PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE.
FACULTAD DE MATEMÁTICAS.
DEPARTAMENTO DE ESTADÍSTICA.

UN MODELO BAYESIANO SEMIPARAMÉTRICO EXPLICATIVO PARA EL ANÁLISIS DE DATOS EDUCACIONALES CHILENOS

POR PAULA FARIÑA.

Tesis presentada a la Facultad de Matemáticas de la
Pontificia Universidad Católica de Chile para optar
al grado académico de Doctor en Estadística

Profesor Guía: Ernesto San Martín.

Enero de 2010.

Santiago. Chile.

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULTAD DE MATEMÁTICAS

DEPARTAMENTO DE ESTADÍSTICA

Título de tesis: Un Modelo Bayesiano Semiparamétrico Explicativo para el Análisis de datos Educacionales Chilenos.

Autor: Paula Fariña.

Fecha: Enero de 2010.

Santiago. Chile.

Jurado Externo 1:	Alejandra Mizala Dep. de Ingeniería Industrial. Univ. de Chile
Jurado Externo 2:	Alejandro Jara Dep. de Estadística. Univ. de Concepción
Jurado Interno:	Fernando Quintana Dep. de Estadística. P. Univ. Católica de Chile
Profesor Guía:	Ernesto San Martín Dep. de Estadística. P. Univ. Católica de Chile

A Lucila y Luis, las dos luces en mi vida.

Paula

Índice general

Agradecimientos

En general, Chile ha sido para mi una puerta abierta a muchas oportunidades, no sólo a través de instituciones como CONICYT, sino gracias a las personas que he tenido la suerte de conocer. Agradezco infinitamente al profesor Ernesto San Martín por todo el entusiasmo y tiempo dedicado en esta tesis. Ha sabido tener confianza en mi cuando yo la había perdido, ingrediente que resultó indispensable. Agradezco también al profesor Fernando Quintana por haber sido un manantial de conocimiento tanto en el transcurso de esta investigación como durante los cursos que he tenido la suerte de asistir. A Alejandro Jara por haber nutrido este trabajo con sugerencias interesantes, y al Departamento de Estadística de la Pontificia Universidad Católica de Chile, profesores y administrativos que me han acompañado en esta etapa.

Quiero además, aprovechar este espacio para agradecer especialmente a Luis por haberme apoyado durante estos 5 años en los que hubo, por cierto, momento duros; y a mis padres - Raquel, Enrique y Eduardo -, quienes desde muy temprano han sabido inculcarme la pasión por el conocimiento. A pesar de la distancia física, su influencia está presente en mí a cada paso, y junto con mi marido han sido el motor que me ha impulsado en este proyecto. Gracias a Mallén, Karla y todos mis compañeros de estudio con quien he compartido y aprendido; y a mis amigos: Andrea, Arnaldo, Estela, Nelson, Pancho, Óscar y la gente del Taller 56 que me han apoyado y han sido testigos del esfuerzo depositado en estas páginas.

Finalmente, debo mencionar que esta Tesis no hubiera podido realizarse sin el apoyo de CONICYT, organismo que ha financiado mis estudios de doctorado mediante las becas de Doctorado 2004 y Término de Tesis Doctoral 2009; y de la buena disposición de SIMCE, organismo que suministró los datos para el caso de estudio.

Resumen

Este trabajo propone una nueva metodología para analizar mediciones educacionales que surgen a partir de instrumentos de medición estandarizada como PISA (Programme for International Student Assessment), TIMSS (Third International Mathematics and Science Study) o SIMCE (Sistema de Medición de la Calidad de la Educación); y que se implementan para monitorear el desempeño de los sistemas educacionales de los países. En estos casos, los esfuerzos estadísticos se focalizan en caracterizar las habilidades de una *población o subpoblación* de estudiantes así como en comprender cómo factores a nivel individual (sexo, nivel educacional de los padres, nivel socioeconómico de la familia, etc.), a nivel colegio (nivel socioeconómico de la familia, etc), y factores ambientales externos, influyen sobre las habilidades. Se tomó un enfoque Bayesiano, proponiendo un *modelo de tipo IRT Bayesiano con covariables*. Las covariables se incorporan indexando una colección de distribuciones de habilidad aleatorias. La distribución para dicha colección se especifica como una mezcla de ANOVA DDP, y hace posible una gran diversidad de formas para las distribuciones dentro de dicha colección. Esta especificación es equivalente a considerar una *Mezcla de Normales* donde la distribución mezclante proviene de un *Proceso Dirichlet Dependiente*. Tiene la ventaja de relajar el supuesto arbitrario de normalidad de las habilidades - comúnmente utilizada en la literatura de modelos IRT - y es nueva dentro del ámbito de la psicometría. También lo es la estrategia para la inferencia: se

considera como parámetros de interés a las *distribuciones* de la habilidad condicionales en covariables, lo que permite inferir acerca de la *forma* de dichas distribuciones, y no sólo sobre su primer y segundo momento, como típicamente ocurre con los modelos lineales jerárquicos mixtos.

La metodología propuesta deja a disposición del investigador estimaciones del valor esperado de la distribución a posteriori de la habilidad condicional a las covariables, junto con una banda de credibilidad de 95 % para dicho valor esperado. El estimador propuesto es acertado para comprender y comparar las habilidades de sub-poblaciones de individuos con determinadas características. Las bandas de credibilidad son empleadas para respaldar las características de forma de las densidades, como la bimodalidad y asimetría. Para cuantificar las diferencias entre densidades se utiliza la divergencia de Kullback-Leibler; y se propone, también, una forma de decidir si las divergencias son relevantes o no.

El caso de estudio se basa en mediciones educativas provistas por SIMCE para el caso chileno. Se restringe a los datos del SIMCE 2004 para 8vo grado de las comunas de Peñalolén, La Florida y Las Condes. Estas comunas son parte del área metropolitana de Santiago y cubren un amplio rango socioeconómico. En cuanto al año, 8vo grado es el final de la educación básica, momento apropiado para evaluar el desempeño global de esta etapa. Los datos disponibles incluyen también covariables provenientes de la encuesta de padres. Específicamente se utiliza una variable categórica que indica la fuente de financiamiento del colegio. Posee tres niveles: colegios municipales, particulares subvencionados y particulares pagados. Se construye un índice socioeconómico disponible a nivel individual y de colegio. Finalmente se considera una variable que indica si el estudiante repitió al menos un año en el pasado.

Desde el punto de vista práctico, una de nuestras principales conclusiones es que, para los estudiantes de colegios particulares subvencionados y que no repiten, el índice de nivel socioeconómico del individuo afecta fuertemente las formas distribucionales. Por otra parte, las diferencias en el índice de nivel socioeconómico del colegio se reflejan como desplazamientos en las distribuciones de habilidad. Las divergencias de K-L son altas al comparar alumnos de niveles socioeconómico individuales bajo y alto. Se reducen considerablemente al comparar estudiantes con niveles socioeconómicos bajo y medio, tanto para la variable individual como de colegio. Sin embargo, estas diferencias siguen siendo relevantes. Las distribuciones de habilidad de alumnos que repiten y que no repiten poseen formas disímiles y divergencias de K-L moderadas, pero relevantes también. Si bien encontramos que las distribuciones de habilidad no cambian su localización y forma por tipo de colegio, este resultado debe tomarse con cautela ya que en el presenta estudio no se tiene en cuenta el sesgo de selección que ocurre debido a que los alumnos no están asignados en forma aleatoria por tipo de colegio.

La presencia de bimodalidad es otro resultado que debe ser destacado, pues indica la presencia de dos tipos de individuos al interior del subgrupo. Encontramos bimodalidad en los subgrupos de alumnos que no repiten, con nivel socioeconómico individual medio, y nivel socioeconómico de colegio medio y alto. También es destacable los resultados acerca de la presencia de asimetría, pues indica que la masa probabilidad está más cargada hacia valores bajos de habilidad (simetría a la izquierda) o altos de habilidad (simetría a la derecha). A partir del estudio se detecta una marcada asimetría hacia la izquierda para alumnos de colegios municipales y subvencionados que no repiten con nivel socioeconómico individuales y de colegio bajos, y en la mayoría de los subgrupos de repitentes. La presencia de bimodalidad y asimetría al mismo tiempo indica que el subgrupo en cuestión contienen dos tipos distintos de individuos

en su interior, donde uno de los tipo de individuo es menos frecuente que el otro. Tal es el caso de los subgrupos de alumnos que no repiten con nivel socioeconómico individuales y de colegio altos. La asimetría a la derecha junto con la bimodalidad implican que el tipo de alumnos con menor habilidad es menos frecuente que el de mayor habilidad. Se contrasta la metodología propuesta con los métodos alternativos estándar, mediante estimaciones de un modelo IRT explicativo paramétrico y un modelo jerárquico de componentes de varianza.

Desde el punto de vista teórico esta tesis contribuye a introducir la estadística Bayesiana No Paramétrica dentro del área de la Psicometría. Se amplía, así, la batería de modelos existentes para el análisis de datos educacionales. Típicamente, los modelo empleados en el área son modelo IRT paramétricos y modelos lineales jerárquicos. La riqueza particular de la especificación propuesta se halla en permitir la inferencia sobre la forma de distribuciones *enteras*, no sólo sobre el primer y segundo momento; extrayendo más información de los datos que los modelos paramétricos alternativos. Su implementación práctica implicó el desarrollo de un software próximamente disponible al público en http://www.paulaestadistica.blogspot.com/. También será incluido en el DPpackage de Jara (2007) para usuarios de R.

Capítulo 1

Contexto de esta investigación: A modo de Introducción

Las mediciones educacionales a gran escala como PISA (Programme for International Student Assessment), TIMSS (Third International Mathematics and Science Study), etc. han adquirido gran relevancia en los últimos años. A diferencia de los exámenes de admisión, que buscan ordenar individuos de acuerdo a sus habilidades, el propósito detrás de estos exámenes es monitorear el desempeño del sistema educacional en su conjunto. En este caso los esfuerzos estadísticos se focalizan en caracterizar la *población* en lugar de las habilidades individuales. Además, se busca comprender cómo algunos factores socioeconómicos a nivel individual, de colegio y de país determinan las habilidades de los estudiantes.

Típicamente los estudios sobre el tema suponen una relación lineal para vincular las covariables socioeconómicas con los puntajes observados. Más precisamente, la esperanza del puntaje de una prueba estandarizada condicional a las covariables es lineal en las mismas, aunque a veces se incluyen interacciones. Los modelos jerárquicos (HLM), también conocidos como modelos multinivel (MM), son los más comúnmente

usados (ver Goldstein, 1995). En este trabajo se propone una metodología distinta para describir las habilidades de los estudiantes con determinadas características socioeconómicas. La misma utiliza un modelo basado en la teoría de respuesta al ítem (IRT)[1] Bayesiano. Los modelos IRT suponen la existencia de una variable latente a nivel individual, la habilidad, que si bien no es observable, es posible hacer inferencia sobre ella dado que las respuestas al instrumento de medición son manifestaciones observables de dicha habilidad. El modelo desarrollado aquí va mas allá del modelo IRT clásico: modelamos la distribución de habilidad en forma no paramétrica y proponemos que las habilidades individuales se vean influenciadas por las características socioeconómicas y de entorno del individuo. De esta forma, logramos inferir sobre la forma distribucional de las habilidades utilizando la información observable tanto de la prueba como de los datos socioeconómicos, sin hacer suposiciones restrictivas (como por ejemplo suponer una distribución normal para las habilidades). Se propone estimar el valor esperado de la función de distribución a posteriori de la habilidad condicional en covariables, el cual coincide con la distribución de habilidad de un nuevo individuo hipotético dadas las características socioeconómicas. Esto es apropiado para el presente estudio donde lo importante es comprender y comparar las habilidades de un individuo genérico con determinadas características, y no la habilidad de un individuo en particular de la muestra. Si bien este modelo fue desarrollado para analizar el caso chileno, puede ser utilizado sin inconvenientes en cualquier sistema educacional que posea la información necesaria, e incluso para comparar países. En este sentido, nuestra propuesta es un avance con respecto al trabajo de Wo (2005), que plantea hacer inferencia a partir de *valores factibles*, simulaciones (5 valores por cada individuo) de la distribución a posteriori de la habilidad individual.

[1]Los modelos IRT son tipo especial de modelo lineal generalizado que se describe en detalle en la sección 3.1.

En lo que resta de este capítulo se exponen brevemente las características del sistema educacional chileno y las investigaciones existentes hasta el momento sobre este país. Se deja para los restantes capítulos la explicación detallada del nuevo modelo y su aplicación al caso chileno.

1.1. El Sistema Educacional chileno

El sistema educacional chileno se divide en 8 años de educación básica y 4 de educación media. Los últimos 4 años son también obligatorios desde 2003. Su actual estructura se conformó después de la reforma de 1981, llevada a cabo durante el Régimen Militar. Las particularidades centrales de dicha reforma son la descentralización de la administración de colegios públicos, los cuales pasaron a depender de los municipios; y el establecimiento de un programa de cupones[2], ofreciendo una experiencia sui generis para América Latina en este tipo de programas. Los sistemas de cupones fueron propuestos por primera vez por Friedman[3]. En ellos el gobierno asigna a los padres un cupón que puede ser usado en un colegio privado o público, según la preferencia de los padres. La hipótesis detrás de este esquema es que los colegios se verán sometidos a una situación competitiva que producirá mejoras en la educación.

El programa de cupones chileno tiene la particularidad de ser *universal*, es decir se aplica a todo los estudiantes del sistema educativo. Aparte de Chile, otros países como Dinamarca, Países Bajos, Corea del Sur y Suecia también poseen programas de cupones universales. Por otro lado, también existen programas de cupones dirigidos a una población específica de estudiantes, como en Estados Unidos (Milwaukee Parental Choice Program), Colombia, Guatemala y Pakistan, entre otros[4].

[2]Traducción libre de la autora del término en inglés *voucher system*.
[3]Ver Friedman (1955).
[4]West (1997) y Lara et al. (2009) presentan mayores detalles sobre estos casos.

El sistema de cupones más conocido se denomina *fondos que acompañan al niño*[5] en el cual las escuelas reciben un subsidio por parte del Estado proporcional al número de niños inscritos. El caso chileno es un ejemplo del esquema con fondos que acompañan al niño. El sistema incluye tres tipos de colegio: *municipales* financiados por el estado y administrado por las comunas; colegios *particulares subvencionados* financiados en forma mixta por el estado y los padres, y administrado por el sector privado; y colegios *particulares pagados*, que funcionan únicamente mediante la mensualidad pagada por los padres. El subsidio que otorga el gobierno a los colegio es el mismo tanto para colegios públicos como para subsidiados. En 2006 consistió en US$61.5 mensuales por alumno inscrito en la educación básica, y US$73.3 mensuales para colegios de educación media (ver Lara et al., 2009). En ese mismo año, las participaciones de inscritos en los distintos tipos de colegio fue de 47.7 % para colegios colegios municipales, 44 % para particulares subvencionados y 8.3 % para privados, lo que implica una importante transferencia de alumnos de colegios públicos a los particulares subvencionados desde la conformación del sistema[6].

El esquema se completa con la información provista por SIMCE (Sistema de Medición de la Calidad de la Educación) - una prueba de carácter censal aplicado a los estudiantes anualmente - usada por los padres para elegir correctamente los colegios, y por el gobierno para monitorear el desempeño del sistema y evaluar políticas públicas.

Los colegios públicos y particulares subvencionados se ven sometidos a diferentes reglas. En primer lugar los colegios particulares subvencionados seleccionan a los alumnos mediante exámenes, entrevistas a padres, etc., mientras que los colegios

[5]Traducción libre de la autora del término en inglés *funds-follow-the-child*.
[6]En 1981 el 78 % de los alumnos asistían a colegios públicos.

público deben admitir a todos los alumnos que se inscriben. En segundo lugar, la reglamentación de contratos docentes difieren. Mientras que los colegios particulares subvencionados funcionan como empresas privadas y siguen las leyes establecidas en el Código Laboral para contratar y despedir empleados; los contratos para los docentes del sector público están sometidos a una legislación especial, el Estatuto Docente, que impone negociaciones colectivas y restricciones para despidos. Finalmente la forma en que estos tipos de colegio obtienen financiamiento adicional también difiere. Los colegios particulares subvencionados pueden cobrar una cuota adicional a los padres. En el caso de los colegios municipales, sólo los colegios de educación media tiene permitido cobrar una cuota adicionales a los padres, aunque rara vez lo hacen. La financiación adicional de colegios públicos surge generalmente de los municipios, dependiendo, entonces, de la capacidad económica de cada municipio el poder asignar recursos extras al colegio. Para mayores detalles sobre el sistema educacional chilenos se puede ver Mizala y Romaguera (2000), Lara et al. (2009) y Manzi et al. (2008).

1.2. Bibliografía Previa

Una basta bibliografía ha surgido con el objetivo de analizar los factores que contribuyen a explicar las habilidades de los estudiantes. Antes de 1997 la información socioeconómica provista por SIMCE estaba disponible sólo a nivel de colegio, por lo que los primeros estudios emplearon al colegio y no al individuo como unidad de análisis, como por ejemplo Mizala y Romaguera (2000). Algunos estudios se focalizan en medir el impacto de políticas públicas, especialmente la efectividad del sistema de cupones. Hsieh y Urquiola (2006), por ejemplo, estudia los efectos de la introducción del sistema de cupones en el desempeño de los estudiantes y en la estratificación social.

Por otro lado, otro grupo de estudios empíricos evalúan el sistema de cupones en

términos de las diferencias en la calidad educativa impartida por los distintos tipos de colegio, sobre todo en lo que respecta a las diferencias entre colegios particulares subvencionado y públicos. Los trabajos en esta línea se encuentran con problemas propios de la característica observacional de los datos. En efecto, como no es posible observar el desempeño de un estudiante de un determinado tipo de colegio en caso de haber asistido a otro tipo de colegio; sumado a que los alumnos no está asignados en forma aleatoria a los distintos tipos de colegios[7], se genera un problema de identificación, conocido como problema de selección. En esta línea se pueden mencionar trabajos que proponen distintas estrategias para tratar el sesgo de selección como Lara et al. (2009), Mizala y Romaguera (2001), Sapelli y Vial (2002) y Sapelli y Vial (2005)[8] y Sapelli y Vial (2002).

El aumento de información estadística sobre aspectos relevantes pare determinar la habilidad de los estudiantes chilenos dio origen a nuevos trabajos como Manzi et al. (2008), que aprovecha datos del Sistema Nacional de Evaluación Docente (DocenteMás) para estudiar el efecto de variables relacionadas al colegio y a los docentes en el desempeño de los alumnos. Finalmente, resta mencionar que Chile cuenta con información proveniente de una serie de mediciones internacionales en las que ha participado (TIMSS, PISA y LLECE). Estos datos han sido utilizados en trabajos como Ramírez (2002), Manzi et al. (2008), PISA (2006) entre otros. En la siguiente sección se presentan algunos resultados generales obtenidos en la prueba PISA 2006 con el objetivo de ubicar la educación chilena dentro de Latinoamérica y el mundo.

[7]Como se mencionaba en la sección anterior, los colegios particulares subvencionados realizan selecciones de alumnos a través de tests y entrevistas al alumno y su familia, etc. Además los padres también seleccionan los colegios de acuerdo a criterios que no necesariamente están relacionados con la calidad de la educación.

[8]Para una discusión más detallada de la literatura empírica sobre este tema puede verse Lara et al. (2009)

1.3. Informe PISA 2006

PISA 2006 mide el desempeño de estudiantes en ciencia, matemática y lectura; junto con características socioeconómicas a nivel individual[9] para 57 países. Chile obtuvo la posición 40 en el ranking, un desempeño global bajo, pero el primer lugar entre los países Latinoamericanos que participaron: Uruguay, Mexico, Brasil, Argentina, Colombia y Chile.

El reporte sobre la prueba PISA se documenta en PISA (2006). En su capítulo 4 trata sobre la relación entre calidad y equidad de la educación. Se analizan comparativamente los países presentando información sobre la variabilidad del puntaje al interior y entre colegios, se compara el desempeño de los estudiantes según su condición de inmigrante o no, se analiza la relación entre desempeño y entorno socioeconómico individual y del colegio, entre otros temas. La Tabla 1.1 resume los estadísticos empleados en PISA y la interpretación que se les da a los mismos. Los valores de los estadísticos corresponden a la prueba de ciencia, pero los resultados son muy similares para los otros dos exámenes. El entorno socioeconómico de los individuos se mide a través de un *índice de estatus económico, social y cultural* (ESCS), que se construye a partir de variables vinculadas al estatus ocupacional y la escolaridad de los padres, y de un índice de posesión en el hogar de elementos relevantes para el estudio (computador, internet, libros, entre otras). El puntaje del índice obtenido por cada estudiante se determina a partir de un análisis de Componentes Principales y se utiliza una escala tal que la media del ESCS sea 0 y el error estándar 1. Las variables explicativas de interés presentadas en el capítulo 4 de PISA (2006) son $ESCS_{est}$, el índice socioeconómico a nivel individual; y $ESCS_{col}$, el índice socioeconómico a nivel colegio. La variable respuesta utilizada es Y, el puntaje en ciencia en la escala PISA

[9] A diferencia de la prueba SIMCE, que colecta información socioeconómica a partir del reporte de padres y apoderados, en PISA la dicha información se basa en el reporte de los alumnos.

2006.

Estadístico	Descripción	Total*	Prom. OECD	Latinoamérica**	Chile
$\bar{Y}$	Puntaje Promedio	475.36	500.00	407.50	438.00
s.e.		(3.00)	(0.50)	(3.67)	(4.30)
$\hat{Y}_{aju}$	Puntaje Promedio si todos los países tuvieran el mismo valor de ESCS igual al promedio de los países OECD	481.00	500.00	432.83	465.00
s.e.		(2.81)	(0.50)	(3.25)	(3.30)
R^2	Medida de la **fuerza** de la relación entre el puntaje y ESCS	13.70	14.40	17.73	23.30
s.e.		(1.51)	(0.26)	(1.78)	(1.92)
α	Cambio en el puntaje ante un cambio en una unidad en el índice ESCS (Coeficiente de Regresión)	36.00	40.00	31.33	38.00
s.e.		(2.15)	(0.40)	(1.73)	(1.80)
$q_{0.05}$	percentil de 5 % del índice ESCS	-1.71	-1.43	-2.75	-2.55
$q_{0.95}$	percentil de 95 % del índice ESCS	1.36	1.50	1.19	1.30
$q_{0.95}$ - $q_{0.05}$	Largo de la recta de regresión	3.07	2.93	3.94	3.85
SCT	Suma de Cuadrados Totales	8387.65	8971	8198.33	8446
$\frac{100SCT}{SCT_{OECD}}$	**Suma de Cuadrados Totales** expresada como porcentaje de la Suma de Cuadrados Totales promedio de la OECD	93.49	100.00	91.38	94.10
$\frac{100SCT_b}{SCT_{OECD}}$	**Variabilidad entre colegios** como porcentaje de la Suma de Cuadrados Totales promedio de la OECD	33.28	33.00	39.65	53.00
$\frac{100SCT_w}{SCT_{OECD}}$	**Variabilidad al interior de col.** como porcentaje de la Suma de Cuadrados Totales promedio de la OECD	61.14	68.10	51.68	52.20
$\frac{SCE(ESCS_{est})_b}{SCT_{OECD}}$	**Variabilidad entre col. explicada por el índice $ESCS_{est}$** como porcentaje de la Suma de Cuadrados Totales promedio de la OECD	6.90	7.20	9.68	14.20
$\frac{SCE(ESCS_{est})_w}{SCT_{OECD}}$	**Variabilidad al interior de col. explicada por el índice $ESCS_{est}$** como porcentaje de la SCT de la OECD	2.65	3.80	1.08	0.80

Tabla 1.1: Estadísticos Propuestos en el informe PISA para analizar la relación entre puntaje y entorno socioeconómico.

* : esta columna se construyó promediando los estadísticos de todos los países.
**: esta columna se construyó promediando los estadísticos de los países Latinoamericanos.

Se desprende de la Tabla 1.1 que, para el caso chileno, la relación entre el entorno socioeconómico y la variabilidad del puntaje (R^2) es alta. El 23.3 % de la variabilidad del puntaje se explica por factores socioeconómico, la segunda mayor relación después de Bulgaria. Además esta relación supera tanto al promedio general, como

los promedios para los países de la OECD y latinoamericanos. El estadístico α indica que un aumento en una unidad del índice $ESCS$ para un individuo se traduce en 38 puntos de la prueba PISA. Las filas $q_{0.05}$ y $q_{0.95}$ muestran los percentiles empíricos de la variable $ESCS$. Estos estadísticos informan sobre las disparidades socioeconómicas que se observan en cada país. Cuanto mayor es la diferencia entre ambas $(q_{0.95} - q_{0.05})$, mayor la disparidad. El caso chileno resalta también por sus altos niveles de disparidad, lo que sugiere que el elemento socioeconómico es clave para analizar el Sistema Educacional Chileno y, por ende, es un punto que hay que analizar en detalle. Finalmente, otra característica que salta a la vista del caso chileno es la gran variabilidad de puntaje entre colegios, y la baja variabilidad intra- colegio; dejando en evidencia la estructura segmentada de su sistema educativo. Este panorama se hace transparente gracias al análisis estadístico de tipo multinivel.

1.4. Modelos jerárquicos para el caso chileno

Existe en Chile un debate público y académico sobre cómo mejorar la calidad de la educación en todo los niveles socioeconómicos. Desde el punto de vista académico, la mayoría de los estudios utilizan modelos lineales. Manzi et al. (2008), por ejemplo, utilizan el enfoque multinivel para estudiar las características individuales y de colegio sobre los puntajes SIMCE. Mizala y Romaguera (2000), emplean una regresión lineal con los puntajes promedio por colegio como variable dependiente. Los modelos *jerárquicos o multinivel*(HLM), se caracterizan por modelar datos anidados - es decir, datos que aparecen conformando subgrupos o niveles. En pruebas educacionales, por ejemplo, los individuos se encuentran dentro de cursos, que a su vez pertenecen a colegios y que finalmente se ubican dentro de países. El supuesto de independencia de las variable respuesta, propio del modelo lineal simple (OLS) puede no ser realista en este tipo de datos. Los modelos HLM suponen correlación entre las observaciones

al interior de los grupos. Este vinculo intra-grupo se logra al incorporar efectos aleatorios a distinto nivel. Para exponer los modelos HLM más claramente, se presenta un modelo simple de dos niveles conocido como *Modelo de Componentes de Varianza* en las ecuaciones (1.1) y (1.2). El sub-índice i representa al individuo (nivel 1) y el j, al colegio (nivel 2). y_{ij} se emplea para denotar el puntaje del individuo i del colegio j, y u_{0j} son los efectos aleatorios, generalmente especificados como indica la ecuación (1.2). Finalmente, b es un vector con parámetros fijos desconocidos, y x_i es un vector con los valores de las covariables.

$$y_{ij}|(x_i, u_{0j}) \overset{i.}{\sim} N(b'x_i + u_{0j}, \sigma^2) \tag{1.1}$$

$$u_{0j}|x_i \overset{i.i.d.}{\sim} N(0, \tau), \tag{1.2}$$

El error de medición está dado por $\epsilon_{0ij} = y_{ij} - E(y_{ij}|x_i, u_{0j})$ que, por construcción es no correlacionado con $E(y_{ij}|x_i, u_{0j})$, provisto que $E(u_{0j}^2) < \infty$ y $E(y_{0j}^2) < \infty$. La distribución de y_{ij}, una vez que se marginaliza con respecto a u_{0j}, es también normal. Utilizando propiedades de esperanza condicional, se pueden obtener sus parámetros:

$$E(y_{ij}|x_i) = b'x_i, \text{ y } Var(y_{ij}|x_i) = \tau + \sigma^2. \tag{1.3}$$

Es decir que el modelo HLM, a diferencia del modelo OLS, descompone la variabilidad de los puntajes y_{ij} en componentes propios de cada nivel. Además, y lo que es más importante, la covarianza entre dos observaciones provenientes del mismo grupo está dada por $Cov(y_{ij}, y_{i'j}|x_i, x_{i'}) = \tau$; mientras que dos observaciones de distinto nivel son independientes, es decir $Cov(y_{ij}, y_{i'j'}|x_i) = 0$ si $j \neq j'$.

El modelo se puede complejizar incluyendo coeficientes aleatorios (adicionando un término $u_{1j}x_{ij}$ con u_{1j} aleatorio a la media en la ecuación (1.1)), agregando más niveles, etc[10]. A su vez, las variables explicativas se pueden incorporar afectando al

[10]Ver Goldstein (1995) para mayores detalles sobre modelos de multinivel.

puntaje a distinto nivel. Como ejemplo notemos que la variable $ESCS$ en los datos PISA influye a nivel individual $ESCS_{est}$ y a nivel colegio $ESCS_{col}$. Al presentarse a distinto nivel, la variable $ESCS$ se interpreta distinto: en el primer caso representa el nivel socioeconómico del individuo, pero en el segundo se trata del efecto sobre el puntaje del individuo que tienen sus *pares o compañeros de estudio*.

En lo que se refiere a PISA 2006, los efectos de las características de los colegios y del sistema educativo sobre el desempeño de los estudiantes se analiza a través de un modelo jerárquico de 3 niveles, donde el primer nivel corresponde al individuo, el segundo al colegio y el tercero al país (ver PISA, 2006, capítulo 5). De esta manera, se tratan comparativamente temas como la política de admisión de los colegios, las diferencias entre colegios públicos y privados, las formas de financiamiento de los establecimientos educativos, el rol de los padres tanto en la elección del colegio como en su participación en el colegio, entre otros temas.

1.5. Esquema del documento

Hasta aquí hemos presentado someramente los trabajos aplicados para el caso chileno, desarrollados con anterioridad a la investigación que se reporta en este documento. Este estudio tomó un enfoque diferente para analizar los datos, pero antes de pasar al modelo elegido, en el próximo capítulo se presentan las bases teóricas sobre las que se sustenta nuestro modelo. En el capítulo 2 se introducen nociones de Estadística Bayesiana No Paramétrica, más específicamente los Procesos Dirchlet (DP), mezclas de procesos Dirichlet (MDP), y Procesos Dirichlet Dependientes (DDP). Se presentan sus propiedades, y las técnicas de simulación y de estimación a posteriori disponibles para mezclas de procesos Dirichlet. Además se expone una

breve explicación acerca de los modelos basados en la Teoría de Respuesta al Item (Modelos IRT). Al final del capítulo se explica el concepto de identificación - tanto desde el enfoque Bayesiano como clásico. Generalmente se presta poca atención a los problemas de identificación. Sin embargo, sólo es posible interpretar los parámetros de interés de un modelo en el caso en que estén identificados. En otras palabras: es una condición necesaria para que el modelo pueda ser usado en la práctica. Es por ello que incluimos un estudio de identificación de nuestro modelo[11].

En el capítulo 3 se expone el modelo propuesto. Se presentan los parámetros de interés y el estudio de identificación correspondiente. También se explica la estrategia utilizada para hacer inferencia. Al final del capítulo se muestra un estudio de simulación para el modelo propuesto. El capítulo 4 presenta la aplicación al caso chileno. Se explican en detalle la características de los datos disponibles, los detalles sobre el procedimiento MCMC empleado y los resultados obtenidos. Finalmente, el capítulo 5 expone las conclusiones.

[11]En muchas oportunidades, se pueden presentar modelos con parámetros no identificados que no son de interés. Esto no representa un problema en la medida en que no se pretenda hacer inferencia a partir de ellos.

Capítulo 2

Ingredientes de esta Investigación: Estadística Bayesiana No Paramétrica y modelos IRT

La estructura del modelo IRT Semiparamétrico Explicativo presentado en el capítulo 3 se nutre de dos vertientes teóricas: la Estadística Bayesiana No Paramétrica (NPB), y los modelos basados en la Teoría de Respuesta al Ítem (IRT). En este capítulo se presentan las nociones básicas de ambas ramas teóricas con la finalidad de hacer de este documento, un trabajo autocontenido y accesible a lectores no expertos. En lo que respecta a la estadística NPB, se definen los Procesos Dirichlet, (Ferguson, 1973), la herramienta básica para definir una distribución en el espacio de distribuciones discretas, y se exponen sus propiedades. Se extiende la exposición a modelos de mezclas de Procesos Dirichlet (MDP), (Lo, 1984), que permiten establecer una distribución sobre el espacio de distribuciones continuas; y Procesos Dirichlet Dependientes (DDP), (MacEachern, 1999), que asignan una distribución a una colección de distribuciones aleatorias indexadas por valores de covariables. El modelo propuesto en el capítulo 3 emplea una especificación denominada mezclas de ANOVA DDP, (De

Iorio et al., 2004), la cual reúne las condiciones de ser un caso particular de DDP y un modelo de mezcla al mismo tiempo. Se dedica especial atención a esta especificación. También se exponen los algoritmos disponibles para hacer estimación a posteriori de MDP. Finalmente, se presenta un apartado introduciendo modelos IRT desde la perspectiva Bayesiana, y una última sección con conceptos empleados para la identificación de los parámetros de interés del modelo propuesto.

2.1. Estadística Bayesiana No Paramétrica

Una particularidad de la metodología Bayesiana es que el parámetro de interés - aquí notado como ϑ - se modela como un elemento aleatorio. La naturaleza aleatoria de ϑ representa la incertidumbre a priori acerca del parámetro, que es caracterizado en un modelo Bayesiano, a través de *la distribución a priori para* ϑ. Cuando se especifica una distribución a priori particular para ϑ, se incorpora al modelo un cierto grado de subjetividad. Primero, se adiciona una creencia a priori sobre la familia de distribuciones a la que pertenece (Normal, t, Cauchy, Uniforme, etc.). Segundo, una vez que la familia es escogida, un elemento particular de la familia debe ser seleccionado. Como ejemplo consideremos que se elige la distribución normal $N(\mu, \sigma^2)$. Los valores de μ y de σ^2 deben ser también determinados[1].

En el ejemplo previo, μ representa la adivinanza a priori, mientras que σ^2 puede ser usado para medir el grado de incertidumbre (un valor alto de σ^2 representa mayor incertidumbre). La elección de μ y σ^2 le da cierta flexibilidad al modelo al permitir al investigador elegir la localización y escala de la distribución de ϑ. Incluso se puede agrandar el modelo incorporando distribuciones para μ y σ^2, dando aún más flexibilidad al modelo. Sin embargo, este cambio puede seguir siendo muy rígido para algunas

[1]En lo que sigue los elementos *no aleatorios* (en el ejemplo μ y σ^2) se los denominará **hiperparámetros**.

aplicaciones: deja previamente establecido que la distribución pertenece a la familia normal - con su forma específica - es decir unimodal, simétrica, etc-. Esto puede no ser apropiado en algunos casos. La estadística *Bayesiana no Paramétrica* (NPB) es la herramienta apropiada cuando no hay una creencia a priori acerca de la familia de distribuciones de donde proviene la distribución a priori.

2.1.1. De la estadística Bayesiana Paramétrica a la no Paramétrica: la distribución Dirichlet

Llamemos Θ al espacio al cual pertenece ϑ, y supongamos que estamos interesados en la distribución P_ϑ asociada a ϑ. La estadística NPB considera a P_ϑ como el *parámetro de interés*; entonces el *espacio paramétrico* pasa a ser el conjunto de todas las posibles distribuciones de probabilidad sobre Θ, que será denotado por $M(\Theta)$. Para ilustrar la inferencia NPB con un caso simple, consideremos $\Theta = \{1, 2\}$. Se tiene que:

$$M(\Theta) = \{\mathbf{p} = (p_1, p_2) : p_1 \geq 0, p_2 \geq 0, p_1 + p_2 = 1\} = \{\mathbf{p} = (p_1, 1 - p_1) : 0 \leq p_1 \leq 1\}.$$

Cualquier valor en el intervalo $[0, 1]$ define una distribución a priori en $M(\Theta)$. La simplicidad de Θ permite indexar a cada distribución sobre Θ con un parámetro unidimensional p_1. En particular, si p_1 distribuye Beta(α_1,α_2), la inferencia Bayesiana consiste en obtener las distribuciones a posteriori y/o predictiva . Si $\vartheta_1, ..., \vartheta_n$ es una muestra de $\mathbf{p}$, entonces la distribución a posteriori para p_1 está dada por:

$$p(p_1|\vartheta_1, ..., \vartheta_n) \propto p_1^{\sum_{i=1}^n \mathbb{1}\delta_{\{\vartheta_i=1\}}+\alpha_1-1}(1-p_1)^{\sum_{i=1}^n \delta_{\{\vartheta_i=2\}}+\alpha_2-1},$$

donde δ representa la medida de Dirac. La distribución $p(p_1|\vartheta_1, ..., \vartheta_n)$ corresponde al kernel de una Beta($\sum_{i=1}^n \delta_{\{\vartheta_i=1\}}+\alpha_1$, $\sum_{i=1}^n \delta_{\{\vartheta_i=2\}}+\alpha_2$). Por su parte, la distribución posteriori predictiva, es decir, la distribución de un nuevo elemento de la muestra ϑ_{n+1}

condicional en la muestra, $\vartheta_1, ..., \vartheta_n$, está dada por:

$$p(\vartheta_{n+1} = 1|\vartheta_1, ..., \vartheta_n) = \frac{\sum_{i=1}^{n} \delta_{\{\vartheta_i=1\}} + \alpha_1}{\alpha_1 + \alpha_2 + n},$$

resultado que se desprende del esquema de urna de Polya de la secuencia $\vartheta_1, \vartheta_2, ...$

El ejemplo previo se puede generalizar modificando Θ por un espacio más complejo. Consideremos el caso en que $\Theta = \{1, 2, ..., k\}$. Aquí

$$M(\Theta) = \{\mathbf{p} = (p_1, p_2, ..., p_{k-1}) : p_i \geq 0 \ \forall i = 1, 2, ...k-1, \sum_{i=1}^{k-1} p_i \leq 1\}$$

$p_k = 1 - \sum_{i=1}^{k-1} p_i$. La distribución a priori se especifica usando la distribución Dirichlet $D(\alpha_1, ..., \alpha_k)$, que es la generalización de la distribución Beta al caso multivariado. Aquí la distribución a posteriori es también Dirichlet con parámetros $D(\sum_{i=1}^{n} \delta_{\{\vartheta_i=1\}} + \alpha_1, ..., \sum_{i=1}^{n} \delta_{\{\vartheta_i=k\}} + \alpha_k)$ y la distribución a posteriori predictiva también de describe de acuerdo a un esquema de urna de Polya como:

$$p(\vartheta_{n+1} = j|\vartheta_1, ..., \vartheta_n) = \frac{\alpha_j + \sum_{i=1}^{n} \delta_{\{\vartheta_i=j\}}}{\sum_{i=1}^{k} \alpha_i + n}, \ j = 1, ..., k.$$

El caso donde $\Theta = \{1, 2, ...\} = \mathbb{N}$ requiere el uso de *procesos estocásticos* pues

$$M(\Theta) = \{\mathbf{p} = (p_1, p_2, ...) : p_i \geq 0 \ \forall i = 1, 2, ..., \sum_{i=1}^{\infty} p_i = 1\},$$

y por ende una distribución para el proceso p debe especificarse. El proceso Dirichlet (DP) es la extensión natural de los ejemplo de arrriba. El DP como priori se propuso por primera vez por Ferguson (1973) y sus propiedades han sido ampliamente estudiadas desde entonces. Sethuraman (1994), Blackwell y MacQueen (1973), Antoniak (1974), Korwar y Hollander (1973), Walker et al. (1999), Rolin (1992) son algunos artículos relevantes. Hoy en día, el estado del arte en el área NPB deja disponible un gran número de distribuciones a priori diferentes para procesos estocásticos como

árboles de Polya, Polinomios de Bernstein, etc. Este trabajo se concentra en DP; el lector interesado en mayores detalles sobre otros procesos puede ver Gosch y Ramamoorthi (2003) y Müller y Quintana (2004b).

