &= \Pr(\theta_{g-1}^{(j)} \leq Z_j < \theta_g^{(j)}|\mathbf{H}, Y). \end{aligned} \tag{4.44}$$

Análogamente, para el segundo conjunto de parámetros,

$$\begin{aligned} Z_j &= \nu_{\mathbf{z}_j} + \mathbf{B}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \boldsymbol{\zeta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) + \epsilon_j, \quad j = 1, \ldots, p \\ \Pr(X_j = g|\mathbf{H}, Y) &= \Pr(\omega_{g-1}^{(j)} \leq Z_j < \omega_g^{(j)}|\mathbf{H}, Y). \end{aligned} \tag{4.45}$$

Luego, de (4.44) y (4.45), tenemos que

$$\begin{aligned} \Pr(X_j = g|\mathbf{H}, Y) = \Pr(&\theta_{g-1}^{(j)} - \mu_{\mathbf{z}_j} - \mathbf{A}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) \leq \epsilon_j \\ &< \theta_g^{(j)} - \mu_{\mathbf{z}_j} - \mathbf{A}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})), \end{aligned} \tag{4.46}$$

y

$$\Pr(X_j = g|\mathbf{H}, Y) = \Pr(\omega_{g-1}^{(j)} - \nu_{\mathbf{z}_j} - \mathbf{B}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\zeta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) \leq \epsilon_j < \omega_g^{(j)} - \nu_{\mathbf{z}_j} - \mathbf{B}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\zeta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})), \tag{4.47}$$

para $g = 1, \ldots, G_j$ y $j = 1, \ldots, p$. Como consecuencia de (4.46) y (4.47), con $g = 1$ (teniendo en cuenta que $\omega_0^{(j)} = \theta_0^{(j)} = -\infty$), y utilizando el supuesto de normalidad para ϵ_j, obtenemos

$$\Phi\left(\frac{\theta_1^{(j)} - \mu_{\mathbf{z}_j} - \mathbf{A}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})}{\delta_{jj}}\right) = \Phi\left(\frac{\omega_1^{(j)} - \nu_{\mathbf{z}_j} - \mathbf{B}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\zeta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})}{\lambda_{jj}}\right), \tag{4.48}$$

donde $\delta_{jj} \equiv [\boldsymbol{\Delta}]_{jj}$ y $\lambda_{jj} = [\boldsymbol{\Lambda}]_{jj}$, siendo $\Phi(\cdot)$ la función de distribución acumulada de una normal estándar. De (4.48) tenemos que

$$\frac{\theta_1^{(j)} - \mu_{\mathbf{z}_j} - \mathbf{A}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})}{\delta_{jj}} = \frac{\omega_1^{(j)} - \nu_{\mathbf{z}_j} - \mathbf{B}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\zeta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})}{\lambda_{jj}} \tag{4.49}$$

o equivalentemente,

$$\frac{1}{\delta_{jj}}\left(\theta_1^{(j)} - \mu_{\mathbf{z}_j} - [\mathbf{A}_{\mathbf{z}_j}, \boldsymbol{\beta}_{\mathbf{z}_j}]\begin{bmatrix} \mathbf{f}_Y - \bar{\mathbf{f}}_Y \\ \mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}} \end{bmatrix}\right) = \frac{1}{\lambda_{jj}}\left(\omega_1^{(j)} - \nu_{\mathbf{z}_j} - [\mathbf{B}_{\mathbf{z}_j}, \boldsymbol{\zeta}_{\mathbf{z}_j}]\begin{bmatrix} \mathbf{f}_Y - \bar{\mathbf{f}}_Y \\ \mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}} \end{bmatrix}\right) \tag{4.50}$$

para $j = 1, \ldots, p$. De la condición *1.* $\delta_{jj} = \lambda_{jj} = 1$ para todo $j = 1, \ldots, p$ y por *2.* $\mu_{\mathbf{z}_j} = \nu_{\mathbf{z}_j} = 0$. Luego si tomamos esperanza en (4.49), obtenemos que $\theta_1^{(j)} = \omega_1^{(j)}$ para todo $j = 1, \ldots, p$. Con esto, para $g > 1$, es fácil ver que $\theta_g^{(j)} = \omega_g^{(j)}$. En particular, si tomamos $g = 2$, de (4.46) y (4.47) obtenemos

$$\Phi\left(\frac{(\theta_2^{(j)} - \mu_{\mathbf{z}_j} - \mathbf{A}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}))}{\delta_{jj}}\right) = \Phi\left(\frac{(\omega_2^{(j)} - \nu_{\mathbf{z}_j} - \mathbf{B}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\zeta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}))}{\lambda_{jj}}\right), \tag{4.51}$$

puesto que por(4.48),

$$\Phi\left((\theta_1^{(j)} - \mu_{\mathbf{z}_j} - \mathbf{A}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}))/\delta_{jj}\right) = \Phi\left((\omega_1^{(j)} - \nu_{\mathbf{z}_j} - \mathbf{B}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\zeta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}))/\lambda_{jj}\right).$$

Por lo tanto, usando las condiciones *1.* a *3.* como antes obtenemos $\theta_2^{(j)} = \omega_2^{(j)}$. Luego, tomando $g = 3$ y usando (4.51) vamos a tener que $\theta_3^{(j)} = \omega_3^{(j)}$ y continuando de esta manera, obtendremos $\theta_g^{(j)} = \omega_g^{(j)}$ para todo k y por lo tanto $\boldsymbol{\Theta} = \boldsymbol{\Omega}$. Con todo esto en (4.50) tenemos,

$$[\mathbf{A}_{\mathbf{z}_j}, \boldsymbol{\beta}_{\mathbf{z}_j}] \begin{bmatrix} \mathbf{f}_Y - \bar{\mathbf{f}}_Y \\ \mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}} \end{bmatrix} = [\mathbf{B}_{\mathbf{z}_j}, \boldsymbol{\zeta}_{\mathbf{z}_j}] \begin{bmatrix} \mathbf{f}_Y - \bar{\mathbf{f}}_Y \\ \mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}} \end{bmatrix}, \tag{4.52}$$

Multiplicando por $[(\mathbf{f}_Y - \bar{\mathbf{f}}_Y)^T, (\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})^T]$, tomando esperanza y usando la condición *2.*, llegamos a que $\mathbf{A}_{\mathbf{z}_j} = \mathbf{B}_{\mathbf{z}_j}$ y $\boldsymbol{\beta}_{\mathbf{z}_j} = \boldsymbol{\zeta}_{\mathbf{z}_j}$ para $j = 1, \ldots, p$, y por lo tanto $\mathbf{A}_{\mathbf{z}} = \mathbf{B}_{\mathbf{z}}$ y $\boldsymbol{\beta}_{\mathbf{z}} = \boldsymbol{\zeta}_{\mathbf{z}}$.

