



COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

POSTGRADO EN SOCIOECONOMÍA, ESTADÍSTICA E
INFORMÁTICA

ESTADÍSTICA

ESTIMACIÓN DE LOS MODELOS DE REGRESIÓN LOG-NORMAL SESGADO Y PROBIT SESGADO LATENTE EN ÁREAS PEQUEÑAS MEDIANTE INFERENCIA BAYESIANA VARIACIONAL

SAUL ARTURO ORTIZ MUÑOZ

T E S I S
PRESENTADA COMO REQUISITO PARCIAL
PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS

MONTECILLO, TEXOCO, ESTADO DE MÉXICO

2026

ESTIMACIÓN DE LOS MODELOS DE REGRESIÓN LOG-NORMAL SESGADO Y PROBIT SESGADO LATENTE EN ÁREAS PEQUEÑAS MEDIANTE INFERNERIA BAYESIANA VARIACIONAL

Saul Arturo Ortiz Muñoz, M.C.
Colegio de Postgraduados, 2026

RESUMEN

Se proponen tres modelos de regresión Bayesiana variacional en áreas pequeñas con errores normales asimétricos: uno para respuesta continua y otro para respuesta binaria/ordinal. En estos dos últimos, el ajuste es posible gracias al uso de variables latentes. Se siguió un enfoque Bayesiano objetivo: se empleó la distribución de referencia para los parámetros de forma de la distribución normal sesgada y una *a priori* para realizar búsqueda estocástica de variables. Al resto de parámetros se asignaron densidades planas. Se desarrollaron estudios de simulación para cada modelo y se comparó contra el método Hamiltoniano Monte Carlo, basado en *Markov chain Monte Carlo*. Se encontró que el método variacional es bastante competitivo cuando se ajusta a respuestas continuas. Se estudió un conjunto de datos sobre el ingreso corriente total per cápita en los hogares de la Ciudad de México, con información proveniente de fuentes oficiales y siguiendo los criterios de procesamiento establecidos por el Coneval. El modelo continuo señala que la variable respuesta presenta distintos grados de sesgo por alcaldía. Se identificaron las covariables más relevantes para los modelos de regresión. Se concluye que el método variacional es una alternativa viable para acelerar la inferencia sin pérdidas abrumadoras de precisión, con respecto a los algoritmos Bayesianos usuales basados en muestreo. La aproximación variacional mostró ser especialmente útil con las respuestas continuas.

Palabras clave: Distribución normal asimétrica, estimación en áreas pequeñas, inferencia Bayesiana variacional, regresión probit.

**ESTIMATION OF SKEWED LOG-NORMAL AND LATENT SKEWED
PROBIT REGRESSION MODELS IN SMALL AREAS USING
VARIATIONAL BAYESIAN INFERENCE**

Saul Arturo Ortiz Muñoz, M.C.
Colegio de Postgraduados, 2026

ABSTRACT

Key words: skew-normal distribution, small-area estimation, variational bayes, probit regression.

AGRADECIMIENTOS

A la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) por la ayuda financiera.

Al Colegio de Postgraduados, campus Montecillo (COLPOS), por la formación y desarrollo profesional.

A los miembros del consejo particular, gracias por sus comentarios valiosos y acertados.

El regalo más hermoso es el perdón, lo más maravilloso del mundo es el amor, la felicidad mas dulce es la paz.

CONTENIDO

RESUMEN	ii
ABSTRACT	iii
LISTA DE FIGURAS	x
LISTA DE CUADROS	xv
CAPÍTULO 1. INTRODUCCIÓN	1
1.1 Objetivos	2
1.2 Hipótesis	3
CAPÍTULO 2. REVISIÓN DE LITERATURA	4
2.1 Distribución normal asimétrica y truncamiento oculto	4
2.1.1 Distribución normal asimétrica	5
2.1.2 Distribución normal asimétrica multivariada	7
2.1.3 Distribución normal asimétrica: parametrización centrada	9
2.1.4 Distribución normal asimétrica multivariada: parametrización centrada	11
2.1.5 Proceso de truncamiento oculto	12
2.1.6 Representación alternativa	15
2.2 Inferencia Bayesiana	15
2.2.1 Estimadores Bayesianos	17
2.2.2 Métodos de Monte Carlo	18
2.2.3 Cadenas de Markov en un espacio de estados general	20
2.2.4 Métodos Markov Chain Monte Carlo	21
2.2.5 Algoritmo Hamiltonian Monte Carlo	22
2.2.6 Ejemplo 1: aproximando una distribución normal bivariada	25
2.3 Inferencia Bayesiana variacional	26
2.3.1 Divergencia Kullback-Leibler	28
2.3.2 Ejemplo: divergencia KL entre las densidades normal y normal asimétrica	29
2.3.3 Límite inferior de la evidencia	31
2.3.4 Restricción Campo Medio	33