2.1.2. Los procesos Dirichlet y sus propiedades

La Definición 2.1 presenta la definición de Ferguson de un proceso Dirichlet. Nótese que se encuentra enunciado en una forma general donde Θ puede ser cualquier conjunto, como $\mathbb{N}$ o $\mathbb{R}$.

Definición 2.1. *Sea Θ un conjunto y $\mathcal{A}$ una $\sigma-$álgebra asociada con Θ. Sea α una medida no nula, finita, no negativa y sigma aditiva sobre (Θ,$\mathcal{A}$). Decimos que P_ϑ es un* Proceso Dirichlet *(DP) sobre (Θ,$\mathcal{A}$) con parámetro α si para cada partición finita $A_1, ..., A_n$ de Θ, la distribución conjunta de la probabilidad aleatoria $(P_\vartheta(A_1), ...P_\vartheta(A_n))$ es Dirichlet con parámetro $(\alpha(A_1), ..., \alpha(A_n))$. Se denota como $P_\vartheta|\alpha \sim DP(\alpha)$*

Ferguson (1973) probó la existencia y unicidad de esa medida de probabilidad. Además, muchas propiedades han sido estudiadas. A continuación se presenta un resumen de las propiedades que serán útiles en este trabajo. En lo que sigue, $M = \alpha(\Theta)$; $\bar{\alpha}$ representa α/M, la distribución normalizada a partir de α; y se deja δ para referirse a la medida de Dirac.

Propiedad 2.1. *Si P_ϑ es $DP(\alpha)$ y $A \in \mathcal{A}$, entonces $E(P_\vartheta(A)) = \bar{\alpha}(A)$.*

Propiedad 2.2. *Si P_ϑ es $DP(\alpha)$ y $A \in \mathcal{A}$, entonces $V(P_\vartheta(A)) = \frac{\bar{\alpha}(A)(1-\bar{\alpha}(A))}{(M+1)}$.*

Propiedad 2.3. *Si P_ϑ es $DP(\alpha)$ y $\vartheta_1, ..., \vartheta_n$ son i.i.d. P_ϑ, la distribución a posteriori es también un DP, es decir $P_\vartheta|\vartheta_1, ..., \vartheta_n$ es $DP(\alpha + \sum_{i=1}^{n} \delta_{\vartheta_i})$.*

Propiedad 2.4. *Si P_ϑ es $DP(\alpha)$, entonces P_ϑ es casi seguramente una distribución discreta.*

Propiedad 2.5. *Si P_ϑ es $DP(\alpha)$ y $\vartheta_1, ..., \vartheta_n$ son i.i.d. P_ϑ, entonces la distribución de $\vartheta_1, ..., \vartheta_n$ sigue un esquema de urna de de Polya con parámetro α, es decir:*

$P_\vartheta(\vartheta_l \in A|\vartheta_1, ..., \vartheta_{l-1}) = \frac{\alpha(A)+\sum_{i=1}^{l-1}\delta_{\vartheta_i}(A)}{M+(l-1)}$, *para* $l = 1, ..., n$.

Propiedad 2.6. *Sethuraman (1994) propone una definición constructiva de un DP como sigue. Sea $(x_1, x_2, ...)$ i.i.d Beta$(1, M)$. Sea $(y_1, y_2, ...)$ i.i.d. $\bar{\alpha}$ e independiente de $(x_1, x_2, ...)$. Sean $w_1 = x_1$ y $w_n = x_n \prod_{1\leq i\leq n-1}(1 - x_i)$ para $n = 2, 3, ...$ Entonces $P_\vartheta = \sum_{i=1}^{\infty} w_i \delta_{y_i}$ es $DP(\alpha)$.*

Propiedad 2.7. *Sea $\alpha_n = M_n\bar{\alpha}$ una secuencia de medidas en Θ. Entonces $DP(\alpha_n)$ es una secuencia de medidas de probabilidad sobre el espacio $M(\Theta)$.*

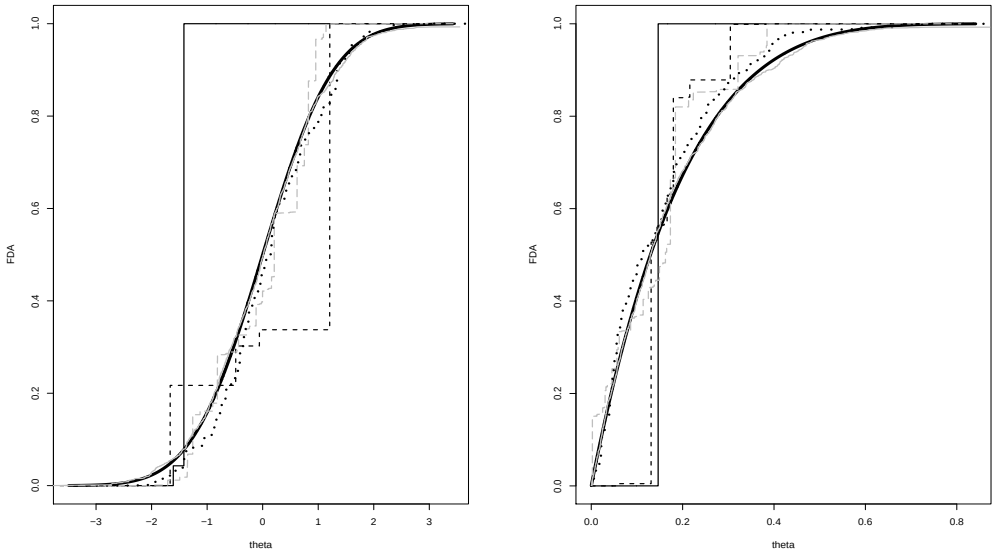
1. *Si $M_n \to 0$, entonces $DP(\alpha_n) \to \mu$, donde el soporte de μ son las medidas degeneradas en puntos de Θ, es decir $\mu(\{P_\vartheta \equiv \delta_{\vartheta'}, \vartheta' \in \Theta\}) = 1$, y si $A \subset \Theta$, $\mu(\{\delta_{\vartheta'} : \vartheta' \in A\}) = \bar{\alpha}(A)$. Este resultado se debe a Sethuraman y Tiwari (1982)*

2. *Si $M_n \to \infty$, entonces $DP(\alpha_n) \to \mu$, donde μ es la medida degenerada en $\bar{\alpha}$.*

La Propiedad 2.1 señala que $\bar{\alpha}$ es la media del proceso. Por otra parte, a través de la Propiedad 2.2 se puede notar que M tiene una relación inversa con la varianza del proceso (la varianza crece cuando M decrece). Sin embargo, hay que tener cuidado en la interpretación de M, pues cuando $M \to 0$ no corresponde con un caso de alta variabilidad (no informativo). Esto se evidencia observando la Propiedad 2.7(1), que indica que en tal caso P_ϑ es casi seguramente una distribución concentrada en un punto. Generalmente, el parámetro del proceso Dirichlet, α, se expresa en dos partes, $(M\bar{\alpha})$. De esta manera la notación $DP(\alpha)$ se modifica por $DP(M\bar{\alpha})$. $\bar{\alpha}$ se la denomina *medida basal*, mientras que a M se lo llama *parámetro de masa total*.

Para clarificar las ideas, la Figura 2.1.2 presenta simulaciones de Procesos Dirichlet. El gráfico 2.1.2(a) muestra simulaciones de un DP con medida basal, $\bar{\alpha}$, normal

estándar, mientras que el gráfico 2.1.2(b) hace lo propio para un DP con medida basal Beta. En los dos casos, a medida que M crece, las simulaciones se van acercando cada vez más a la medida basal. Sin embargo, para valores pequeños de M, la masa de probabilidad se concentra en pocos puntos, lo que no coincide con un caso no informativo.



(a) Medida basal Normal Estándar (b) Medida basal Beta(1,5)

Figura 2.1: Para ambos gráficos, la línea continua gruesa corresponde a la medida basal del proceso, mientras que las restantes líneas finas son simulaciones del proceso para distintos valores de M. El gráfico (a) corresponde a un PD con medida basal Normal Estándar ($PD(M, N(0,1))$), mientras que el gráfico (b) presenta resultados a partir de un PD con medida basal Beta ($PD(M, Beta(1,5))$). Para ambos gráficos la curva continua negra es una simulación del proceso tomando $M = 0.1$, la cortada negra presenta una simulación para el caso de $M = 1$. La línea cortada gris hace lo propio para $M = 10$, la punteada negra corresponde a $M = 100$ y finalmente la línea continua gris presenta el caso de $M = 1000$. Se puede apreciar que M es el parámetro de precisión en los Procesos Dirichlet ya que cuánto mayor es el valor de M las simulaciones se acercan más a la medida basal.

2.1.3. Modelos de Mezcla de Procesos Dirichlet

En muchas aplicaciones ϑ se considera una variable continua. En tales casos, la especificación de P distribuyendo DP no es apropiada, ya que los DP asignan probabilidad uno al conjunto de distribuciones discretas (Propiedad 2.4). En esas situaciones

pueden definirse distribuciones aleatorias continuas considerando un Kernel continuo y convolucionando dicho kernel mediante una distribución aleatoria P, que a su vez distribuye DP. Si el kernel de la mezcla es continuo, entonces la mezcla resultante preserva la continuidad. Estos modelos los denominaremos en este trabajo Mezcla de Procesos Dirichlet (MDP) y fueron estudiados por Lo (1984). Cabe aclarar que no se trata de los mismos modelos de mezcla planteados por Antoniak (1974), quien considera como kernel $DP(\alpha_u)$ y los mezcla de acuerdo a una distribución mezclante $H(u)$. Para ser más específico, el modelo MDP se presenta en las ecuaciones de (2.1).

$$\theta_i|G \overset{i.i.d.}{\sim} G, \qquad G(\theta_i) = \int p(\theta_i|\vartheta)P(d\vartheta), \qquad P|M,\bar{\alpha} \sim DP(M\bar{\alpha}). \tag{2.1}$$

Aquí $p(\theta_i|\vartheta)$ es un Kernel, que puede especificarse como una distribución continua, normal por ejemplo, induciendo una distribución a priori para G cuyo soporte son distribuciones continuas. Este modelo también suele expresarse rompiendo la mezcla como muestran las ecuaciones en (2.2).

$$\theta_i|\vartheta_i \overset{i.}{\sim} p(.|\vartheta_i), \qquad \vartheta_i|P \overset{i.i.d.}{\sim} P, \qquad P|M,\bar{\alpha} \sim DP(M\bar{\alpha}), \qquad i = 1,...,n \tag{2.2}$$

Para expresar el modelo MDP en la forma (2.2) se introducen variables aleatorias latentes, ϑ_i. Condicional en ϑ_i , θ_i distribuye $p(.|\vartheta_i)$. La variable ϑ_i no necesariamente es de interés, pero se introduce porque en algunas situaciones es más sencillo expresar el modelo de esta forma. Por ejemplo, esta especificación es útil a la hora de estimar el modelo, como se verá en la sección 2.1.6.

$$\theta_i \perp\!\!\!\perp (P,M,\bar{\alpha})|\vartheta_i, \qquad \vartheta_i \perp\!\!\!\perp (M,\bar{\alpha})|P, \qquad P|M,\bar{\alpha} \sim DP(M\bar{\alpha}), \qquad i = 1,...,n \tag{2.2bis}$$

Las ecuaciones en (2.2 bis) muestran una forma alternativa de expresar (2.2) a través de relaciones de independencia condicional. Aquí también se evidencia la estructura jerárquica del modelo.

2.1.4. Procesos Dirichlet Dependientes

Los procesos Dirichlet Dependientes (DDP) se utilizan para modelar situaciones en las que, si bien no se conoce la distribución de un parámetro, se sabe que esta distribución *depende* de ciertas covariables conocidas. Desde el punto de vista matemático, esto implica generar una colección de distribuciones aleatorias indexadas por covariables denotadas por d, $\{P_d, d \in D\}$. Se han planteado varios caminos para lograr este fin. Uno de ellos se presenta en MacEachern (1999). El autor propone emplear la representación de Sethuraman (Propiedad 2.6). En esta representación, la distribución aleatoria se expresa como una mezcla numerable de distribuciones degeneradas: $P = \sum_{i=1}^{\infty} w_i \delta_{\phi_i}$. Los componentes aleatorios en esta representación son w_i y ϕ_i, $i = 1, ..., \infty$. Los DDP incorporan dependencia haciendo depender ϕ_i y/o w_i de las covariables d. El caso en que la secuencia $\{w_i\}_{i \in \mathbb{N}}$ no depende de las covariables se lo denomina modelo DDP con w único[2]. El modelo DDP con w único logra la dependencia con respecto a las covariables reemplazando cada ϕ_i por un proceso estocástico en D, denotado ϕ_{iD}. Condicional en d, la distribución de P_d es un DP. La representación de Sethuraman en el caso del modelo DDP con w único queda expresada como: $P_d = \sum_{i=1}^{\infty} w_i \delta_{\phi_{id}}$. Dado que la secuencia de w's es común, el parámetro de masa total, M, es también el mismo para todo d. Por otra parte, la medida basal puede, o no, especificarse dependiendo de d. El caso más simple se da cuando es la misma para todos los valores de la covariable.

Aplicaciones del model DDP se pueden encontrar en De Iorio et al. (2004) que propone un modelo llamado mezclas de ANOVA DDP; y en Gelfand et al. (2005) en una aplicación con datos espaciales. Recientemente ha surgido una basta bibliografía sobre el tema con estrategias alternativas para generar dependencia como Müller y Quintana (2004a) con un modelo de mezclas de DP jerárquico; Griffin y Steel (2006) y

[2]Traducción libre de la autora del nombre propuesto en MacEachern (1999) *single-w DDP model*.

Dunson y Park (2008), que utilizan procesos stick-breaking, Dunson (2007) y Dunson et al. (2007) entre otros.

En este trabajo nos concentraremos en mezclas de ANOVA DDP propuesto en De Iorio et al. (2004). Brevemente, el modelo ANOVA DDP es un caso simple de modelo DDP con w único y medida basal, α, igual para todos los valores de d. Su particularidad es que el proceso $\phi_{i\mathbb{D}}$ se define mediante: $\phi_{id} = d\phi_i^*$ con $\phi_i^* \overset{i.i.d.}{\sim} \bar{\alpha}$. Si el conjunto de distribuciones indexadas por d, $\{P_d, d \in D\}$ distribuye ANOVA DDP, se escribe $\{P_d, d \in D\} \sim$ ANOVA DDP$(M\bar{\alpha})$. La simplicidad de este modelo radica en que si $\vartheta'|P_d \sim P_d$, con $\{P_d : d \in D\} \sim$ ANOVA DDP$(M\bar{\alpha})$, entonces $\vartheta' = d\vartheta$, donde $\vartheta|P \sim P$ y $P \sim DP(M\bar{\alpha})$. Las mezclas de modelos ANOVA DDP es un modelo que mezcla distribuciones normales cuyas localizaciones, ϑ', condicionadas en P_d, distribuyen P_d; y P_d es una distribución desconocida tal que $\{P_d, d \in D\} \sim$ ANOVA DDP$(M\bar{\alpha})$. El modelo se escribe como:

$$\theta_i|(d_i = d, F_d) \sim F_d, \text{ con } F_d(\theta) = \int N(\theta; \vartheta', \Sigma) P_d(d\vartheta')$$

$$\{P_d : d \in D\} \sim \text{ ANOVA DDP}(M\bar{\alpha})$$

Pero también puede expresarse utilizando DP como se escribe a continuación:

$$\theta_i|(d_i = d, F_d) \sim F_d, \text{ con } F_d(\theta) = \int N(\theta; \vartheta d_i, \Sigma) P(d\vartheta)$$

$$P|(M, \bar{\alpha}) \sim DP(M\bar{\alpha})$$

Los autores consideran ejemplos donde las covariables son factores o tratamientos. En este caso d_i es un vector con la fila de la matriz de diseño correspondiente al individuo i. Sin embargo, el modelo permite trabajar con covariables continuas sin ninguna dificultad. Por último cabe mencionar que el modelo de mezcla de ANOVA DDP puede ser visto como una colección de modelos MDP indexados por los valores de d.

2.1.5. Técnicas de Simulación

Aunque la teoría vinculada a Procesos Dirichlet se empezó a desarrollar desde 1973, fue recién durante la década del 90 que empezó a usarse en casos de aplicación. Los avances en la velocidad de procesamiento de computadores, junto con el desarrollo de los métodos Monte Carlo Markov Chain (MCMC), hicieron posible su implementación práctica. En lo que sigue consideraremos variables aleatorias ϑ_i tal que $\vartheta_i|P \overset{i.i.d.}{\sim} P$ para $i = 1, ..., n$; y $P \sim DP(M\bar{\alpha})$. Blackwell y MacQueen (1973) han dado la clave sobre la forma de simular observaciones provenientes de la distribución marginal de ϑ_i. La estrategia está basada en el método de composición y aprovecha la representación de urna de Polya (Propiedad 2.5). Los pasos a seguir para simular de un $DP(M\bar{\alpha})$ se detallan a continuación:

- Simular ϑ_1 a partir de la medida basal, $\bar{\alpha}$.
- Para j=2,...,n, secuencialmente simular de $\vartheta_j|\vartheta_1, ..., \vartheta_{j-1}$ con la siguiente ley de probabilidad. $\vartheta_j = \vartheta_i$ con $i = 1, ..., j-1$ con probabilidad $\frac{1}{M+n}$, y ϑ_j proviene de $\bar{\alpha}$ con probabilidad $\frac{M}{M+n}$. Nótese que en cada paso j, los valores anteriormente simulados: $\vartheta_1, ..., \vartheta_{j-1}$ tienen una probabilidad positiva de volver a ser simulados.

Como resultados se obtienen simulaciones provenientes de la distribución marginal de ϑ, es decir, una vez que se marginaliza con respecto de P. Por otro lado, la representación constructiva de Sethuraman (1994), Propiedad 2.6, permite obtener una aproximación de P_ϑ. Los pasos para simular son:

- Simular de $x_1, ..., x_n$ i.i.d. $Beta(1, M)$.
- A partir de estos valores calcular $w_1, ..., w_n$ mediante $w_1 = x_1$ y para $j = 2, ..., n$ $w_j = x_j \prod_{1 \leq i \leq j-1}(1 - x_i)$. Para elegir n se puede usar el hecho de que

$E(\sum_{j=1}^{n} w_j) = 1 - (M/(M+1))^n$. Se toma ϵ suficientemente pequeño, por ejemplo $\epsilon = 0.0001$, y se escoge n tal que $(M/(M+1))^n = \epsilon$. Luego w_n se reemplaza por $1 - \sum_{j=1}^{n-1} w_j$.

- Simular de $y_1, ..., y_n$ i.i.d. $\bar{\alpha}$
- Luego se construye la aproximación $\hat{P}_\vartheta$ de P_ϑ como $\hat{P}_\vartheta = \sum_{i=1}^{n} w_i \delta_{y_i}$

En lo que se refiere a la estimación de distribuciones a posteriori de los parámetros de interés en un modelo que incorpora DP, actualmente existe una batería de algoritmos. Típicamente en estadística NPB, los DP aparecen flexibilizando la forma funcional de una distribución a priori. Cuando se trata de modelos jerárquicos - como es el caso del modelo estudiado en esta tesis - , el muestreo de Gibbs suele ser una buena alternativa. Este método fue introducido al campo de la estadística por Gelfand y Smith (1990). Requiere poder simular alternadamente de las condicionales completas de los parámetros del modelo. Estas distribuciones suelen conocerse en modelos jerárquicos por la forma en que éstos se enuncian. A partir de las distribuciones condicionales se construye una cadena de Markov cuya distribución estacionaria es la distribución conjunta a posteriori de los parámetros involucrados en el muestreo. Si la cadena es irreducible, aperiódica y recurrente positiva, entonces converge a la distribución conjunta a posteriori (ver Tierney, 1994). Para más detalles sobre el muestreo de Gibbs, puede verse también Gelman et al. (1995a).

Generalmente, la inferencia Bayesiana se lleva a cabo resumiendo las distribuciones a posteriori mediante medias, desviaciones estándar, intervalos de confianza, etc. Cuando una cadena de Markov es recurrente positiva y aperiódica, se dice que la cadena es ergódica; e implica que su media empírica converge casi seguramente a la esperanza matemática (Teorema Ergódico). Este teorema es muy importante en la

práctica de muestreo de Gibbs, y MCMC en general, pues asegura una buena estimación de la esperanza a posteriori a través de la media empírica, una vez que la cadena converge a la distribución estacionaria. A diferencia de la estimación de medias, la estimación del error de estimación de dichas medias suele verse afectada por la presencia de autocorrelación. La implementación del muestreo de Gibbs requiere, entonces, utilizar criterios para determinar el período de quema, es decir la iteración (N_{burn}) a partir de la cual la cadena alcanzó la convergencia a la distribución estacionaria; y la cantidad (N_{thin}) de iteraciones que se deben descartar entre una iteración y otra para eliminar la autocorrelación. Existe una serie de tests de convergencia habitualmente usados como el test de Gelman y Rubin (1992), el de Geweke (1992) y el test de Hedelberger y Welch (1983). Una buena revisión de diagnósticos de convergencia MCMC se puede ver en Cowles y Carlin (1996). Para implementarlos se dispone del paquete BOA, Smith (2004). Es conveniente complementar el diagnóstico mediante tests, con controles de tipo gráficos, como chequear que las cadenas de todos los parámetros simulados no tengan tendencia, chequear que las muestras extraídas sean independientes mediate funciones de autocorrelación de cada parámetro, y observar que las medias móviles converjan.

2.1.6. Algoritmos para modelos de mezcla de Procesos Dirichlet

La presente sección ofrece un panorama sobre los algoritmos alternativos, existentes para estimar el modelo propuesto. Éste, contiene una mezcla de ANOVA DDP para especificar las distribuciones de habilidad, la cual, puede ser entendida como una colección de MDP (ve secciones 2.1.3 y 2.1.4). Se expone en detalle la parte de los algoritmos que resuelven la estimación de MDP. En su mayoría se tratan de muestreos de Gibbs. El modelo MDP expresado en forma jerárquica, se especifica en la

ecuación (2.2). Sea $\vartheta = (\vartheta_1, ..., \vartheta_n)$ y $\theta = (\theta_1, ..., \theta_n)$. La implementación del muestreo de Gibbs se lleva a cabo marginalizando con respecto a P, es decir que en lugar de obtener simulaciones de $(\vartheta, P)|\theta$, se excluye P y se simula de $\vartheta|\theta$. Esto simplifica el muestreo de Gibbs, eludiendo la difícil tarea de simular de P. Cabe aclarar que si bien el algoritmo no arroja estimaciones de P, sino de la variable ϑ una vez que se marginaliza con respecto a P; es posible obtener simulaciones de P y de funcionales de P, después del ajuste del modelo. Para ello se puede utilizar la metodología propuesta por Gelfand y Kottas (2002), que aproxima simulaciones de la distribución a posteriori de P mediante la representación de Sethuraman (1994), como se menciona en la sección anterior.

Denotando $\vartheta_{-i} = (\vartheta_1, ..., \vartheta_{i-1}, \vartheta_{i+1}, ..., \vartheta_n)$, las distribuciones condicionales completas, $p(\vartheta_i|\vartheta_{-i}, \theta)$, son fáciles de obtener cuando se trabaja con DP, gracias a la estructura de urna de Polya de la distribución marginal de ϑ (Propiedad 2.5). En efecto, $p(\vartheta_i|\vartheta_{-i}, \theta) \propto p(\theta_i|\vartheta_i)p(\vartheta_i|\vartheta_{-i})$. El segundo término del lado derecho de la relación proporcional sigue un esquema de urna de Polya, es decir que $p(\vartheta_i|\vartheta_{-i}) \propto M\bar{\alpha}(\vartheta_i) + \sum_{j\neq i} \delta_{\vartheta_j}(\vartheta_i)$. Por su parte, el primer término toma dos formas posibles: $p(\theta_i|\vartheta_i = \vartheta_j) = p(\theta_i|\vartheta_j)$ para todo $j \neq i$, y $p(\theta_i|\vartheta_i \neq \vartheta_j, j \neq i) \propto p(\theta_i|\vartheta_i) \int p(\theta_i|\vartheta)\bar{\alpha}(d\vartheta)$. La distribución condicional completa una vez que se integra con respecto a P se expresa en la ecuación (2.3).

$$p(\vartheta_i|\vartheta_{-i}, \theta) \propto q_0 G_i(\vartheta_i) + \sum_{j=1, j\neq i}^{n} q_j \delta_{\vartheta_j}(\vartheta_i) \tag{2.3}$$

En la ecuación (2.3), $q_0 = M \int p(\theta_i|\vartheta)\bar{\alpha}(d\vartheta)$, $q_j = p(\theta_i|\vartheta_j)$ y $G_i(\vartheta_i) \propto p(\theta_i|\vartheta_i)\bar{\alpha}(\vartheta_i)$. En caso de que $p(\theta_i|\vartheta)$ y $\bar{\alpha}(\vartheta)$ sean conjugados se dice que el modelo MDP es conjugado. Esto simplifica considerablemente el muestreo gracias a que q_0 puede ser obtenido analíticamente. En caso contrario se trata de un modelo no conjugado.

Los primeros algoritmos resolvieron el problema de estimación para modelos conjugados. Se puede mencionar la propuesta de Escobar y West (1995). Basándose en trabajos previos de West (1990) y Escobar (1994), los autores desarrollaron un método computacional que permite evaluar distribuciones a posteriori y predictivas para un caso particular de (2.2). Se trata de un modelo de mezcla de distribuciones normales donde $\vartheta_i = (\mu_i, \sigma_i^2)$, $p(\theta_i|\vartheta_i) \equiv N(\mu_i, \sigma_i^2)$, P es una distribución a priori sobre $\mathbb{R} \times \mathbb{R}^+$, $\bar{\alpha}$ es la forma conjugada Normal/Gama invertida, es decir: $(\sigma_i^2)^{-1} \sim Gama(\frac{s}{2}, \frac{S}{2})$ y $\mu_i|\sigma_i^2 \sim N(\mu_0, \tau\sigma_i^2)$, con s, S conocidos, $\mu_0 \sim N(\mu_a, \sigma_a^2)$ y $\tau^{-1} \sim Gama(\frac{w}{2}, \frac{W}{2})$. Proponen un algoritmo basado en el siguiente esquema de muestreo de Gibbs:

Algoritmo 2.1. *(Escobar y West, 1995):*

- Paso 1*: Elegir valores iniciales para* $\vartheta = (\vartheta_1, ..., \vartheta_n)$
- Paso 2*: Muestrear de* ϑ *secuencialmente mediante* $(\vartheta_1|\vartheta_{-1}, \theta)$*, luego* $(\vartheta_2|\vartheta_{-2}, \theta)$ *hasta* $(\vartheta_n|\vartheta_{-n}, \theta)$[3]*, incorporando los elementos recientemente muestreados al vector* ϑ_{-i}*.*
- Paso 3*: Muestrear de* $(\mu_0|\theta, \tau)$ *y* $(\tau|\theta, \mu_0)$*, que al tratarse de un caso conjugado distribuyen Normal y Gama invertida respectivamente.*
- Paso 4*: volver al paso 2 hasta la convergencia*

Una vez que se dispone de una muestra $(\vartheta_1, \tau_1, \mu_{0_1}), ..., (\vartheta_N, \tau_N, \mu_{0_N})$ se puede aproximar la distribución predictiva como $p(\theta_{n+1}|\theta_1, ..., \theta_n) \simeq N^{-1} \sum_{i=1}^{N} p(\theta_{n+1}|\vartheta_i, \tau_i, \mu_{0_i}; s, S, M)$.

Nótese, que el esquema de urna de Polya de la distribución marginal de ϑ, genera una estructura de cluster implícita en los DP. En efecto, la ecuación (2.3) asigna una probabilidad proporcional a q_j a que ϑ_i sea el mismo valor que ϑ_j, por lo que la

[3] *ver ecuación 7 de Escobar and West (1995)*

probabilidad de que se repitan valores en ϑ es positiva. Además, si dos observaciones tienen verosimilitud alta en el valor ϑ_j, entonces la probabilidad de estar asociadas al mismo valor ϑ_j es alta. De manera que las observaciones *similares* tienden a juntarse en un mismo *cluster o grupo*, caracterizado con el mismo valor de ϑ. Llamaremos localizaciones de clusters a los parámetros $\{\vartheta_1^*, ..., \vartheta_L^*\}$, $L < n$, de valores distintos de $\{\vartheta_1, ..., \vartheta_n\}$, y $\{n_1, ..., n_L\}$ al número de observaciones asociadas a cada cluster. Esta estructura de clusters permite simplificar la ecuación (2.3). Si notamos L^- al número de valores distintos en ϑ_{-i}, y n_j^- al número de observaciones del cluster j en ϑ_{-i}. La ecuación (2.3) puede re-expresarse como:

$$p(\vartheta_i|\vartheta_{-i}, y) \propto q_0 G_i + \sum_{j=1}^{L^-} n_j^- q_j \delta_{\vartheta_j^*} \tag{2.4}$$

Bush y MacEachern (1996), presentan una estrategia computacional que explota esta estructura de cluster. La ventaja de su propuesta, se logra al reparametrizar el modelo considerando los parámetros de localización de clusters ϑ^* y un vector $s = (s_1, ..., s_n)$ con etiquetas de cluster, es decir s_i toma los valores entre 1 y L para indicar la pertenencia a un determinado cluster. Una vez conocido s, la secuencia ϑ^* es independiente y proveniente de la medida basal, resultado presentado en Korwar y Hollander (1973). El algoritmo de Bush y MacEachern (1996) implica adicionar un paso al muestreo de Gibbs, pero éste no demanda esfuerzo computacional extra, y la convergencia se agiliza, permitiendo que la cadenas se muevan más rápidamente. Los autores aplican esta idea en un modelo de diseño de bloques aleatorizados, otro caso particular de (2.2). Las particularidades son: $p(\theta|\vartheta, \mu, \tau, \sigma^2) \equiv N(1\mu + X_1\tau + X_2\vartheta, \sigma^2 I)$ donde X_1 es una matriz de indicadores de tratamiento, X_2 es una matriz de indicadores de bloques, $G_0 \equiv N(0, \rho_2^2)$. El modelo se completa con las distribuciones: $\mu|\omega_0, \rho_0^2 \sim N(\omega_0, \rho_0^2)$, $\tau|\rho_1^2 \sim N\{0, \rho_1^2(I - T^{-1}J)\}$, $\sigma^{-2}|\alpha_0, \lambda_0 \sim \Gamma(\alpha_0, \lambda_0)$, $\rho_2^{-2}|\alpha_2, \lambda_2 \sim \Gamma(\alpha_2, \lambda_2)$ con J una matriz con todas sus componentes iguales a uno. Como se puede observar, también se trata de un modelo conjugado. El algoritmo propuesto se esboza a continuación:

Algoritmo 2.2. *(Bush y MacEachern, 1996):*

- Paso 1*: determinar valores iniciales para* $\vartheta, \mu, \sigma^2, \rho_2^2, \tau$.
- Paso 2'*: simular de* $\vartheta_j|\vartheta_{-j}, \theta, \tau, \mu, \sigma^2, \rho_2^2$*, lo que implícitamente determina una configuración de clusters* s.
- Paso 2"*: simular de las localizaciones de los clusters* $\vartheta_l^*|\vartheta_{-l}^*, s, \theta, \tau, \mu, \sigma^2, \rho_2^2$.
- Paso 3*: generar simulaciones de las condicionales completas de* $\tau, \mu, \sigma^2, \rho_2^2$.
- Paso 4*: volver al paso* 2' *hasta la convergencia.*

El algoritmo 2.2 se puede esbozar de la siguiente manera: se descompone el vector ϑ en un vector con localizaciones de cluster, ϑ^*, y otro vector de etiquetas de cluster s; y se simula alternadamente de las condicionales completas de ϑ^* y s. Este esquema puede hacerse más eficiente cuando se está en presencia de una caso conjugado, marginalizando con respecto a las localizaciones de clusters al momento de muestrear de la condicionales de s, (ver MacEachern, 1998). El modelo desarrollado en este trabajo utiliza esta estrategia ya que se trata de un caso conjugado. Los detalles del algoritmos se discuten en profundidad en la sección 3.2.

Si se deseara modificar la especificación de la medida basal y caer en un caso no conjugado, la disponibilidad de algoritmos para MDP no conjugados permitiría obtener estimaciones a posteriori sin dificultad. En lo que resta de este apartado, se esbozan los algoritmos alternativos para casos no conjugados.

MacEachern y Müller (1998) generalizan las propuestas anteriores presentando un algoritmo tanto para casos conjugados como no conjugados[4]. El algoritmo se conoce

[4]West, Müller y Escobar (1994) presentan un método alternativo

como *algoritmo no gaps* y puede aplicarse a cualquier modelo de la forma (2.2). La innovación consiste en utilizar una nueva parametrización reemplazando ϑ^* definido arriba, por una versión aumentada: $\vartheta^{**} = \{\vartheta_F^*, \vartheta_E^*\}$, donde $\vartheta_F^* = \{\vartheta_1^*, ..., \vartheta_L^*\}$ son las localizaciones de los clusters no vacíos, y $\vartheta_E^* = \{\vartheta_{L+1}^*, ..., \vartheta_n^*\}$ son cluster potenciales que se encuentran vacíos. Los pasos se esboza a continuacion[5].

Algoritmo 2.3. *(MacEachern y Müller, 1998):*

- Paso 1*: Se generan valores iniciales para los parámetros del modelo.*
- Paso 2'*: Se muestrea de* $(\vartheta_F^*, s|\vartheta)$ *que en realidad es una permutación de las etiquetas de clusters.*
- Paso 2"*: Sólo en caso de que* s_i *sea la etiqueta del único elemento del cluster, se muestrea de* $(s_i|s_{-i}, \vartheta^{**})$
- Paso 2"'*: Se muestrea de* $(\vartheta^{**}|s)$
- Paso 3*: Se muestrea en la forma tradicional de las distribuciones condicionales completas de los restantes parámetros del modelo.*
- Paso 4*: volver al paso* 2' *hasta la convergencia.*

El algoritmo 8 de Neal (2000), es otra alternativa cuando se está en presencia de un caso no conjugado. Al igual que el de MacEachern y Müller (1998), este algoritmo aumenta el muestreo de Gibbs con parámetros auxiliares. En esta propuesta sin embargo, los parámetros auxiliares existen sólo temporalmente. El algoritmo se describe en los siguientes pasos:

Algoritmo 2.4. *(Neal, 2000, Algoritmo número 8):*

[5]El artículo detalla la forma de las distribuciones condicionales usadas en el algoritmo así como sugerencias para eficientizarlo.

- Paso 1*: Para $i = 1, ..., n$, sea $h = L^- + m$. Si s_i no es un singleton, simular valores independientes a partir de $\bar{\alpha}$ para ϑ_s^* con $L^- < s \leq h$. Si s_i es un singleton, entonces hacer $s_i = L^- + 1$ y simular valores independientes a partir de $\bar{\alpha}$ para ϑ_s^* con $L^- + 1 < s \leq h$. Simular un nuevo valor de s_i de entre $\{1, ..., h\}$ con las probabilidades: $p(s_i = s | s_{-i}, \theta, \vartheta_1^*, ..., \vartheta_h^*) = b \frac{n_s^-}{n-1+M} p(\theta_i | \vartheta_s^*)$ para $1 \leq s \leq L^-$; y $p(s_i = s | s_{-i}, \theta, \vartheta_1^*, ..., \vartheta_h^*) = b \frac{M/m}{n-1+M} p(\theta_i | \vartheta_s^*)$ para $L^- \leq s \leq h$, donde b es la constante de normalización. Modificar ϑ^* para que sólo contenga aquellos valores asociados a una o más observaciones.*

- Paso 2*: Para todo $s \in \{s_1, ..., s_n\}$, simular un nuevo valor de $(\vartheta_s^* | \theta_i)$ tal que $s_i = s$*

- Paso 3*: volver al paso 1 hasta la convergencia.*

Cuando $m = 1$, este algoritmo se asemeja al algoritmo no gaps. La diferencia está en que aquí se hace más probable, que un s_i que es parte de un cluster numeroso, pase a conformar un nuevo cluster. Neal (2000) propone también el siguiente algoritmo de tipo M-H para casos no conjugados (algoritmo 7).