Adicionalmente, para cualquier par de variables ordinales X_i y X_j $(i, j = 1, \ldots, p,\ i \neq j)$ tenemos que,

$$\begin{aligned} \Pr(X_i = g, X_j = l | \mathbf{H}, Y) &= \Pr(\theta_{g-1}^{(i)} < Z_i \leq \theta_g^{(i)}, \theta_{l-1}^{(j)} < Z_j \leq \theta_l^{(j)}) \\ &= \Phi_2\left(\boldsymbol{\Delta}_{ij}^{-1/2} \begin{pmatrix} \theta_g^{(i)} - \mu_{\mathbf{z}_i} - \mathbf{A}_{\mathbf{z}_i}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{z}_i}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) \\ \theta_l^{(j)} - \mu_{\mathbf{z}_j} - \mathbf{A}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) \end{pmatrix}\right) \\ &\quad - \Phi_2\left(\boldsymbol{\Delta}_{ij}^{-1/2} \begin{pmatrix} \theta_{g-1}^{(i)} - \mu_{\mathbf{z}_i} - \mathbf{A}_{\mathbf{z}_i}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{z}_i}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) \\ \theta_{l-1}^{(j)} - \mu_{\mathbf{z}_j} - \mathbf{A}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) \end{pmatrix}\right). \end{aligned} \tag{4.53}$$

Teniendo en cuenta que (4.53) se cumple para el otro conjunto de parámetros $(\boldsymbol{\Lambda}, \boldsymbol{\nu}, \mathbf{B}, \boldsymbol{\zeta}, \boldsymbol{\Omega})$, luego si tomamos $g = l = 1$ (y por lo tanto $\theta_{g-1}^{(i)} = \theta_{l-1}^{(j)} \equiv \theta_0 = -\infty$ y $\omega_{g-1}^{(i)} = \omega_{l-1}^{(j)} \equiv \omega_0 = -\infty$) llegamos a que

$$\begin{aligned} &\Phi_2\left(\boldsymbol{\Delta}_{ij}^{-1/2} \begin{pmatrix} \theta_1^{(i)} - \mu_{\mathbf{z}_i} - \mathbf{A}_{\mathbf{z}_i}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{z}_i}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) \\ \theta_1^{(j)} - \mu_{\mathbf{z}_j} - \mathbf{A}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) \end{pmatrix}\right) \\ &\qquad = \Phi_2\left(\boldsymbol{\Lambda}_{ij}^{-1/2} \begin{pmatrix} \omega_1^{(i)} - \nu_{\mathbf{z}_i} - \mathbf{B}_{\mathbf{z}_i}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\zeta}_{\mathbf{z}_i}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) \\ \omega_1^{(j)} - \nu_{\mathbf{z}_j} - \mathbf{B}_{\mathbf{z}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\zeta}_{\mathbf{z}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) \end{pmatrix}\right), \end{aligned} \tag{4.54}$$

donde $\mathbf{\Delta}_{ij} = \begin{pmatrix} \delta_{ii} & \delta_{ij} \\ \delta_{ji} & \delta_{jj} \end{pmatrix}$ y $\mathbf{\Lambda}_{ij} = \begin{pmatrix} \lambda_{ii} & \lambda_{ij} \\ \lambda_{ji} & \lambda_{jj} \end{pmatrix}$ son las matrices de covarianza para el par i, j de sus respectivas variables latentes (i.e. (Z_i, Z_j)) y $\Phi_2(.)$ representa la función de distribución de una normal bivariada.

Puesto que $\delta_{jj} = \lambda_{jj} = 1$, $\mu_{\mathbf{z}_j} = \nu_{\mathbf{z}_j} = 0$, $\mathbf{A_z} = \mathbf{B_z}$ y $\boldsymbol{\beta}_\mathbf{z} = \boldsymbol{\zeta}_\mathbf{z}$, de (4.54) obtenemos que $\delta_{ij} = \lambda_{ij}$ (y $\delta_{ji} = \lambda_{ji}$) para todo $i \neq j$ y $i, j = 1, \ldots, p$. Por lo tanto, se tiene que $\mathbf{\Delta_{zz}} = \mathbf{\Lambda_{zz}}$.

Por otra parte, para las variables continuas observadas $\mathbf{W}$, usando la normalidad de sus distribuciones marginales, tendremos que

$$\mathbf{W}|(\mathbf{H}, Y) \sim N(\boldsymbol{\mu}_\mathbf{w} + \mathbf{A_w}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \boldsymbol{\beta}_\mathbf{w}(\mathbf{H} - \boldsymbol{\mu}_\mathbf{H}), \mathbf{\Delta_{ww}}) \tag{4.55}$$

y

$$\mathbf{W}|(\mathbf{H}, Y) \sim N(\boldsymbol{\nu}_\mathbf{w} + \mathbf{B_w}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \boldsymbol{\zeta}_\mathbf{w}(\mathbf{H} - \boldsymbol{\mu}_\mathbf{H}), \mathbf{\Lambda_{ww}}) \tag{4.56}$$

y, usando la condición *3.* vamos a obtener, $\boldsymbol{\mu}_\mathbf{w} = \boldsymbol{\nu}_\mathbf{w}$, $\mathbf{A_w} = \mathbf{B_w}$, $\boldsymbol{\beta}_\mathbf{w} = \boldsymbol{\zeta}_\mathbf{w}$ and $\mathbf{\Delta_{ww}} = \mathbf{\Lambda_{ww}}$. Por lo tanto, resta probar que $\mathbf{\Delta_{zw}} = \mathbf{\Lambda_{zw}}$. Tomando una variable ordinal X_i $(i = 1, \ldots, p)$ y una variable continua observada W_j $(j = p+1, \ldots, t)$, tenemos que

$$\begin{aligned} \Pr(X_i \leq k, W_j \leq l|\mathbf{H}, Y) = \Pr\Big(&\epsilon_i \leq \theta_g^{(i)} - \mathbf{A}_{\mathbf{z}_i}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{z}_i}(\mathbf{H} - \boldsymbol{\mu}_\mathbf{H}), \\ &\epsilon_j \leq W_j - \mu_{\mathbf{w}_j} - \mathbf{A}_{\mathbf{w}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{w}_j}(\mathbf{H} - \boldsymbol{\mu}_\mathbf{H})\Big) \\ = \Phi_2\left(\mathbf{\Delta}_{ij}^{-1/2}\begin{pmatrix} \theta_g^{(i)} - \mathbf{A}_{\mathbf{z}_i}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{z}_i}(\mathbf{H} - \boldsymbol{\mu}_\mathbf{H}) \\ W_j - \mu_{\mathbf{w}_j} - \mathbf{A}_{\mathbf{w}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\beta}_{\mathbf{w}_j}(\mathbf{H} - \boldsymbol{\mu}_\mathbf{H}) \end{pmatrix}\right), \end{aligned} \tag{4.57}$$

donde en este caso, bajo las restricciones de identificación, $\mathbf{\Delta}_{ij} = \begin{pmatrix} 1 & \delta_{ij} \\ \delta_{ji} & \delta_{jj} \end{pmatrix}$.

De la misma manera,

$$\Pr\left(X_i \leq k, W_j \leq l|\mathbf{H}, Y\right) = \Pr\Bigg(\epsilon_i \leq \omega_g^{(i)} - \mathbf{B}_{\mathbf{z}_i}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\zeta}_{\mathbf{z}_i}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}),$$
$$\epsilon_j \leq W_j - \nu_{\mathbf{w}_j} - \mathbf{B}_{\mathbf{w}_j}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\zeta}_{\mathbf{w}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}})\Bigg)$$
$$= \Phi_2\left(\boldsymbol{\Lambda}_{ij}^{-1/2}\begin{pmatrix} \omega_g^{(i)} - \mathbf{B}_{\mathbf{z}_i}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) - \boldsymbol{\zeta}_{\mathbf{z}_i}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) \\ W_j - \nu_{\mathbf{w}_j} - \mathbf{B}_{\mathbf{w}_j} - \boldsymbol{\zeta}_{\mathbf{w}_j}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}) \end{pmatrix}\right), \quad (4.58)$$

donde $\boldsymbol{\Lambda}_{ij}^{-1/2} \doteq \begin{pmatrix} 1 & \lambda_{ij} \\ \lambda_{ji} & \lambda_{jj} \end{pmatrix}^{-1/2}$. Como $\theta_g^{(i)} = \omega_g^{(i)}$, $\boldsymbol{\mu}_{\mathbf{w}} = \boldsymbol{\nu}_{\mathbf{w}}$, $\boldsymbol{\Delta}_{\mathbf{ww}} = \boldsymbol{\Lambda}_{\mathbf{ww}}$, $\mathbf{A} = \mathbf{B}$ y $\boldsymbol{\beta} = \boldsymbol{\zeta}$, luego de (4.57) y (4.58) tenemos que $\boldsymbol{\Delta}_{ij} = \boldsymbol{\Lambda}_{ij}$, y por lo tanto$\delta_{ij} = \lambda_{ij}$. De esta manera, para todo $i = 1, \ldots, p$ y $j = 1, \ldots, s$, $\delta_{ij} = \lambda_{ij}$ (y $\delta_{ji} = \lambda_{ji}$), por ende $\boldsymbol{\Delta}_{\mathbf{zw}} = \boldsymbol{\Lambda}_{\mathbf{zw}}$ (y $\boldsymbol{\Delta}_{\mathbf{wz}} = \boldsymbol{\Lambda}_{\mathbf{wz}}$). Con esto hemos probado que $\boldsymbol{\Delta} = \boldsymbol{\Lambda}$, quedando así demostrado el teorema. □

Capítulo 5

Trabajos Posteriores

5.1. Introducción

En este breve capítulo se presentan algunos puntos sobre los que se planea seguir trabajando como continuidad de la tesis. En particular, presentamos dos temas para trabajos posteriores que están en la misma línea de la tesis. El primero constituye una ampliación de lo presentado en el Capítulo 4 sobre reducción suficiente en un contexto de predictores de naturaleza mixta. En particular, se expone una parametrización alternativa del modelo de regresión inversa $\mathbf{X}, \mathbf{W}, \mathbf{H}|Y$, para luego mostrar que de hecho el modelo presentado en el Capítulo 4 es un caso particular del modelo aquí presentado. Es decir, mostraremos que la factorización desarrollada en el Capítulo 4 está anidada a esta otra factorización que proponemos en el presente capítulo. A su vez, la estimación de los parámetros de los modelos que surgen de dicha factorización, pueden obtenerse vía máxima verosimilitud con las mismas metodologías anteriormente presentadas. Sin embargo, resta encontrar una reducción suficiente para esta factorización.

El segundo tema presentado, tiene que ver más con una extensión del enfoque general presentado en la tesis, motivado a partir de la aplicación de índices SES a modelos de regresión en economía y ciencias sociales. Específicamente, por muchos modelos presentes en la literatura relacionada a desarrollo económico y economía de la salud, entre otras, hay una necesidad de separar el índice obtenido vía reducción suficiente de otras covariables que son relevantes para la regresión de Y sobre el conjunto de predictores, pero que no se desea reducir. Por ende, una extensión hacia un método de reducción suficiente de dimensiones parcial para variables

mixtas, tendría un alcance mayor para la aplicabilidad académica y práctica (i.e. orientado a políticas) del enfoque de SDR.

5.2. Factorización Alternativa

En el Capítulo 4 para el modelo de regresión $Y|\mathbf{X}, \mathbf{W}, \mathbf{H}$, en base al enfoque de regresión inversa, se factorizó la distribución conjunta condicionada de $\mathbf{X}, \mathbf{W}, \mathbf{H}|Y$ a fin de obtener la correspondiente reducción suficiente del modelo de regresión original; a saber

$$f(\mathbf{X}, \mathbf{W}, \mathbf{H}|Y) = f(\mathbf{X}, \mathbf{W}|\mathbf{H}, Y)f(\mathbf{H}|Y). \tag{5.1}$$

Con esta factorización modelamos a $(\mathbf{X}, \mathbf{W})$ como función de Y y $\mathbf{H}$, dejando $\mathbf{H}$ como función solamente de Y. En términos de la variable latente $\mathbf{Z}$ subyacente de $\mathbf{X}$, por (**??**), teníamos

$$f(\mathbf{V}, \mathbf{H}|Y) = f(\mathbf{V}, |\mathbf{H}, Y)f(\mathbf{H}|Y) \tag{5.2}$$

donde $\mathbf{V} \equiv (\mathbf{Z}, \mathbf{W})$. A esta factorización, desarrollada en el capítulo previo, la denominaremos Modelo 1.

Por otra parte, la distribución conjunta condicionada de $\mathbf{X}, \mathbf{W}, \mathbf{H}|Y$ puede factorizarse como

$$f(\mathbf{X}, \mathbf{W}, \mathbf{H}|Y) = f(\mathbf{X}, \mathbf{W}|Y)f(\mathbf{H}|\mathbf{X}, \mathbf{W}, Y), \tag{5.3}$$

o, en términos de la variable latente,

$$f(\mathbf{V}, \mathbf{H}|Y) = f(\mathbf{V}|Y)f(\mathbf{H}|\mathbf{V}, Y). \tag{5.4}$$

Esta factorización alternativa implica modelar por un lado las variables $(\mathbf{X}, \mathbf{W})$ como función de Y (similar, al caso del Capítulo 3, pero incluyendo las continuas), y por otro lado, modelar las variables dicotómicas $\mathbf{H}$ en función de $(\mathbf{X}, \mathbf{W}, Y)$. A esta factorización la denominaremos Modelo 2.