Algoritmo 2.5. *(Neal, 2000, Algoritmo número 7):*

- Paso 1*: Para $i = 1, ..., n$, actualizar los clusters s_i como sigue. Si s_i no es un singleton, se crea un nuevo cluster, s_i^*, y se simula $\vartheta_{s_i^*}$ a partir de $\bar{\alpha}$. Se acepta s_i^* con probabilidad $a(s_i^*, s_i) = min\{1, \frac{M}{n-1} \frac{p(\theta_i | \vartheta_{s_i^*}^*)}{p(\theta_i | \vartheta_{s_i}^*)}\}$. Si s_i es un singleton, se simula s_i^* a partir de s_{-i} eligiendo $s_i^* = s$ con probabilidad $\frac{n_s^-}{n-1}$. Se acepta s_i^* con probabilidad $a(s_i^*, s_i) = min\{1, \frac{n-1}{M} \frac{p(\theta_i | \vartheta_{s_i^*}^*)}{p(\theta_i | \vartheta_{s_i}^*)}\}$.*

- Paso 2*: Para $i = 1, ..., n$. Si s_i es un singleton, no hacer nada. Si no, elegir un nuevo valor de s_i a partir de $\{s_1, ..., s_n\}$ usando la probabilidad $p(s_i = s | s{-}i, \theta_i, \vartheta^*, s_i \in \{s_1, ..., s_n\}) = b \frac{n_s^-}{n-1} p(\theta_i | \vartheta_s^*)$, donde b es la constante de normalización.*

- Paso 3*:Para todo* $s \in \{s_1, ..., s_n\}$*, simular un nuevo valor de* $\vartheta_s^*|\theta_i$ *tal que* $s_i = s$

- Paso 4*: volver al paso 1 hasta la convergencia.*

Las investigaciones para acelerar el paso 2 de los algoritmos previos - es decir la parte en que se tiene que simular de las configuraciones y localizaciones de clusters - dieron lugar a una segunda generación de algoritmos. Dentro de las propuestas más actuales, se encuentra la de Dahl (2005), quien plantea un algoritmo con un paso de Metropolis-Hasting para actualizar las configuraciones. El algoritmo - llamado algoritmo asignación de Mezclas-Separaciones secuenciales (SAMS) - tiene reminiscencias del trabajo de Liu (1996) y es válido tanto para casos conjugados como no conjugados. Los pasos a seguir para actualizar la configuraciones son:

Algoritmo 2.6. *(Dahl, 2005):*

- Paso 1*: Seleccionar un par de indices distintos* i, j

- Paso 2*: En caso de que* i *y* j *pertenezcan al mismo cluster, se propone una nueva configuración que consiste en separar el cluster en cuestión en dos. Primero* i *y* j *se ubican en dos nuevos clusters, luego los restantes índices se ubican en alguno de los dos nuevos clusters con una probabilidad de transición dada por la ecuación (12) o (13) de Dahl (2005), según se trate del caso conjugado o no conjugado. En caso contrario, se propone unir los dos clusters conformando un único grupo.*

- Paso 3: *Se calcula la razón de Metropolis-Hasting que está explicitado en la ecuación (14) o (15) de Dahl (2005) según se trate del caso conjugado o no conjugado.*

- Paso 4*: (sólo para el caso no conjugado). En caso de que* i, j *pertenezcan al mismo cluster originalmente,* ϑ_i^* *se actualiza ya sea de* $\bar{\alpha}$ *o de una caminata*

aleatoria. En caso contrario la localización del nuevo cluster toma el valor $\vartheta^* = \vartheta_j^*$

- Paso 5*: aceptar o rechazar la propuesta siguiendo la lógica del algoritmo de M-H.*

Autor	**Año**	**Nombre**	**Algoritmo**	**Aporte**
Caso Conjugado				
Escobar y West	1995		Gibbs	Los primeros m. de Gibbs
Bush y MacEachern	1996		Gibbs	mejora velocidad de convergencia
Caso no Conjugado				
MacEachern y Müller	1998	algoritmo no gap	Gibbs	permite estimar modelos no conj.
Neal	2000	m. de Gibbs con parámetros auxiliares	Gibbs	mejora velocidad de convergencia
Neal	2000	Actualizaciones M-H y m. de Gibbs parcial	Gibbs	mejora velocidad de convergencia
Dahl	2006	Mezclas-Separaciones secuenciales (SAMS)	MCMC	mejora velocidad de convergencia

Tabla 2.1: Resumen de algoritmos

La Tabla 2.1 presenta un resumen de los algoritmos mencionados. Como se mencionó previamente, el algoritmo empleado en este trabajo, está explicado con detalles en la sección 3.2. Se trata de una adaptación al modelo IRT del código presentado en De Iorio et al. (2004), que a su vez se basa en la propuesta de Bush y MacEachern (1996). El código estará disponible en http://www.paulaestadistica.blogspot.com/. Otra versión de este algoritmo se presentará próximamente en el paquete DPpackage para usuarios de R, ver Jara (2007).

2.2. Modelos IRT

Los modelos basados en la Teoría de Respuesta al Ítem (modelos IRT) son un tipo especial de modelos lineales generalizados ampliamente utilizados en el campo

de la psicometría. Esta familia de modelos representa un avance con respecto a la teoría Clásica al incorporar no sólo parámetros latentes para caracterizar la *habilidad* del individuo, sino que también parámetros vinculados a los *items*. Generalmente, los parámetros de items y habilidades se consideran efectos fijos y aleatorios respectivamente. Aquí mencionaremos la versión bayesiana, donde los parámetros vinculados al ítem, usualmente interpretados como la dificultad de los items, son también considerados aleatorios. También restringimos la exposición a modelos IRT unidimensionales, es decir, modelos en los que la habilidad es unidimensional[6]. Los modelos IRT tienen la ventaja de poner la localización de habilidades y dificultades en la misma escala, permitiendo comparar esos parámetros. Se trata, por tanto, de una escala intervalar, para detalles ver Fisher y Molenaar (1995). Entre los modelos unidimensionales más típicos se encuentra el modelo Rasch (RM), presentado en Rasch (1960).

2.2.1. Modelos Rasch paramétricos

Para ser más explícitos, supongamos datos provenientes de una prueba con J items aplicados a I individuos. Cada celda de la matriz de datos Y de dimensión $I \times J$, notada Y_{ij}, puede tomar los valores 1 (si la respuesta fue correcta) o 0 (si no). Cada individuo i está dotado de un parámetro unidimensional θ_i que captura su habilidad latente. Cada ítem j tiene asociado un parámetro β_j que representan la dificultad del ítem. El conjunto de respuestas de una persona es entendida como medidas repetidas de la misma habilidad latente. Sea $h(x) = \frac{e^x}{1+e^x}$. El modelo estadístico del (RM) se enuncia de la siguiente forma:

$$P(Y_{ij} = y_{ij}|\theta_i, \beta_j) \sim Bern(h(\theta_i - \beta_j)) \tag{2.5}$$

[6]Los modelo multidimensionales buscan evaluar las distintas operaciones cognitivas que realiza el individuo al responder un ítem.

para todo individuo i e ítem j. Cuando las habilidades individuales son consideradas como parámetros, entonces los parámetros de interés son θ y β, donde $\beta = (\beta_1, ..., \beta_J)$ y $\theta = (\theta_1, ..., \theta_I)$. Desde el punto de vista Bayesiano, el modelo requiere la especificación de las distribuciones a priori de θ´s y β´s. Generalmente se emplean distribuciones normales para tal fin, es decir:

$$\theta_i|\mu_\theta, \sigma_\theta^2 \overset{i.i.d.}{\sim} N(\mu_\theta, \sigma_\theta^2) \tag{2.6}$$

$$\beta_j|\mu_\beta, \sigma_\beta^2 \overset{i.i.d.}{\sim} N(\mu_\beta, \sigma_\beta^2) \tag{2.7}$$

También se supone $\theta \perp\!\!\!\perp \beta$. La identificación de los parámetros requiere fijar uno de los β's, por ejemplo $\beta_1 = 0$, o bien fijar μ_θ, en por ejemplo el valor 0. En tal caso, los parámetros de interés son (β, σ_θ^2) (ver San Martín y Rolin, 2009). Dado que θ y β son parámetros latentes (no observables), no hay ningún argumento a priori para creer que las habilidades y las dificultades se comporten de acuerdo a una distribución normal. Generalmente, se emplea esta especificación por sencillez y por ser distribuciones ampliamente conocidas. Para darle más flexibilidad al modelo desde el enfoque paramétrico, es posible emplear mezclas de normales. Por ejemplo si se desea relajar el supuesto de habilidades normales se puede reemplazar (2.6) por (2.8), que especifica una mezcla de normales como distribución a priori de las habilidades.

$$\theta_i|\sigma_\theta^2 \overset{i.i.d.}{\sim} F \ , \ f(\theta_i) = \int \phi(\frac{\theta_i - x}{\sigma_\theta}) G_0(dx), \tag{2.8}$$

En la ecuación (2.8) f es la densidad asociada a la distribución F, $G_0 \equiv N(\mu_0, \sigma_0^2)$ y ϕ representa la densidad de la distribución normal estándar. Este modelo es equivalente a aumentar un grado de jerarquía, introduciendo una distribución normal con media μ_0 y error estándar σ_0 para μ_θ. En este caso, el parámetro de interés es también (β, σ_θ^2). De aquí en más, al modelo conformados por las ecuaciones (2.5), (2.6) y

(2.7) se lo denominará *modelo Rasch Normal nulo con parámetros de interés* (θ, β), mientras que al conformado por las ecuaciones (2.5), (2.8) y (2.7), *modelo basal nulo con parámetros de interés* (θ, β). Ambos modelos requieren la restricción $\beta_1 = 0$, y se completan con los supuestos de independencia condicional 2.1 y 2.2.

Supuesto 2.1. *Sea* $Y_i = (Y_{i1}, ..., Y_{iJ})$ *el vector de respuestas del individuo* i. *Se supone que* $Y_1, ..., Y_I$ *son vectores independientes condicional en* θ *y* β.

Supuesto 2.2. *Las respuestas del mismo estudiante a diferentes items son independientes condicional a la habilidad de la persona, es decir* $Y_{i1}, ..., Y_{iJ}|\theta_i, \beta$ *son variables independientes. Este supuesto se conoce como* ***independencia local***.

El Supuesto 2.1 implica que los individuos no se copian durante la prueba o bien que las respuestas de un estudiante no está relacionada con la respuesta de cualquier otro estudiante. El Supuesto 2.2 indica que el individuo no *aprende* durante la prueba. Para clarificar esta idea notemos que $E(Y_{ij}|Y_{i1}, ..., Y_{ij-1}, Y_{ij+1}, ..., Y_{iJ}, \theta_i) = E(Y_{ij}|\theta_i)$, es decir que θ_i es la única información relevante para explicar la respuesta de un individuo a un determinado ítem. Las respuestas de ese mismo individuo a los demás ítems no influye.

Los modelos normal y basal nulos se pueden extender especificando distribuciones a priori para μ_β, σ^2_β y σ^2_θ. Esta ampliación permite flexibilizar más la forma de las distribuciones a priori, pero siempre dentro de un enfoque paramétrico. La curva característica del ítem (ICC) - que se obtiene fijando β_j y dejando variar θ - es estrictamente creciente con forma logística. Este resultado es deseable pues significa que la probabilidad condicional de una respuesta correcta siempre aumenta con la habilidad del individuo. Otra ventaja particular de estos modelos es que el puntaje observado es decir $\sum_{j=1}^{J} Y_{ij}$ es un estadístico suficiente para θ_i dado β. Esta propiedad es útil en materia de cálculos. Además, ambos modelos pueden ser modificados para

incorporar covariables que caractericen la habilidad del individuo. El modelo Rasch de regresión latente, propuesto por Verhelst y Eggen (1989), por ejemplo, plantea introducir una regresión de θ_i en covariables d_i. En tal caso a la ecuación (2.6) se reemplaza por las ecuación (2.9) y (2.10)

$$\theta_i|\mu_\theta^{d_i}, \sigma_\theta^2 \overset{i.i.d.}{\sim} N(\mu_\theta^{d_i}, \sigma_\theta^2) \tag{2.9}$$

$$\mu_\theta^{d_i} = \vartheta d_i' \tag{2.10}$$

El modelo que surge de considerar ϑ fijo (no aleatorio) se llamará aquí *modelo Rasch Normal con parámetros de interés* $(\beta_2, ..., \beta_J, \mu_\theta^1, ..., \mu_\theta^K)$, fijando β_1 en cero y siendo K el número de valores distintos de la covariable. Por otra parte, si ϑ es considerado aleatorio y se le asigna una distribución normal es equivalente a reemplazar la ecuación (2.8) por (2.11)

$$\theta_i|\sigma_\theta^2 \overset{i.i.d.}{\sim} F^{d_i} \ , \ f^{d_i}(\theta_i) = \int \phi(\frac{\theta_i - \vartheta d_i'}{\sigma_\theta}) G_0(d\vartheta), \tag{2.11}$$

y se denominará aquí *modelo basal con parámetros de interés* $(\beta_2, ..., \beta_J, \mu_\theta^1, ..., \mu_\theta^K)$. Aquí, G_0 es también normal, pero posiblemente multivariada, dependiendo de la dimensión de d_i.

2.2.2. Modelos basales desde el enfoque Semiparamétricos

Los modelo Rasch basales permiten más variedad de formas para la distribución de habilidades que los modelos Rasch normales. Sin embargo, su flexibilidad sigue siendo limitada si se compara con modelos que emplean herramientas no paramétricas. En efecto, suponiendo que los parámetros de interés son ahora la distribución de habilidades F y los parámetros de dificultad $\beta_2, ..., \beta_J$, fijando β_1 en cero. En tal caso

el modelo estadístico - ecuación (2.5) -, puede re-escribirse en término de los nuevos parámetros mediante las ecuaciones (2.12) y (2.13).

$$Y_{ij}|(\beta_j, F) \overset{i.}{\sim} F_j(\beta_j, F) \tag{2.12}$$

$$F_j(y_{ij}; \beta_j, F) = \int h(x - \beta_j)^{y_{ij}} (1 - h(x - \beta_j))^{1-y_{ij}} F(dx) \tag{2.13}$$

Es decir que ahora el modelo estadístico es una mezcla de de distribuciones Bernoulli. Considerando esta reparametrización en el modelo basal nulo, la distribución a priori para β continúa siendo la misma, definida en la ecuación (2.7). Por su parte, F puede ser considerada un elemento aleatorio introduciendo una distribución para σ_θ^2. En tal caso, la distribución a priori para F queda implícitamente determinada por la distribución de σ_θ^2 mediante la ecuación (2.14). Sea f la densidad *aleatoria* asociada a F, que puede ser entendida como una función de σ_θ^2: $f = g(\sigma_\theta^2) = \int \phi(\frac{\theta - x}{\sigma_\theta}) G_0(dx)$. El soporte para la distribución inducida es $g(\mathbb{R}^+)$, el conjunto de todas las mezclas de normales de la forma $f(\theta) = \int \phi(\frac{\theta - x}{\sigma_\theta}) G_0(dx)$; que es un subconjunto del espacio de funciones de distribución sobre Θ, $M(\theta)$. Consideremos la σ-álgebra $\mathcal{A}$, partes de $M(\Theta)$ y sea $A \in \mathcal{A}$, entonces la distribución se escribe como:

$$P(f \in A) = P(\sigma_\theta^2 \in g^{-1}(A)) \tag{2.14}$$

Nótese que considerando esta distribución a priori, el soporte para el nuevo parámetro, F, está dado *sólo* por las distribuciones mezcla de dos normales de la forma especificada arriba, un soporte bastante limitado. Claramente, la probabilidad de que F tome una forma distinta que una mezcla de normales es cero, es decir que si A' es el conjunto de todas las distribuciones cuyas densidades no toman la forma dada por g, entonces:

$$P(f \in A') = 0,$$

En este sentido el modelo basal nulo impone una distribución a priori muy restrictiva para F, si se lo compara con otras posibles distribuciones a priori como los DP, MDP, etc. Algo similar ocurre al considerar al modelo basal desde el enfoque no paramétrico. En este caso, el modelo estadístico queda dado por las ecuaciones (2.15) y (2.16), y los parámetros de interés son $(\beta_2, ..., \beta_J, F_1, ..., F_K)$.

$$Y_{ij}|(\beta_j, F_k) \overset{i.}{\sim} F_{jk}(\beta_j, F_k) \tag{2.15}$$

$$F_{jk}(y_{ij}; \beta_j, F_k) = \int h(x - \beta_j)^{y_{ij}} (1 - h(x - \beta_j))^{1-y_{ij}} F_k(dx) \tag{2.16}$$

La distribución a priori para la colección $(F_1, ..., F_K)$ también queda implícitamente definida mediante la distribución de σ_θ^2. El soporte de dicha distribución está dada por el conjunto de todas las colecciones $(f_1, ..., f_K)$, con $f_k = g_k(\sigma_\theta^2) = \int \phi(\frac{\theta_i - \vartheta d^k}{\sigma_\theta}) G_0(d\vartheta)$, un subconjunto del espacio producto $M(\Theta)^K$. Análogamente, para A perteneciente a la σ-álgebra de partes de $M(\Theta)^K$, la medida de probabilidad se muestra en la ecuación (2.17).

$$P(\{(f_1, ..., f_K) \in A) = P(\sigma_\theta^2 \in (g_1^{-1}, ..., g_K^{-1})(A)), \tag{2.17}$$

y nuevamente, para el conjunto A' de todas las colecciones que no toman la forma dada por $(g_1, ...g_K)$ la probabilidad de ocurrencia es nula:

$$P((f_1, ..., f_K) \in A') = 0.$$

La ecuaciones (2.12), (2.13), (2.7) y (2.14) conforman el *modelo basal nulo* expresando desde un enfoque no paramétrico; mientras que las ecuaciones (2.15), (2.16), (2.7) y (2.17) hacen lo propio para el *modelo basal*. La explicación para los nombre de los modelos quedará claro en el capítulo 3, ya que las f_k's serán las densidades asociadas a las medidas basales del modelo propuesto.

2.2.3. Modelos Rasch Semiparamétricos

La estadística no paramétrica ofrece herramientas para flexibilizar la forma de F mucho más allá de lo que permiten los modelos basales. En este sentido, una gran variedad de modelos - denominados **Modelos IRT Semi-Paramétricos** - han surgido con la idea de flexibilizar el supuesto asociado a la distribución de los parámetros aleatorios θ y/o β. Desde el punto de vista clásico, el trabajo de Woods y Thissen (2006), por ejemplo, propone estimar la distribución de θ a partir de métodos spline y menciona numerosos trabajos en esa línea. Pero siguiendo con el enfoque semiparamétrico Bayesiano, recordemos que en este caso se concentra la atención en la distribución de habilidades, considerándola un elemento aleatorio y proponiendo, por tanto, una distribución a priori para ella. Roberts y Rosenthal (1998) y Duncan y MacEachern (2008) han introducido la estadística NPB en modelos IRT. Ambos artículos proponen un modelo donde F proviene de un proceso Dirichlet. El problema con esta especificación es que las distribuciones que surgen de un DP son casi seguramente discretas (ver Propiedad 2.4). Una forma de sortear este obstáculo es especificando un mezcla de distribuciones continuas, con distribución mezclante proveniente de un DP (MDP). Esta opción ofrece mayor flexibilidad en las formas de habilidades, permitiendo, a su vez, que la habilidad sea una variable continua. Las ecuación (2.18) muestra la densidad de F cuando se especifica como MDP.

$$f(\theta) = \int \phi(\frac{\theta - \vartheta}{\sigma_\theta^2}) G(d\vartheta) \ , \ G \sim DP(MG_0) \tag{2.18}$$

Aquí la aleatoriedad de F está inducida por el par (σ_θ^2, G) con $\sigma_\theta^2 \perp\!\!\!\perp G$. En lo que sigue de este documento, el modelo conformado por las ecuaciones (2.12), (2.13), (2.18) y (2.7); y los Supuestos 2.1 y 2.2, se denominará *modelo Rasch Semiparamétrico nulo* o simplemente *modelo nulo*. Es importante destacar que en este modelo los vectores de respuestas del individuo Y_i's son i.i.d. condicional en (F, β); característica que se desprende del Supuesto 2.1 y el hecho de que $\theta_i | F \overset{i.i.d.}{\sim} F$; ver Mouchart y San

Martín (2003). La importancia de esta propiedad radica en que, como los Y_i´s son observables, es posible confrontar esta propiedad con los datos reales y decidir hasta qué punto el modelo nulo es apropiado en una situación dada. Concretamente, si los Y_i's fueran idénticamente distribuidos, deberían ser intercambiables por lo que 2 submuestras disjuntas cualesquiera de los datos deberían presentar histogramas similares.

2.2.4. Otras variantes de modelos Rasch

En poblaciones de estudiantes muy heterogéneas, donde no se verifica que los patrones de respuestas provengan de una misma distribución, la incorporación de covariables al análisis puede ser una solución. DeBoeck y Wilson (2004) presentan una gran variedad de modelos llamados **Modelos IRT Explicativos**, que incorporan covariables al análisis, las cuales pueden modificar a los items y/o a las habilidades del individuo. Como se verá en el capítulo 3, nuestro modelo reúne las condiciones de ser un modelo IRT semiparamétrico e incorporar covariables al mismo tiempo.

Los modelos antes mencionados no agotan la variedad de especificaciones que han surgido a partir del RM. Existe una serie de trabajos, por ejemplo, que buscan flexibilizar el modelo IRT modificando la curva ICC. Estos modelos se conocen en la literatura como **Modelos IRT No Paramétrico**. Se puede mencionar los esfuerzos realizados por Ramsay (ver por ejemplo Ramsay, 1991). Duncan y MacEachern (2008), y Miyazaki y Hoshino (2009) proponen modelos no paramétrico donde la curva ICC se estima mediante métodos propios de la estadística NPB. Recientemente, Karabatsos y Walker (2009), han extendido los trabajo desde el enfoque NPB en psicometría al área de equating. Por último es interesante mencionar que existen modelos conocidos como **Modelos de Funcionamiento Diferencial de Items (DIF)**. Estos exploran situaciones en los que algunos items actúan en forma diferencial para

distintas sub-poblaciones. Estos modelo pueden incorporar covariables para determinar esos grupos poblacionales o bien hacer un estudio con clases latentes o clusters, ver por ejemplo Cohen y Bolt (2005).

2.3. El concepto de identificación Bayesiana

El presente trabajo incluye un estudio de identificación del modelo desarrollado. Generalmente, en estadística aplicada se dedica poco espacio a la identificación de nuevos modelos. Sin embargo, los estudios de identificación son de suma importancia pues no es posible hacer inferencia sobre parámetros no identificados. El hecho de que gran parte de los fenómenos que se intenta medir en psicometría no son observables y por tanto deben modelarse con variables latentes, invita a prestar más atención a este tema. Si un parámetro es no identificado, la combinación de la estructura del modelo junto con los datos no proveen la información suficiente para decir algo acerca de dicho parámetro. Esto es claro desde el punto de vista clásico pues al intentar estimar un parámetro no identificado ocurre que hay más de un posible valor estimado, poniendo al investigador en una disyuntiva. Desde el punto de vista Bayesiano, tampoco tiene sentido hacer inferencia sobre parámetros no identificados. Sin embargo, el hecho de que siempre es posible obtener una única distribución a posteriori del parámetro de interés ha traído confusión sobre el tema. Hay que recalcar entonces que, dentro de este último enfoque, importa que la distribución a posteriori realmente se esté actualizando con los datos, lo que no ocurre si el parámetro no está identificado[7].

Brevemente, el enfoque clásico considera un modelo estadístico como una familia de distribuciones de muestreo indexada por un parámetro. De esta forma, llamamos

[7]Aunque las distribuciones a priori y a posteriori difieran, esto no necesariamente implica que haya una actualización real.

Ω al espacio muestral, $\mathfrak{X}$ a la σ-ágebra asociada a Ω, y $P^{\gamma}(\omega)$ una familia de distribuciones indexada por el parámetro γ; se tiene que el trío $(\Omega, \mathfrak{X}, P^{\gamma})$ con $\gamma \in \Gamma$ define un modelo estadístico.

Definición 2.2. *El parámetro γ se dice identificado en forma clásica o* ***c-identificado*** *si la aplicación $\gamma \to P^{\gamma}$ es inyectiva*

Mientras que en la teoría clásica, la inyectividad de la aplicación $\gamma \to P^{\gamma}$ determina la identificación de γ, en el caso Bayesiano la identificación se define a partir del concepto de *mínima suficiencia*. A diferencia de los modelos clásicos, para describir un modelo Bayesiano basta con definir la distribución conjunta de (Y, γ) y no una familia de distribuciones. Consideremos entonces un modelo Bayesiano definido por la distribución conjunta sobre (Y, γ). A continuación se definen los conceptos de suficiencia paramétrica y suficiencia paramétrica mínima necesarias para caracterizar la identificación Bayesiana. Para mayores detalles, ver el capítulo 4 de Florens et al. (1990).

Definición 2.3. *Una función $g(\gamma)$ del parámetro γ es un* ***parámetro suficiente*** *para Y si la distribución condicional de la muestra Y dado γ es la misma distribución que la distribución de la muestra Y dado $g(\gamma)$, esto es,*

$$Y \perp\!\!\!\perp \gamma | g(\gamma) \tag{2.19}$$

Nótese que la condición (2.19) implica que la distribución de Y está completamente determinada por $g(\gamma)$, o en otras palabras, γ es redundante una vez que se conoce $g(\gamma)$. Gracias a la simetría de la relación de independencia condicional, se puede concluir también que $g(\gamma)$ es un parámetro suficiente si la distribución condicional

de la parte redundante γ dado el parámetro suficiente $g(\gamma)$ no se actualiza por la muestra, es decir $p(\gamma|Y, g(\gamma)) = p(\gamma|g(\gamma))$.

Definición 2.4. *Una función $g(\gamma)$ del parámetro γ es un* ***parámetro mínimo suficiente*** *si $g(\gamma)$ es un parámetro suficiente y es función de cualquier otro parámetro suficiente.*

La definición 2.4 abre las puertas para caracterizar la identificación Bayesiana. La definición se detalla a continuación:

Definición 2.5. *El parámetro γ se dice identificado en forma Bayesiana por Y o* ***b-identificado*** *por Y, que denotamos $\gamma \prec Y$, si es un parámetro mínimo suficiente. El símbolo "$\prec$"se utilizará para denotar la relación "b-identificado por"*

Para clarificar estos conceptos, consideremos el siguiente ejemplo simple. Sea X una variable aleatoria con distribución normal y tomemos la familia de distribuciones normales:

$$P^{\gamma} \in \{P : P \overset{d}{=} N(\mu_1 + \mu_2, 1), \gamma = (\mu_1, \mu_2)\}$$

El parámetro γ no está identificado desde el punto de vista clásico. En efecto, la aplicación $\gamma \to P^{\gamma}$ no es inyectiva, lo que se puede confirmar fácilmente tomando 2 valores distintos de γ, como por ejemplo $\gamma' = (\mu_1', \mu_2')$ y $\gamma'' = (\mu_1' + c, \mu_2' - c)$ con $c \neq 0$, que determinan una misma distribución $P^{\gamma'} \overset{d}{=} P^{\gamma''} \overset{d}{=} N(\mu_1' + \mu_2', 1)$. Para analizar este ejemplo a partir del enfoque Bayesiano, se debe especificar una distribución conjunta de (X, γ), que es equivalente a definir un modelo condicional para $X|\gamma$ junto con una distribución a priori para γ (pues $p(X, \gamma) = p(X|\gamma)p(\gamma)$). Consideremos $\gamma = (\mu_1, \mu_2)$. La distribución condicional $X|(\mu_1, \mu_2) \sim N(\mu_1 + \mu_2, 1)$, y las prioris $\mu_1 \sim N(0, 1)$, $\mu_2 \sim N(0, 1)$, con $\mu_1 \perp\!\!\!\perp \mu_2$. Este ejemplo tampoco es b-identificado. Para verificarlo notemos que $\gamma' = g(\gamma) = \mu_1 + \mu_2$ es un parámetro suficiente para X:

$p(X, \gamma|g(\gamma)) = p(X|\gamma, g(\gamma))p(\gamma|g(\gamma)) = \phi(X - g(\gamma))p(\gamma|g(\gamma)) = p(X|(g(\gamma))p(\gamma|g(\gamma))$ donde ϕ es la densidad normal estándar. De manera que $X \perp\!\!\!\perp \gamma|g(\gamma)$. Ahora bien, como γ no puede ser expresada como función de γ' se tiene que γ no es un parámetro mínimo suficiente y por tanto γ no está identificado por X.

El ejemplo permite clarificar que el problema de identificación Bayesiana consiste en encontrar un parámetro suficiente que no posea información redundante, que equivale a hallar el parámetro mínimo suficiente (ver Florens et al. (1990)). Es importante aclarar que los resultados de b-identificación no dependen de las distribuciones a priori elegidas. Si los parámetros de interés son identificados para una distribución a priori μ, también lo estarán para cualquier otra distribución μ', siempre y cuando los conjunto nulos que determinan μ y μ' coincidan. Si bien el ejemplo anterior no verifica la identificación desde ninguno de los dos enfoques, lo conceptos no son idénticos. En general c-identificación es *más fuerte* que la b-identificación, es decir que c-identificación implica b-identificación para toda distribución a priori, pero la relación inversa no se verifica. Para mayores detalles sobre la comparación de estos dos conceptos, ver la sección 4.6.2 de Florens et al. (1990).

Finalmente, se presentan dos Teoremas que permiten combinar resultados de identificación en distintos modelos. Serán útiles en la demostración de identificación del modelo propuesto, pues como veremos en el capítulo 3, la especificación propuesta puede ser entendida como una unión de modelos distintos.

Teorema 2.1. *Sean $Y_1, ..., Y_N$, $\gamma_1, ..., \gamma_N$ y D variables aleatorias que satisfacen:*

1. $\perp\!\!\!\perp_{1 \leq i \leq N} Y_i|\gamma_1, ..., \gamma_N, D$
2. $Y_i \perp\!\!\!\perp \gamma_1, ..., \gamma_N|\gamma_i, D$ *para* $i = 1, ..., N$
3. $\gamma_i \prec Y_i|D$ *para* $i = 1, ..., N$

entonces $(\gamma_1, ..., \gamma_N) \prec Y_1, ...Y_N|D$

Teorema 2.2. *Sean* Y, γ, D_1 *y* D_2 *variables aleatorias que satisfacen:*

1. $D_2 \perp\!\!\!\perp Y|\gamma, D_1$

2. $\gamma \prec Y|D_1$

entonces $\gamma \prec Y|D_1, D_2$

El **Teorema 2.1** es el teorema 2 de Mouchart y San Martín (2003). Supongamos que Y representa los valores muestrales, γ los parámetros y D las covariables en un modelo bayesiano. Entonces, la primera condición indica que $Y_1, ..., Y_N$ es una secuencia de variables aleatorias independientes condicional a $\gamma_1, ...\gamma_N$ y D. La segunda condición de Teorema indica que γ_i es un parámetro *suficiente* para Y_i condicional en D, y la tercera establece que cada γ_i está b-identificado por Y_i condicional en D. En tal caso, el teorema confirma la b-identificación del conjunto de γ's por el conjunto de Y's condicional en D, a partir de la identificación de cada γ_i por cada Y_i dado D. Heurísticamente, este teorema confirma la identificación de un parámetro γ si sus componentes γ_i también lo están, bajo las condiciones 1 y 2. El **Teorema 2.2** es el teorema 4.5.3 de Florens et al. (1990). Establece que si γ está identificado por Y condicional a D_1, seguirá estándolo condicional a D_1 junto con cualquier variable aleatoria que sea independiente de Y condicional en γ y D_1.

Capítulo 3

Modelo Rasch Bayesiano Semiparamétrico Explicativo

3.1. El Modelo

Este trabajo, siguiendo la línea de los trabajos de Quin (1998), Duncan (2004), Duncan y MacEachern (2008), presenta un aporte para introducir la estadística Bayesiana NP dentro del área de la psicometría. El trabajo está focalizado en estudiar la habilidad de distintas subpoblaciones de estudiantes chilenos.

Como se mencionó en la sección 3.1, los modelos Rasch Semiparamétricos Bayesianos existentes hasta el momento presuponen que los patrones de respuesta de los individuos son intercambiables, condicional en la distribución de habilidades y los parámetros de habilidad. Muchas veces, en poblaciones heterogéneas, como es el caso de los estudiante chilenos, este supuesto no se cumple[1]. El camino elegido aquí para modelar este fenómeno es incorporando variables explicativas al modelo. Esta estrategia tiene sentido pues la heterogeneidad de los individuos se origina precisamente

[1]Ver la sección 4.1 y particularmente la Tabla 4.3 y la Figura 4.1 para mayor claridad sobre este tema.

por diferencias en las covariables, que a su vez reflejan variabilidad socioeconómica y cultural. La modelación se llevó a cabo teniendo en cuenta dos elementos básicos: primero, parece lógico suponer que un individuo responde a la prueba usando su habilidad y que dicha habilidad no es un "ente aislado" sino que está vinculado al capital sociocultural del individuo. Segundo, dado que la habilidad del individuo es no observable, no hay ningún argumento para suponer que la distribución de habilidades tenga la forma de una distribución normal. Es por eso que consideramos razonable inferir sobre la forma de la distribución de habilidad a partir de la información provista por los datos, a saber los patrones de respuesta.

3.1.1. Especificación del Modelo

En lo que sigue se mantiene la notación del capítulo 2, es decir I representa el número de individuos, J es el número de items, $Y = (Y_{ij})$ es un matriz de $(I \times J)$ con respuestas (correcta o incorrecta) del individuo i al ítem j. $\beta = (\beta_1, ..., \beta_J)'$ es un vector con parámetros latentes que representan la dificultad del ítem y se define $h(x) = \frac{e^x}{1+e^x}$. Se utiliza además, D para nombrar a la matriz de diseño de tamaño $(I \times R)$, y d_i' para representar cada fila de D, es decir, las covariables correspondientes al individuo i. Proponemos utilizar como *parámetros* a las distribuciones de habilidad indexadas por covariables, en lugar de las habilidades. Es decir que los individuos con el mismo valor de covariable poseen la misma distribución, pero la distribución cambia para alumnos con distintas características socioeconómicas. Suponiendo que hay K, $K \leq I$ patrones de covariables en la muestra, entonces $G_1, ..., G_K$ son los parámetros de interés; y el parámetro de habilidad condicional en el patrón de covariables distribuye G_k, es decir $\theta_i|(d_i = d_k, G_k) \sim G_k$. Marginalizando con respecto a θ_i en la ecuación (2.5) se llega al modelo estadístico propuesto. Éste se expresa en las ecuaciones (3.1) y (3.2).

$$Y_{ij}|(\beta_j, G_k, d_k = d_i) \overset{i.}{\sim} F_{kj}(\beta_j, G_k) \tag{3.1}$$

$$F_{kj}(y_{ij};\beta_j,G_k)=\int h(x-\beta_j)^{y_{ij}}(1-h(x-\beta_j))^{1-y_{ij}}G_k(dx) \tag{3.2}$$

Las respuestas Y_{ij} condicionadas en los parámetros de interés $(\beta_j, G_k : d_k = d_i)$ provienen de una distribución F_{kj}, que es una mezcla de distribuciones Bernoulli con distribución mezclante G_k. En esta expresión el parámetro de habilidad ha desaparecido ya que el interés está ahora en la distribución que genera la habilidad para el individuo i. G_k está indexada por el subíndice k para indicar que el valor de la covariable del individuo i es igual a d_k. Si bien la habilidad no está explícita en las ecuaciones (3.1) y (3.2), en esta parte del modelo se especifica que el individuo responde a los items usando su habilidad, cuya distribución depende de sus características socioeconómicas.

Desde un enfoque Bayesiano, los parámetros son considerados aleatorios y se les asigna una distribución. Para especificar G_k se considera una mezca de distribuciones normales como indica la ecuación (3.3), donde la aleatoriedad de G_k está dada por el par (G, σ_θ^2), que son considerados elementos aleatorios.

$$G_k(\theta)=\int \Phi(\frac{\theta-d_k'\vartheta}{\sigma_\theta})G(d\vartheta) \tag{3.3}$$

Aquí es necesario hacer tres comentarios. En primer lugar, las distribuciones aleatorias $G_1, ..., G_K$ no son independientes entre si ya que todas dependen del mismo par de elementos aleatorios (G, σ_θ^2). La diferencia entre ellas está dada por el patrón de covariables d_k. En segundo lugar las covariables entran en juego a través de la media de la distribución normal al interior de la mezcla. Esto hace que $G_k \to G_{k_0}$ cuando $d_k \to d_{k_0}$ (ver MacEachern, 1999). Esta propiedad es deseable pues es de esperar que si los valores de las covariables son parecidos, también los sean las distribuciones que indexan. Tercero, los parámetros (G, σ_θ^2) no son de particular interés para la inferencia y sólo se introducen para inducir una distribución a priori sobre G_k que incluya una

gran diversidad de formas de distribuciones continuas.