Para la factorización (5.1)-(5.2), suponemos los modelos

$$\mathbf{V}|Y \sim N(\tilde{\boldsymbol{\mu}} + \mathbf{A}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \boldsymbol{\beta}(\mathbf{H} - \boldsymbol{\mu}_{\mathbf{H}}), \boldsymbol{\Delta}), \tag{5.5}$$

donde $\mathbf{A} \doteq \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}$, con $\boldsymbol{\mu}_{\mathbf{z}} = 0$ y $\text{diag}(\boldsymbol{\Delta}_{\mathbf{zz}}) = 1$ por identificabilidad (ver Capítulo 4), mientras que las variables dicotómicas fueron modeladas a través de

$$\mathbf{H}|Y = y \sim Bernoulli(\boldsymbol{\Gamma}^y), \tag{5.6}$$

en el marco de modelos Ising (i.e. incluyendos solo interacciones dobles), donde $\boldsymbol{\Gamma}^y$ es una matriz $q \times q$ simétrica, tal que para cada $[\boldsymbol{\Gamma}^y]_{ij} = \gamma_{ij}^y$ se asumió el el modelo lineal

$$\gamma_{ij}^Y = \tau_{ij,0}^* + \boldsymbol{\tau}_{ij}^T \left(\mathbf{f}_Y - \bar{\mathbf{f}}_Y\right), \qquad i, j = 1, \ldots, q \tag{5.7}$$

donde $\boldsymbol{\tau}_{ij}^T = (\tau_{ij,1}, \ldots, \tau_{ij,r})$ es un vector de parámetros (independientes de Y) y $\tau_{ij,0}^*$ es el intercepto para cada (i, j). También introdujimos la siguiente notación: definimos las matrices de orden $q \times q$, $\boldsymbol{\tau}_0$ y $\boldsymbol{\tau}_k$ $(k = 1, \ldots, r)$, de la siguiente manera $[\boldsymbol{\tau}_0^*]_{ij} = \tau_{ij,0}^*$, y $[\boldsymbol{\tau}_k]_{ij} = \tau_{ij,k}$ con $i, j = 1, \ldots, q$ y $k = 1, \ldots, r$, y luego definimos el vector $\boldsymbol{\tau}_0 \doteq \text{vech}(\boldsymbol{\tau}_0^*)$ de dimensión $q(q+1)/2$, y las matriz $\boldsymbol{\tau} = [\,\text{vech}(\boldsymbol{\tau}_1), \ldots, \text{vech}(\boldsymbol{\tau}_r)] \in \mathbb{R}^{q(q+1)/2 \times r}$. También asumimos un modelo de rango reducido, de manera que teníamos $\boldsymbol{\tau}$ de rango $c \leq \min(r, q(q+1)/2)$ de forma tal que especificamos $\boldsymbol{\tau}$ de la forma $\boldsymbol{\tau} = \boldsymbol{\kappa}\boldsymbol{\iota}$, donde $\boldsymbol{\kappa} \in \mathbb{R}^{q(q+1)/2 \times c}$ y $\boldsymbol{\iota} \in \mathbb{R}^{c \times r}$ son matrices de rango completo.

De forma equivalente, para la factorización (5.3)-(5.4) podemos suponer que

$$\mathbf{V}|Y \sim N(\tilde{\boldsymbol{\mu}} + \tilde{\mathbf{A}}(\mathbf{f}_Y - \bar{\mathbf{f}}_Y), \tilde{\boldsymbol{\Delta}}), \tag{5.8}$$

donde $\tilde{\mathbf{A}} \doteq \tilde{\boldsymbol{\Delta}}\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\xi}}$, con $\tilde{\boldsymbol{\mu}}_{\mathbf{z}} = 0$ y $\text{diag}(\tilde{\boldsymbol{\Delta}}_{\mathbf{zz}}) = 1$. Ahora, para las variables binarias suponemos el modelo,

$$\mathbf{H}|\mathbf{V}, Y \sim Bernoulli(\boldsymbol{\Gamma}^{Y,\mathbf{v}}), \tag{5.9}$$

y para cada parámetro de la matriz simétrica $\boldsymbol{\Gamma}^{Y,\mathbf{v}}$ $q \times q$, suponemos también un modelo lineal de la forma

$$\gamma_{ij}^{y,\mathbf{v}} = \tilde{\tau}_{ij,0}^* + \tilde{\boldsymbol{\tau}}_{ij}^T(\mathbf{f}_Y - \bar{\mathbf{f}}_Y) + \boldsymbol{\ell}_{ij}^T(\mathbf{V} - \boldsymbol{\mu}) \tag{5.10}$$

donde $\tilde{\boldsymbol{\tau}}_{ij}^T = (\tilde{\tau}_{ij,1}, \ldots, \tilde{\tau}_{ij,r})$ y $\boldsymbol{\ell}_{ij}^T = (\ell_{ij,1}, \ell_{ij,2} \ldots, \ell_{ij,t})$ son vectores de parámetros, independientes de Y. Definimos $\tilde{\boldsymbol{\tau}}_0$ y $\tilde{\boldsymbol{\tau}}$ de forma análoga al modelo anterior, y $\boldsymbol{\ell} = [\,\mathrm{vech}(\boldsymbol{\ell}_1), \ldots,\, \mathrm{vech}(\boldsymbol{\ell}_t)] \in \mathbb{R}^{q(q+1)/2 \times t}$, con $t = p + q$. También en este caso podemos asumir $\tilde{\boldsymbol{\tau}} = \tilde{\boldsymbol{\kappa}}\tilde{\boldsymbol{\iota}}$.

Con estas parametrizaciones, tenemos que para el Modelo 1 la función de densidad conjunta condicionada a Y, está dada por

$$\begin{aligned} f(\mathbf{V}, \mathbf{H}|Y = y) = {} & \frac{(2\pi)^{-t/2}|\boldsymbol{\Delta}|^{-1/2}}{G(\boldsymbol{\Gamma}^y)} \\ & \times \exp\Big\{ -\frac{1}{2}(\mathbf{V} - \boldsymbol{\mu})^T\boldsymbol{\Delta}^{-1}(\mathbf{V} - \boldsymbol{\mu}) + (\mathbf{f}_y - \bar{\mathbf{f}}_y)^T\mathbf{A}\boldsymbol{\Delta}^{-1}(\mathbf{V} - \boldsymbol{\mu}) \\ & - \frac{1}{2}(\mathbf{f}_y - \bar{\mathbf{f}}_y)^T\mathbf{A}^T\boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{f}_y - \bar{\mathbf{f}}_y) + (\mathbf{H} - \boldsymbol{\mu}_H)^T\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}(\mathbf{V} - \boldsymbol{\mu}) \\ & - (\mathbf{H} - \boldsymbol{\mu}_H)^T\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{f}_y - \bar{\mathbf{f}}_y) - \frac{1}{2}(\mathbf{H} - \boldsymbol{\mu}_H)^T\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}(\mathbf{H} - \boldsymbol{\mu}_H) \\ & + \mathrm{vech}^T(\mathbf{H}\mathbf{H}^T)\boldsymbol{\tau}_0) + \mathrm{vech}^T(\mathbf{H}\mathbf{H}^T)\boldsymbol{\tau}(\mathbf{f}_y - \bar{\mathbf{f}}_y)\Big\}, \end{aligned} \tag{5.11}$$

donde $G(\tilde{\boldsymbol{\Gamma}}^y) = \sum_{\{H_j : j=1,\ldots,q\}} \exp(\sum_{j=1}^q \gamma_{jj}^y H_j + \sum_{1 \le j < j' \le q} \gamma_{jj'}^y H_j H_{j'})$.