La especificación de la parte del modelo vinculada a la habilidad requiere, entonces, definir distribuciones a priori para G y σ_θ^2. Las ecuaciones (3.4) y (3.5)muestran dichas distribuciones. La ecuación (3.4) indica que G proviene a priori de un DP, y junto con la ecuación (3.3) describen una mezcla de ANOVA DDP para la distribución de habilidades (ver la sección 2.1.4 para más detalles sobre DDP). La distribución a priori de σ_θ^2 es estándar: se emplea una distribución chi cuadrado invertida escalada, que es denotada con el símbolo $inv-\chi^2_\nu(s)$ donde ν son los grados de libertad y s el parámetro de escala.

$$G|M, G_0 \sim DP(M, G_0) \tag{3.4}$$

$$\sigma_\theta^2 \sim inv-\chi^2_{\nu_\theta}(s_\theta^2) \tag{3.5}$$

G_0 se considera $N_R(\mu_\vartheta, \Sigma_\vartheta)$, la distribución típicamente usada en la versión paramétrica del modelo, y que permite, además, trabajar con un caso conjugado como se explica en la sección 2.1.6. Finalmente los últimos parámetros concernientes a la habilidad en la jerarquía son μ_ϑ y Σ_ϑ que también se especifican en forma estándar.

$$\mu_\vartheta|\Sigma_\vartheta \sim N(0, \Sigma_\vartheta) \tag{3.6}$$

$$\Sigma_\vartheta \sim inv-Wishart_{\nu_\vartheta}(\Lambda_\vartheta^{-1}) \tag{3.7}$$

La distribuciones a priori vinculadas con los parámetros de dificultad se detallan en las ecuaciones (3.8), (3.9) y (3.10) y son equivalentes a los utilizados generalmente junto con el modelo estadístico Rasch presentado en la ecuación (2.5).

$$\beta_j|\mu_\beta, \sigma_\beta^2 \overset{i.i.d.}{\sim} N(\mu_\beta, \sigma_\beta^2) \text{ para } j = 2, ..., J \tag{3.8}$$

$$\mu_\beta|\sigma_\beta^2 \sim N(0, \sigma_\beta^2) \tag{3.9}$$

$$\sigma_\beta^2 \sim inv - \chi_{\nu_\beta}^2(s_\beta^2) \tag{3.10}$$

β_1 debe ser fijado (aquí se consideró $\beta_1 = 0$), lo que permite identificar los parámetros $(G_1, ..., G_K, \beta_2, ..., \beta_J)$ como se explicará en detalle en la sección 3.1.3. El modelo queda completamente especificado con las ecuaciones (3.1)-(3.10) y los supuestos (3.1)-(3.6) de independencia condicional (para mayores detalles ver el Apéndice C).

Para el model estadístico se supone:

Supuesto 3.1. *Los parámetros* $(G_1, ...G_K, \beta_2, ..., \beta_J)$ *son suficientes en el modelo estadístico, es decir que*

$$Y \perp\!\!\!\perp (\mu_\beta, \sigma_\beta^2, \mu_\vartheta, \Sigma_\vartheta) | (G_1, ..., G_K, \beta_2, ..., \beta_J)$$

Supuesto 3.2. $(Y_1, ..., Y_I)$ *son independientes condicional en* $(G_1, ...G_K, \beta_2, ...\beta_J, D)$

Supuesto 3.3. *Se cumple la independencia local:* $(Y_{i1}, ..., Y_{iJ})$ *son independientes condicional en* $(G_k, \beta_2, ...\beta_J, d_i = d_k)$

Para la distribución a priori conjunta de los parámetros se supone:

Supuesto 3.4. *Los parámetros vinculados con la habilidad son independientes con los relacionados a la dificultad:* $(G_1, ..., G_K, \mu_\vartheta, \Sigma_\vartheta) \perp\!\!\!\perp (\beta_2, ..., \beta_J, \mu_\beta, \sigma_\beta^2)$

Supuesto 3.5. *Condicional en* $(\mu_\vartheta, \Sigma_\vartheta)$*, la varianza de la distribución normal en la ecuación (3.3) es independiente de la distribución mezclante G:* $G \perp\!\!\!\perp \sigma_\theta^2 | (\mu_\vartheta, \Sigma_\vartheta)$

Supuesto 3.6. *La varianza de la distribución normal en la ecuación (3.3) es independiente de los parámetros de la medida basal de G:* $\sigma_\theta^2 \perp\!\!\!\perp (\mu_\vartheta, \Sigma_\vartheta)$

Los parámetros de interés en el modelo definido por las ecuaciones (3.1)-(3.10) y los supuestos 3.1-3.6 son $(G_1, ..., G_K, \beta_1, ..., \beta_J)$, es decir las distribuciones de habilidad y los parámetros de dificultad. El modelo Rasch establece una misma escala para medir

habilidades y dificultades, lo que hace posible su comparación. Este modelo, a su vez, permite hacer inferencia acerca de la forma distribucional de la habilidad para cada patrón de covariables.

3.1.2. Elección de hiperparámetros

El parámetro σ_θ^2, junto con el hiperparámetro M están involucrados en el *suavizamiento* de G_k. Valores relativamente pequeños para σ_θ^2 y moderados para M ($M \simeq 1$) generan una cantidad reducida de clusters junto con normales concentradas sobre su media al interior de la mezcla, lo que tiende a generar densidades para G_k multimodales. Si σ_θ^2 es grande, las normales al interior de la mezcla de expanden suavizando la forma de la densidad de G_k. La ecuación (3.5) establece la distribución a priori para σ_θ^2. Los valores de ν_θ y s_θ^2 se fijan en 5 y 0.1 respctivamente, lo que implica una media para σ_θ^2 de $E(\sigma_\theta^2|\nu_\theta, s_\theta^2) = 0.17$ y un error estándar de $se(\sigma_\theta^2|\nu_\theta, s_\theta^2) = 0.23$, para suponer a priori un valor relativamente pequeño. Condicional en σ_θ^2, G_0 induce una distribución basal para la mezcla, a saber:

$$G_{k0}(\theta) = E(G_k|\sigma_\theta^2) = \int N(\theta, d_k'\vartheta, \sigma_\theta^2) G_0(d\vartheta),$$

que es una mezcla de normales con distribución mezclante G_0 (ver Lo, 1984). Se escogió la distribución normal para G_0, lo que hace que el modelo basal - que surge de reemplazar G_k por G_{k0} -, coincida con el modelo basal definido en la sección 3.1. Se trata de la versión paramétrica de nuestro modelo y su competidor estándar. Una característica interesante de nuestra especificación es que el modelo basal es un caso particular cuando $M \to \infty$. En efecto, la Propiedad 2.7 implica que $p(\{G = G_0\}) = 1$ cuando M tiende a ∞, por lo que $p(\{G_k = G_{k0}\}) = 1$. Este punto es importante a la hora de comparar modelos y determinar hasta qué punto es mejor la estimación del modelo semiparamétrico. En el presente estudio se escogió $M = 1$, un valor moderado que permite a G_k alejarse de G_{k0} (ver sección 2.1.2 para más detalles).

Las distribuciones a priori para los parámetros de G_0 - μ_ϑ y Σ_ϑ - se escogieron para indicar poca información sobre ellos: ν_ϑ se especificó igual a $R+2$, ya que valores menores implican una distribución Wishart impropia. Por su parte Λ_ϑ se consideró igual a la matriz identidad de dimensión R, que es la esperanza para Σ_ϑ. Para mayores detalles ver Gelman et al. (1995b). Los hiperparámetros vinculados con la parte del ítem - $\mu_\beta, \nu_\beta, s^2_\beta$ - se fijaron también en 0, 5 y 0.1 respectivamente. Esto es, quizás, una distribución a priori un tanto informativa para σ^2_β. Sin embargo, la cantidad de observaciones utilizadas en la aplicación hacen que no haya grandes diferencias con un modelo menos informativo. Además, como las habilidades θ_i y la dificultades de ítems β_j se miden con la escala logística, sumado a que β_1 se fija en 0; esto hace que el soporte relevante de valores para estos dos parámetros raramente se escapan del intervalo $[-3, 3]$. En otras palabras, la estructura del modelo hace que sepamos a priori que σ^2_β no puede ser tan grande. En la sección 3.3 se presenta un estudio de sensibilidad, el cual muestra que el modelo es robusto, es decir que sigue proveyendo buenas estimaciones de G_k ante cambios en los valores de los hiperparámetros.

3.1.3. Los parámetros de interés y su identificación

En esta sección se estudian las condiciones para que los parámetros de interés, $(G_1, ..., G_K, \beta_2, ..., \beta_J)$, estén identificados. Recordemos que todos los individuos con igual valor en las covariables poseen la misma distribución de habilidad a priori, de manera que hay tantas distribuciones G_k como valores distintos de patrones de covariables. Aquí llamaremos *bloque* a cada grupo de individuos con igual patrón de covariables. Supongamos que hay K bloques distintos en la muestra. Cada bloque tiene asociado un valor de la covariable d_k con $k = 1, ..., K$ que indexa a la distribución de habilidad del bloque k: G_k. Como las habilidades provienen de distribuciones distintas según el bloque, los patrones de respuesta pierden la propiedad de ser i.i.d..

Esto impide extender en forma directa el resultado de identificación para el modelo Rasch Semiparamétrico *sin* covariables, que se demuestra en el Teorema 2 de San Martín et al. (2008). Utilizando la condición i.i.d. para los patrones de respuesta, San San Martín et al. (2008) muestran que la identificación del modelo Rasch Semiparamétrico *sin covariables*[2] se da cuando el número de items tiende a infinito, si se fija el valor de uno de los parámetro de dificultad.

Nuestro modelo describe la heterogeneidad entre estudiantes permitiendo que individuos con covariables diferentes tengan también distribuciones distintas. Esto hace que la condición de respuestas i.i.d. propias del modelo Rasch, se relaje a una situación con respuestas independientes, pero no idénticamente distribuidas. Sin embargo, al interior de cada bloque las observaciones son idénticamente distribuidas. Se demuestra la identificación en 2 pasos. Primero se usa el resultado de San Martín et al. (2008) en cada bloque de observaciones i.i.d.. Segundo, los bloques se combinan usando el Teorema 2.1.

Antes de comenzar con la demostración se presenta la notación. La demostración requiere de una única observación por bloque. Esto es debido a que cada bloque conforma un proceso i.i.d. condicional en (β, G_k) con $k = 1, ..., K$. En tal caso, el análisis de identificación es equivalente si se considera una secuencia infinita de individuos por bloque, o sólo un individuo por bloque. En otras palabras en un proceso i.i.d. el tamaño muestral no cumple un rol en el análisis de identificación, para detalles ver Teorema 9.3.12 de Florens et al. (1990). Consideraremos entonces una colección de patrones de respuesta representativo de cada bloque denotada $(Y_1^*, ..., Y_K^*)$. También usamos el símbolo "$\prec$"para notar la relación *"b-identificado por"*, concepto definido

[2]El resultado es válido para cualquier distribución a priori para H, la distribución de habilidades común a todos los individuos de la muestra

en la sección 2.3.

Demostración: Como se mencionó previamente, el Teorema 2.2 de San Martín et al. (2008) asegura que los parámetros $(G_k, \beta_2, ...\beta_J)$ están identificados por Y_k^* condicional en d_k y β_1, cuando $J \to \infty$, o matemáticamente:

$$(G_k, \beta_{2:\infty}) \prec Y_k^* | \beta_1, d_k \tag{3.11}$$

Para unir las K ecuaciones en (3.11) se puede aprovechar el Teorema 2.1 con $Y_k = Y_k^*$, $\gamma_k = (G_k, \beta_{2:\infty})$ y $D = (\beta_1, d_1, ..., d_K)$. Las tres condiciones que requiere el Teorema 2.1 se satisfacen en nuestro modelo pues:

1. $\perp\!\!\!\perp_{1 \leq k \leq K} Y_k^* | G_1, ..., G_K, \beta_{1:\infty}, d_1, ..., d_N$.

   Es decir que las observaciones son mutuamente independientes en el modelo que surge después de integrar con respecto a θ.

2. $Y_k^* \perp\!\!\!\perp G_1, ..., G_K, \beta_{2:\infty} | G_k, \beta_{1:\infty}, d_1, ...d_K$ para $k = 1, ..., K$,

   que significa que Y_1^* sólo depende de $(G_1, \beta_{1:\infty}, d_k)$. En otras palabras, las respuestas de un individuo con covariables d_k dependen sólo de los parámetros de dificultad de la prueba y de la distribución de habilidad asociada a la covariables d_k.

3. $(G_k, \beta_{2:\infty}) \prec Y_k^* | \beta_1, d_1, ..., d_K$ para $k = 1, ..., K$

La última condición surge de aplicar del Teorema 2.2 a la ecuación (3.11) con $\gamma = (G_k, \beta_{2:\infty})$,$Y = Y_k^*$, $D_1 = (\beta_1, d_k)$ y $D_2 = (d_1, ..d_{i-1}, d_{i+1}, ..., d_K)$. La condición 2 se satisface por la ecuación (3.11) y la condición 1 es parte de la estructura del modelo pues suponemos que las respuestas Y_k^* no se ven afectadas por los patrones de covariables de los otros individuos $d_1, ..., d_{k-1}, d_{k+1}, ...d_K$, es decir:

$$d_1, ..., d_{i-1}, d_{i+1}, ..., d_K \perp\!\!\!\perp Y_k^* | G_k, \beta_{1:\infty}, d_k.$$

Luego, el Teorema 2.1 se puede aplicar para concluir:

$$(G_1, ..., G_K, \beta_{2:\infty}) \prec Y_1^*, ..., Y_K^* | \beta_1, d_1, ..., d_K. \tag{3.12}$$

Es importante recalcar que este resultado de identificación es asintótico, es decir que se aplica para el caso en que en que haya infinitos items. Obviamente en la práctica los tests poseen una cantidad finita de items. En San Martín et al. (2008) (Corolario 5.2) también se ha demostrado la consistencia de la distribución de habilidad en el modelo Rasch Semiparamétrico cuando $J \to \infty$ e $I \to \infty$. Este resultado permite afirmar que, cuando se poseen muchos items e individuos, la distribución a posteriori se acerca al verdadero valor de la distribución.

3.2. El algoritmo y la estrategia para la inferencia

Como se ha mencionado en la sección anterior, los parámetros de interés son $(G_1, ..., G_K, \beta_2, ..., \beta_K)$. Recordemos que la ecuación (3.3) especifica a G_k como un elemento aleatorio, ya que es una función de G y σ_θ^2. La especificación aleatoria del parámetro de interés es propia de la estadística Bayesiana. Este enfoque, a su vez, propone hacer inferencia a través de la distribución a posteriori del parámetro de interés. En este caso se trata de obtener $p(G_1, ...G_K, \beta_2, ...\beta_J | \beta_1, Y, D)$. La estrategia para hacer inferencia es la siguiente: en el caso de los parámetros de dificultad, se resume la distribución a posteriori $p(\beta_2, ...\beta_J | \beta_1, Y, D)$ mediante la esperanza a posteriori de cada uno de los parámetros de dificultad como estimador puntual, es decir, $\hat{\beta}_j = E(\beta_j | \beta_1, Y, D)$; e intervalos de 95 % de confianza para los β's, es decir que se estiman los cuantiles $(q_{\alpha/2} - q_{1-\alpha/2})$, con $\alpha = 0.05$. En el caso de las distribuciones de habilidad, se propone utilizar como estimador $\hat{G}_k = E(G_k | \beta_1, Y, D)$. El algoritmo desarrollado es una adaptación del muestreo de Gibbs presentado en Bush y MacEachern (1996). Para implementarlo, se re-expresa el modelo propuesto rompiendo la

mezcla en la ecuación (3.3), e introduciendo variables aleatorias auxiliares ϑ_i, como muestran las ecuaciones (3.13) y (3.14).

$$\theta_i|(\vartheta_i, d_i = d_k, \sigma_\theta^2) \sim N(d_k'\vartheta_i, \sigma_\theta^2) \tag{3.13}$$

$$\vartheta_i|G \sim G \tag{3.14}$$

Aquí ϑ_i proviene de una distribución desconocida G, que a su vez se genera de un proceso Dirichlet. La estructura de muestreo de Gibbs requiere simular de las distribuciones condicionales completas de cada elemento aleatorio en forma alternada, hasta llegar a la convergencia. Una de las formas de optimizar el muestreo de Gibbs consiste en colapsar el espacio estado, es decir, integrar sobre elementos aleatorios que no son de interés. En este caso, es posible evitar la difícil tarea de simular de G marginalizando con respecto a G (ver la sección 2.1.6). Un segundo procedimiento que mejora la eficiencia del muestreo de Gibbs, es reemplazar $\vartheta = (\vartheta_1, ..., \vartheta_I)$ por el par (ϑ^*, s), donde ϑ^* es un vector de tamaño L, $L \leq I$, con los valores o clusters diferentes de ϑ; y s es un vector de tamaño I que indica la pertenencia a un cierto cluster[3]. Sea $a = (Y, \theta_1, ..., \theta_I, \beta_2, ..., \beta_J, \sigma_\theta^2, \mu_\beta, \sigma_\beta^2, \vartheta^*, s, L, \mu_\vartheta, \Sigma_\vartheta)$, el vector con todos los elementos aleatorios relevantes en el muestreo de Gibbs, y sea $\eta = (M, \nu_\beta, s_\beta^2, \nu_\theta, s_\theta^2, \nu_\vartheta, \Lambda_\vartheta)$ el vector con todos los hiperparámetros. Se utiliza , además, la notación a_{-x} para indicar el vector a extrayéndole el componente x. La distribución condicional completa para cada s_i está dada por:

$$p(s_i|a_{-s_i}, \eta, \beta_1, Y, D) \propto q_0\delta_{L^-+1}(s_i) + \sum_{l=1}^{L^-} n_l^- q_l \delta_{s_l}(s_i), \tag{3.15}$$

[3]Nótese que ϑ está relacionado con $L!$ pares diferentes (ϑ^*, s). Esta falta de identificación no es un problema pues no estamos interesados en estimar s.

donde L^- es el número de clusters diferentes una vez que el individuo i es descartado, n_l^- es el número de individuos dentro del cluster l una vez que el individuo i es descartado, $q_0 = M \int \phi(\frac{\theta_i - d_i' x}{\sigma_\theta}) G_0(dx)$, y $q_l = \int \phi(\frac{\theta_i - d_i' x}{\sigma_\theta}) g_l(x) dx$ donde $g_l(x) = [\prod_{i' \in B^-} \phi(\frac{\theta_{i'} - d_{i'}' x}{\sigma_\theta})] g_0(x)\} / \int [\prod_{i' \in B^-} \phi(\frac{\theta_{i'}' - d_{i'}' x}{\sigma_\theta})] G_0(dx)$ y $B^- = \{i' : 1 \leq i' \leq I, s_{i'} = l, i' \neq i\}$. Una vez que se conocen las configuraciones de clusters s, la distribución condicional completa para ϑ_l^* es normal:

$$\vartheta_l^* | (a_{-\vartheta_l^*}, \eta, \beta_1, Y, D) \sim N(\mu_{\vartheta_l^*}, \Sigma_{\vartheta_l^*}), \tag{3.16}$$

donde $\mu_{\vartheta_l^*} = \mu_\vartheta - \Sigma_\vartheta \tilde{d}_l \Sigma_{\tilde{\theta}_l}^{-1} (\tilde{\theta}_l - \tilde{d}_l' \mu_\theta)$, con $\Sigma_{\tilde{\theta}_l} = \sigma_\theta^2 I_{n_l} + \tilde{d}_l' \Sigma_\vartheta \tilde{d}_l$, $\tilde{\theta}_l$ es el vector con todos los elementos de θ que pertenecen a cluster l, y $\tilde{d}_l$ es una matriz de tamaño $R \times n_l$ cuyas columnas son los patrones de covariables de los individuos del cluster l, en el mismo orden que se presentan en $\tilde{\theta}_l$; y $\Sigma_{\vartheta_l^*} = \Sigma_\vartheta - \Sigma_\vartheta \tilde{d}_l \Sigma_{\tilde{\theta}_l}^{-1} \tilde{d}_l' \Sigma_\vartheta$. La mejora al reemplazar (2.4) por (3.15) y (3.16) se refleja en q_l, donde se integra sobre la localización de los clusters ϑ^*, aprovechando la estructura conjugada de esta parte del modelo. Una referencia más detallada sobre este punto se puede ver en MacEachern (1998). Los tiempos de ejecución para calcular q_l, $\mu_{\vartheta_l^*}$ y $\Sigma_{\vartheta_l^*}$ pueden ser reducidos resolviendo las formulas cuadráticas que surgen de la mezcla de dos normales. De esta manera se evita hacer cálculos matriciales, que pueden demorarse en clusters numerosos. Nótese, además, que esta estrategia para colapsar el espacio estado del muestreo Gibbs es dinámico, es decir que debe hacerse en cada iteración del muestreo.

Lamentablemente, el modelo estadístico no es conjugado con respecto a las distribuciones de los parámetros β y θ. Se propone introducir un algoritmo Metropolis Hastings al interior del muestreo de Gibbs para cada uno de estos parámetros θ_i y β_j. La distribución que se muestra en la ecuación (3.17) es la función objetivo en el algoritmo M-H para θ_i. Se trata de la distribución condicional completa de θ_i, que se puede expresar en términos de $t_i = \sum_{j=1}^{J} Y_{ij}$, el estadístico suficiente para θ_i, para

optimizar los cálculos.

$$p(\theta_i|(a_{-\theta_i},\eta,\beta_1,Y,D) \propto exp\{\theta_i t_i - 0.5\sigma_\theta^{-2}(\theta_i - d_i'\vartheta_{s_i}^*)^2\}/\prod_{j=1}^{J}(1+exp\{\theta_i-\beta_j\}) \quad (3.17)$$

La transición elegida en este caso es $q(\theta_n,\theta_{n+1}) \equiv N(\theta_n,\tau_{\theta_n})$, donde $\tau_{\theta_n} = 1/l''(\theta_n)$ y l'' es la segunda derivada de la log-verosimilitud con respecto a θ. La simulación de la condicional completa de β_j es análoga al caso anterior. Se presenta en la ecuación (3.18), también en términos de $t_j = \sum_{i=1}^{I} Y_{ij}$, el estadístico suficiente para β_j.

$$p(\beta_j|(a_{-\beta_j},\eta,\beta_1,Y,D) \propto \exp\{-\beta_j t_j - 0.5\sigma_\beta^{-2}(\beta_j-\mu_\beta)^2\}/\prod_{i=1}^{I}(1+exp\{\theta_i-\beta_j\}) \quad (3.18)$$

En este caso, $q(\beta_n,\beta_{n+1}) \equiv N(\beta_n,\tau_{\beta_n})$, con $\tau_{\beta_n} = 1/l''(\beta_n)$, y l'' la derivada segunda de la log-verosimilitud con respecto a β. El empleo de los estadísticos suficientes t_i y t_j contribuye, también, en la eficiencia del algoritmo. Su uso evita trabajar con la base de datos completa de patrones respuestas Y que es una matriz de $I \times J$, reemplazándola por dos vectores, uno de tamaño I con las sumas por fila de Y, y otro de largo J con las sumas por columna. Las condicionales completas para σ_θ^2, μ_β, σ_β^2, μ_ϑ, Σ_ϑ se muestran en el Apéndice A. Son casos conjugados estándar por lo que no requieren aclaraciones especiales.

Para los fines del presente estudio, se le agrega un paso al algoritmo de MacEachern (1998) para simular de $E(G_k|\beta_1,\eta,Y,D)$. Nótese que la esperanza a posteriori de G_k coincide con la distribución a posteriori de habilidad de un nuevo individuo hipotético con características socioeconómicas d_k. En efecto, como $p(\theta_{I+1}|G_k, d_{I+1} = d_k, Y,\eta,D,\beta_1) = G_k|(Y,\eta,D,\beta_1)$, se tiene que $E(G_k|Y,\eta,D,\beta_1) = E_{G_k}(p(\theta_{I+1}|G_k, d_{I+1} = d_k, Y,\eta,D,\beta_1)) = p(\theta_{I+1}|d_{I+1} = d_k, Y,\eta,D,\beta_1)$. Luego, para obtener estimaciones de la esperanza deseada, basta con estimar la distribución de habilidad de un nuevo individuo de la muestra. Se propone armar una grilla de valores que cubran el soporte de la densidad deseada, para luego estimar el valor de la densidad en cada punto de la

grilla. Siguiendo la notación previa, sea K el número de patrones distintos de covariables. Una vez que se tiene una muestra de $a|(\eta, \beta_1, Y, D)$, la distribución a posteriori de un nuevo individuo de la muestra condicional en $(a, \eta, Y, d_{I+1} = d_k)$ es conocida:

$$p(\theta_{I+1}|a, \eta, Y, d_{I+1} = d_k) = \sum_{j=1}^{N} \frac{n_j}{I+M} \phi(\frac{\theta_{I+1} - d_k' \vartheta_j^*}{\sigma_\theta}) + \int \frac{M}{I+M} \phi(\frac{\theta_{I+1} - d_k' x}{\sigma_\theta}) G_0(dx) \tag{3.19}$$

Se define una grilla de valores de θ_{I+1}. Luego se evalúa la ecuación (3.19) en esos valores. Denotando $f_{I+1}^k = (f_{I+1}^{1k}, f_{I+1}^{2k}, ...f_{I+1}^{Nk})$ los valores evaluados de la ecuación (3.19) en una grilla de tamaño N, el muestreo de Gibbs arroja una muestra de f_{I+1}^k para cada patrón de covariable deseado d_k. Para hacer inferencia sobre $p(\theta_{I+1}|\beta_1, \eta, Y, d_{I+1} = d_k)$ se estima el valor esperado y los cuantiles de 0.025 y 0.975 de f_{I+1}^k en cada punto de la grilla. Los cuantiles se emplean para armar una banda de credibilidad de 95 %.

Finalmente, el algoritmo presentado incorpora, también, un paso adicional para proveer divergencias de Kullback-Leibler (KL), ver Kullback y Leibler (1951). Esta se utiliza como una medida cuantitativa de las diferencias entre densidades. La divergencia de KL se define como:

$$KL(f_1, f_2) = \int f_1 \log\left(\frac{f_1}{f_2}\right) d\mu, \tag{3.20}$$

donde f_1 y f_2 son densidades con respecto a una medida dominante μ. La incorporación de KL dentro del muestro de Gibbs guarda coherencia con el análisis Bayesiano pues, dada la aleatoriedad de las funciones que se desean comparar, las divergencias entre ellas también deben ser consideradas aleatorias. La ecuación (3.20) muestra la divergencia de K-L como un funcional. El enfoque de estimación planteado por Gelfand y Kottas (2002) permite fácilmente estimar funcionales a partir de las densidades estimadas. No sólo divergencias de K-L, sino también medias, medianas, intervalos de confianza, etc. Concretamente, el cómputo de la ecuación (3.20) se realiza a partir de

los valores de dos grillas con distinto patrón de covariable, $f_{I+1}^{k_1}$ y $f_{I+1}^{k_2}$ por ejemplo; evaluando $y_1^{k_1,k_2}, ..., y_N^{k_1,k_2}$, donde $y_n^{k_1,k_2} = log(f_{I+1}^{k_1 n}/f_{I+1}^{k_2 n}) f_{I+1}^{k_1 n}$ para todo $n = 1, ..., N$. Luego se estima la integral en (3.20) mediante:

$$KL^{k_1,k_2} = \sum_{n=1}^{N-1} (\theta_{n+1} - \theta_n) y_n^{k_1,k_2}, \tag{3.21}$$

donde θ_n es un punto de la grilla. Si bien este paso se incorporó dentro del muestreo, puede ser realizado una vez finalizado el muestreo. Este método para evaluar integrales es también empleado para obtener muestras de esperanzas y varianzas de las densidades estimadas. En el caso de las esperanzas, se reemplaza $y_n^{k_1,k_2}$ por $\theta_n f_{I+1}^{nk}$ en la ecuación (3.21); y en el caso de las varianzas, se utiliza $\theta_n^2 f_{I+1}^{nk}$ en lugar de $y_n^{k_1,k_2}$.

Una vez finalizado el muestreo de Gibbs, se dispone de una muestra de f_{I+1}^{k} para cada patrón de covariable de manera que es posible definir KL^{kk}, la divergencia de KL entre densidades al interior de la muestra de f_{I+1}^{k}. En este caso, la divergencia KL^{kk} compara dos realizaciones de una misma densidad, por lo que da una idea de la variabilidad que hay al interior de la muestra de densidades correspondiente al patrón de covariable d_k. El conocer KL^{kk} es útil para determinar si la divergencia entre dos densidades con distinto valor de covariables son relevantes. Para clarificar esta idea, consideremos la comparación entre dos densidades correspondientes al patrón de covariables d_{k_1} y d_{k_2}; y supongamos que las divergencia de KL^{k1k1} y KL^{k2k2} fueran mayores que KL^{k_1,k_2} con alta probabilidad. En tal caso no hay argumentos para afirmar que $f_{I+1}^{k_1}$ y $f_{I+1}^{k_2}$ son diferentes, ya que las divergencias indican que las dos muestras de densidades podrían ser, en realidad, realizaciones de una misma muestra. Concretamente, si se están comparando las densidades correspondientes al patrón de covariables d_{k_1} y d_{k_2}, se propone construir la variable $C^{k_1k_2} = max\{KL^{k_1k_1}, KL^{k_2k_2}\}$, luego se estima $P(KL^{k_1k_2} < C^{k_1k_2})$. Para un cierto valor de esa probabilidad, por ejemplo 0.60, las distribuciones se consideran diferentes. El umbral establecido - 0.60 -, es por cierto

arbitrario. Por esa razón se propone dejar explícito el valor de $P(KL^{k_1k_2} < C^{k_1k_2})$ para permitir al lector la elección de otro umbral.

Para medir el ajuste del modelo se usa el estimador CPO (conditional predictive ordinate) propuesto en Geisser y Eddy (1979). El estadístico CPO es una herramienta muy útil para la selección de modelos que ha sido ampliamente utilizado en la literatura en muchos contextos. Una descripción detallada de cómo calcular el CPO se puede hallar en en Gelfand et al. (1992) o en Chen et al. (2000). En el presente contexto, el estadístico CPO para el i-ésimo individuo y el j-ésimo ítem se define como:

$$CPO_{ij} = E_{\theta,\beta|Y_{(-i)}}(P(Y_{ij} = y_{ij}|\theta_i, \beta_j))$$

donde la esperanza es tomada con respecto a la distribución a posteriori de (θ, β) condicional en $Y_{(-i)}$, es decir las respuestas excluyendo el i-ésimo individuo. Heurísticamente, si θ_i es un valor altamente probable a posteriori para el parámetro de habilidad del individuo i, y β_j un valor altamente probable a posteriori para el parámetro de dificultad de ítem j, un buen modelo debería predecir una alta probabilidad de que el individuo i responda al ítem j como efectivamente lo hizo, si su habilidad fuera θ_i y la dificultad del ítem fuera β_j. De esta forma, valores altos del estadístico CPO sugieren un mejor ajuste. Un resumen estadístico útil del estadístico CPO es el logaritmo de la verosimilitud pseudomarginal (LPML), definido como LPML $= \sum_{i=1}^{I}\sum_{j=1}^{J}\log(\text{CPO}_{ij})$. Para seleccionar el modelo se propone un esquema *paso a paso hacia adelante*[4] basado en los valores LPML para comparar modelos y seleccionar el mejor, considerando todas las covariables disponibles. Para calcular el estadístico LPML se calculan los valores CPO_{ij} usando la simplificación: $\bar{CPO}_{ij} = \{\frac{1}{T}\sum_{t=1}^{T} P(Y_{ij} = y_{ij}|\theta_i^t, \beta_j^t)^{-1}\}^{-1}$, donde T indica el número de simulaciones

[4]Traducción libre de la autora del término en inglés *forward stepwise*.

MCMC. Luego el LPML se calcula como $LPML = \sum_{i=1}^{I} \sum_{j=1}^{J} log(C\bar{P}O_{ij})$.

Por último, al algoritmo esbozado arriba, también se le adicionó un paso para obtener simulaciones de la distribución a posteriori predictiva de los puntajes para cada patrón de covariables $(Y_{I+1}^1, ..., Y_{I+1}^K)$. Serán empleadas en el presente trabajo de aplicación para contrastar el ajuste de nuestro modelo con un modelo lineal jerárquico de componentes de varianza. El paso adicional se describe a continuación. Se simula de una nueva etiqueta s_{I+1} con probabilidad: $P(s_{I+1} = l) = n_l/(I + M)$, si $l = 1, ..., L$; y $P(s_{I+1} = L + 1) = M/(I + M)$. Si $s_{I+1} = l, l \leq L$, entonces $\vartheta_{I+1} = \vartheta_l^*$, si no se simula de ϑ_{I+1} con distribución $N(\mu_\vartheta, \Sigma_\vartheta)$. Luego, para $k = 1, ..., K$, se simula de θ_{I+1}^k con distribución $N(d'^k \vartheta_{I+1}, \sigma_\theta^2)$. Finalmente, se simula de Y_{I+1}^k con distribución $Bern(h(\theta_{I+1}^k - \beta_j))$. El código del algoritmo estará disponible a la brevedad en http://www.paulaestadistica.blogspot.com. Otra versión de este algoritmo se presentará próximamente en el paquete DPpackage para usuarios de R, ver Jara (2007).

3.3. Estudio de simulación

El objetivo de esta sección es probar si el modelo propuesto es capaz de obtener estimaciones satisfactorias de las distribuciones de habilidad. Notemos que la habilidad y su distribución de probabilidad son parámetros *latentes*, es decir que nunca son observados. El Modelo Rasch Explicativo Semiparamétrico extrae información acerca de la habilidad del individuo, una variable continua, a través de una secuencia de respuestas binarias 0-1 individuales. Como se mencionó en la sección 3.1.3, las densidades de interés están identificadas sólo en la situación teórica de una prueba con infinitos items. En este sentido, es un gran desafío lograr reproducir las distribuciones de habilidad.

Por simplicidad se considerará una única covariable categórica con 3 niveles. Los datos poseen 1000 individuos por cada nivel de la covariable, conformando una muestra de 3000 observaciones. Los individuos pertenecientes al nivel 1,2 y 3 están dotados con una habilidad proveniente de las distribuciones:

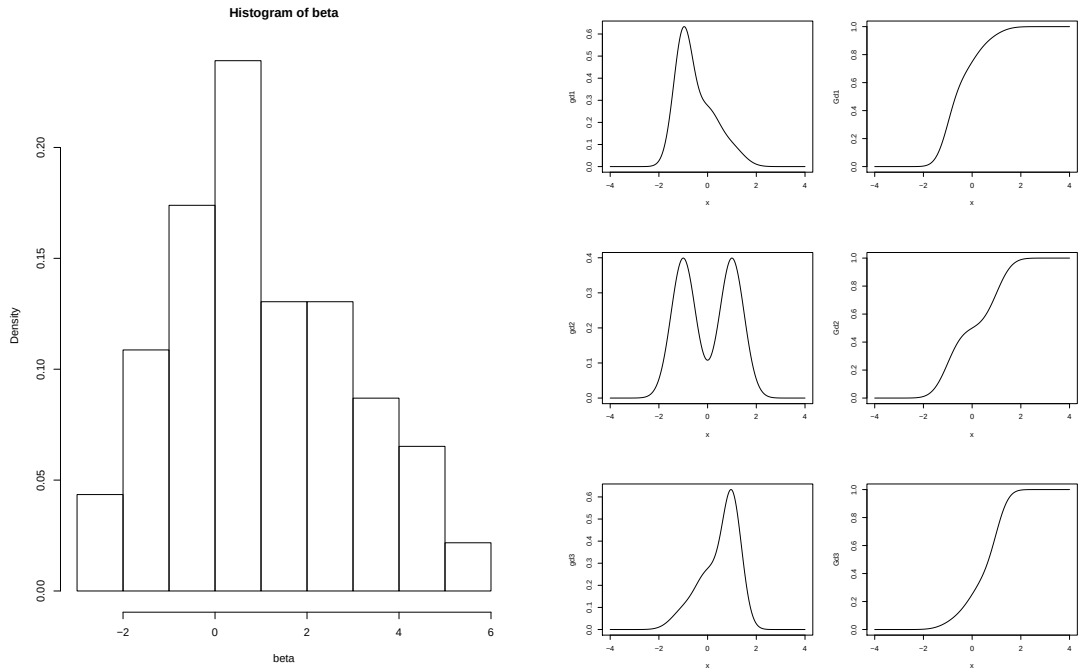
$$f_1(\theta) = 0.6\frac{1}{0.4}\phi(\frac{\theta-(-1)}{0.4}) + 0.3\frac{1}{0.5}\phi(\frac{\theta-0}{0.5}) + 0.1\frac{1}{0.5}\phi(\frac{\theta-1}{0.5})$$
$$f_2(\theta) = 0.5\frac{1}{0.5}\phi(\frac{\theta-(-1)}{0.5}) + 0.5\frac{1}{0.5}\phi(\frac{\theta-1}{0.5})$$
$$f_3(\theta) = 0.1\frac{1}{0.5}\phi(\frac{\theta-(-1)}{0.5}) + 0.3\frac{1}{0.5}\phi(\frac{\theta-0}{0.5}) + 0.6\frac{1}{0.4}\phi(\frac{\theta-1}{0.4})$$

respectivamente. La Figura 3.1(b) presenta las tres densidades. Las distribuciones elegidas difieren entre sí: la primera mezcla de normales es asimétrica a la izquierda, la segunda es bimodal mientras que la tercera es asimétrica a la derecha. El objetivo de las simulaciones es corroborar que nuestro modelo es capaz de detectar estas diferencias distribucionales.

Se considera una prueba con 46 items, el mismo tamaño que en el caso de la aplicación basada en datos SIMCE. Los parámetros β_j se simulan de las prioris definidas por $\sigma^2_\beta \sim inv - \chi^2_5(2)$ y (3.9). La muestra resultante cubre el rango $[-2.7, 5.2]$[5] con un valor medio de 1. La Figura 3.1(a) muestra el histograma de β. Finalmente, (θ, β) se usan para simular una base de datos con las respuestas binarias.

En primer lugar, es de interés estudiar la **sensibilidad del modelo**, es decir el efecto que tiene la elección de hiperparámetros sobre la inferencia. Para este fin, se armaron 50 sub-muestras a partir de la muestra original de 3000 individuos. Cada una de las sub-muestras de tamaño 300 (con 100 observaciones por cada nivel de

[5]Se simularon varias veces hasta obtener una muestra de β que cubriera el rango relevante completo en la escala logit, $[-3, 3]$. Como se mostrará abajo esto es relevante para recuperar las densidades de habilidad.



(a) Histograma de beta.