Por su parte, para el Modelo 2, la densidad conjunta condicionada a Y será:

$$\begin{aligned} f(\mathbf{V}, \mathbf{H}|Y = y) = {} & \frac{(2\pi)^{-t/2}|\tilde{\boldsymbol{\Delta}}|^{-1/2}}{G(\boldsymbol{\Gamma}^{y,\mathbf{v}})} \\ & \times \exp\Big\{ -\frac{1}{2}(\mathbf{V} - \tilde{\boldsymbol{\mu}})^T\tilde{\boldsymbol{\Delta}}^{-1}(\mathbf{V} - \tilde{\boldsymbol{\mu}}) + (\mathbf{f}_y - \bar{\mathbf{f}}_y)^T\tilde{\mathbf{A}}\tilde{\boldsymbol{\Delta}}^{-1}(\mathbf{V} - \tilde{\boldsymbol{\mu}}) \\ & - \frac{1}{2}(\mathbf{f}_y - \bar{\mathbf{f}}_y)^T\tilde{\mathbf{A}}^T\tilde{\boldsymbol{\Delta}}^{-1}\tilde{\mathbf{A}}(\mathbf{f}_y - \bar{\mathbf{f}}_y) + \mathrm{vech}^T(\mathbf{H}\mathbf{H}^T)\tilde{\boldsymbol{\tau}}_0 \\ & + \mathrm{vech}^T(\mathbf{H}\mathbf{H}^T)\tilde{\boldsymbol{\tau}}(\mathbf{f}_y - \bar{\mathbf{f}}_y) + \mathrm{vech}^T(\mathbf{H}\mathbf{H}^T)\boldsymbol{\ell}(\mathbf{V} - \tilde{\boldsymbol{\tau}}_0)\boldsymbol{\mu})\Big\}, \end{aligned} \tag{5.12}$$

donde $G(\tilde{\boldsymbol{\Gamma}}^{y,\mathbf{v}}) = \sum_{\{H_j : j=1,\ldots,q\}} \exp(\sum_{j=1}^q \gamma_{jj}^{y,\mathbf{v}} H_j + \sum_{1 \le j < j' \le q} \gamma_{jj'}^{y,\mathbf{v}} H_j H_{j'})$.

Vamos a mostrar que el Modelo 1 está anidado en el Modelo 2. Específicamente vamos a ver que, si en el Modelo 2 consideramos que $\boldsymbol{\ell}_s$ es una matriz diagonal, luego hay una correspondencia uno a uno entre los parámetros de ambos modelos. Por ende, el Modelo 1 estará anidado en el Modelo 2. Dados los parámetros del Modelo 1, definimos los parámetros del Modelo 2

como:

$$\begin{aligned}
\tilde{\boldsymbol{\mu}} &= \boldsymbol{\mu} - \boldsymbol{\beta}\boldsymbol{\mu}_H && (5.13)\\
\tilde{\boldsymbol{\Delta}} &= \boldsymbol{\Delta}, && (5.14)\\
\tilde{\mathbf{A}} &= \mathbf{A}, && (5.15)\\
\tilde{\boldsymbol{\tau}}_0 &= \boldsymbol{\tau}_0 + \text{ vech}\left(-\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta} + \text{diag}([\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}}])\right), && (5.16)\\
\text{vech}(\tilde{\boldsymbol{\tau}}_k) &= \text{vech}\left(\boldsymbol{\tau}_k - \text{diag}([\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A}]_{\bullet k})\right), k = 1, \ldots, r, && (5.17)\\
\text{vech}(\boldsymbol{\ell}_s) &= \text{vech}(\text{diag}([\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}]_{\bullet s}), s = 1, \ldots, t. && (5.18)
\end{aligned}$$

donde $[\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A}]_{\bullet k}$ representa la k-ésima columna de la matriz $q \times r$ $\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A}$, $[\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}]_{\bullet s}$ es la s-ésima columna de la matriz $q \times t$ $\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}$ y $\text{diag}(\mathbf{c})$ es una matriz diagonal con $[\text{diag}(\mathbf{c})]_{ii} = c_i$, para un cierto vector $\mathbf{c}$. Ademas, de (5.16)-(5.18), tenemos que $\boldsymbol{\Gamma}^{y,\mathbf{v}} = \boldsymbol{\Gamma}^y$, y $G(\boldsymbol{\Gamma}^{y,\mathbf{v}}) = G(\boldsymbol{\Gamma}^y)$.

Recíprocamente, dado que $\boldsymbol{\ell}_s$ es una matriz diagonal, y dado $\tilde{\boldsymbol{\mu}}$, $\tilde{\mathbf{A}}$, $\tilde{\boldsymbol{\Delta}}$, $\tilde{\boldsymbol{\tau}}_0, \tilde{\boldsymbol{\tau}}_1, \ldots, \tilde{\boldsymbol{\tau}}_r$ y $\boldsymbol{\ell}_s$ definimos

$$\begin{aligned}
\mathbf{A} &= \tilde{\mathbf{A}}, && (5.19)\\
\boldsymbol{\Delta} &= \tilde{\boldsymbol{\Delta}}, && (5.20)\\
& && (5.21)
\end{aligned}$$

Ahora, podemos obtener β de la ecuación

$$\text{vech}(\boldsymbol{\ell}_s) = \text{vech}(\text{diag}([\boldsymbol{\beta}^T\tilde{\boldsymbol{\Delta}}^{-1}]_{\bullet s}), s = 1, \ldots, t. \quad (5.22)$$

Dado $\boldsymbol{\beta}$ podemos obtener $\boldsymbol{\mu}_{\mathbf{H}}$ de la ecuación

$$\boldsymbol{\mu} = \tilde{\boldsymbol{\mu}} - \boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}}. \quad (5.23)$$

Luego, dado $\boldsymbol{\beta}$ y $\boldsymbol{\mu}_{\mathbf{H}}$, $\boldsymbol{\Delta}$ y $\tilde{\boldsymbol{\tau}}$ podemos obtener $\boldsymbol{\tau}_0$ y $\boldsymbol{\tau}_k$ de la ecuaciones

$$\begin{aligned}
\boldsymbol{\tau}_0 &= \tilde{\boldsymbol{\tau}}_0 - \text{ vech}\left(-\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta} + \text{diag}([\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{H}}])\right),\\
\text{vech}(\boldsymbol{\tau}_k) &= \text{vech}\left(\tilde{\boldsymbol{\tau}}_k + \text{diag}([\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\mathbf{A}]_{\bullet k})\right), k = 1, \ldots, r.
\end{aligned}$$

De esta manera, vemos que el Modelo 1 está anidado en uno más general, el Modelo 2. Con este resultado, como trabajo futuro queda por encontrar la reducción suficiente para esta factorización alternativa, y así poder evaluar qué parametrización se adapta mejor al conjunto de datos, vía por ejemplo, cociente de verosimilitud.