(b) Densidades y Distribuciones Acumuladas verdaderas. Arriba f_1. En el medio f_2 y abajo f_3

Figura 3.1: Parámetros simulados

la covariable). La Tabla 3.1 columna 9 presenta los resultados de LPML[6] para la muestra completa (3000 individuos), modificando los valores de los hiperparámetros. Por su parte la columna 10 presenta los resultados promedio de LPML de las 50 sub-muestras. En las primeras 4 filas se modifica el valor de M dejando los restantes hiperparámetros fijos. M es el parámetro de masa total del procesos Dirichlet. Cuando $M = \infty$ es equivalente a un modelo paramétrico donde G es reemplazado por G_0, la medida basal del DP. Si bien generalmente las modificaciones de M suelen traer aparejadas alteraciones en la estimación, vemos que esto no ocurre para esta simulación: los cambios en el estadístico LPML son leves tanto para la muestra total como para el promedio de las sub-muestras. La Figura 3.2a,b y c grafica las estimacioes de

[6]LPML es una medida de ajuste del modelo que se explica en la sección 3.2

densidad para el caso de $M = 1$. La línea negra corresponde a la verdadera distribución de donde provienen los datos, mientras que las restantes curvas son estimaciones de la densidad con $M = 1$. Todas las estimaciones son cercanas entre si y cercanas a la densidad verdadera, rescatando exitosamente las formas asimétricas de f_1 y f_3, y la bimodalidad de f_2. Alteraciones en los hiperparámetros ν_θ, ν_β, ν_ϑ, s^2_θ, s^2_θ y Λ_ϑ tampoco producen grandes cambios en la estimación como se ejemplifica las Figuras 3.3 y 3.3. En resumen, observando los valores de LPML y los gráficos se puede afirmar para esta simulación el modelo resulta robusto ante cambios en los hiperparámetros. Debido a restricciones de tiempo no se realizaron más simulaciones, pero sería interesante estudiar la sensibilidad con una simulación que incluya una covariable contínua también.

Modelo	**M**	s^2_θ	ν_θ	s^2_β	ν_β	Λ_ϑ	ν_ϑ	**LPML**	**LPML (pr.)**
(no informativo)	**1**	1	2	1	2	I	5		-5239 (67.50)
	1	0.1	5	0.1	5	I	8	-52207	-5238 (67.30)
	5	0.1	5	0.1	5	I	8	-52210	-5238 (67.18)
	10	0.1	5	0.1	5	I	8	-52212	-5238 (56.18)
	100	0.1	5	0.1	5	I	8	-52217	-5247 (77.24)
	1	**0.5**	5	0.1	5	I	8	-	-5239 (67.33)
	1	**1**	5	0.1	5	I	8	-	-5241 (67.20)
	1	0.1	5	**0.5**	5	I	8	-	-5241 (67.2)
	1	0.1	5	**1**	5	I	8	-	-5238 (67.25)
	1	0.1	5	0.1	5	**0.1I**	8	-	-5239 (67.46)
	1	0.1	5	0.1	5	**10I**	8	-	-5238 (67.31)

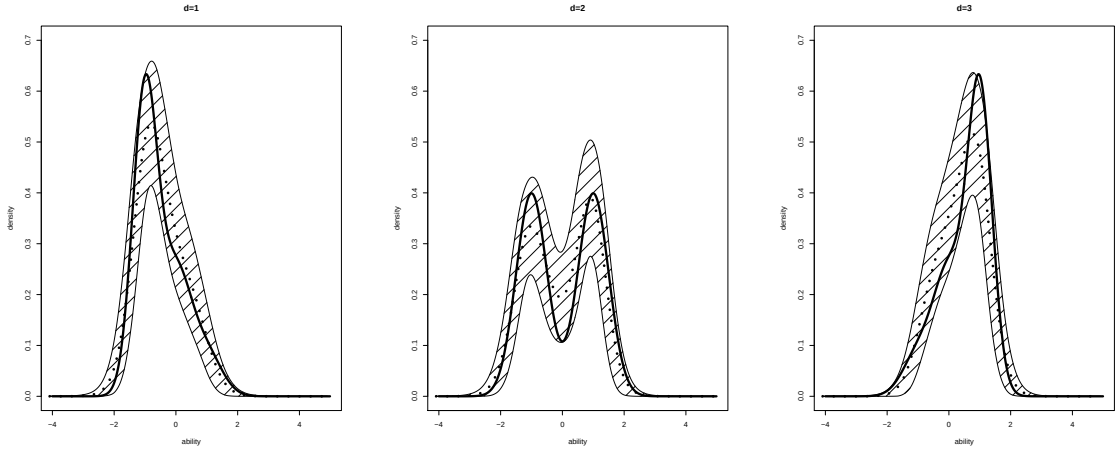
Tabla 3.1: Sensibilidad del modelo. La columna LPML muestra resultados para la base completa de 3000 observaciones. La columna LPML pr. muestra el LPML promedio de estimar las 50 sub-muestras. Entre paréntesis se expresa en error estándar.

En segundo lugar, se presta atención a una situación especial generalmente presente en datos educacionales, esto es, el número de individuos con un patrón específico de covariables difiere. En particular, para los datos chilenos estudiados este fenómeno se verifica, como deja en evidencia la columna 5 de la Tabla 4.8. De hecho sólo 31 niveles de un total de 54 niveles de covariables presentan observaciones. La segunda simulación se denomina aquí estudio de **consistencia del modelo**, y consiste en

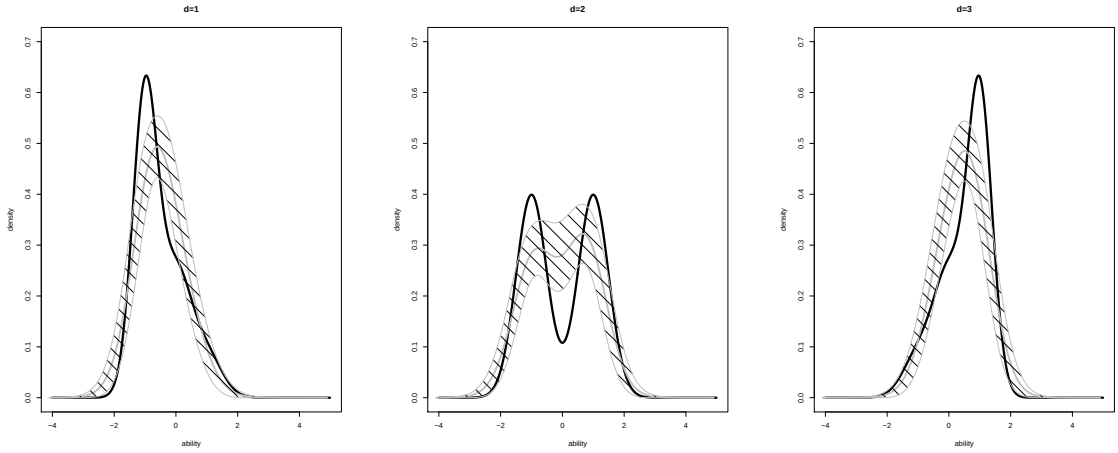
tomar sub-muestras con diferente cantidad de individuos por patrón de covariable y estimar el modelo con las sub-muestras. Como este ejercicio implica variación muestral, también repetimos la estimación 50 veces para cada sub-muestra. La Tabla 3.2 presenta el LPML medio y la Figuras 3.4 muestra algunas las densidades estimadas. En la Figura 3.4a, b y c se presenta un casos con una base de datos con 300 individuos. En este caso, el modelo reconoce las formas funcionales de las densidades a pesar de tener distinta cantidad de individuos por nivel de covariable. La Figura 3.4 d, e y f muestra el caso de 10 observaciones por cada nivel de covariable. La distribución bimodal no se puede reproducir tan fácilmente en el segundo caso.

n° de obs. en nivel 1	**n° de obs. en nivel 2**	**n° de obs. en nivel 3**	**LPML pr.**
10	10	10	-836 (24.34)
100	100	100	-5305 (67.30)
200	90	10	-5195 (53.57)
10	90	200	-5368 (55.34)
200	100	0	-5234 (51,04)
0	100	200	-5231 (64,13)
150	0	150	-5272 (67,16)

Tabla 3.2: Consistencia del modelo. Efecto de un número diferente de individuos por patrón de covariables

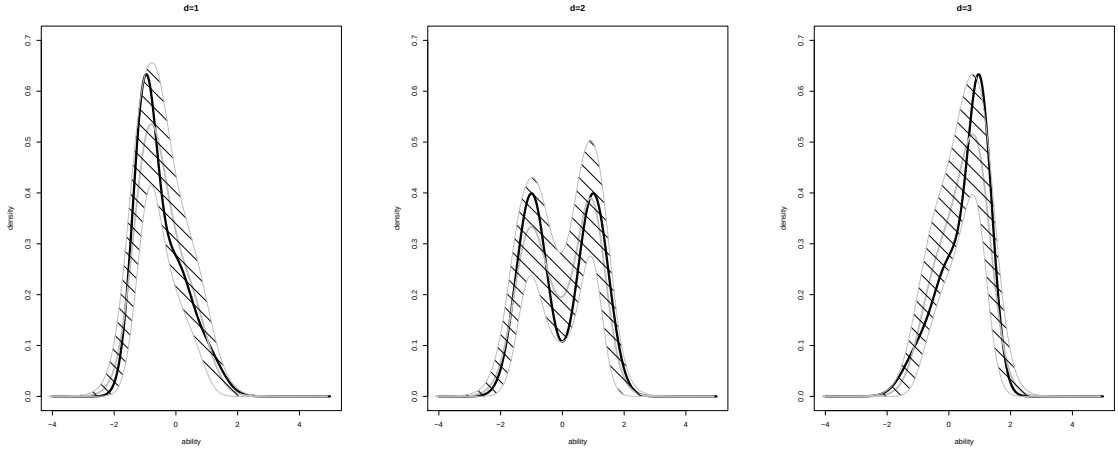


(a) Medias a posteriori de f_1. (b) Medias a posteriori de f_2. (c) Medias a posteriori de f_3.

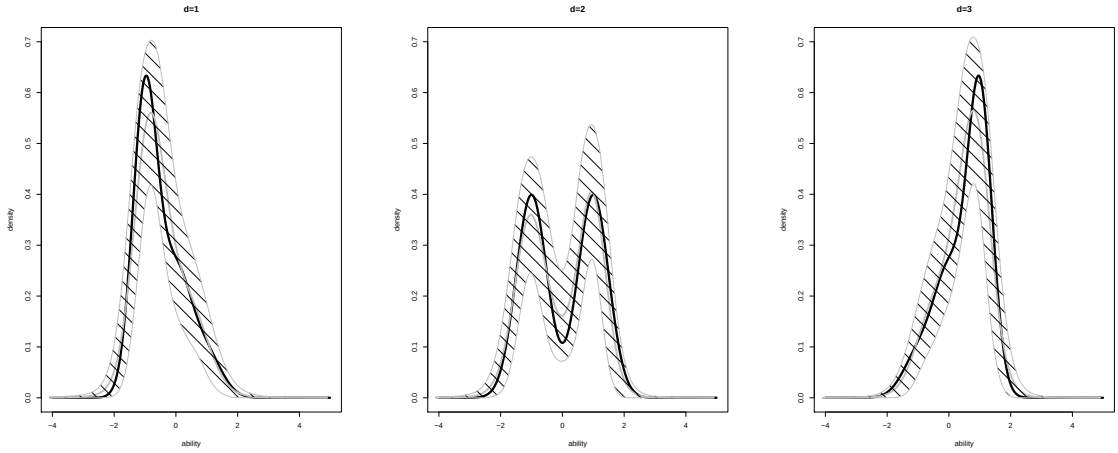


(d) Bandas a posteriori para f_1. (e) Bandas a posteriori para f_2. (f) Bandas a posteriori para f_3.

Figura 3.2: Estudio de sensibilidad. En todos los gráficos las curvas negras gruesas son las distribuciones verdaderas. Los gráficos (a), (b) y (c) muestran los promedios de las 50 cuando M es igual a 1 (curva negra punteada). Los gráficos de abajo: (c), (d) y (e), muestran los promedios de las 50 cuando $s_\theta^2 = 1$ (línea entera gris). Las bandas corresponden a +- 1 error estándar, calculado mediante medias grupales.

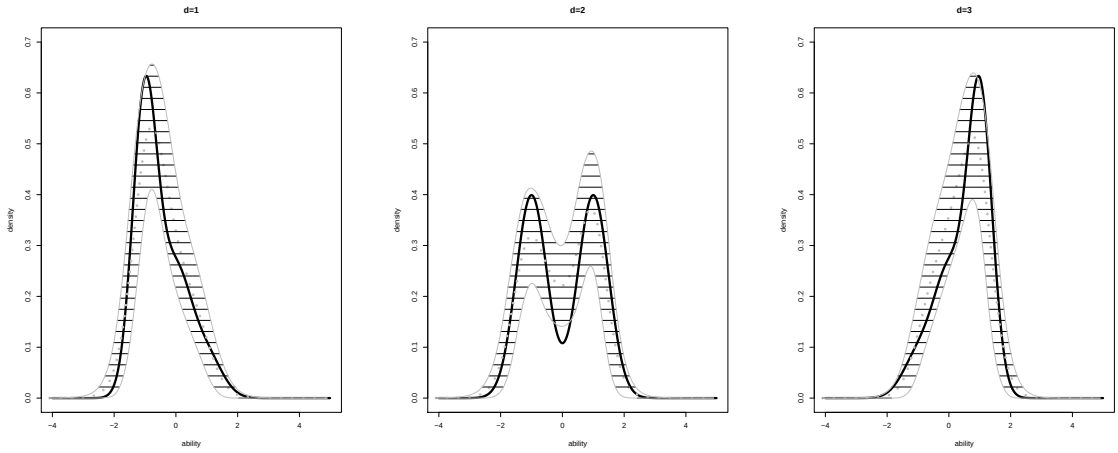


(a) Medias a posteriori de f_1. (b) Medias a posteriori de f_2. (c) Medias a posteriori de f_3.

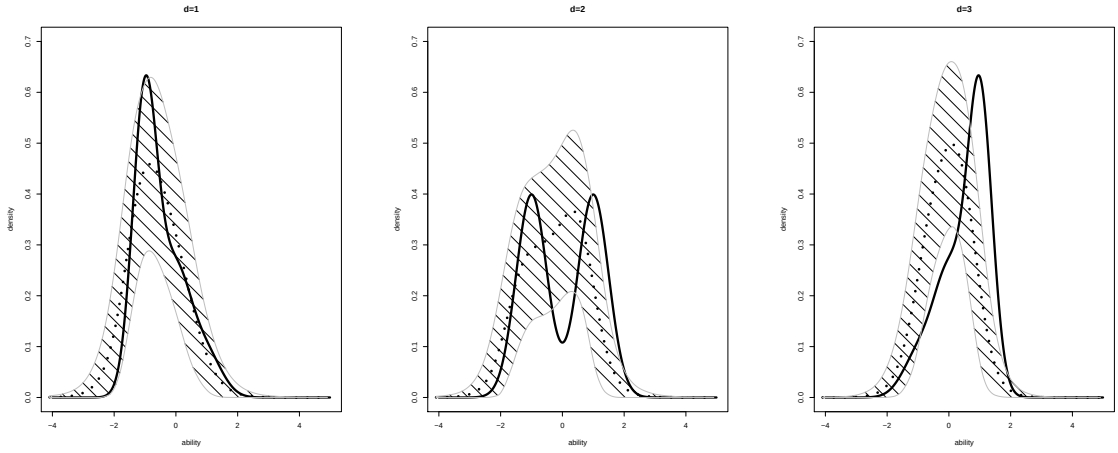


(d) Bandas a posteriori para f_1. (e) Bandas a posteriori para f_2. (f) Bandas a posteriori para f_3.

Figura 3.3: Estudio de sensibilidad. En todos los gráficos las curvas negras gruesas son las distribuciones verdaderas. Los gráficos (a), (b) y (c) muestran los promedios de las 50 muestras cuando s^2_β es igual a 1 (línea gris entera). Los gráficos de abajo: (c), (d) y (e), muestran los promedios de las 50 cuando $\Lambda_\vartheta = 10$ (línea gris entera). Las bandas corresponden a +- 1 error estándar, calculado mediante medias grupales.



(a) Medias a posteriori de f_1. (b) Medias a posteriori para f_2. (c) Medias a posteriori para f_3.



(d) Bandas a posteriori para f_1. (e) Bandas a posteriori para f_2. (f) Bandas a posteriori para f_3.

Figura 3.4: Estudio de consistencia. En todos los gráficos las líneas negras gruesas son las distribuciones verdaderas. Los gráficos (a), (b) y (c) muestran los promedios de las 50 muestras para la estimación con 200 en el nivel 1 de la covariable, 90 observaciones en el nivel 2, y 10 en el nivel 3 (curva gris punteada). Los gráficos de abajo: (c), (d) y (e), muestran los promedios de las 50 muestras para la estimación con 10 observaciones por cada patrón de covariables (linea punteada negra). Las bandas corresponden a +- 1 error estándar, calculado mediante medias grupales.

Capítulo 4

El modelo Rasch Semiparamétrico Explicativo y la prueba SIMCE: Una herramienta a la medida del problema

4.1. Los datos

Esta tesis usa datos del Sistema de Medición de la Calidad de la Educación (SIMCE), un instrumento anualmente aplicado a nivel nacional. Consiste en una prueba de logros de tipo censal que mide hasta qué punto los estudiantes alcanzan los requerimientos básicos establecidos para las diferentes áreas del conocimiento. La prueba SIMCE evalúa a los estudiantes en tres áreas: Lenguaje, Matemáticas y Ciencias. Hasta 2005, la prueba se aplicó en forma alternada a 4to y 8vo grado del nivel básico (9 y 13 años respectivamente); y 2do año de la educación media (16 años). Desde 2006, 4to grado se evalúa todos los años. SIMCE también reúne información sobre profesores y el ambiente familiar a través de una encuesta a profesores y otra a padres. La prueba SIMCE fue creada para medir hasta qué punto los estudiantes logran alcanzar

los objetivos propuestos por las currículos. Esto representa una diferencia conceptual con respecto a los exámenes de tipo PISA que miden conocimientos y habilidades relevantes para alcanzar los desafíos de la vida real sin hacer referencia a un curriculum.

Nuestro estudio se restringe a los datos del SIMCE 2004 para 8vo grado de las comunas de Peñalolén, La Florida y Las Condes. Estas comunas son parte del área metropolitana de Santiago y cubren un amplio rango socioeconómico. En cuanto al año, 8vo grado es el final de la educación básica por lo que es el momento apropiado para evaluar el desempeño global de esta etapa. A cada estudiante se le asignó aleatoriamente una de las dos diferentes formas de la prueba, C y D. En este trabajo nos concentramos sólo en la forma C. Durante la prueba, los examinados debían responder 47 preguntas (de aquí en adelante se denominan *items*) en 90 minutos. Sólo uno de los items era una pregunta abierta. Los restantes 46 items eran de selección múltiple, cada uno con 4 posibles opciones. Cada ítem poseía sólo una respuesta correcta. Se utiliza una variable binaria para cada persona y cada ítem, para indicar si la respuesta es correcta o no. El puntaje observado del individuo se mide aquí a través de la suma de respuestas correctas.

Los datos disponibles incluyen también covariables provenientes de la encuesta de padres. *Tipo de colegio* es una variable categórica que indica la fuente de financiamiento del colegio. Posee tres niveles: colegios municipales, particulares subvencionados y particulares pagados. Junto con tipo de colegio se incluye una covariable categórica llamada *selección* para controlar la selectividad de los colegios. Se construye un índice socioeconómico basado en la escolaridad del padre y la madre (de 14 niveles cada una), e ingreso familiar (15 niveles). Esta covariable está disponible a nivel individual (SES) y de colegio (SESp). Finalmente la variable *repite* indica si el estudiante repitió al menos un año. Detalles sobre la definición de las covariables se puede ver en la Tabla 4.1.

Se consideraron solamente colegios con al menos 20 observaciones (estudiantes) para tener una medida aceptable de la variable SESp. El número total de examinados resultó de 6.909 pertenecientes a un total de 228 colegios: 127 en La Florida, 61 en Las Condes y 40 en Peñalolén. Sin embargo, a pesar que se pidió contestar la totalidad del formulario sólo el 81 % de los estudiantes lo completó en su totalidad. Esto es un escenario normal en los exámenes SIMCE pues el resultado de la prueba no tiene ninguna influencia para los estudiantes. Por otra parte, los datos faltantes en el cuestionario de padres asciende a 25 % de la muestra, incluso después de implementar una práctica habitual de imputar datos faltantes de SES con los datos de SESp. Los datos restantes suman 3.863 individuos, la subpoblación considerada aquí.

La Figura 4.1 presenta histogramas suavizados de los puntajes SIMCE para cada tipo de colegio. El desempeño de los estudiantes provenientes de colegio municipales es claramente más pobre que el desempeño de estudiantes de colegios particulares pagados, los de colegios particulares subvencionados ocupan una posición intermedia. Las formas de los histogramas también cambian en gran manera: los estudiantes de colegios públicos presentan un histograma asimétrico a la izquierda sugiriendo que la prueba fue difícil para esta sub-población. Los examinados de colegios subsidiados son un caso intermedio con un histograma levemente asimétrico a la izquierda. Finalmente, la prueba fue relativamente fácil para los estudiantes de colegios privados que presentan un histograma asimétrico a la derecha. La Tabla 4.2 provee estadísticos descriptivos de los puntajes.

La Tabla 4.3 muestra que el status socioeconómico a nivel individual y de colegio cambia abruptamente entre tipos de colegio. Además, el 44 % de los estudiantes están expuestos a procesos de selección para ingresar a los colegios. La selección es la regla

Variable	tipo	característica
tipo de colegio	categórica (3 niveles)	nivel 1: colegio municipal nivel 2: colegio particular subvencionado nivel 3: colegio particular pagado
SES	índice	promedio de la escolaridad del padre (pescol) y de la madre (mescol), y del ingreso familiar (ingreso)
SESp	índice	promedio a nivel colegio del índice de nivel socioeconómico (SES)
selección	binaria	1 si el colegio selecciona* a los estudiantes basado en una prueba o en comportamiento durante una sesión de juegos y 0 si no
repite	binaria	1 si el estudiante repitió al menos un año y 0 si no
mescol	categórica (14 niveles)	escolaridad de la madre nivel 1: enseñanza básica incompleta nivel 2: enseñanza básica completa nivel 3: enseñanza media incompleta (H) nivel 4: enseñanza media incompleta (TP) nivel 5: enseñanza media completa (H) nivel 6: enseñanza media completa (TP) nivel 7: centro de formación técnica incompleto (CFT) nivel 8: instituto profesional (IP) incompleto nivel 9: CFT completo nivel 10: IP completo nivel 11: Universitario incompleto nivel 12: Universitario completo nivel 13: Diplomado o Postítulo nivel 14: Magister o PHD
pescol	categórica (14 niveles)	escolaridad del padre. Idem mescol
ingreso	categórica (15 niveles)	ingreso familiar nivel 1: menos de CP 100.000 nivel 2: entre CP 100.000 y CP 200.000 nivel 3: entre CP 201.000 y CP 300.000 nivel 4: entre CP 301.000 y CP 400.000 nivel 5: entre CP 401.000 y CP 500.000 nivel 6: entre CP 501.000 y CP 600.000 nivel 7: entre CP 601.000 y CP 800.000 nivel 8: entre CP 801.000 y CP 1.000.000 nivel 9: entre CP 1.001.000 y CP 1.200.000 nivel 10: entre CP 1.201.000 y CP 1.400.000 nivel 11: entre CP 1.401.000 y CP 1.600.000 nivel 12: entre CP 1.601.000 y CP 1.800.000 nivel 13: entre CP 1.801.000 y CP 2.000.000 nivel 14: entre CP 2.001.000 y CP 2.200.000 nivel 15: más de CP 2.200.000

Tabla 4.1: Descripción de las covariables

* Se consideró que el colegio seleccionaba a través de una prueba si más de la mitad de los estudiantes de ese colegio lo afirmaban. Lo mismo en el caso de selección a través de una sesión de juegos.

Comuna	**Tipo de colegio**	**Puntaje Promedio***	**% de inscripción**
La Florida	Total	23.63	55.70
		(8.47)	
	municipal	19.75	26.85
		(7.25)	
	particular subvencionado	24.81	67.57
		(8.38)	
	particular pagado	28	5.58
		(8.55)	
Penalolén	Total	29.73	20.17
		(8.95)	
	municipal	19.56	9.66
		(7.00)	
	particular subvencionado	22.98	11.36
		(8.79)	
	particular pagado	31.94	78.98
		(8.14)	
Las Condes	Total	31.82	24.13
		(8.43)	
	municipal	22.34	44.73
		(7.26)	
	particular subvencionado	29.45	40.47
		(8.20)	
	particular pagado	33.32	14.81
		(7.74)	
Total	Total	25.43	100
		(9.29)	
	municipal	19.91	26.3
		(7.21)	
	particular subvencionado	24.77	48.5
		(8.55)	
	particular pagado	32.49	25.2
		(8.08)	

Tabla 4.2: Puntaje Observado Promedio por tipo de colegio. Entre paréntesis se muestra el error estándar. El total de individuos considerado es 3863

* El puntaje se define como el número de respuestas correctas

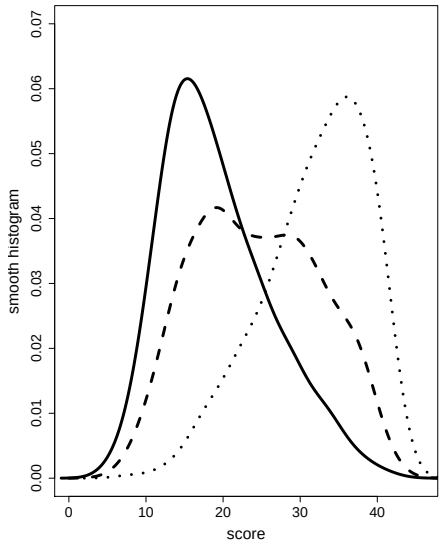


Figura 4.1: Histograma de los puntajes por tipo de colegio. La línea punteada corresponde a colegios particulares pagados, la continua a colegios municipales y la línea cortada muestra a los colegios particulares subvencionados.

para los colegios particulares pagados. Los colegios municipales generalmente no seleccionan, pero el 40 % de los estudiantes de colegios particulares subvencionados son elegidos en la práctica, a pesar de que la selección no está permitida para este tipo de instituciones. La Tabla 4.3 pone en evidencia también que la intecambiabilidad de los patrones de respuesta no se verifica en estos datos: una vez que el tipo de colegio se incluye en el análisis la media de los puntajes totales cambia por lo que no pueden entenderse como provenientes de una misma distribución. Se sigue entonces que una única distribución de habilidad para la población completa de estudiantes es inapropiada.

4.2. Detalles sobre la implementación del algoritmo

Como se explicó en la sección 3.2, la estrategia para hacer inferencia es la siguiente: en el caso de los parámetros de dificultad, se presenta la distribución a posteriori $p(\beta_2, ...\beta_J|\beta_1, Y, D)$, y la esperanza a posteriori de cada uno de los parámetros de dificultad como estimador puntual, es decir, $\hat{\beta}_j = E(\beta_j|\beta_1, Y, D)$. En el caso de las distribuciones de habilidad, se propone utilizar como estimador $\hat{G}_k = E(G_k|\beta_1, Y, D)$. La presencia de dos covariables continuas, SES y SESp, implica una colección muy numerosa de densidades G_k. Para fines expositivos, se evalúan dichas densidades en valores específicos de las variables continuas. La elección de dichos valores se hizo de la siguiente manera: en el caso de la variable SES, se dividen a los estudiantes en tres grupos: nivel socioeconómico bajo (valores de SES hasta 4); nivel socioeconómico medio (valores de SES entre 4 y 9); y nivel socioeconómico alto (valores de SES mayores que 9). En cada grupo, se promedian los valores de SES, y se escoge el promedio al interior de cada grupo para evaluar las densidades a posteriori. El mismo procedimiento se consideró para la variable SESp. Concretamente, los valores donde se evaluaron las distribuciones predictivas son 4.768, 6.227 y 7.686 para SESp bajo, medio y alto respectivamente; y 2.626, 6.204 y 11.657 para SES bajo, medio y alto, respectivamente. Este método resultó en un total de 54 patrones distintos de

Variable	Municipal	Particular Subvencionado	Particular Pagado	Total
N° obs.	1098	1798	967	3863
SES	3.052	5.568	11.739	6.398
	(1.468)	(2.534)	(1.356)	(3.835)
SESp	3.05	5.567	11.736	6.396
	(0.826)	(2.045)	(0.901)	(3.604)
selección	0.056	0.407	0.945	0.442
	(0.231)	(0.491)	(0.228)	(0.497)
repite	0.189	0.099	0.043	0.111
	(0.392)	(0.299)	(0.204)	(0.314)

Tabla 4.3: Número de observaciones, media y error estándar (entre paréntesis) de las covariables

covariables, pero sólo 31 tienen observaciones disponibles[1]. Restringimos el análisis a estas últimas.

La sección 3.2 también presenta el esquema del algoritmo empleado para obtener las estimaciones a posteriori. El código del algoritmo se escribió en lenguaje C para lograr mayor velocidad de ejecución y al mismo tiempo facilitar el traspaso a otros lenguajes y sistemas operativos. Las salidas del código no se limitan a las estimaciones empleadas para esta aplicación $(\hat{G}_1, ..., \hat{G}_{54}, \hat{\beta}_2, ...\hat{\beta}_{46})$, sino que también incluyen simulaciones provenientes de la distribución conjunta a posteriori de los parámetros $(\theta_1, ..., \theta_{3863})$, y de la distribución a posteriori predictiva de los puntajes $(Y^1_{I+1}, ..., Y^{54}_{I+1})$. Si bien en el presente trabajo de aplicación sólo se emplean para contrastar el modelo propuesto con un modelo HLM de componentes de varianza (ver sección 4.3.1), es importante mencionar que la inferencia a posteriori predictiva de los puntajes es posible a través de este enfoque, y el código desarrollado arroja las estimaciones para hacerlo.

El código también deja a disposición del usuario estimaciones a posteriori de los parámetros $\sigma^2_\theta, L, n, s, \vartheta^*_1, ..., \vartheta^*_L, \mu_\vartheta, \Sigma_\vartheta, \mu_\beta$ y σ^2_β, sin embargo, su correcto empleo requiere un estudio previo de identificación para estos parámetros, que excede los alcances de este estudio. Como se mencionó en la sección 2.3, hay que ser cuidadoso en el uso de resultados a posteriori, ya que el algoritmo *siempre* arroja una distribución a posteriori de los parámetros involucrados en el muestreo de Gibbs, aunque éstos no estén identificados. En tal caso, la actualización no es real y no es correcto hacer inferencia sobre esos parámetros.

[1]Los conteos de los individuos por patrón de covariable se presentan en la columna 5 de la Tabla 4.8.

Se corrió una única cadena (usando puntos iniciales provenientes de las distribuciones a priori) con un período de quema de $10,000$ iteraciones y dejando 200 iteraciones entre medio, para producir $2,000$ simulaciones de la distribución a posteriori conjunta de los parámetros. El muestreo de Gibbs debe actualizar 3863 parámetros de habilidad y 45 parámetros de dificultad. La característica condicional de las distribuciones involucradas en el muestreo, implica que haya mucha correlación entre estos parámetros. Antes de descartar 200 iteraciones intermedias, las cadenas tienden a ir juntas, mostrando formas similares. Las tasas de rechazo de los pasos de Metropolis-Hasting para β's y θ's resultaron de 29 %.

En cada iteración del muestreo de Gibbs se actualizan 7802 parámetros más L parámetros auxiliares ϑ_l^*, donde $\bar{L} = 10.75$. La actualización de las etiquetas de cluster - s - es la parte del muestreo que demanda más tiempo. Además, se estiman funciones de los parámetros involucrados en el muestreo: por cada patrón de covariable se evalúan 2000 puntos de la densidad predictiva[2] ($54 \times 2000 = 108000$ puntos totales), y se calculan 1431 divergencias de Kullback-Leibler $KL^{k_1 k_2}$ $k_1 = 1, ..., 53, k_2 = k_1 + 1, ..., 54$. El tiempo de ejecución fue de 6 días, 1 hora y 44 minutos en un procesador xeon(4) 2.80Ghz, 2Gb RAM. Se emplearon criterios estándar para medir la convergencia como los presentados en el paquete BOA, Smith (2004). Los resultados de convergencia para algunos parámetros se presentan en el Apéndice D. Finalmente, el código arroja también el valor del estadístico $LPML$.

[2]Se eligieron $2,000$ puntos en una grilla sobre el intervalo $[-4, 5]$, que cubre el soporte de todas las distribuciones de interés.

4.3. Resultados

4.3.1. Comparación de modelos

Empleando el procedimiento paso a paso, detallado en la sección 3.2, se escoge un modelo con SESp, SES, *repite* y *tipocol* como covariables. La Tabla 4.4 presenta los resultados de dicho procedimiento. Los mejores modelos en cada paso están en negrita. El paso 0 corresponde al modelo semiparamétrico sin covariables. Éste se denomina *modelo nulo* y su estructura se explica en la sección 3.1. En el Paso 1, el índice de nivel socioeconómico a nivel colegio, SESp, es la variable más explicativa. Este resultado, frecuentemente encontrado en sistemas educacionales, significa que el efecto nivel socioeconómico de los pares es más importante para el desempeño individual que sus propio nivel socioeconómico. En este paso se incluyeron dos modelos extra para decidir si las variables continuas, SES y SESp, se incorporaban en el análisis en forma lineal o cuadrática. Los valores de LPML en ambos modelos sugieren que una relación lineal es más apropiada para ajustar los datos. La variable que se incorpora en el paso 2 es SES, el nivel socioeconómico a nivel individual. En el paso 3, la covariable con menor valor de LPML es *tipo de colegio*. Sin embargo, las diferencias entre este modelo y el que incorpora a la covariable *repite*, son muy pequeñas. Por esta razón, se considera un modelo extra en el paso 4, para dejar disponible al lector los resultados del procedimiento si se escogiera la variable repite en el paso 3, en lugar de tipo de colegio. Los resultados de los pasos 4 y 5 son más concluyentes, señalando al modelo que descarta la variable selección, e incluye a *tipo de colegio* y *repite*. En resumen, el procedimiento sugiere un modelo con SESp, SES, *repite* y *tipo de colegio* como covariables.

El paso 6 del procedimiento se agregó para verificar el efecto de la elección de M, - el parámetro de masa total del DP -, sobre la estimación. Se consideran los

valores $M = 1$ (el modelo propuesto), y $M = 5, 100$, y M tendiendo a ∞. Se puede notar que el modelo ajusta mejor cuando $M = 5$, pero lo hace peor para $M = 100$, y marcadamente peor si $M \to \infty$. Este último caso corresponde al *modelo basal*, que surge de reemplazar G por G_0 en la ecuación (3.4). Los detalles sobre el modelo basal se pueden ver en las secciones 3.1 y 3.1.2. La Figura 4.2 compara las densidades estimadas con los 4 distintos valores de M. Se considera el subgrupo de alumnos de colegios particulares subvencionados que no repiten, para simplificar la exposición, pero los restantes subgrupos también muestran estimaciones muy cercanas para los casos en que M es igual a 1 y 5; y difieren para los casos $M = 100$ y $M = \infty$. El modelo paramétrico estima distribuciones simétricas y unimodales en todos los casos, es decir, que no es capaz de distinguir diferencias en las formas de las distribuciones. El modelo con $M = 100$ es un caso intermedio entre el modelo basal y el semiparamétrico propuesto ($M = 1$), pues estima densidades mas suavizadas que el modelo propuesto, pero es más flexible que el modelo basal. Los resultados presentados en la siguiente sección se basan en el modelo con M igual a 1. Si bien el valor de $LPML$ es menor para el modelo con $M = 5$, la cercanía de las estimaciones no generan alteraciones en las conclusiones si se emplea el primer modelo. Las comparaciones entre el modelo basal y los restantes dejan en evidencia que la inferencia basada en comparar densidades sólo tiene sentido hacerla desde una perspectiva semiparamétrica, donde las distribuciones de habilidad son consideradas el parámetro aleatorio a estimar. En un modelo paramétrico, la distribución a priori de la habilidad ejerce una influencia mucho mayor sobre la distribución a posteriori, restringiendo su forma. Nótese, incluso, que el modelo nulo funciona mucho mejor que el modelo basal (valores de LPML menores). Esto sugiere que una estructura semiparamétrica es superior a una paramétrica para ajustar los datos, aún si las covariables no se incluyen.

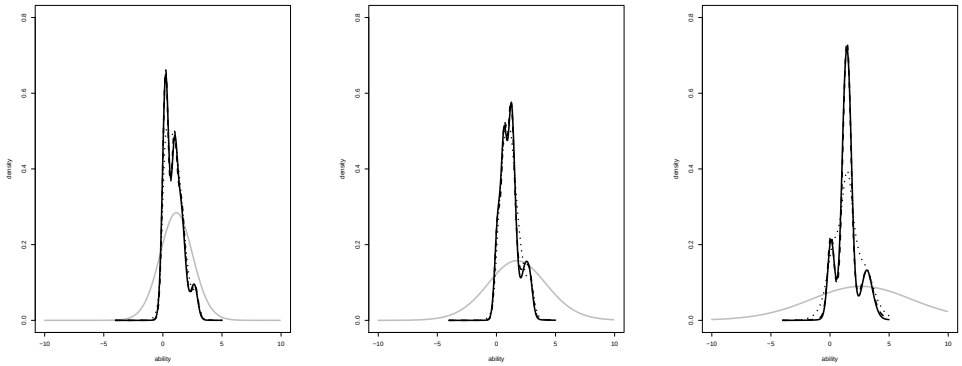
Paso	Modelo	LPML*
0	modelo nulo	155.44
1	tipo de colegio	88.59
	SES	35.35
	SES y SES2	42.12
	SESp	**31.76**
	SESp y SESp2	49.29
	repite	128.73
	selección	109.31
2	SESp tipo de colegio	38.72
	SESp SES	**16.90**
	SESp repite	19.66
	SESp selección	31.83
3	**SESp SES tipo de colegio**	**10.92**
	SESp SES repite	10.97
	SESp SES selección	16.08
4	**SESp SES tipo de colegio repite**	**9.23**
	SESp SES tipo de colegio selección	16.20
	SESp SES repite selección	11.84
5	SESp SES repite tipo de colegio selección	14.06
6	**SESp SES tipo de colegio repite M=5**	**-0.984**
	SESp SES tipo de colegio repite M=100	27.812
	modelo basal SESp SES repite tipo de colegio	239.42

Tabla 4.4: Resultados del procedimiento paso a paso hacia adelante. * Se presenta el estadístico $-(LPML + 96,8000)$ para simplificar la comparación. Valores más pequeños implican mejor ajuste.