5.3. Reducción Parcial

La aplicación considerada a la construcción de índices SES también motiva el desarrollo de métodos de reducción parcial. Específicamente, al trabajar con datos sociales para la construcción de un índice SES, existe información en ciertas variables que son relevantes para explicar la respuesta que se está modelando, pero que no constituyen variables específicamente de índole socio-económica para que sean incluidas en un índice, como podría ser el sexo, variables regionales o temporales. También en muchos casos interesa construir el índice teniendo sólo en cuenta la información puramente económica, esto es a través de los activos y las condiciones habitacionales, y dejar otras dimensiones como la educación o el acceso a la salud, de forma separada del índice. Por ello, estamos interesados en considerar la información de todos los predictores pero sólo queremos reducir un grupo de ellos (en un índice SES), manteniendo los restantes sin alterar. Tal como se expresa comúnmente en la literatura relacionada, para una determinada respuesta Y, el modelo de interés tiene la forma

$$Y = b_0 + b_1 SES + g(\mathbf{U}) + \boldsymbol{\varepsilon}, \tag{5.24}$$

donde SES es el índice de estatus socio-económico, $\mathbf{U}$ es un vector de variables relevantes para el modelo pero que no forman parte del índice SES, $g(\cdot)$ es una cierta función, comúnmente (pero no necesariamente) lineal de la forma $g(\mathbf{U}) = \mathbf{d}^T\mathbf{U}$ y $\boldsymbol{\varepsilon}$ un término de perturbación aleatorio. El modelo (5.24) ya contiene la reducción en la variable SES. Es decir, la regresión primitiva al modelo (5.24) está dada por

$$Y|\mathbf{X}, \mathbf{W}, \mathbf{H}, \mathbf{U} \tag{5.25}$$

donde $(\mathbf{X}, \mathbf{W}, \mathbf{H})$ son las variables que interesa reducir (i.e. ordinales, continuas y dicotómicas) a los fines de contruir la SES, y $\mathbf{U}$ es un vector de variables relevantes para el modelo (5.25)

pero que no interesa reducir. Por ende, el objetivo es encontrar una reducción $\mathbf{R}(\mathbf{X}, \mathbf{W}, \mathbf{H})$ para la regresión de Y sobre $(\mathbf{X}, \mathbf{W}, \mathbf{H}, \mathbf{U})$ de forma tal que

$$Y|\mathbf{R}(\mathbf{X}, \mathbf{W}, \mathbf{H}), \mathbf{U} \stackrel{\mathrm{d}}{=} Y|\mathbf{X}, \mathbf{W}, \mathbf{H}, \mathbf{U}. \tag{5.26}$$

La literatura relacionada a SDR parcial es muy escasa. Chiaromonte et al. (2002) introducen la idea de subespacio central parcial $\mathcal{Y}|\mathbf{W}^{\mathbf{U}}$ definido como la intersección de todos los subespacios $\mathcal{S}$ que satisfacen $Y \perp\!\!\!\perp W|(P_{\mathcal{S}}\mathbf{W}, \mathbf{U})$ donde $P_{\mathcal{S}}$ es la proyección sobre $\mathcal{S}$ con el producto escalar estándar. Chiaromonte et al. (2002) proponen un estimador para el subespacio central parcial solo para el caso en que $\mathbf{U}$ (i.e. las variables que no interesa reducir) son categóricas con $\mathbf{W}$ continuas. Wen and Cook (2007), si bien relaja algunos supuestos de Chiaromonte et al. (2002), también está limitado a reducciones parciales de un conjunto de predictores continuos para cada sub-población que definen la categóricas incluidas en $\mathbf{U}$. Bajo esta forma de abordar reducción parcial, resulta difícil generalizarlo a fin de admitir cualquier naturaleza en el conjunto de predictores $\mathbf{U}$. Feng et al. (2013) propone una extensión que admite $\mathbf{U}$ continua, aplicando un proceso de discretización sobre las mismas a fin de estimar el subespacio central parcial.

Como trabajo futuro se planea abordar el problema de SDR de forma que admita más posibilidades entre los predictores, y tenga una representación más parsimoniosa respecto a estas primeras aproximaciones basadas en estimar el subespacio central parcial. Específicamente, se plantea obtener un método de reducción parcial de dimensiones para mezclas de variables de distinta naturaleza bajo el enfoque de modelar la regresión inversa.

Bibliografía

Adragni, K. & Cook, R. (2009), 'Sufficient dimension reduction and prediction in regression', *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **367**(1906), 4385–4405.

Akaho, S. (2001), *A kernel method for canonical correlation analysis*, Springer, Tokio.

Alkire, S. & Foster, J. (2011), 'Counting and multidimensional poverty measurement', *Journal of Public Economics* **95**, 476–487.

Alkire, S. & Santos, M. (2014), 'Measuring acute poverty in developing world: Robustness and scope of the multidimensional poverty index', *World Development* **59**, 251–274.

Angeles, G. & You, Y. (2007), Vailability of data for estimating ses indices using household surveys, Carolina Population Center, Chapel Hil, NC.

Bobadilla, J., Ortega, F., Hernando, A. & Bernal, J. (2012), 'Generalization of recommender systems: Collaborative filtering extended to groups of users and restricted to groups of items', *Expert Systems with Applications* **39**(1), 172–186.

Bollen, K., Glanville, J. & Stecklov, G. (2001), 'Socioeconomic status and class in studies on fertility and health in developing countries', *Annual Review of Sociology* **27**, 153–185.

Bura, E. & Cook, R. (2001*a*), 'Estimating the structural dimension of regressions via parametric inverse regression', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **63**(2), 393–410.

Bura, E. & Cook, R. D. (2001*b*), 'Estimating the structural dimension of regressions via parametric inverse regression', *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **63**, 393–410.

Bura, E., Duarte, S. & Forzani, L. (2015), 'Sufficient reductions in regressions with exponential family inverse predictors', *Journal of the American Statistical Association (in press)* .

Bura, E. & Forzani, L. (2015), 'Sufficient reductions in regressions with elliptically contoured inverse predictors', *Journal of the American Statistical Association* **110**(509), 420–434.

Burges, C. (2009), 'Dimension reduction: A guided tour', *Foundations and Trends in Machine Learning* **2**(4), 275–365.