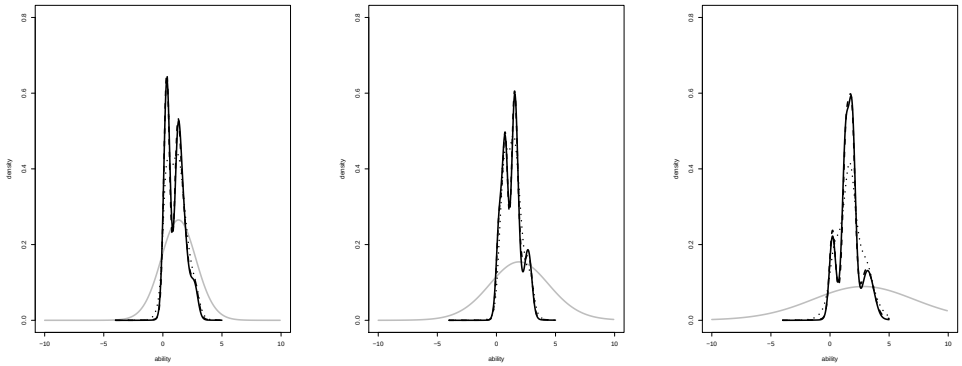
Como una forma de comparar el ajuste de nuestro modelo con respecto a los modelos lineales comúnmente usados, se presenta la Figura 4.3. En ella se comparan las distribuciones de los puntajes observados por patrón de covariables, con la distribución a posteriori predictiva de los puntajes del modelo escogido, y la distribución de los puntajes que se desprenden de un modelo HLM. El modelo HLM escogido es el *Modelo de Componentes de Varianza*, especificado en las ecuaciones (1.1) y (1.2). Los niveles considerados son individuos y colegios , y se emplean las mismas covariables del modelo propuesto. Como se explica en la sección 1.4, los modelo HLM suponen una distribución normal para los puntajes cuya media y varianza están dadas por las ecuaciones en (1.3). Para estimar el modelo jerárquico, se empleó el procedimiento MIXED de SAS, SAS Institute (1999). Las estimaciones resultantes para b, τ y σ^2 son $\hat{b}_{constante} = 13.018$, $\hat{b}_{SESp} = 1.09$, $\hat{b}_{SES} = 0.539$, $\hat{b}_{repite} = 3.592$, $\hat{b}_{tipocolM} = -1.022$, $\hat{b}_{tipocolPP} = -2.235$, $\hat{\tau} = 5.925$ y $\hat{\sigma}^2 = 48.031$[3]. La particularidad de este modelo es que descompone la varianza en dos partes: σ^2 y τ, que son interpretados como la variabilidad intra y entre colegios. Para la submuestra considerada, que se restringe a los colegios de Peñalolén, la Florida y Las Condes; la variabilidad intra colegios es mayor que la variabilidad entre colegios. La distribuciones de los puntajes por patrón de covariables que predice el modelo HLM son normales con media $d_k\hat{b}$, y varianza $\hat{\sigma}^2 + \hat{\tau} = 53.956$. En la Figura 4.3 también se considera, sin pérdida de generalidad, el subgrupo de alumnos de colegios particulares subvencionados que no repiten. Se excluyen los dos casos en que hay una sola observación: el subgrupo con SESp bajo y SES alto, y el subgrupo con SESp alto y SES bajo. Para mostrar la densidad de puntajes observados se utilizó la función `density` de R. Los valores de SES y SESp para el HLM se evaluaron en los mismos puntos empleados para el modelo propuesto, que se describen en la sección 4.2. Se puede observar que los casos en que la distribución de puntajes observados se aleja de la normalidad, el modelo propuesto se acerca

[3]Los restantes detalles de la estimación se encuentran en el Apéndice B.

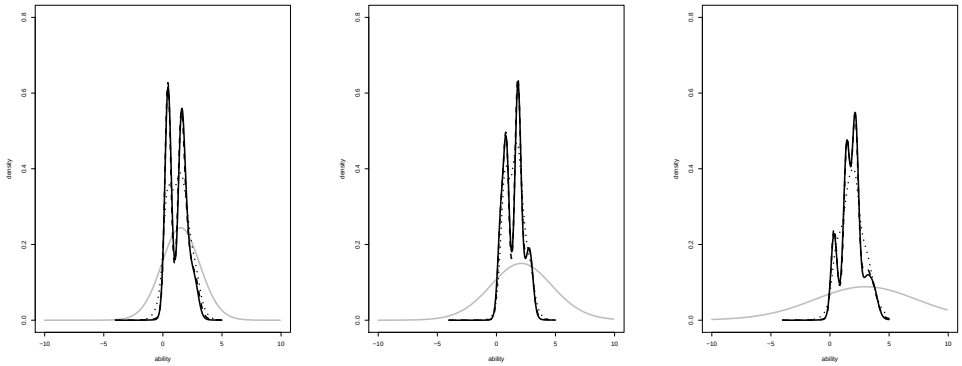
más a la distribución de puntajes observados, que el modelo HLM. La inspección de densidades predichas versus observadas en tales casos (gráficos 4.3c-4.3f), deja en evidencia la superioridad del modelo no paramétrico para ajustar los datos. Por otro lado, en los casos en que la densidad de puntajes observados toma una forma simétrica unimodal, como es el caso de de los gráficos 4.3a y 4.3b, el modelo HLM ajusta con la misma eficacia que el modelo paramétrico.



(a) SESp y SES bajo (b) SESp bajo y SES medio (c) SESp bajo y SES alto



(d) SESp medio y SES bajo (e) SESp y SES medio (f) SESp medio y SES alto

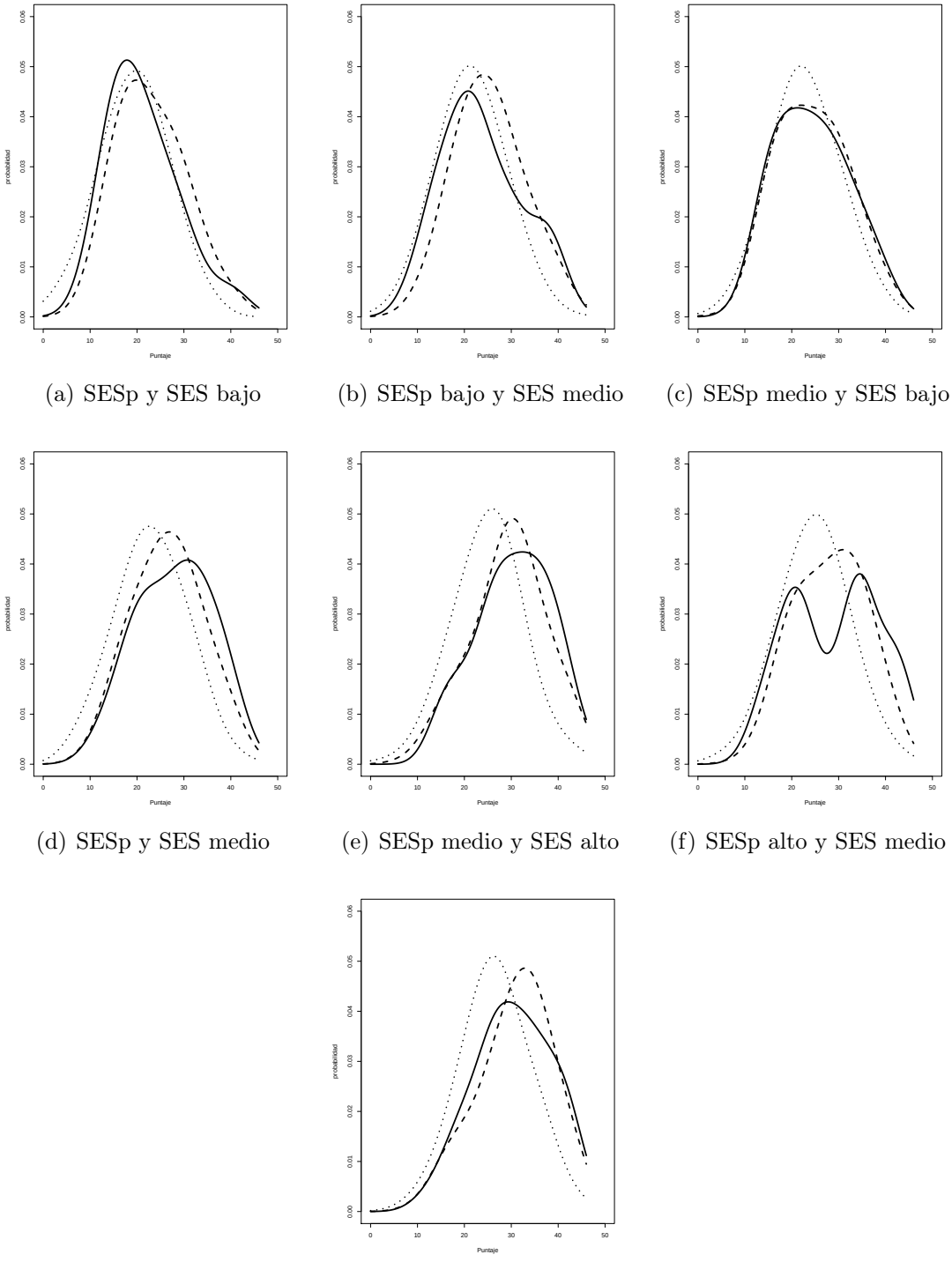


(g) SESp alto y SES bajo (h) SESp alto y SES medio (i) SESp y SES alto

Figura 4.2: Gráficos Comparativos de modelos con diferentes valores de M. Todos los gráficos son para alumnos de colegios particulares subvencionados que no repiten y presentan $\hat{G}_k$, las densidades estimadas para $E(G_k|Y, d_k)$. En todos los gráficos, la curva entera negra es la densidad esperada a posteriori que surge del modelo propuesto ($M = 1$), la curva rayada corresponde al caso de $M = 5$, y la punteada al caso $M = 100$. Finalmente la curva entera gris muestra las estimaciones para el modelo basal ($M = \infty$).

En lo que se refiere a la forma de caracterizar los subgrupos de alumnos, el modelo HLM presentado permite inferir acerca de las medias de los puntajes por grupo, y determinar qué factores son relevantes para determinar dichas medias a través de test de hipótesis. Los tests para el modelo de componentes de varianza se encuentran en la tabla `Solución para efectos fijos` del apéndice B. Indican que *tipo de colegio* es la única covariable no significativa a un nivel de confianza de 95 %. La tabla 4.5 muestra las estimaciones de las medias para cada subgrupo. La columna 8 muestra las medias para el modelo HLM, y la 6 para el modelo propuesto. En general, el modelo propuesto, predice medias para los puntajes más cercana a las medias observadas por grupo, a pesar de que el modelo HLM corre con la ventaja de incorporar la pertenencia al colegio, información que no es utilizada en nuestro modelo. Si bien en el modelo HLM empleado, la varianza estimada es común a todos los subgrupos ($\sigma^2 + \tau$), esta restricción puede sortearse desde el enfoque de los modelos lineales, especificando una matriz de varianzas covarianzas con valores distintos en la diagonal dependiendo del subgrupo. Pero el supuesto de normalidad sólo deja espacio para estimar las medias y varianzas para las distribuciones por grupo. Sin embargo, los datos disponibles poseen más información que la media y la varianza, que no es tenida en cuenta en los modelos lineales[4].

[4]González y San Martín (2009) mejoran estas limitaciones empleando una metodología que complementa estimaciones HLM con estimadores Kernel.



(a) SESp y SES bajo (b) SESp bajo y SES medio (c) SESp medio y SES bajo

(d) SESp y SES medio (e) SESp medio y SES alto (f) SESp alto y SES medio

(g) SESp y SES alto

Figura 4.3: Gráficos Comparativos entre el modelo HLM y el propuesto, para alumnos de colegios particulares subvencionados que no repiten. En todos los gráficos, la curva entera es un histograma suavizado de los puntajes observados, la curva punteada es la distribución de puntajes que se desprende del modelo HLM, y la curva rayada es la densidad posteriori esperada de los puntajes que surge del modelo propuesto.

t. de colegio	repite	SESp	SES	media obs.	media mod.prop.	s.e. mod.prop.	media HLM
municipal	si	bajo	bajo	16.852	18.584	5.719	19.634
					(0.412)	(0.342)	
			medio	16.3	20.716	6.261	21.562
					(0.41)	(0.377)	
		medio	bajo	23.5	20.115	6.099	21.225
					(0.538)	(0.464)	
			medio	22.75	22.258	6.465	23.153
					(0.45)	(0.388)	
	no	bajo	bajo	20.177	22.506	7.273	23.226
					(0.344)	(0.255)	
			medio	23.633	24.618	7.271	25.155
					(0.302)	(0.228)	
			alto	23	27.603	7.735	28.095
					(0.635)	(0.465)	
		medio	bajo	23.667	24.028	7.467	24.817
					(0.472)	(0.349)	
			medio	23.281	26.127	7.342	26.746
					(0.336)	(0.243)	
			alto	26.167	29.034	7.64	29.686
					(0.543)	(0.402)	
particular s.	si	bajo	bajo	17.488	19.508	6.062	17.399
					(0.39)	(0.335)	
			medio	19.167	21.66	6.689	19.327
					(0.38)	(0.374)	
		medio	bajo	19.455	21.05	6.312	18.99
					(0.477)	(0.402)	
			medio	20.906	23.193	6.767	20.918
					(0.362)	(0.354)	
			alto	33	26.251	7.86	23.858
					(0.618)	(0.631)	
		alto	alto	25.429	27.678	7.831	25.45
					(0.549)	(0.574)	
	no	bajo	bajo	21.262	23.376	7.539	20.991
					(0.298)	(0.204)	
			medio	23.696	25.474	7.643	22.92
					(0.261)	(0.176)	
			alto	11	28.37	8.171	25.86
					(0.616)	(0.46)	
		medio	bajo	24.388	24.898	7.649	22.582
					(0.395)	(0.267)	
			medio	27.878	26.971	7.634	24.511
					(0.23)	(0.144)	
			alto	30.441	29.778	8.029	27.451
					(0.495)	(0.372)	
		alto	bajo	31	26.367	7.82	24.173
					(0.516)	(0.352)	
			medio	29.045	28.403	7.694	26.102
					(0.296)	(0.181)	
			alto	30.421	31.111	7.922	29.042
					(0.403)	(0.3)	
particular p.	si	alto	medio	36	23.595	6.219	23.722
					(0.517)	(0.429)	
			alto	28.675	26.721	7.401	26.662
					(0.613)	(0.623)	
	no	medio	medio	32.214	25.997	6.532	25.723
					(0.511)	(0.347)	
			alto	36.5	28.995	7.234	28.663
					(0.668)	(0.463)	
		alto	medio	31.465	27.48	6.679	27.314
					(0.439)	(0.309)	
			alto	33.965	30.378	7.175	30.254
					(0.51)	(0.353)	

Tabla 4.5: Media y error estándar de la distribución de puntajes predicha por los modelos propuesto y HLM. Filas ausentes son niveles vacíos. Para el modelo HLM, el estimador del error estándar es común a todos los subgrupos e igual a 7.345. Los números entre paréntesis son errores estándar de las estimaciones.

En este trabajo se emplea un modelo IRT. A diferencia de los modelos lineales, la estructura de este tipo de modelos permite inferir acerca de las dificultades de los items y las habilidades individuales, en lugar de trabajar con los puntajes. Para contrastar la inferencia paramétrica y semiparamétrica, las Tablas 4.6 y 4.7 comparan las estimaciones de medias y varianzas de los modelos propuesto y basal. En la Tabla 4.6 se muestran los parámetros de habilidad por patrón de covariable, y en la Tabla 4.7, los parámetros de dificultad. Ambos modelos generan estimaciones similares en el sentido de que ordenamiento de los parámetros según su media estimada es muy similar; aunque la especificación paramétrica tiende a subestimar las varianzas de los parámetros. Pero el aporte del modelo propuesto va más allá de presentar mejores estimaciones varianzas: ofrece estimaciones de *toda la distribución de habilidad* por patrón de covariable, y no sólo del primer y segundo momento de dicha distribución. De esta manera, se aprovecha al máximo la información acerca de los subgrupos contenida en los datos.

Para ejemplificar cómo la metodología propuesta aquí es más rica que los análisis convencionales, consideremos la comparación entre alumnos que asisten a distinto tipo de colegio en el subgrupo que no repite con SES y SESp medio[5]. La columna 5 de la Tabla 4.6 presenta las estimaciones de media para el modelo propuesto. La media de la distribución de habilidad en colegios municipalizados tiene un valor esperado de 1.21, mientras que los valores correspondientes a colegios particulares subvencionados y particulares pagados son de 1.31 y 1.18 respectivamente. El error estándar de las medias para estos subgrupos son 0.055, 0.048 y 0.69; lo que indica que las diferencias entre medias de habilidad por tipo de colegio dejan de ser relevantes una vez que se

[5]Para hacer inferencia sobre el tipo de colegios se está suponiendo que los individuos están asignados en forma aleatoria a los distintos tipos de colegio. Este supuesto no es realista para el caso chileno, como se menciona en la sección 1.2, por que los resultados con respecto a tipo de colegio y su significatividad deben ser revisados.

t. de colegio	repite	SESp	SES	media mod.prop.	s.e. mod.prop.	media mod.basal	s.e. mod.basal
municipal	si	bajo	bajo	0.384	0.431	0.699	1.755
				(0.061)	(0.069)	(0.139)	(0.344)
			medio	0.619	0.75	1.304	7.198
				(0.06)	(0.107)	(0.383)	(1.897)
		medio	bajo	0.549	0.639	0.882	2.329
				(0.071)	(0.11)	(0.154)	(0.431)
			medio	0.783	1.012	1.485	7.952
				(0.063)	(0.131)	(0.395)	(2.017)
	no	bajo	bajo	0.81	1.182	0.78	2.113
				(0.054)	(0.107)	(0.215)	(0.553)
			medio	1.043	1.64	1.382	7.835
				(0.052)	(0.127)	(0.463)	(2.28)
			alto	1.382	2.675	1.835	18.191
				(0.087)	(0.303)	(0.657)	(2.604)
		medio	bajo	0.975	1.517	0.963	2.73
				(0.065)	(0.154)	(0.23)	(0.664)
			medio	1.208	2.03	1.562	8.618
				(0.055)	(0.151)	(0.474)	(2.411)
			alto	1.548	3.162	1.958	18.623
				(0.08)	(0.299)	(0.657)	(2.546)
particular s.	si	bajo	bajo	0.485	0.563	1.064	2.889
				(0.06)	(0.084)	(0.14)	(0.456)
			medio	0.72	0.96	1.662	9.001
				(0.059)	(0.124)	(0.379)	(2.097)
		medio	bajo	0.651	0.789	1.247	3.598
				(0.067)	(0.118)	(0.156)	(0.556)
			medio	0.885	1.241	1.841	9.861
				(0.058)	(0.141)	(0.391)	(2.205)
			alto	1.232	2.317	2.137	19.368
				(0.083)	(0.305)	(0.588)	(2.192)
		alto	alto	1.397	2.767	2.244	19.817
				(0.077)	(0.311)	(0.589)	(2.12)
	no	bajo	bajo	0.911	1.427	1.145	3.368
				(0.052)	(0.11)	(0.216)	(0.757)
			medio	1.145	1.964	1.736	9.711
				(0.05)	(0.124)	(0.456)	(2.522)
			alto	1.489	3.148	2.038	19.221
				(0.085)	(0.305)	(0.628)	(2.328)
		medio	bajo	1.076	1.781	1.328	4.121
				(0.059)	(0.145)	(0.232)	(0.876)
			medio	1.31	2.374	1.912	10.589
				(0.048)	(0.131)	(0.466)	(2.63)
			alto	1.655	3.654	2.15	19.633
				(0.074)	(0.282)	(0.628)	(2.266)
		alto	bajo	1.241	2.213	1.511	5.048
				(0.071)	(0.201)	(0.249)	(1.019)
			medio	1.475	2.863	2.085	11.599
				(0.053)	(0.164)	(0.475)	(2.728)
			alto	1.82	4.237	2.252	20.06
				(0.067)	(0.271)	(0.627)	(2.194)
particular p.	si	alto	medio	0.922	1.203	1.6	8.89
				(0.069)	(0.15)	(0.409)	(2.133)
			alto	1.271	2.255	1.996	18.712
				(0.08)	(0.267)	(0.614)	(2.368)
	no	medio	medio	1.182	1.808	1.497	8.615
				(0.069)	(0.181)	(0.475)	(2.388)
			alto	1.529	2.973	1.904	18.548
				(0.087)	(0.299)	(0.657)	(2.524)
		alto	medio	1.347	2.259	1.674	9.57
				(0.064)	(0.185)	(0.486)	(2.516)
			alto	1.694	3.518	2.016	19.011
				(0.074)	(0.27)	(0.656)	(2.442)

Tabla 4.6: Media y error estándar de la distribución a posteriori de habilidades de los modelos propuesto y basal. Filas ausentes son niveles vacíos. Los números entre paréntesis son errores estándar de las estimaciones.

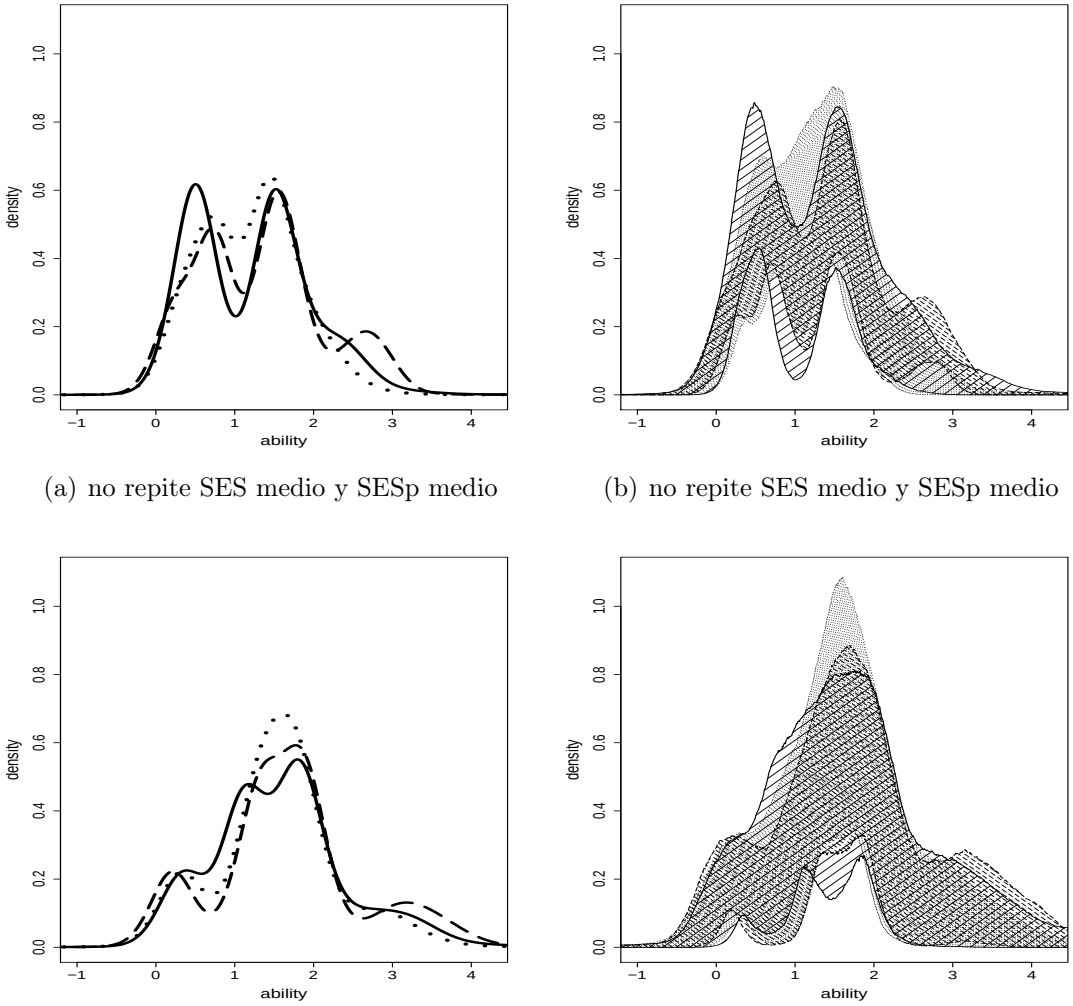
parámetro	**media m.prop.**	**intervalo 95 % m.prop**	**error MC m.prop**	**media m.basal**	**intervalo 95 % m.basal**	**error MC m.basal**
β_2	-1.351	(-1.491 - -1.213)	0.003	-1.324	(-1.461 - -1.189)	0.003
β_{19}	-0.964	(-1.09 - -0.838)	0.003	-0.93	(-1.054 - -0.8)	0.003
β_{29}	-0.361	(-0.474 - -0.243)	0.003	-0.321	(-0.431 - -0.205)	0.003
β_{32}	-0.357	(-0.475 - -0.243)	0.003	-0.317	(-0.425 - -0.208)	0.003
β_{14}	-0.252	(-0.376 - -0.139)	0.003	-0.21	(-0.321 - -0.095)	0.003
β_1	0	(0 - 0)	0	0	(0 - 0)	0
β_6	0.028	(-0.093 - 0.151)	0.003	0.075	(-0.026 - 0.182)	0.003
β_9	0.039	(-0.072 - 0.154)	0.003	0.08	(-0.028 - 0.186)	0.003
β_{13}	0.041	(-0.075 - 0.154)	0.003	0.086	(-0.024 - 0.194)	0.003
β_{18}	0.071	(-0.044 - 0.187)	0.003	0.115	(0.007 - 0.227)	0.003
β_{16}	0.125	(0.009 - 0.243)	0.003	0.17	(0.064 - 0.28)	0.003
β_{33}	0.14	(0.018 - 0.255)	0.003	0.185	(0.082 - 0.293)	0.003
β_{27}	0.192	(0.076 - 0.305)	0.003	0.238	(0.132 - 0.342)	0.003
β_{15}	0.206	(0.095 - 0.32)	0.003	0.251	(0.147 - 0.364)	0.003
β_{21}	0.276	(0.16 - 0.391)	0.003	0.321	(0.215 - 0.424)	0.003
β_{41}	0.284	(0.172 - 0.393)	0.003	0.331	(0.222 - 0.44)	0.003
β_{40}	0.288	(0.175 - 0.4)	0.003	0.332	(0.228 - 0.44)	0.003
β_{10}	0.294	(0.178 - 0.404)	0.003	0.338	(0.235 - 0.448)	0.003
β_{39}	0.717	(0.605 - 0.821)	0.003	0.764	(0.665 - 0.869)	0.003
β_7	0.775	(0.665 - 0.885)	0.003	0.825	(0.721 - 0.928)	0.003
β_8	0.825	(0.715 - 0.931)	0.003	0.873	(0.77 - 0.976)	0.003
β_{37}	0.836	(0.722 - 0.946)	0.003	0.884	(0.781 - 0.988)	0.003
β_5	0.851	(0.737 - 0.961)	0.003	0.901	(0.798 - 1.01)	0.003
β_{35}	0.9	(0.791 - 1.019)	0.003	0.948	(0.842 - 1.056)	0.003
β_4	0.915	(0.802 - 1.027)	0.004	0.963	(0.857 - 1.065)	0.003
β_{25}	0.921	(0.804 - 1.034)	0.003	0.968	(0.859 - 1.075)	0.003
β_{28}	1.176	(1.062 - 1.287)	0.003	1.226	(1.121 - 1.335)	0.003
β_{46}	1.181	(1.066 - 1.291)	0.003	1.23	(1.128 - 1.335)	0.003
β_{42}	1.265	(1.152 - 1.374)	0.004	1.315	(1.211 - 1.421)	0.003
β_{30}	1.286	(1.172 - 1.397)	0.003	1.336	(1.23 - 1.445)	0.003
β_{45}	1.385	(1.27 - 1.494)	0.003	1.434	(1.329 - 1.541)	0.003
β_{24}	1.451	(1.337 - 1.559)	0.003	1.499	(1.398 - 1.61)	0.003
β_{20}	1.493	(1.387 - 1.603)	0.003	1.542	(1.442 - 1.644)	0.003
β_{36}	1.522	(1.406 - 1.634)	0.003	1.572	(1.464 - 1.679)	0.003
β_{17}	1.55	(1.438 - 1.663)	0.003	1.604	(1.499 - 1.71)	0.003
β_{43}	1.553	(1.439 - 1.662)	0.003	1.604	(1.501 - 1.71)	0.003
β_{38}	1.646	(1.528 - 1.763)	0.004	1.697	(1.595 - 1.806)	0.003
β_{34}	1.713	(1.596 - 1.829)	0.003	1.765	(1.666 - 1.872)	0.003
β_{26}	1.735	(1.627 - 1.847)	0.003	1.785	(1.68 - 1.891)	0.003
β_{31}	1.746	(1.628 - 1.856)	0.003	1.798	(1.693 - 1.903)	0.003
β_3	1.836	(1.72 - 1.946)	0.003	1.888	(1.78 - 1.996)	0.003
β_{11}	1.847	(1.732 - 1.957)	0.003	1.899	(1.795 - 2.004)	0.003
β_{22}	1.919	(1.806 - 2.031)	0.003	1.97	(1.857 - 2.082)	0.003
β_{23}	2.081	(1.969 - 2.197)	0.003	2.131	(2.024 - 2.244)	0.003
β_{12}	2.5	(2.38 - 2.62)	0.003	2.553	(2.441 - 2.662)	0.003
β_{44}	2.708	(2.584 - 2.826)	0.004	2.764	(2.654 - 2.879)	0.003

Tabla 4.7: Estimaciones a posteriori de los parámetros de dificultad, para los modelos propuesto y basal.

controla por variables socioeconómicas. Empleando el modelo basal se llegaría al mismo resultado, y utilizando HLM se concluye en forma análoga, pero en términos de puntaje: la variable tipo de colegio no resulta significativa para explicar los puntajes a un nivel de confianza de 95 %. El resultado típicamente encontrado acerca de que la variabilidad *intra-tipo de colegio* es mayor que la variabilidad *entre-tipos de colegios* se verifica para estos subgrupos también (la varianza de las tres medias es de 0.005, mientras que la media de las varianzas de los tres subgrupos es 2.072).

Hasta el momento, la inferencia no es diferente a lo que propone la estadística paramétrica. Sin embargo, la metodología propuesta aquí va más allá al comparar la forma completa de las distribuciones y no sólo su media. La Figura 4.4a muestra las densidades promedio a posteriori para los mismos tres subgrupos de alumnos. En la Figura 4.4b se pueden ver las bandas de confianza para cada densidad estimada, las cuales se calcularon a partir del método propuesto en la sección 3.2. Lo primero que se puede apreciar es que las tres distribuciones analizadas son bimodales, característica sobre la que no es posible hacer inferencia desde el enfoque paramétrico. La bimodalidad puede ser un dato interesante a analizar: estaría indicando que para alumnos dentro del grupo analizado, hay a su vez dos tipos de estudiantes diferentes, siendo unos más hábiles que otros. De esta manera, queda explícita la forma que tiene la variabilidad al interior de los subgrupos, lo que no es posible hacer empleando modelos HLM. También es llamativo que las forma distribucional no cambia por tipo de colegio: tanto las distribuciones para colegios municipales del subgrupo, como las de los otros dos tipos de colegio, reflejan claramente una forma bimodal en sus distribuciones de habilidad.

Para completar la presentación gráfica con información cuantitativa, la Tabla 4.8 presenta estadísticos descriptivos de las densidades estimadas. En la columna 5 se



(a) no repite SES medio y SESp medio

(b) no repite SES medio y SESp medio

(c) no repite SES alto y SESp medio

(d) no repite SES alto y SESp medio

Figura 4.4: Gráficos Comparativos por tipo de colegio. Los gráficos (a) y (c) presentan las densidades medias a posteriori de un nuevo individuo de la muestra, es decir que se grafica $\hat{G}_k$, las densidades estimadas para $E(G_k|Y, d_k)$. La línea entera corresponde a los colegios municipales, la linea rayada a los particulares subvencionados y la línea punteada a los colegios particulares pagados. A la derecha, gráficos (b) y (d), se muestran las bandas de credibilidad de 95 % para $E(G_k|Y, d_k)$. La banda pintada con líneas enteras está asociada a la curva entera del gráfico de la izquierda, la pintada con líneas rayadas se asocia a la curva rayada y la pintada con puntos se vincula con la curva punteada.

presentan los conteos de los individuos por patrón de covariable. Las filas excluidas corresponden a niveles sin observaciones, por lo que se puede apreciar que no hay individuos de colegios municipales con SESp alto, ni tampoco de colegio particulares pagados de SESp bajo. Por su parte, la columna 7 muestra las varianzas de las densidades estimadas. Destaca el hecho de que para todos los tipo de colegio, las densidades para alumnos que no repiten con SES alto presentan las varianzas más altas. Las varianzas de las densidades para alumnos repitentes son siempre más pequeñas que la de alumnos no repitentes, dejando el resto de las covariables constantes. Esto evidencia que los alumnos que repiten son similares entre si al interior de los subgrupos. La varianzas crecen junto con el nivel socioeconómico, tanto individual como de colegio, al dejar constante el resto de las covariables. Valores de la mediana (columna 9) inferior a la media (columna 6) indican la presencia de asimetría hacia la izquierda, como es el caso de las densidades para subgrupos con SES y SESp bajos. Este resultado implica mayor masa de probabilidad en niveles bajos de habilidad para estos subgrupos. Por el contrario, cuando la mediana es mayor que la media, se trata de densidades con asimetría hacia la derecha y reflejan mayor probabilidad acumulada sobre valores altos de habilidad, como ocurre, por ejemplo, para los subgrupos con SESp alto y SES medio que no repiten. Los resultados de asimetría también se pueden observar gráficamente.

4.3.2. Analisis del modelo escogido

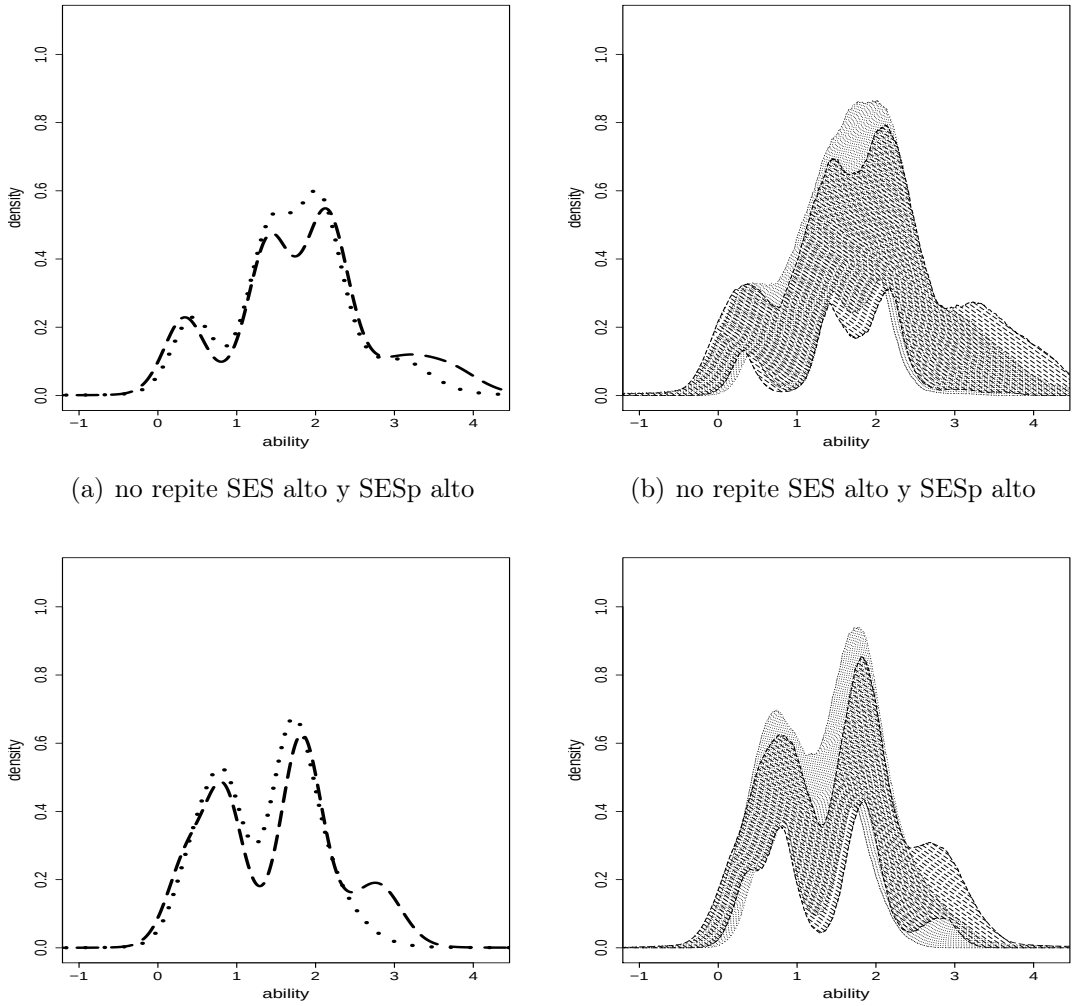
Siguiendo con la Figura 4.4, se puede apreciar que las formas sí cambian entre los gráficos 4.4a y 4.4c. En éste último - que considera al subgrupo de alumnos que no repite, SES alto y SESp medio - se observan distribuciones asimétricas hacia la derecha. Aquí se evidencia la utilidad de las bandas de credibilidad: si bien la estimaciones de las distribuciones promedio presentan una leve bimodalidad, las bandas de

tipo de colegio	repite	SESp	SES	n° de obs.	media	var.	1er cuar.	mediana	3er cuar.
municipal	si	bajo	bajo	176	0.384	0.431	0.005	0.333	0.684
			medio	20	0.619	0.750	0.216	0.540	0.900
		medio	bajo	4	0.549	0.639	0.122	0.509	0.918
			medio	8	0.784	1.013	0.338	0.707	1.134
	no	bajo	bajo	741	0.810	1.182	0.212	0.761	1.287
			medio	98	1.044	1.641	0.450	0.999	1.485
			alto	1	1.387	2.685	0.869	1.328	1.782
		medio	bajo	12	0.975	1.517	0.320	1.004	1.503
			medio	32	1.209	2.032	0.558	1.251	1.715
			alto	6	1.554	3.173	0.990	1.535	2.021
particular s.	si	bajo	bajo	80	0.485	0.563	0.090	0.410	0.765
			medio	12	0.720	0.961	0.315	0.653	0.995
		medio	bajo	11	0.651	0.789	0.207	0.594	1.004
			medio	64	0.885	1.242	0.446	0.828	1.229
			alto	5	1.235	2.322	0.774	1.242	1.616
		alto	alto	7	1.401	2.774	0.914	1.400	1.832
	no	bajo	bajo	370	0.911	1.427	0.288	0.828	1.391
			medio	56	1.146	1.965	0.567	1.071	1.553
			alto	1	1.493	3.157	1.017	1.431	1.845
		medio	bajo	139	1.077	1.781	0.396	1.080	1.602
			medio	852	1.311	2.375	0.675	1.323	1.787
			alto	102	1.660	3.665	1.152	1.629	2.088
		alto	bajo	1	1.241	2.214	0.504	1.323	1.823
			medio	22	1.476	2.866	0.774	1.575	2.025
			alto	76	1.826	4.249	1.269	1.841	2.340
particular p.	si	alto	medio	2	0.922	1.204	0.509	0.873	1.305
			alto	40	1.274	2.260	0.774	1.305	1.701
	no	medio	medio	14	1.182	1.808	0.689	1.206	1.625
			alto	4	1.533	2.980	1.116	1.553	1.949
		alto	medio	43	1.347	2.261	0.797	1.427	1.845
			alto	864	1.698	3.526	1.251	1.737	2.174

Tabla 4.8: Estadísticos descriptivos de las distribuciones a posteriori de habilidad por patrón de covariable. Filas ausentes son niveles vacíos.

confianza (gráfico 4.4d) no confirman dicha forma. Las Figuras 4.5 y 4.6 también presentan comparaciones por tipo de colegio para otros subgrupos. Todos los subgrupos presentados son consistentes en mostrar que el tipo de colegio no genera grandes alteraciones en la forma ni localización de la distribución de habilidad. En la Figura 4.5 sólo se comparan distribuciones de colegios particulares debido a la falta de datos. El gráfico 4.5a tiene la particularidad de mostrar distribuciones bimodales y asimétricas al mismo tiempo. Esto indica que los subgrupos de alumnos que no repiten con SES y SESp altos contienen dos tipos distintos de individuos en su interior, donde el tipo de individuo con menor habilidad es menos frecuente que el de mayor habilidad. Para el subgrupo que no repite con SES medio y SESp alto la bimodalidad se mantiene, pero la asimetría no es tan marcada. La Figura 4.6 sólo incluye distribuciones para colegios municipales y particulares pagados. En este caso se observan densidades asimétricas a la izquierda para los subgrupos de repitentes.

Aprovechando la heterogeneidad que se observa en estudiantes que asisten a los colegios particulares subvencionados, se utilizó este subgrupo para examinar el efecto de las variables socioeconómicas sobre la habilidad. Las Figuras 4.7-4.9 muestran comparaciones entre estudiantes que no repitieron. Los gráficos 4.7a y 4.7b, por ejemplo, consideran las estimaciones de distribuciones a posteriori para diferentes categorías de SES, en alumnos con SESp bajo. Encontramos que a diferencia de la variable tipo de colegio, el factor SES tiene una gran influencia en la forma de la habilidad. La masa de probabilidad de los estudiantes con SES bajo está más cargada hacia la izquierda del rango de habilidad. Por otra parte, los estudiantes con SES alto tienen mayor probabilidad de desempeñarse mejor que los estudiantes con SES medio y bajo. Los gráficos 4.7c y d comparan SESp para alumnos con SES bajo. Las diferencias en forma son menos notorias en este caso, y las curvas resultantes son desplazamientos de izquierda a derecha desde el SESp bajo al alto. Este resultado también es válido



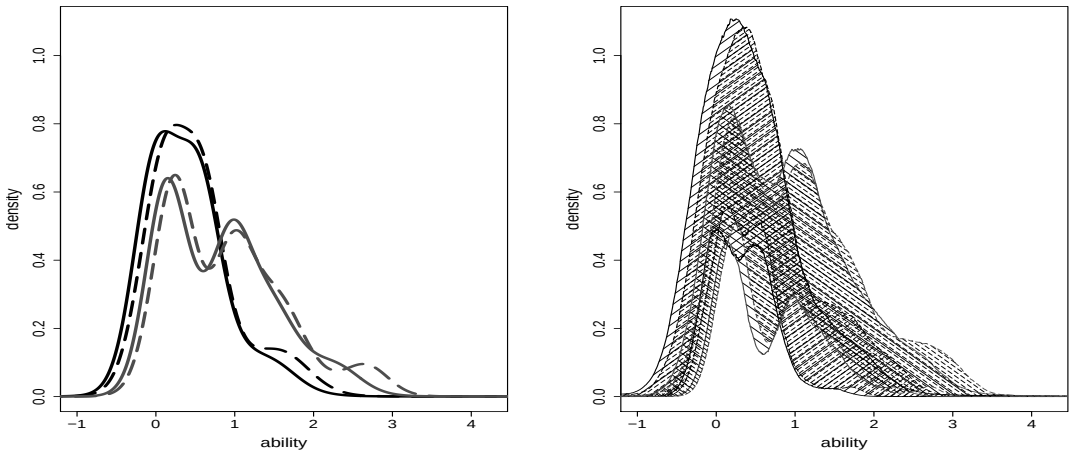
(a) no repite SES alto y SESp alto

(b) no repite SES alto y SESp alto

(c) no repite SES medio y SESp alto

(d) no repite SES medio y SESp alto

Figura 4.5: Gráficos Comparativos por tipo de colegio. Los gráficos (a) y (c) presentan las densidades medias a posteriori de un nuevo individuo de la muestra, es decir que se grafica $\hat{G}_k$, las densidades estimadas para $E(G_k|Y, d_k)$. La línea rayada corresponde a los colegios particulares subvencionados y la línea punteada a los particulares pagados. A la derecha, gráficos (b) y (d), se muestran las bandas de credibilidad de 95 % para $E(G_k|Y, d_k)$. La banda pintada con líneas rayadas está asociada a la curva rayada del gráfico de la izquierda y la pintada con puntos se vincula con la curva punteada.



(a) SES bajo y SESp bajo

(b) SES bajo y SESp bajo

(c) SES bajo y SESp medio

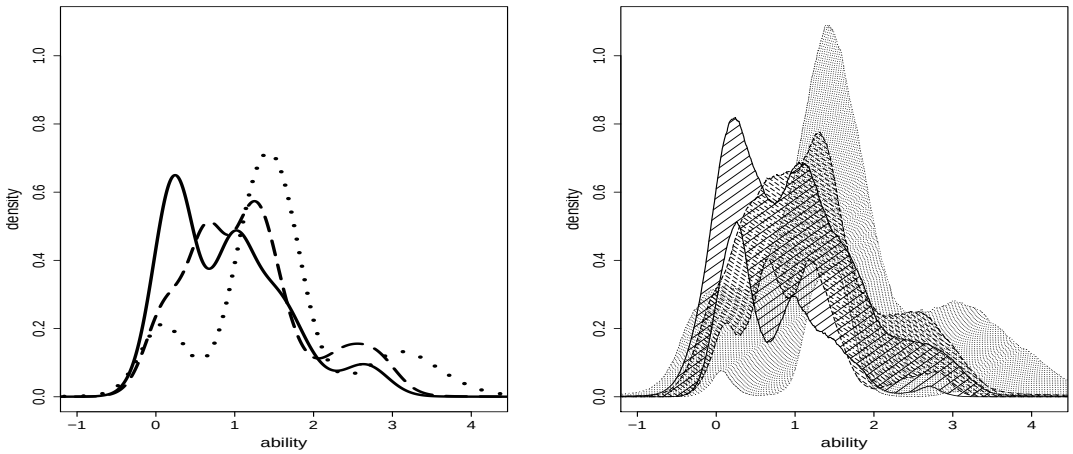
(d) SES bajo y SESp medio

Figura 4.6: Gráficos Comparativos por tipo de colegio. Los gráficos (a) y (c) presentan las densidades medias a posteriori de un nuevo individuo de la muestra, es decir que se grafica $\hat{G_k}$, las densidades estimadas para $E(G_k|Y, d_k)$. La línea entera corresponde a los colegios municipales y la rayada a los particulares subvencionados. El color gris corresponde a individuos que no repiten mientras que el color negro a los que repiten. A la derecha, gráficos (b) y (d), se muestran las bandas de credibilidad de 95 % para $E(G_k|Y, d_k)$. La banda pintada con líneas enteras está asociada a la curva entera , la pintada con líneas enteras se asocia a la curva entera del gráfico de la izquierda y la pintada con rayas se vincula con la curva rayada. La inclinación de las rayas marcan la diferencia para asociar las bandas a las curvas negras y grises.

para los subgrupos considerados en las Figuras 4.8 y 4.9. Es interesante notar que si bien la variable SESp fue la primera en ser incluida en el procedimiento paso a paso, es SES la que genera más diferencias entre distribuciones, fenómeno que se desprende específicamente de esta metodología.

Finalmente, las Figuras 4.10 y 4.11 muestran las diferencias entre estudiantes que repitieron y que no repitieron, para varias combinaciones de SES y SESp. Las distribuciones de alumnos repitentes tienden a ser asimétricas a la izquierda indicando alta probabilidad de poseer una baja habilidad. Las distribuciones para alumnos que no repiten suelen ser bimodales para los subgrupos considerados. En general, repetir o no, parece no estar vinculado con factores socioeconómicos, al menos no para estos datos.

El análisis gráfico tiene, sin embargo, algunas limitaciones. Las bandas de confianza dan idea del error y sirven para confirmar una forma determinada de las distribuciones como la bimodalidad, pero es muy difícil decidir si dos densidades difieren o no a partir de ellas. En efecto, las Figuras 4.4-4.11 muestran que las bandas para distintas densidades se superponen en algunas partes y en otras no, pero no es claro cómo concluir si esas diferencias son relevantes. Para suplir esta falencia y tener una medida cuantitativa de las diferencias entre densidades, se exponen las divergencias de KL en la Tabla 4.9. La divergencia de KL es tratada como un elemento aleatorio de manera que se presentan estadísticos de su distribución. La columna 7 presenta estimaciones de la probabilidad de que la divergencia KL entre dos densidades sea mayor que un cierta cota. Dicha cota es también aleatoria y está construida a partir de las divergencias de KL entre simulaciones de una misma densidad, siguiendo la metodología propuesta en la sección 3.2. Para decidir si dos densidades son diferentes se puede establecer un límite para dicha probabilidad, por ejemplo 0.6. Valores menores se consideran diferencias irrelevantes. Utilizando este valor límite, todas las



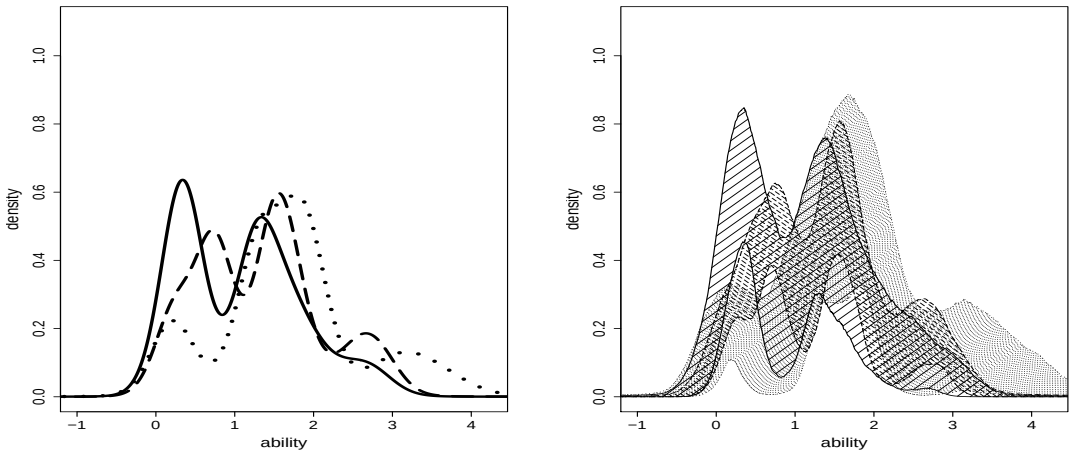
(a) SESp bajo

(b) SESp bajo

(c) SES bajo

(d) SES bajo

Figura 4.7: Estimaciones de las densidades medias a posteriori de habilidad para estudiantes de colegios particulares subvencionados que no repiten. Se grafica $\hat{G}_k$, las densidades estimadas para $E(G_k|Y, d_k)$. En el gráfico (a) se comparan alumnos con SESp bajo. La línea entera corresponde a estudiantes con SES bajo, la linea rayada a los estudiantes con SES medio y la línea punteada a los estudiantes con SES alto. En el gráfico (c) se comparan alumnos con SES bajo. La línea entera corresponde a estudiantes con SESp bajo, la linea rayada a los estudiantes con SESp medio y la línea punteada a los estudiantes con SESp alto. A la derecha, gráficos (b) y (d), se muestran las bandas de credibilidad de 95 % para $E(G_k|Y, d_k)$. La banda pintada con líneas enteras está asociada a la curva entera del gráfico de la izquierda, la pintada con líneas rayadas se asocia a la curva rayada y la pintada con puntos se vincula con la curva punteada.



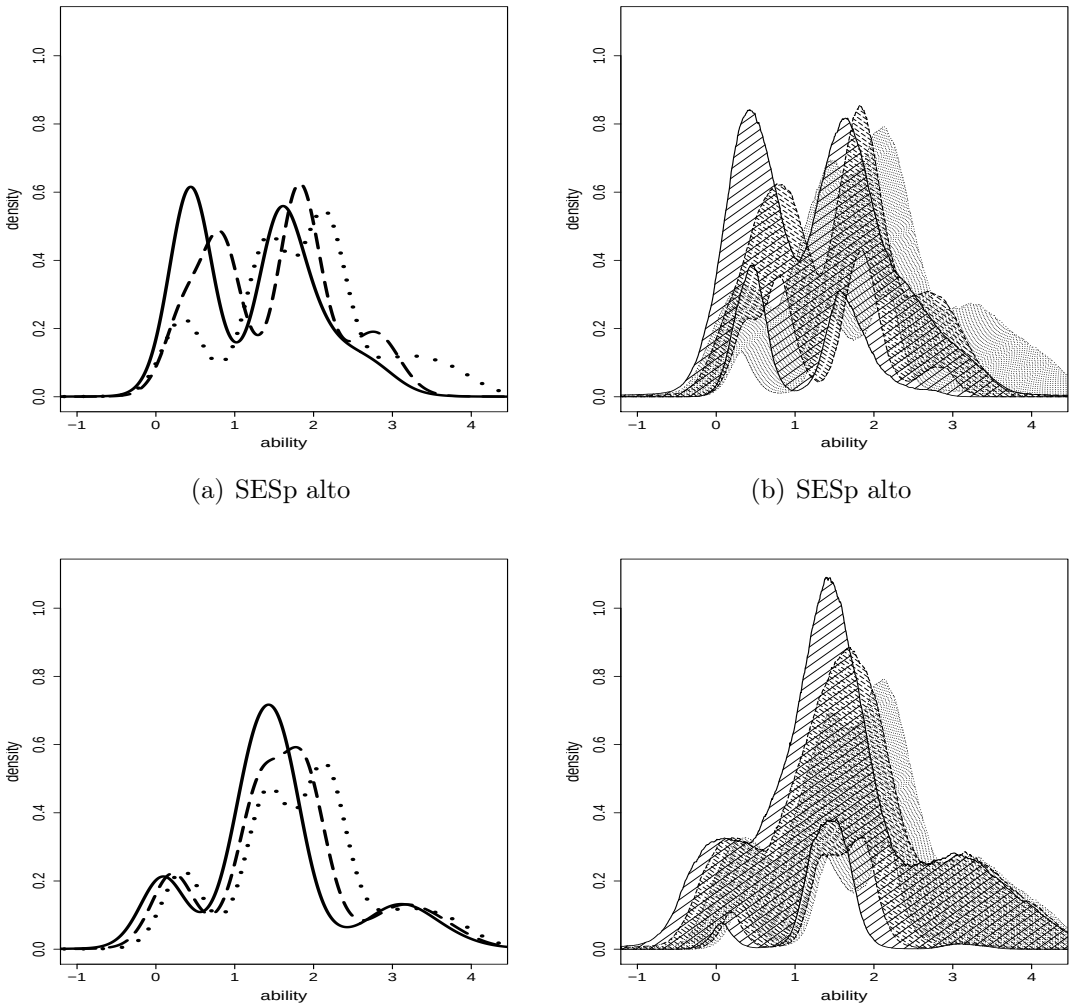
(a) SESp medio

(b) SESp medio

(c) SES medio

(d) SES medio

Figura 4.8: Estimaciones de las densidades medias a posteriori de habilidad para estudiantes de colegios particulares subvencionados que no repiten. Se grafica $\hat{G}_k$, las densidades estimadas para $E(G_k|Y, d_k)$. En el gráfico (a) se comparan alumnos con SESp medio. La línea entera corresponde a estudiantes con SES bajo, la linea rayada a los estudiantes con SES medio y la línea punteada a los estudiantes con SES alto. En el gráfico (c) se comparan alumnos con SES medio. La línea entera corresponde a estudiantes con SESp bajo, la linea rayada a los estudiantes con SESp medio y la línea punteada a los estudiantes con SESp alto. A la derecha, gráficos (b) y (d), se muestran las bandas de credibilidad de 95 % para $E(G_k|Y, d_k)$. La banda pintada con líneas enteras está asociada a la curva entera del gráfico de la izquierda, la pintada con líneas rayadas se asocia a la curva rayada y la pintada con puntos se vincula con la curva punteada.



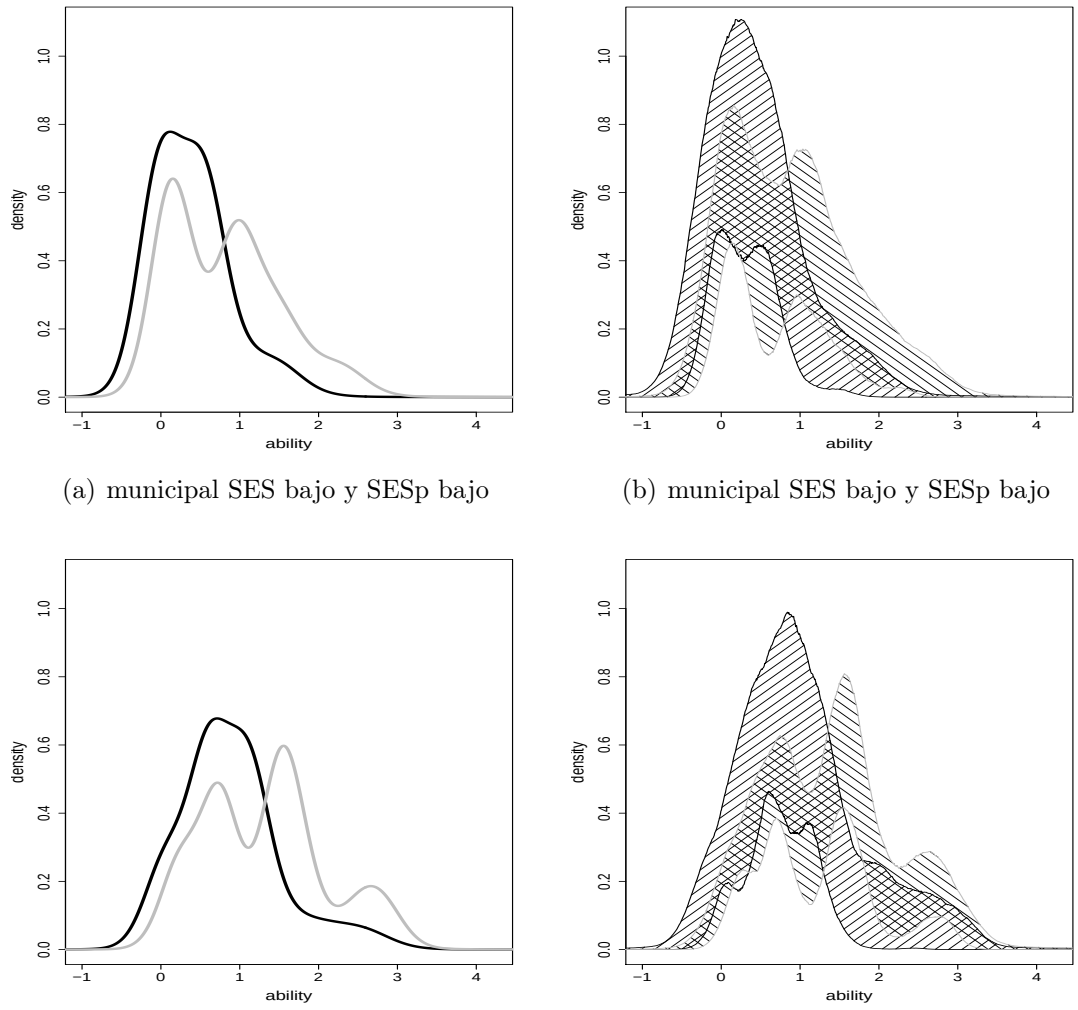
(a) SESp alto

(b) SESp alto

(c) SES alto

(d) SES alto

Figura 4.9: Estimaciones de las densidades medias a posteriori de habilidad para estudiantes de colegios particulares subvencionados que no repiten. Se grafica $\hat{G}_k$, las densidades estimadas para $E(G_k|Y, d_k)$. En el gráfico (a) se comparan alumnos con SESp alto. La línea entera corresponde a estudiantes con SES bajo, la linea rayada a los estudiantes con SES medio y la línea punteada a los estudiantes con SES alto. En el gráfico (c) se comparan alumnos con SES alto. La línea entera corresponde a estudiantes con SESp bajo, la linea rayada a los estudiantes con SESp medio y la línea punteada a los estudiantes con SESp alto. A la derecha, gráficos (b) y (d), se muestran las bandas de credibilidad de 95 % para $E(G_k|Y, d_k)$. La banda pintada con líneas enteras está asociada a la curva entera del gráfico de la izquierda, la pintada con líneas rayadas se asocia a la curva rayada y la pintada con puntos se vincula con la curva punteada.



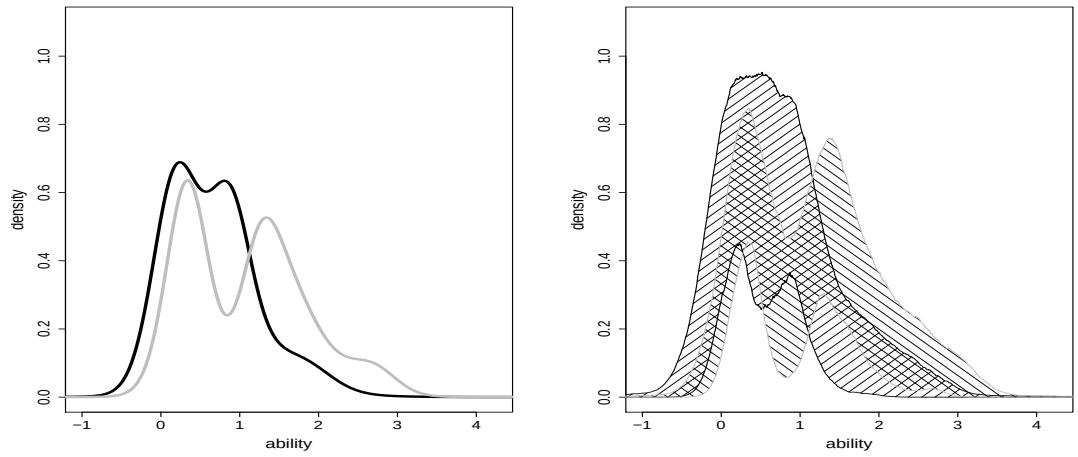
(a) municipal SES bajo y SESp bajo

(b) municipal SES bajo y SESp bajo

(c) particular subvencionado SES medio y SESp medio

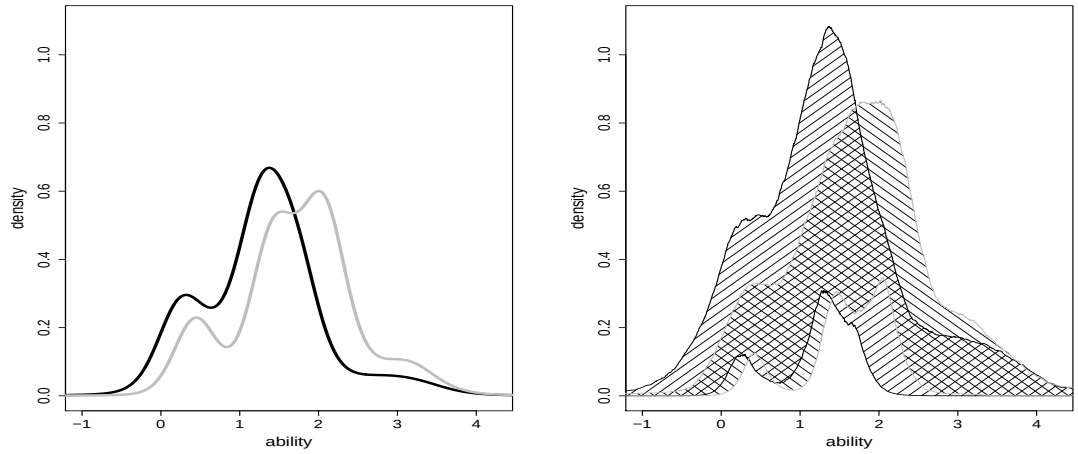
(d) particular subvencionado SES medio y SESp medio

Figura 4.10: Gráficos Comparativo de repite. Se grafica $\hat{G}_k$, las densidades estimadas para $E(G_k|Y, d_k)$. En los gráficos (a) y (c) el color gris corresponde a individuos que no repiten mientras que el color negro a los que repiten. A la derecha, gráficos (b) y (d), se muestran las bandas de credibilidad de 95 % para $E(G_k|Y, d_k)$. La inclinación de las rayas marcan la diferencia para asociar las bandas a las curvas negras y grises.



(a) particular subvencionado SES bajo y SESp medio

(b) particular subvencionado SES bajo y SESp medio



(c) particular pagado SES alto y SESp alto

(d) particular pagado SES alto y SESp alto

Figura 4.11: Gráficos Comparativo de repite. Se grafica $\hat{G}_k$, las densidades estimadas para $E(G_k|Y, d_k)$. En los gráficos (a) y (c) el color gris corresponde a individuos que no repiten mientras que el color negro a los que repiten. A la derecha, gráficos (b) y (d), se muestran las bandas de credibilidad de 95 % para $E(G_k|Y, d_k)$. La inclinación de las rayas marcan la diferencia para asociar las bandas a las curvas negras y grises.

comparaciones de tipo de colegio (Figuras 4.4-4.6), resultan irrelevantes a excepción de las diferencias entre colegios particulares subvencionados y particulares pagados de alumnos que no repiten con SES y SESp medio, y SES medio y SESp alto. Contrariamente, las comparaciones de SES y SESp (Figura 4.7-4.9), excepcionalmente presentan valores menores a 0.6: en el caso de alumnos con nivel socioeconómico alto.

Figura	**Gráfico**	**Comparación**	**media**	**intervalo 95 %**	**error MC**	**P(kl>c)**
4.4	a	municipal-p.subvencionado	0.12	(0.024-0.326)	0.004	0.429
		municipal-p.pagado	0.168	(0.028-0.521)	0.005	0.432
		p.subvencionado-p.pagado	0.15	(0.051-0.323)	0.003	0.635
	c	municipal-p.subvencionado	0.187	(0.033-0.537)	0.005	0.216
		municipal-p.pagado	0.195	(0.034-0.57)	0.005	0.21
		p.subvencionado-p.pagado	0.171	(0.054-0.411)	0.003	0.193
4.5	a	p.subvencionado-p.pagado	0.155	(0.049-0.371)	0.003	0.28
	c	p.subvencionado-p.pagado	0.156	(0.05-0.347)	0.003	0.624
4.6	a	municipal-p.subvencionado, repite	0.078	(0.017-0.222)	0.002	0.254
		municipal-p.subvencionado, no repite	0.084	(0.018-0.223)	0.002	0.325
	c	municipal-p.subvencionado, repite	0.091	(0.018-0.283)	0.002	0.176
		municipal-p.subvencionado, no repite	0.095	(0.019-0.256)	0.002	0.238
4.7	a	bajo-medio SES	0.151	(0.056-0.349)	0.004	0.754
		bajo-alto SES	0.598	(0.254-1.335)	0.009	0.857
		medio-alto SES	0.524	(0.158-1.235)	0.011	0.751
	c	bajo-medio SESp	0.11	(0.037-0.275)	0.003	0.466
		bajo-alto SESp	0.351	(0.124-0.801)	0.008	0.832
		medio-alto SESp	0.139	(0.039-0.341)	0.004	0.401
4.8	a	bajo-medio SES	0.169	(0.055-0.415)	0.004	0.675
		bajo-alto SES	0.469	(0.23-0.953)	0.006	0.875
		medio-alto SES	0.449	(0.145-1.051)	0.01	0.807
	c	bajo-medio SESp	0.11	(0.04-0.247)	0.003	0.699
		bajo-alto SESp	0.336	(0.134-0.704)	0.007	0.968
		medio-alto SESp	0.139	(0.042-0.331)	0.004	0.739
4.9	a	bajo-medio SES	0.216	(0.065-0.507)	0.005	0.636
		bajo-alto SES	0.435	(0.228-0.794)	0.006	0.891
		medio-alto SES	0.435	(0.148-1.014)	0.01	0.857
	c	bajo-medio SESp	0.141	(0.054-0.285)	0.003	0.156
		bajo-alto SESp	0.401	(0.171-0.758)	0.007	0.682
		medio-alto SESp	0.136	(0.049-0.285)	0.003	0.238
4.10	a	repite-no repite	0.336	(0.196-0.624)	0.004	0.957
	c	repite-no repite	0.318	(0.183-0.552)	0.004	0.887
4.11	a	repite-no repite	0.366	(0.195-0.732)	0.006	0.911
	c	repite-no repite	0.416	(0.203-0.807)	0.006	0.624

Tabla 4.9: Estimaciones a posteriori de la divergencia de Kullback-Leibler.

Las divergencias de KL también reflejan que hay algunas distribuciones más disímiles que otras. Observando los valores medios (columna 4 de la Tabla 4.9), se

confirma lo que se observa a partir de los gráficos. Para las Figuras 4.4-4.6 los valores son bajos. Es notorio, además, que todas las comparaciones entre alumnos con SES bajo y medio; y SESp bajo y medio, presentan valores medios pequeños para KL, aunque estas diferencias resultan relevantes. Los mayores valores de $\bar{KL}$ se ven para las distribuciones que comparan SES bajo y alto, en los gráficos 4.7a, 4.8a y 4.9a, que llegan a 0.598 en el gráfico 4.7a. Finalmente, las divergencias para los gráficos que comparan SESp - 4.7b, 4.8b y 4.9b - presentan valores intermedios. Son en general mayores que los valores medios de KL que comparan tipo de colegio, y menores en relación a los valores que compran SES. Lo mismo ocurre para las comparaciones entre alumnos que repiten y no repiten.

	repite	SES	SESp	tipo de colegio	selección
repite	1.00				
SES	-0.18	1.00			
SESp	-0.18	0.94	1.00		
tipo de colegio	-0.17	0.82	0.87	1.00	
selección	-0.16	0.66	0.71	0.65	1.00

Tabla 4.10: Matriz de Correlación de las covariables

4.4. Conclusiones

Un conocimiento claro de los factores que explican las diferencias en el desempeño de los estudiantes es un tema relevante para ayudar a los tomadores de decisiones a mejorar el sistema educativo. Desde una perspectiva aplicada, este trabajo propone un modelo que permite estimar y comparar densidades a posteriori de las habilidades latentes. Las herramientas empleadas para el análisis son las densidades medias a posteriori condicionadas en covariables, junto con bandas de credibilidad; y divergencias de Kullback-Leibler para cuantificar las diferencias entre densidades.

Los estudiantes de colegios particulares subvencionados y que no repiten fueron analizados en detalle. Mostramos que en estos grupos, la variable SES afecta fuertemente las formas distribucionales. Por otra parte, las diferencias en SESp se reflejan como desplazamientos en las distribuciones de habilidad. La inferencia acerca de las formas de las distribuciones es respaldada mediante el uso de bandas de credibilidad que permiten corroborar características de las densidades como la bimodalidad y asimetría para algunos subgrupos. La presencia de bimodalidad es un resultado que debe ser destacado pues estaría indicando la presencia de dos tipos de individuos al interior del subgrupo. Encontramos bimodalidad en los subgrupos de alumnos que no repiten, con SES medio y SESp medio y alto. Por su parte, la presencia de asimetría

indica que la masa probabilidad está más cargada hacia valores bajos de habilidad (simetría a la izquierda) o altos de habilidad (simetría a la derecha). A partir del estudio se detecta una marcada asimetría hacia la izquierda para alumnos de colegios municipales y subvencionados que no repiten con SES y SESp bajo, y en la mayoría de los subgrupos de repitentes. La presencia de bimodalidad y asimetría al mismo tiempo indica que el subgrupo en cuestión contienen dos tipos distintos de individuos en su interior, donde uno de los tipo de individuo es menos frecuente que el otro. Tal es el caso de los subgrupos de alumnos que no repiten con SES y SESp altos. La asimetría a la derecha junto con la bimodalidad implican que el tipo de alumnos con menor habilidad es menos frecuente que el de mayor habilidad.

Si bien encontramos que las distribuciones de habilidad no cambian su localización y forma por tipo de colegio, este resultado debe tomarse con cautela ya que en el presente estudio no se tiene en cuenta el sesgo de selección que ocurre debido a que los alumnos no están asignados en forma aleatoria por tipo de colegio.

Desde el punto de vista teórico, este trabajo representa un avance a la colección de modelos que han sido usados para caracterizar la educación chilena porque incluye covariables y al mismo tiempo relaja el dudoso supuesto de normalidad. La comparación con modelos HLM sugieren que el modelo propuesto ajusta mejor a los datos, al predecir mejor la distribución de puntajes observados por patrón de covariables. El modelo puede ser potencialmente usado para analizar sistemas educacionales de otros países, y lo que es más interesante, para comparar las habilidades entre países diferentes si se dispone de información tipo PISA.

Capítulo 5

Conclusiones Generales

El trabajo de investigación realizado en esta tesis tuvo como punto de partida una pregunta aplicada, a saber: ¿Es posible mejorar la metodología preexistente en Chile para caracterizar las habilidades de estudiantes y para comparar alumnos de distintos tipos de colegio, con determinadas características socioeconómicas?. Consideramos interesante trabajar desde una perspectiva Bayesiana. Para hacer inferencia elegimos las distribuciones a posteriori condicionadas en covariables. Ésta acierta en dar una respuesta a la interrogante inicial, pues informa sobre la habilidad de un individuo *hipotético* proveniente de un cierto tipo de colegio y estrato social.

Metodológicamente, este trabajo propone estimar y comparar densidades a posteriori de las habilidades latentes. Las herramientas empleadas para el análisis son las densidades medias a posteriori condicionadas en covariables junto con bandas de credibilidad y divergencias de Kullback-Leibler. Ésta última se utiliza para cuantificar las diferencias entre densidades, mientras que las bandas de credibilidad permiten respaldar las características de forma de las densidades medias, como la bimodalidad y asimetría para algunos subgrupos.

Desde el punto de vista práctico, una de nuestras principales conclusiones es que, para los estudiantes de colegios particulares subvencionados y que no repiten, la variable SES afecta fuertemente las formas distribucionales. Por otra parte, las diferencias en SESp se reflejan como desplazamientos en las distribuciones de habilidad. La presencia de bimodalidad es otro resultado que debe ser destacado, pues indica la presencia de dos tipos de individuos al interior del subgrupo. Encontramos bimodalidad en los subgrupos de alumnos que no repiten, con SES medio y SESp medio y alto. También es destacable los resultados acerca de la presencia de asimetría, pues indica que la masa probabilidad está más cargada hacia valores bajos de habilidad (simetría a la izquierda) o altos de habilidad (simetría a la derecha). A partir del estudio se detecta una marcada asimetría hacia la izquierda para alumnos de colegios municipales y subvencionados que no repiten con SES y SESp bajo, y en la mayoría de los subgrupos de repitentes. La presencia de bimodalidad y asimetría al mismo tiempo indica que el subgrupo en cuestión contienen dos tipos distintos de individuos en su interior, donde uno de los tipos de individuo es menos frecuente que el otro. Tal es el caso de los subgrupos de alumnos que no repiten con SES y SESp altos. La asimetría a la derecha junto con la bimodalidad implican que el tipo de alumnos con menor habilidad es menos frecuente que el de mayor habilidad.