Caruso, G., Sosa-Escudero, W. & Svarc, M. (2015), 'Deprivation and the dimensionality of welfare: A variable-selection cluster-analysis approach', *Review of Income and Wealth* **61**(4), 702–722.

Casella, G. & Berger, R. L. (2002), *Statistical inference*, Duxbury advanced series, Brooks Cole, Pacific Grove (Calif.).

Chang, Y.-C. & Liang, J.-Y. (1994), 'Estimation in the exponential family in the precense of nuisance parameters: Compromise between bias and precision', *Statistica Sinica* **4**, 169–185.

Chen, X., Zou, C. & Cook, R. (2010), 'Coordinate-independent sparse sufficient dimension reduction and variable selection', *The Annals of Statistics* **38**, 3696–3723.

Cheng, J., Levina, E., Wang, P. & Zhu, J. (2014), 'A sparse ising model with covariates', *Biometrics* **70**(4), 943–953.

Cheng, J., Li, T., Levina, E. & Zhu, J. (2016), 'High-dimensional mixed graphical models', *Journal of Computational and Graphical Statistics* **(in press)**.

Chiaromonte, F., Cook, R. & Li, B. (2002), 'Sufficient dimensions reduction in regressions with categorical predictors', *The Annals of Statistics.* **30**(2), 475–497.

Cook, R. (1994), Using dimension reduction subspaces to identify important inputs in models of physical systems, *in* 'Proceedings of the Section on Physical and Engineering Sciences, American Statistical Association', pp. 18–25.

Cook, R. (1998*a*), 'Principal hessian directions revisited', *Journal of the American Statistical Association* **93**(441), 84–94.

Cook, R. (1998*b*), *Regression Graphics*, Wiley, New York.

Cook, R. (2007), 'Fisher lecture: Dimension reduction in regression (with discussion)', *Statistical Science* **22**, 1–26.

Cook, R. D., Forzani, L. & Zhang, X. (2015), 'Envelopes and reduced-rank regression', *Biometrika* **102**(2), 439–456.

Cook, R. D. & Li, B. (2002*a*), 'Dimension reduction for conditional mean in regression', *The Annals of Statistics* **30**, 455–474.

Cook, R. D. & Ni, L. (2005*a*), 'Sufficient dimension reduction via inverse regression: a minimum discrepancy approach', *Journal of the American Statistical Association* **100**, 410–428.

Cook, R. D. & Weisberg, S. (1991*a*), 'Discussion of sliced inverse regression for dimension reduction', *Journal of the American Statistical Association* **86**, 328–332.

Cook, R. & Forzani, L. (2008), 'Principal fitted components for dimension reduction in regression', *Statistical Science* **23**, 485–501.

Cook, R. & Forzani, L. (2009), 'Likelihood-Based sufficient dimension reduction', *Journal of the American Statistical Association* **104**(485), 197–208.

Cook, R. & Lee, H. (1999), 'Dimension reduction in binary response regression', *Journal of the American Statistical Association* **94**(448), 1187–1200.

Cook, R. & Li, B. (2002*b*), 'Dimension reduction for conditional mean in regression', *The Annals of Statistics* **30**(2), 455–474.

Cook, R. & Ni, L. (2005*b*), 'Sufficient dimension reduction via inverse regression: A minimum discrepancy approach', *Journal of the American Statistical Association* **100**(470), 410–428.

Cook, R. & Weisberg, S. (1991*b*), 'Discussion of sliced inverse regression for dimension reduction', *Journal of the American Statistical Association* **86**, 328–332.

Cook, R. & Yin, X. (2001), 'Dimension reduction and visualization in discriminant analysis (invited, with discussion)', *Australia & New Zeland Journal of Statistics* **43**, 147–200.

Dai, B. (2013), 'Mvb: Multivariate bernoulli log-linear model'. R package version 1.1.

Dai, B., Ding, S. & Wahba, G. (2013), 'Multivariate Bernoulli distribution', *Bernoulli* **19**, 1465–1483.

Doocy, S. & Burnham, G. (2006), 'Assessment of socio-economic status in the context of food insecurity: Implications for field research', *World Health and Population* **8**(3), 32–42.

Duarte, S. (2016), Modelos lineales generalizados: Regresión de rango reducido y reducción suficiente de dimensiones, dissertation, Universidad Nacional del Litoral.

Eaton, M. (1983), *Multivariate Statistics*, Wiley, New York.

Feeny, S., Mcdonald, L. & Posso, A. (2014), 'Are poor people less happy? findings from melanesia', *World Development* **64**, 448–459.

Feng, Z., Wen, X. M., Yu, Z. & Zhu, L. (2013), 'On partial sufficient dimension reduction with applications to partially linear multi-index models', *Journal of the American Statistical Association* **108**(501), 237–246.

Fernald, L. (2007), 'Socio-economic status and body mass index in low-income mexican adults', *Social Science and Medicine* **64**, 2030–2042.

Filmer, D. & Pritchett, J. (1998), 'Estimating wealth effect without expenditure data -of tears: An application to educational enrollments in states of india', *World Bank Policy Research Working Paper* (4).

Filmer, D. & Pritchett, J. (2001), 'Estimating wealth effect without expenditure data -of tears: An application to educational enrollments in states of india', *Demography* **38**(4), 115–132.

Friedman, M. (1957), *A theory of the consumption function*, Princeton University Press, Princeton.

Fukumizu, K., Bach, F. R. & Jordan, M. I. (2009), 'Kernel dimension reduction in regression', *The Annals of Statistics* **37**, 1871–1905.

Fukumizu, K. & Leng, C. (2014), 'Gradient-based kernel dimension reduction for regression', *Journal of the American Statistical Association* **109**(505), 359–370.

Gertheiss, J. & Tutz, G. (2010), 'Sparse modelling of categorical explanatory variables', *The Annals of Applied Statistics* **4**, 2150–2180.

Greene, W. & Hensher, D. (2010), *Modeling Ordered Choices: A Primer*, Cambridge University Press.

Guo, J., Levina, E., Michailidis, G. & Zhu, J. (2015), 'Graphical models for ordinal data', *Journal of Computational and Graphical Statistics* **24**(1), 183–204.

Hoque, S. (2014), 'Asset-based poverty analysis in rural bangladesh: A comparison of principal component analysis and fuzzy set theory', *SRI Papers, Sustainability Research Institute* **59**.

Hsing, T. & Ren, H. (2009), 'An RKHS formulation of the inverse regression dimension-reduction problem', *The Annals of Statistics* **37**, 726–755.

INDEC (2003), *Calidad de los materiales de la vivienda -CALMAT-Hábitat y vivienda por medio de datos censales. Dirección Nacional de Estadísticas Sociales y de Población, Dirección de Estadísticas Poblacionales, área de Información Derivada.*, DNESyP/DEP/P5/PID Serie Hábitat y Vivienda DT No. 13, Buenos Aires.