Si bien encontramos que las distribuciones de habilidad no cambian su localización y forma por tipo de colegio, este resultado debe tomarse con cautela ya que en el presente estudio no se tiene en cuenta el sesgo de selección que ocurre debido a que los alumnos no están asignados en forma aleatoria por tipo de colegio.

Desde el punto de vista teórico, esta tesis contribuye a introducir la estadística Bayesiana No Paramétrica dentro del área de la Psicometría. Es frecuente en esta disciplina que el interés se centre en variables a nivel individual que no son directamente

observables. Tal es el caso de la habilidad de estudiantes en los exámenes educativos. En estas circunstancias, el investigador debe tomar decisiones, muchas veces arbitrarias, acerca de la distribución de los parámetros de interés. Es ahí donde la estadística NPB puede ser explotada por la psicometría. Concretamente, se desarrolló un modelo IRT semiparamétrico con covariables, lo cual representa una innovación en sí misma. La covariables se introducen en el modelo a nivel de la distribución a priori a través de mezclas de ANOVA DDP, un tipo especial de Procesos Dirichlet Dependientes.

La disponibilidad de softwares para trabajar con estadística NPB es cada vez mayor: hoy día existe el paquete DPpackage de R de Jara (2007), que ofrece funciones para la inferencia Bayesiana de una gran variedad de modelos Bayesianos semiparamétricos. En lo que se refiere al modelo desarrollado aquí, su implementación práctica implicó el desarrollo de un software próximamente disponible al público en http://www.paulaestadistica.blogspot.com/.

Resta mencionar que este trabajo puede ser el comienzo para nuevos trabajos. Una posible extensión surge al aplicar el modelo a otros datos como por ejemplo la prueba PISA 2006. De hecho, algunos trabajos vinculados con los datos PISA han intentado hacer estudios dentro de esta línea, aunque no exactamente de la misma manera, ver Wo (2005). Este último paper representa el cálculo oficial que hace la OECD para reportar puntajes PISA. También sería importante trabajar en avances computacionales para mejorar la velocidad con que se realizan las estimaciones, probando otros algoritmos de nueva generación, como el de Dahl (2005). Finalmente, desde el punto de vista teórico, también es posible comparar otro tipo de distribuciones a priori no paramétricas para la habilidad.

Capítulo 6

Apéndices

Apéndice A

Restantes Distribuciones Condicionales Completas para el muestreo de Gibbs: σ^2_θ, μ_β, σ^2_β, Σ_ϑ y μ_ϑ

En este apéndice se exponen las restantes distribuciones condicionales completas del algoritmo desarrollado. Para una idea completa de la estructura del muestreo Gibbs, el lector debe remitirse a la sección 3.2. La notación de dicha sección se mantiene en este apéndice. Para actualizar σ^2_θ, su distribución condicional completa puede obtenerse analíticamente, y tiene la forma: $\sigma^2_\theta|(a_{-\sigma^2_\theta}, \eta, \beta_1, Y, D) \sim inv - \chi^2_{\nu_{\theta 1}}(s^2_{\theta 1})$, donde $\nu_{\theta 1} = (I + \nu_\theta)$, y $s^2_{\theta 1} = \{\sum_{i=1}^{I}(\theta_i - d_i'\vartheta^*_{s_i})^2 + \nu_\theta(s^2_\theta)\}/\nu_{\theta 1}$.

Por su parte, la distribución condicional completa para σ^2_β es: $\sigma^2_\beta|(a_{-\sigma^2_\beta}, \eta, \beta_1, Y, D) \sim inv-\chi^2_{\nu_{\beta 1}}(s^2_{\beta 1})$, donde $\nu_{\beta 1} = (J+\nu_\beta)$, y $s^2_{\beta 1} = \{\sum_{j=1}^{J}[(\beta_j - \bar{\beta})^2 + \frac{J}{J+1}\bar{\beta}^2 + \nu_\beta(s^2_\beta)]\}/\nu_{\beta 1}$. Para actualizar μ_β se emplea la distribución: $\mu_\beta|(a_{-\mu_\beta}, \eta, \beta_1, Y, D) \sim N(\mu_{\beta 1}, \sigma^2_\beta)$, con

$\mu_{\beta 1} = \frac{J}{J+1}\bar{\beta}$.

La distribución condicional completa de Σ_ϑ es , también, un caso conjugado. Está dada por: $\Sigma_\vartheta|(a_{-\Sigma_\vartheta}, \eta, \beta_1, Y, D) \sim inv - Wishart_{\nu_{\vartheta 1}}(\Lambda_{\vartheta 1}^{-1})$, donde $\nu_{\vartheta 1} = (R + 2 + L)$, $\Lambda_{\vartheta 1} = \Lambda_\vartheta + S + \frac{L}{L+1}(\bar{\vartheta^*} - \mu_\vartheta)(\bar{\vartheta^*} - \mu_\vartheta)'$, y $S = \sum_{l=1}^{L}(\vartheta_l^* - \bar{\vartheta^*})(\vartheta_l^* - \bar{\vartheta^*})'$. Finalmente, para actualizar μ_ϑ se utiliza la distribución: $\mu_\vartheta|(a_{-\mu_\vartheta}, \eta, \beta_1, Y, D) \sim N(\mu_{\vartheta 1}, \Sigma_\vartheta/(L+1))$, donde $\mu_{\vartheta 1} = \frac{1}{L+1}\mu_\vartheta + \frac{L}{L+1}\bar{\vartheta^*}$.

Apéndice B

Entradas y Salidas del procedimiento MIXED de SAS

En este apéndice se presentan los códigos y salidas para la estimación del modelo lineal jerárquico (HLM) mencionado en el capítulo 4. El código está basado en el procedimiento MIXED de SAS (SAS Institute, 1999). Esta herramienta se desarrolló para estimar modelos lineales mixtos, siendo útil para ajustar modelo lineales jerárquicos (HLM), pues éstos son casos particulares de modelos lineales mixtos. Para mayor claridad sobre este aspecto, y sobre como estimar HLM mediante el procedimiento MIXED ver Singer (1998). El código se presenta a continuación:

```
PROC MIXED data=l_rasch.p_lf_lc noclprint covtest;
class colegio repite tipocol;
model score = SESp SES repite tipocol/solution ddfm=bw;
random intercept/sub=colegio;
run;
```

En la declaración `class` se introducen las variables categóricas. Aquí están especificadas *repite*, *tipo de colegio*, y una variables más: *colegio*. Ella indica la pertenencia

a un determinado colegio y define el segundo nivel en la jerarquía del modelo. Las salidas del procedimiento anterior se exponen a continuación:

Procedimiento Mixed

Información del modelo

Conj. datos	L_RASCH.P_LF_LC
Variable dependiente	score
Estructura de covarianza	Variance Components
Efecto de asunto	colegio
Método de estimación	REML
Método de varianza del residual	Perfil
Método SE de efectos fijos	Basado en el modelo
Método de grados de libertad	Between-Within

Dimensiones

Parámetros de covarianza	2
Columnas en X	8
Columnas en Z por asunto	1
Asuntos	115
Obs máx por asunto	80

Número de observaciones

Number of Observations Read	3863
Number of Observations Used	3863
Number of Observations Not Used	0

Historia de iteración

Iteración	Evaluaciones	-2 Res Log Like	Criterio
0	1	26328.37926168	
1	2	26101.11880478	0.00000768
2	1	26101.04254375	0.00000004
3	1	26101.04218013	0.00000000

Se ha cumplido el criterio de convergencia.

Estimadores de parámetro de covarianza

Parm Cov	Asunto	Estimador	Error estándar	Valor Z	Pr Z
Intercept	colegio	5.9253	1.0240	5.79	<.0001
Residual		48.0311	1.1103	43.26	<.0001

Estadísticos de ajuste

Verosimilitud	-2 Res Log	26101.0
AIC (mejor más pequeño)		26105.0
AICC (mejor más pequeño)		26105.0
BIC (mejor más pequeño)		26110.5

Solución para efectos fijos

Efecto	tipocol	repite	Estimador	Error estándar	DF	Valor t	Pr > |t|
Intercept			13.0185	0.9858	111	13.21	<.0001
SESp			1.0905	0.1813	111	6.02	<.0001
SES			0.5391	0.08531	3746	6.32	<.0001
repite		1	3.5923	0.3708	99	9.69	<.0001
repite		2	0	.	.	.	.
tipocol	M		-1.0224	0.7168	111	-1.43	0.1566
tipocol	PP		-2.2350	1.1897	111	-1.88	0.0629
tipocol	PS		0	.	.	.	.

Tests de tipo 3 de efectos fijos

Efecto	Num DF	Den DF	F-Valor	Pr > F

SESp	1	111	36.19	<.0001
SES	1	3746	39.94	<.0001
repite	1	99	93.86	<.0001
tipocol	2	111	3.80	0.0252

Apéndice C

Supuestos de independencia condicional del modelo propuesto

En este apéndice se estudia el modelo propuesto, para dejar en evidencia los supuestos de independencia condicional necesarios para especificarlo completamente. Se muestra que los Supuestos 3.1-3.6 son suficientes para este fin. Siguiendo la notación de la sección 3.1, los elementos aleatorios del modelo son las respuestas Y, junto con los parámetros a, con $a = (G, \sigma^2_\theta, \beta_2, ..., \beta_J, \mu_\beta, \sigma^2_\beta, \mu_\vartheta, \Sigma_\vartheta)$. El modelo Bayesiano queda completamente especificado a través de la distribución conjunta de todos los elementos aleatorios, es decir la distribución de (Y, a).

Para dejar en claro cuáles son los supuestos de independencia condicional implícitos en el modelo propuesto, hay que dilucidar las condiciones para que $p(Y, a)$ pueda ser expresada como la multiplicación de las ecuaciones (3.1), (3.4), (3.5), (3.6), (3.7), (3.8), (3.9) y (3.10). Es decir, hay que encontrar las condiciones para que $p(Y, a)$ pueda ser expresada como:

$$p(Y,a) = \prod_{ij} p(Y_{ij}|\beta_j,(\sigma_\theta^2,G),d_i)p(G|\mu_\vartheta,\Sigma_\vartheta)p(\sigma_\theta^2)p(\mu_\vartheta|\Sigma_\vartheta)\times$$

$$p(\Sigma_\vartheta)\prod_j p(\beta_j|\mu_\beta,\sigma_\beta^2)p(\mu_\beta|\sigma_\beta^2)p(\sigma_\beta^2),$$

donde se reemplazó G_k por el par (σ_θ^2, G) en la primer densidad del lado derecho de la igualdad, ya que no implica ningún cambio en la probabilidad definida en (3.1). No se deja explícito en las densidades que se está condicionando con respecto a los hiperparámetros para simplificar la exposición, ya que no son relevantes en este análisis.

La distribución conjunta de (Y, a) siempre puede ser descompuesta en un modelo estadístico y un modelo marginal para los parámetros, es decir $p(Y,a) = p(Y|a)p(a)$. Para que $p(Y|a)$ sea equivalente al modelo estadístico presentado en la ecuación (3.1) deben cumplirse las siguientes relaciones de independencia condicional:

- $(G, \sigma_\theta^2, \beta_2, ..., \beta_J)$ son parámetros suficientes, es decir, $Y \perp\!\!\!\perp (\mu_\beta, \sigma_\beta^2, \mu_\vartheta, \Sigma_\vartheta)|(G, \sigma_\theta^2, \beta_2, ..., \beta_J)$ (**Supuesto 3.1**)
- $Y_1, ..., Y_I$ son independientes condicional en $((G, \sigma_\theta^2), \beta_2, ..., \beta_J, D)$ (**Supuesto 3.2**)
- $Y_{i1}, ..., Y_{iJ}$ son independientes condicional en $((G, \sigma_\theta^2), \beta_2, ..., \beta_J, d_i)$ (**Supuesto 3.3**)

Para analizar la distribución marginal de los parámetros, se denota $a^\theta = (G, \sigma_\theta^2, \mu_\vartheta, \Sigma_\vartheta)$ los parámetros vinculados a las habilidades, y $a^\beta = (\beta_2, ..., \beta_J, \mu_\beta, \sigma_\beta^2)$, los relacionados con las dificultades. Si se supone que los parámetros que generan la dificultad son independientes de los que generan las distribuciones de habilidad (**Supuesto 3.4**), se puede descomponer $p(a)$ como: $p(a) = p(a^\theta)p(a^\beta)$.

Ahora bien, la distribución de los parámetros vinculados a la habilidad puede a su vez descomponerse como $p(a^{\theta}) = p(G|a^{\theta}_{-G})p(a^{\theta}_{-G})$. Para que el primer término del lado derecho coincida con la ecuación (3.4) debe verificarse $G \perp\!\!\!\perp \sigma^2_{\theta}|(\mu_{\vartheta}, \Sigma_{\vartheta})$ (**Supuesto 3.5**). A su vez, para expresar $p(a^{\theta}_{-G})$ como indican las ecuaciones (3.5), (3.6) y (3.7), se requiere la condición $\sigma^2_{\theta} \perp\!\!\!\perp (\mu_{\vartheta}, \Sigma_{\vartheta})$ (**Supuesto 3.6**). En tal caso $p(a^{\theta}) = p(G|\mu_{\vartheta}, \Sigma_{\vartheta})p(\sigma^2_{\theta})p(\mu_{\vartheta}|\Sigma_{\vartheta})p(\Sigma_{\vartheta})$. Haciendo un razonamiento análogo para la parte vinculada a los parámetros de dificultad se tiene que $p(a^{\beta}) = p(\beta|a^{\beta}_{-\beta})p(a^{\beta}_{-\beta}) = \prod_{j=2}^{J} p(\beta_j|\mu_{\beta}, \sigma^2_{\beta})p(\mu_{\beta}|\sigma^2_{\beta})p(\sigma^2_{\beta})$ si se supone que $\perp\!\!\!\perp_{2 \leq j \leq J} \beta_j|(\mu_{\beta}, \sigma^2_{\beta})$.

Apéndice D

Diagnóstico de convergencia para el modelo propuesto

La primer tabla de este apéndice muestra estadísticas descriptivas de algunas cadenas involucradas en el muestreo de Gibbs. Se considera un solo parámetro de dificultad y uno de habilidad. La segunda tabla hace lo propio para las cadenas de 3 valores de la grilla de las densidades estimadas. Se exponen sólo las densidades de alumnos de colegios particulares subvencionados que no repiten, en el orden presentado en la tabla 4.8. Los valores elegidos de la grilla son -1.75, 0.49 y 2.74, denotadas con el subíndice 1,2 y 3 respectivamente. Se empleo el paquete boa de R (Smith, 2004).

```
boa.init()
cadenas=boa.importMatrix("cadenas")
boa.stats(cadenas,probs=c(0.025,0.5,0.975),batch.size=40)
```

	Mean	SD	Naive SE	MC Error	Batch SE	Batch ACF	0.025	0.5	0.975	MinIter	MaxIter	Sample
beta2	-1.351	0.071	0.002	0.004	0.003	0.120	-1.491	-1.351	-1.213	1	2000	2000
theta1	1.434	0.283	0.006	0.007	0.007	0.126	0.926	1.420	1.994	1	2000	2000

mu_beta	0.827	0.132	0.003	0.004	0.004	-0.066	0.571	0.828	1.082	1	2000	2000
sig_beta	0.750	0.152	0.003	0.003	0.003	0.168	0.510	0.730	1.109	1	2000	2000
sig_theta	0.060	0.018	0.000	0.001	0.001	0.116	0.032	0.058	0.099	1	2000	2000
mu_vartheta1	-0.143	0.235	0.005	0.006	0.006	-0.072	-0.581	-0.149	0.318	1	2000	2000
mu_vartheta2	0.502	0.233	0.005	0.006	0.006	-0.039	0.066	0.493	0.989	1	2000	2000
mu_vartheta3	0.097	0.124	0.003	0.003	0.003	-0.039	-0.139	0.095	0.358	1	2000	2000
mu_vartheta4	0.091	0.121	0.003	0.003	0.003	-0.226	-0.139	0.089	0.331	1	2000	2000
mu_vartheta5	0.145	0.192	0.004	0.005	0.005	0.003	-0.214	0.140	0.543	1	2000	2000
mu_vartheta6	0.237	0.200	0.004	0.005	0.005	-0.018	-0.149	0.232	0.639	1	2000	2000
sig_vartheta1_1	0.294	0.196	0.004	0.005	0.005	-0.236	0.096	0.240	0.776	1	2000	2000
sig_vartheta1_2	0.043	0.149	0.003	0.004	0.004	-0.120	-0.243	0.037	0.370	1	2000	2000
sig_vartheta1_3	0.002	0.075	0.002	0.002	0.002	0.170	-0.129	0.002	0.147	1	2000	2000
sig_vartheta1_4	-0.024	0.072	0.002	0.001	0.001	-0.078	-0.184	-0.018	0.097	1	2000	2000
sig_vartheta1_5	0.009	0.114	0.003	0.003	0.003	-0.002	-0.222	0.012	0.226	1	2000	2000
sig_vartheta1_6	0.039	0.118	0.003	0.003	0.003	-0.244	-0.181	0.033	0.291	1	2000	2000
sig_vartheta2_2	0.378	0.284	0.006	0.007	0.007	0.065	0.111	0.307	1.010	1	2000	2000
sig_vartheta2_3	-0.006	0.083	0.002	0.002	0.002	0.060	-0.173	-0.002	0.145	1	2000	2000
sig_vartheta2_4	-0.011	0.081	0.002	0.002	0.002	-0.113	-0.180	-0.007	0.141	1	2000	2000
sig_vartheta2_5	0.044	0.131	0.003	0.003	0.003	-0.047	-0.197	0.036	0.325	1	2000	2000
sig_vartheta2_6	0.068	0.138	0.003	0.003	0.003	-0.086	-0.172	0.056	0.390	1	2000	2000
sig_vartheta3_3	0.128	0.078	0.002	0.002	0.002	-0.068	0.051	0.110	0.316	1	2000	2000
sig_vartheta3_4	-0.010	0.051	0.001	0.001	0.001	-0.016	-0.106	-0.007	0.076	1	2000	2000
sig_vartheta3_5	0.009	0.062	0.001	0.001	0.002	-0.168	-0.109	0.008	0.139	1	2000	2000
sig_vartheta3_6	0.002	0.067	0.002	0.002	0.002	-0.013	-0.131	0.001	0.133	1	2000	2000
sig_vartheta4_4	0.122	0.066	0.001	0.002	0.002	-0.053	0.048	0.106	0.295	1	2000	2000
sig_vartheta4_5	-0.002	0.065	0.001	0.001	0.001	0.040	-0.135	-0.001	0.124	1	2000	2000
sig_vartheta4_6	-0.008	0.065	0.001	0.002	0.002	-0.207	-0.154	-0.005	0.110	1	2000	2000
sig_vartheta5_5	0.229	0.161	0.004	0.004	0.004	-0.071	0.074	0.187	0.620	1	2000	2000
sig_vartheta5_6	0.061	0.119	0.003	0.003	0.003	0.060	-0.116	0.043	0.343	1	2000	2000
sig_vartheta6_6	0.248	0.159	0.004	0.003	0.003	-0.135	0.082	0.205	0.672	1	2000	2000
L	10.757	1.975	0.044	0.075	0.075	-0.058	7.000	11.000	15.000	1	2000	2000

	Mean	SD	Naive SE	MC Error	Batch SE	Batch ACF	0.025	0.5	0.975	MinIter	MaxIter	Sample
th_pred_17_1	0.00006	0.00038	0.00001	0.00001	0.00001	0.00197	0.00000	0.00002	0.00033	1	2000	2000
th_pred_17_2	0.46793	0.10423	0.00233	0.00392	0.00345	0.04174	0.24976	0.47309	0.66331	1	2000	2000
th_pred_17_3	0.08941	0.02977	0.00067	0.00118	0.00117	0.04829	0.02842	0.09042	0.14705	1	2000	2000
th_pred_18_1	0.00006	0.00030	0.00001	0.00001	0.00001	0.02051	0.00001	0.00002	0.00037	1	2000	2000
th_pred_18_2	0.45071	0.07583	0.00170	0.00300	0.00289	0.25265	0.28730	0.45927	0.57979	1	2000	2000
th_pred_18_3	0.14484	0.03759	0.00084	0.00119	0.00105	-0.04245	0.07232	0.14471	0.22173	1	2000	2000
th_pred_19_1	0.00050	0.00214	0.00005	0.00004	0.00004	0.05155	0.00001	0.00002	0.00516	1	2000	2000

```
th_pred_19_2 0.11383 0.08849  0.00198  0.00268  0.00275   0.08493 0.00451 0.09574 0.32714       1    2000    2000
th_pred_19_3 0.09693 0.06577  0.00147  0.00213  0.00194  -0.04133 0.00358 0.09074 0.25764       1    2000    2000
th_pred_20_1 0.00014 0.00079  0.00002  0.00002  0.00002  -0.03769 0.00000 0.00002 0.00124       1    2000    2000
th_pred_20_2 0.53905 0.10524  0.00235  0.00498  0.00424   0.08746 0.31262 0.54158 0.73360       1    2000    2000
th_pred_20_3 0.09492 0.03576  0.00080  0.00105  0.00090  -0.02994 0.02174 0.09576 0.16645       1    2000    2000
th_pred_21_1 0.00003 0.00009  0.00000  0.00000  0.00000   0.12943 0.00001 0.00002 0.00028       1    2000    2000
th_pred_21_2 0.40106 0.06984  0.00156  0.00320  0.00300   0.09605 0.24634 0.40284 0.52819       1    2000    2000
th_pred_21_3 0.18057 0.04436  0.00099  0.00173  0.00183  -0.11317 0.09622 0.17917 0.27177       1    2000    2000
th_pred_22_1 0.00032 0.00115  0.00003  0.00002  0.00002  -0.12270 0.00001 0.00002 0.00351       1    2000    2000
th_pred_22_2 0.14344 0.07584  0.00170  0.00216  0.00215   0.12839 0.01733 0.13859 0.29769       1    2000    2000
th_pred_22_3 0.09187 0.06289  0.00141  0.00202  0.00201   0.02281 0.00615 0.08424 0.25302       1    2000    2000
th_pred_23_1 0.00023 0.00088  0.00002  0.00002  0.00002  -0.02841 0.00001 0.00002 0.00241       1    2000    2000
th_pred_23_2 0.60385 0.11376  0.00254  0.00444  0.00430   0.20569 0.37431 0.60589 0.82632       1    2000    2000
th_pred_23_3 0.11032 0.04985  0.00111  0.00134  0.00119  -0.00253 0.01826 0.10719 0.22245       1    2000    2000
th_pred_24_1 0.00005 0.00021  0.00000  0.00000  0.00000   0.39436 0.00001 0.00002 0.00041       1    2000    2000
th_pred_24_2 0.36857 0.07134  0.00160  0.00269  0.00256  -0.13514 0.22614 0.36673 0.51595       1    2000    2000
th_pred_24_3 0.19067 0.05558  0.00124  0.00209  0.00219   0.02206 0.08611 0.19033 0.30253       1    2000    2000
th_pred_25_1 0.00024 0.00093  0.00002  0.00002  0.00002  -0.15564 0.00001 0.00002 0.00252       1    2000    2000
th_pred_25_2 0.19456 0.06419  0.00144  0.00172  0.00174   0.05877 0.06209 0.19818 0.31177       1    2000    2000
th_pred_25_3 0.12665 0.07015  0.00157  0.00211  0.00208  -0.01168 0.01987 0.11776 0.28833       1    2000    2000
```

A continuación se presentan los resultados de los tests de Geweke, y de Hedelberger y Welch. Para el segundo, todas las cadenas pasan el test de estacionaridad, pero algunas fallan en el test Halfwidth (de correlación). El test de Geweke indica falta de consistencia (p valores menores a 0.05) sólo para 8 cadenas. Se listan sólo las cadenas con problemas.

```
boa.geweke(cadenas,p.first=0.1,p.last=0.5)
```

```
                   Z-Score      p-value
sig_vartheta1_2  2.10429586 0.0353526472
sig_vartheta1_3 -2.74794070 0.0059970855
sig_vartheta4_4 -2.23787285 0.0252293480
th_pred_9_1     -3.32209098 0.0008934556
th_pred_21_2    -2.22448262 0.0261159903
th_pred_24_1    -2.65364691 0.0079627100
th_pred_24_2    -2.29260193 0.0218709331
th_pred_30_1    -2.71087217 0.0067106493
```

```
boa.handw(cadenas,error=0.1,alpha=0.05)
```

	Stationarity Test	Keep	Discard	C-von-M	Halfwidth Test	Mean	Halfwidth
sig_vartheta1_2	passed	2000	0	0.38354805	failed	0.0429613160	8.238530e-03
sig_vartheta1_3	passed	1600	400	0.26203006	failed	0.0035856737	3.587857e-03
sig_vartheta1_4	passed	2000	0	0.12843763	failed	-0.0240181165	2.890606e-03
sig_vartheta1_5	passed	2000	0	0.06112080	failed	0.0090219225	5.270321e-03
sig_vartheta1_6	passed	2000	0	0.14890270	failed	0.0393799720	6.109576e-03
sig_vartheta2_3	passed	2000	0	0.44255616	failed	-0.0057821910	3.422464e-03
sig_vartheta2_4	passed	2000	0	0.08961150	failed	-0.0110825255	3.527494e-03
sig_vartheta2_5	passed	1800	200	0.30327146	failed	0.0456334211	6.252213e-03
sig_vartheta3_4	passed	2000	0	0.08302397	failed	-0.0104877220	2.199651e-03
sig_vartheta3_5	passed	2000	0	0.06090976	failed	0.0089354240	2.782937e-03
sig_vartheta3_6	passed	2000	0	0.14902891	failed	0.0022768200	3.493280e-03
sig_vartheta4_5	passed	2000	0	0.03172647	failed	-0.0018254545	2.875669e-03
sig_vartheta4_6	passed	2000	0	0.06816478	failed	-0.0083317510	3.279363e-03
th_pred_17_1	passed	2000	0	0.22352615	failed	0.0000556000	1.626795e-05
th_pred_18_1	passed	2000	0	0.25543872	failed	0.0000550625	1.266515e-05
th_pred_19_1	passed	2000	0	0.03945125	failed	0.0004987640	8.810018e-05
th_pred_20_1	passed	2000	0	0.09596846	failed	0.0001354305	3.692676e-05
th_pred_21_1	passed	2000	0	0.15815529	failed	0.0000336530	3.802829e-06
th_pred_22_1	passed	2000	0	0.08383770	failed	0.0003216355	4.415185e-05
th_pred_23_1	passed	2000	0	0.08140114	failed	0.0002255330	3.770340e-05
th_pred_24_1	passed	2000	0	0.25012612	failed	0.0000513065	8.936627e-06
th_pred_25_1	passed	2000	0	0.04932641	failed	0.0002378315	4.056040e-05

Bibliografía

Antoniak, C. (1974), "Mixtures of Dirichlet Processes with Applications to Bayesians Nonparametric Problems," *The Annals of Statistics*, 2(6), 1152–1174.

Blackwell, D. y MacQueen, J. B. (1973), "Ferguson Distributions Via Polya Urn Schemes," *The Annals of Statistics*, 1(2), 353–355.

Bush, C. y MacEachern, S. (1996), "A Semiparametric Bayesian Model for Randomised Block Designs," *Biometrika*, 83, 275–85.

Chen, M.-H., Shao, Q.-M., y Ibrahim, J. G. (2000), *Monte Carlo methods in Bayesian computation*, Springer Series in Statistics, New York: Springer-Verlag.

Cohen, A. y Bolt, D. (2005), "A Mixture Model Analysis of Differential Item Functioning," *Journal of Educational Measurement*, 42, 133–148.

Cowles, M. y Carlin, B. (1996), "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review." *Journal of American Statistical Association*, 91(434), 883–904.

Dahl, D. (2005), "Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 1, 281–301.

De Iorio, M., Müller, P., Rosner, G., y MacEachern, S. (2004), "An ANOVA model for dependent random measures," *Journal of the American Statistical Association*, 99(465), 205–215.

DeBoeck, P. y Wilson, M. (2004), *Explanatory item response models : a generalized linear and nonlinear approach*, New York, Springer.

Duncan, K. (2004), "Case and Covariate Influence: Implications for Model Assessment," Ph.D. thesis, Ohio State University.

Duncan, K. y MacEachern, S. (2008), "Nonparametric Bayesian modeling for item response," *Statistical Modelling*, 8(1), 41–66.

Dunson, D. (2007), "Empirical Bayes density regression," *Statistica Sinica*, 17, 481–504.

Dunson, D. y Park, J.-H. (2008), "Kernel stick-breaking processes," *Biometrika*, 95(2), 307–323.

Dunson, D., Pillai, N., y Park, J.-H. (2007), "Bayesian density regression," *Journal of the Royal Statistical Society*, B(69), 163–83.

Escobar, M. (1994), "Estimating Normal Means with a Dirichlet Process Prior," *Journal of the American Statistical Association*, 89, 267–278.

Escobar, M. y West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Ferguson, T. S. (1973), "A Bayesian Analysis of some Nonparametric Problems," *The Annals of Statistics*, 1(2), 209–230.

Fisher, G. y Molenaar, I. (1995), *Rasch Models. Foundations, Recent Developments, and Applications*, New York: Springer-Verlag.

Florens, J., Mouchart, M., y Rolin, J.-M. (1990), *Elements of Bayesian Statistics*, New York: Marcel Dekker.

Friedman, M. (1955), *Economics and the Public Interest*, New Brunswick, NY: Rutgers University Press, chap. The Role of Government in Education, pp. 123–144.

Geisser, S. y Eddy, W. (1979), "A Predictive Approach to Model Selection," *Journal of the American Statistical Association*, 74(365), 153–160.

Gelfand, A., Dey, D., y Chang, H. (1992), "Model determination using predictive distributions with implementation via sampling-based methods," in *Bayesian statistics, 4 (Peñíscola, 1991)*, New York: Oxford Univ. Press, pp. 147–167.

Gelfand, A., Kottas, A., y Maceachern, S. (2005), "Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing," *Journal of the American Statistical Association*, 100, 1021–35.

Gelfand, A. y Smith, A. (1990), "Sampling based approaches to calculating marginal densities." *Journal of the American Statistical Association.*, 85, 398–409.

Gelfand, A. E. y Kottas, A. (2002), "A Computational Approach for Full Nonparametric Bayesian Inference under Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 11(2), 289–305.

Gelman, A., Carlin, J., y Rubin, D. (1995a), *Bayesian Data Analysis*, Great Britain.

Gelman, A. y Rubin, D. (1992), "Inference from iterative simulation using multiple sequences (with discussion)." *Statist. Sci*, 7, 457–511.

Gelman, G., Carlin, J., Stern, H., y D.B., R. (1995b), *Bayesian Data Analysis*, London: Chapman-Hall.

Geweke, J. (1992), *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments.*, in Bayesian Statistics 4: Proceedings of the Fourth Valencia International Conference on Bayesian Statistics. Bernardo, J.M. and Berger, J.O. and Dawid, A.P. and Smith A.F.M. (eds): Oxford. Oxford University Press.

Goldstein, H. (1995), *Multilevel statistical models*, London: Edward Arnold, 2nd ed.

González, J. y San Martín, E. (2009), "Rendimiento de la prueba PISA: ¿Es posible entender los alcances y límites de las comparaciones entre países?." *Estudios Internacionales, SIMCE. Unidad de Curriculum y Evaluación, MINEDUC.*

Gosch, J. y Ramamoorthi, R. (2003), *Bayesian Non Parametrics*, New York: Springer-Verlag, 1st ed.

Griffin, J. y Steel, M. (2006), "Order-based dependent Dirichlet processes," *Journal of the American Statistical Association*, 101, 179–94.

Hedelberger, P. y Welch, P. (1983), "Simulation Run Length Control in the Presence of an Initial Transient." *Operations Research*, 31(6), 1109–1144.

Hsieh, C.-T. y Urquiola, M. (2006), "The effects of generalized school choice on achievement and stratification: Evidence from Chile's voucher program," *Journal of Public Economics*, 90, 1477–1503.

Jara, A. (2007), "Applied Bayesian Non- and Semi-parametric Inference using DPpackage," *Rnews*, 7(3), 17–26.

Karabatsos, G. y Walker, S. (2009), "A Bayesian Non Parametric Approach to Test Equating." *Psychometrika*, 74(2), 211–232.

Korwar, R. y Hollander, M. (1973), "Contributions to the Theory of Dirichlet Processes," *The Annals of Statistics*, 1(4), 705–711.

Kullback, S. y Leibler, R. A. (1951), "On information and sufficiency," *Annals of Mathematical Statistics*, 22, 79–86.

Lara, B., Mizala, A., y A., R. (2009), "The Effectiveness of Private Voucher Education: Evidence from Structural School Switches," *Centro de Economía Aplicada. Ingeniería Industrial. Univesidad de Chile.*, 263.

Liu, J. (1996), "Non Parametric Hierarchical Bayes Via Sequential Imputation," *The Annals of Statistics*, 24, 911–30.

Lo, A. Y. (1984), "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates," *The Annals of Statistics*, 12(1), 351–357.

MacEachern, S. (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA*, American Statistical Association.

MacEachern, S. y Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–38.

MacEachern, S.Ñ. (1998), "Computational Methods for Mixture of Dirichlet Process Models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. Dey, D., Müller, P., y Sinha, D., pp. 23–43.

Manzi, J., Strasser, K., San Martín, E., y Contreras, D. (2008), "Quality of education in Chile," Tech. rep., MIDE. Centro de Medición UC.

Miyazaki, K. y Hoshino, T. (2009), "A Bayesian Semiparametric Item Response Model with Dirichlet Process Prior." *Psychometrika*, 74(1), 1–19.

Mizala, A. y Romaguera, P. (2000), "School Performance and Choice: The Chilean Experience," *The Journal of Human Resources*, 35, 392–417.

— (2001), "Factores explicativos de los resultados escolares en la educación secundaria en Chile," *El Trimestre Económico*, 272, 515–549.

Mouchart, M. y San Martín, E. (2003), "Specification and identification issues in models involving a latent hierarchical structure," *Journal of Statistical Planning and inference*, 111, 143–163.

Müller, P. y Quintana, F. (2004a), "A Method for Combining Inference across Related Nonparametric Bayesian Models," *Journal of the Royal Statistical Society*, B66(3), 735–749.

— (2004b), "Non Parametric Bayesian Data Analysis," *Statistical Science*, 19(1), 95–110.

Neal, R. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9(2), 249–265.

PISA (2006), "Science competencies for tomorrow's world," Tech. rep., OECD.

Quin, L. (1998), "Non Parametric Bayesian Models for Item Response Data," Ph.D. thesis, Ohio State University.

Ramírez, M. J. (2002), "Perfil de rendimiento de Chile en la sub-escala de representación de datos TIMSS 1999." *Estudios Pedagógicos*, 28, 89–107.

Ramsay, J. (1991), "Kernel smoothing approaches to non parametric item characteristic curve estimation," *Psychometrika*, 56, 611–613.

Rasch, G. (1960), *Probabilistic models for some intelligence and attainment tests*, Copenhagen: Danish Institute for Educational Research.

Roberts, G. O. y Rosenthal, J. (1998), "Markov-chain Monte Carlo: Some practical implications of theoretical results," *The Canadian Journal of Statistics*, 26, 5–31.

Rolin, J.-M. (1992), "Some Useful Properties of the Dirichlet Process," Discussion Paper 92-02, Institut de statistique, Université catholique de Louvain.

San Martín, E., Jara, A., Rolin, J.-M., y Mouchart, M. (2008), "On the Analysis of Bayesian Semiparametric IRT-type Models," Tech. rep., Departamento de Estadística, Pontificia Universidad Católica de Chile.

San Martín, E. y Rolin, J.-M. (2009), "Identification of Generalized Linear Models for Binary Outcomes." *Biometrika*, en prensa.

Sapelli, C. y Vial, B. (2002), "The performance of private and public schools in the Chilean voucher system," *Cuadernos de Economía*, 39(118), 423–454.

— (2005), "Private vs public voucher schools in Chile: New evidence on efficiency and peer effects," *Working paper N° 289 Instituto de Economía. Catholic University of Chile.*

SAS Institute (1999), *SAS OnlineDoc (Version 8)*, Cary, NC, USA.: SAS Institute Inc.

Sethuraman, J. (1994), "A constructive definition of Dirichlet process prior," *Statistica Sinica*, 2, 639–650.

Sethuraman, J. y Tiwari, R. (1982), *Convergence of Dirichlet measures and the interpretation of their parameter.*, New York-London:Academic Press: Statistical decision theory and related topics, III, Vol. 2 (West Lafayette, Ind., 1981) 305-315.

Singer, J. (1998), "Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models." *Journal of Educational and Behavioral Statistics.*, 24(4), 323–355.

Smith, B. J. (2004), *Bayesian Output Analysis Program (BOA). Version 1.1.2 for S-PLUS and R*, Available at http://www.public-health.uiowa.edu/boa.

Tierney, L. (1994), "Markov Chains for exploring posterior distributions." *The Annals of Statistics.*, 22(4), 1701–1762.

Verhelst, N. y Eggen, T. (1989), "Psychometrische en statistische aspecten van peilingsonderzoek (PPON rapport 4)," Tech. rep., Arnhem: Cito.

Walker, S., Damien, P., Laud, P., y Smith, A. (1999), "Bayesian Nonparametric Inference for Random Distributions and Related Functions," *Journal of the Royal Statistical Society*, 61(3), 485–527.

West, E. (1997), "Education Vouchers in Principle and Practice: A Survey," *The World Bank Research Observer*, 12, 83–103.

West, M. (1990), *Bayesian Kernel Density Estimation*, Discussion paper 90-A02, Duke University, Institute of Statistics and Decision Sciences.

Wo, M. (2005), "The Role of Plausible Values in Large-Scale Surveys," *Studies in Educational Evaluation*, 31, 114–128.

Woods, C. M. y Thissen, D. (2006), "Item Response Theory with estimation of the latent population distribution using Spline-Based densities," *Psychometrika*, 71(2), 281–301.