Jackman, S. (2009), *Bayesian Analysis for the Social Sciences*, Wiley Series in Probability and Statistics, Wiley.

Jørgensen, B. & Labouriau, R. (2012), *Exponential Family and Statistical Inference*, Monografias de Matematica, no. 54, IMPA, Rio de Janeiro.

Kamakura, W. & Mazzon, J. (2013), 'Socioeconomic status and consumption in an emerging economy', *International Journal of Research in Marketing* **30**, 4–18.

Kolenikov, S. & Angeles, G. (2009), 'Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer?', *The Review of Income and Wealth* **55**(1), 128–165.

Lee, G. & Scott, C. (2012), 'Em algorithms for multivariate gaussian mixture models with truncated and censored data', *Computational Statistics and Data Analysis* **56**(9), 2816–2829.

Lehmann, E. & Casella, G. (2003), *Theory of Point Estimation*, Springer Texts in Statistics, Springer New York.

Li, B., Artemiou, A. & Li, L. (2011), 'Principal support vector machines for linear and nonlinear sufficient dimension reduction', *The Annals of Statistics* **39**(6), 3182–3210.

Li, B. & Wang, S. (2007), 'On directional regression for dimension reduction', *Journal of the American Statistical Association* **102**(479), 997–1008.

Li, B., Zha, H. & Chiaromonte, C. (2005*a*), 'Contour regression: a general approach to dimension reduction', *The Annals of Statistics* **33**(4), 1580–1616.

Li, B., Zha, H. & Chiaromonte, F. (2005*b*), 'Contour regression: a general approach to dimension reduction', *The Annals of Statistics* **33**, 1580–1616.

Li, K. (1992*a*), 'On principal hessian directions for data visualization and dimension reduction: Another application of steinś lemma', *Journal of the American Statistical Association* **87**(420), 1025–1039.

Li, K.-C. (1991*a*), 'Sliced inverse regression for dimension reduction', *Journal of the American Statistical Association* **86**, 316–342. With discussion and a rejoinder by the author.

Li, K. C. (1991*b*), 'Sliced inverse regression for dimension reduction (with discussion)', *Journal of the American Statistical Association* **86**, 316–342.

Li, K.-C. (1992*b*), 'On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma', *Journal of the American Statistical Association* **87**, 1025–1039.

Linting, M. & van der Kooij, A. J. (2009), 'Nonlinear principal components analysis with catpca: A tutorial.', *Journal of Personality Assessment* **94**(1), 12–25.

Long, J. (1997), *Regression models for categorical and limited dependent variables*, 2 edn, SAGE Publications.

Mazzonna, F. (2014), 'The long-lasting effects of family background: A european cross-country comparison', *Economics of Education Review* **40**, 25–42.

Meinshausen, N. & Bühlmann, P. (2006), 'High-dimensional graphs and variable selection with the lasso', *The Annals of Statistics* **34**(3), 1436–1462.

Merola, G. & Baulch, B. (2014), Using sparse categorical principal components to estimate asset indices new methods with an application to rural south east asia, Rome, Italy.

Mokomane, Z. (2012), 'Social protection as a mechanism for family protection in sub-saharan africa', *International Journal of Social Welfare* **22**(3), 248–259.

Murasko, J. (2007), 'Socioeconomic status, height and obesity in children', *Economics and Human Biology* **7**(3), 376–386.

O'Donnell, O., van Doorslaer, E., Wagstaff, A. & Lindelow, M. (2007), *Analyzing health equity using household survey data : a guide to techniques and their implementation*, World Bank, Washington, DC.

Rao, C. (1973), *Linear statistical inference and its applications, 2nd. Edition.*, Wiley, New York.

Richardson, D. & Bradshaw, J. (2012), *Family-oriented anti-poverty policies in developed countries*, Department of Economic and Social Affairs, Division for Social Policy and Development, United Nations, New York, New York.

Roberts, W. (2014), 'Factor analysis parameter estimation from incomplete data', *Computational Statistics & Data Analysis* **70**(0), 61–66.

Roy, K. & Chaudhuri, A. (2009), 'Influence of socioeconomic status, wealth and financial empowerment on gender differences in health and healthcare utilization in later life: Evidence from india', *Social Science and Medicine* **66**, 1951–1962.

Sen, A. (1984), 'The living standard', *Oxford Economics Papers* **36**(supp), 74–90.

Sen, A. (1993), Capability and well-being, *in* M. C. Nussbaum & A. K. Sen, eds, 'The quality of life', Clarendon Press.

Sen, A. (1999), *Development as freedom*, Oxford University Press, New York.

Skrondal, A. & Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Chapman & Hall/CRC.

Székely, G. & Rizzo, M. (2009), 'Brownian distance covariance', *The Annals of Applied Statistics* **3**(4), 1236–1265.

Székely, G., Rizzo, M. & Bakirov, N. (2007), 'Measuring and testing dependence by correlation of distances', *The Annals of Statistics* **35**(6), 2769–2794.

Vyas, S. & Kumaranayake, L. (2006), 'Constructing socio-economic status indices: How to use principal components analysis', *Health Policy and Planning* **21**(6), 459–468.

Wang, X., Feng, H., Xia, Q. & Alkire, S. (2016), 'On the relationship between income poverty and multidimensional poverty in china', *OPHI working Paper* (101), 1–21.

Wen, X. & Cook, D. (2007), 'Optimal sufficient dimension reduction in regressions with categorical predictors.', *Journal of Statistical Planning and Inference* **137**, 1961–1978.

World Bank (2000), *World Development Report 2000/2001: Atacking Poverty*, World Bank, Washington, DC.

Wu, Q., Liang, F. & Mukherjee, S. (2008), 'Regularized sliced inverse regression for kernel models'. Technical report, Duke Univ., Durham, NC.

Xia, Y., Tong, H., Li, W. K. & Zhu, L.-X. (2002*a*), 'An adaptive estimation of dimension reduction space', *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**, 363–410.

Xia, Y., Tong, H., Li, W. & Zhu, L. (2002*b*), 'An adaptative estimation of dimension reduction space', *Journal of the Royal Statistical Society, Series B* **64**, 363–410.

Yeh, Y. R., Huang, S. Y. & Lee, Y. Y. (2009), 'Nonlinear dimension reduction with kernel sliced inverse regression', *IEEE Transactions on Knowledge and Data Engineering* **21**, 1590–1031.

Zhu, H. & Li, L. (2011), 'Biological pathway selection through nonlinear dimension reduction', *Biostatistics* **12**, 429–444.

Zhu, Y. & Zeng, P. (2006*a*), 'Fourier methods for estimating the central subspace and the central mean subspace in regression', *Journal of the American Statistical Association* **101**(476), 1638–1651.

Zhu, Y. & Zeng, P. (2006*b*), 'Fourier methods for estimating the central subspace and the central mean subspace in regression', *Journal of the American Statistical Association* **101**, 1638–1651.