2.3.5	Algoritmo Inferencia Variacional por Ascenso de Coordenadas	33
2.3.6	Ejemplo 1: aproximando una distribución normal bivariada (revisitado)	35
2.3.7	Ejemplo 2: inferencia para la distribución normal	37
2.3.8	Ejemplo 3: inferencia para la distribución normal asimétrica	42
2.3.9	Restricción Forma Fija.....	47
2.3.10	Ejemplo 3: inferencia para la distribución normal asimétrica (revisitado)	48
2.3.11	<i>ADVI: Automatic Differentiation Variational Inference</i>	53
	Aspectos del algoritmo ADVI	58
	Algoritmo ADVI	61
CAPÍTULO 3. METODOLOGÍA	63
3.1	Descripción de los modelos en áreas pequeñas	63
3.1.1	Modelo a nivel unidad con error anidado.....	67
3.2	Modelo log-normal sesgado	71
3.3	Modelo probit sesgado latente	74
3.4	Modelo probit ordenado sesgado latente	78
3.5	Distribución <i>a priori</i> de referencia	81
3.6	Distribución <i>a priori</i> para la búsqueda estocástica de variables	85
3.7	Distribución <i>a posteriori</i>	90
3.8	Predicción de nuevas observaciones ($n_i > 0$ y $n_i = 0$).....	93
3.9	Implementación con métodos <i>Markov Chain Monte Carlo</i>	95
3.10	Implementación en Stan	96
3.10.1	Modelo log-normal sesgado centrado.....	100
3.10.2	Modelo probit ordenado sesgado latente centrado	104
3.10.3	Modelo probit sesgado latente centrado	107
3.11	Simulación y cantidades de interés	107
3.12	Caso de estudio: medición de la pobreza	113
3.12.1	Antecedentes.....	114
3.12.2	Fuentes de información y procesamiento	116
3.12.3	Ruta de trabajo	118
CAPÍTULO 4. RESULTADOS	121

4.1	Efecto de ρ_i con μ_i	122
4.2	Ajuste de los modelos de regresión sin restricción en ρ_i	123
4.2.1	Modelo log-normal sesgado	124
4.2.2	Modelo probit sesgado con variable latente	124
4.2.3	Modelo probit ordenado sesgado con variable latente	125
4.3	Ajuste del Modelo log-normal sesgado, datos simulados	126
4.3.1	Modelos con interceptos en cada área pequeña	126
4.4	Ajuste del Modelo probit sesgado con variable latente, datos simulados	130
4.4.1	Modelos sin interceptos en cada área pequeña	130
4.4.2	Modelos con interceptos en cada área pequeña	133
4.4.3	Modelos con único intercepto en todas las áreas pequeñas	137
4.5	Ajuste del Modelo probit ordenado sesgado con variable latente, datos simulados	141
4.5.1	Modelos sin interceptos en cada área pequeña	141
4.5.2	Modelos con interceptos en cada área pequeña	145
4.5.3	Modelos con único intercepto en todas las áreas pequeñas	148
4.6	Análisis descriptivo del conjunto de datos del ICTPC	151
4.7	Ajuste del Modelo log-normal asimétrico, datos del ICTPC	154
4.7.1	Validación (entrenamiento-prueba)	159
4.7.2	Pronóstico de nuevas observaciones ($n_i = 0$)	160
4.8	Ajuste del Modelo probit sesgado con variable latente, datos del ICTPC	162
4.8.1	Validación (entrenamiento-prueba)	165
4.9	Ajuste del Modelo probit ordenado sesgado con variable latente, datos del ICTPC	166
4.9.1	Validación (entrenamiento-prueba)	170
4.10	Estimaciones de β en los tres modelos, datos del ICTPC	170
CAPÍTULO 5. DISCUSIÓN DE LOS RESULTADOS	173
5.1	Estudio de simulación	173
5.1.1	Interacción entre ρ_i y μ_i	173
5.1.2	Identificación de ρ_i	173
5.1.3	Modelo log-normal sesgado	175

5.1.4	Modelo probit sesgado latente.....	175
5.1.5	Modelo probit ordenado sesgado latente.....	176
5.2	Datos del ICTPC	177
5.2.1	Modelo log-normal sesgado	178
5.2.2	Modelo probit sesgado latente.....	179
5.2.3	Modelo probit ordenado sesgado latente.....	180
CAPÍTULO 6.	CONCLUSIONES Y RECOMENDACIONES	181
6.1	Recomendaciones	183
ANEXOS	190	
Anexo A	190	
Anexo B	204	
Anexo C	205	
Anexo D	209	

LISTA DE FIGURAS

Figura 2.1	Densidad normal sesgada estándar para varios valores de λ	6
Figura 2.2	Contornos analíticos y estimados de la densidad normal sesgada en dos dimensiones.....	9
Figura 2.3	Densidad normal sesgada estándar con parámetros centrados para varios valores de λ	11
Figura 2.4	Superficie de la densidad normal sesgada con parámetros centrados en dos dimensiones.....	12
Figura 2.5	Gráfico de ρ y λ	14
Figura 2.6	Exploración mediante tres métodos MCMC de una densidad normal en dos dimensiones.....	26
Figura 2.7	Divergencia KL entre las densidades normal sesgada y normal conforme el valor absoluto del parámetro de correlación/forma crece.....	30
Figura 2.8	Representación del límite inferior de la evidencia.	32
Figura 2.10	Inferencia Bayesiana variacional para los parámetros de la distribución normal.....	41
Figura 3.2	Distribución <i>a priori</i> SSVS.	87
Figura 3.3	Distribución <i>a priori</i> SSVS con distintas proporciones de mezcla....	90
Figura 3.4	Grafo dirigido acíclico del que corresponde a la representación jerárquica de los modelos log-normal y probit sesgados con parámetros centrados.....	93
Figura 4.1	Proporción de valores $y_{ij} = 1$ simulados.	122
Figura 4.2	Función liga del modelo probit ordenado sesgado.	123
Figura 4.4	125
Figura 4.5	126

Figura 4.6	Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC	127
Figura 4.7	Modelo log-normal sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.	127
Figura 4.8	Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC	128
Figura 4.9	Modelo log-normal sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.	129
Figura 4.10	Matriz de confusión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC.....	130
Figura 4.11	Modelo probit sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 5%.	131
Figura 4.12	Matriz de confusión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC.....	132
Figura 4.13	Modelo probit sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 25%.	133
Figura 4.14	Matriz de confusión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC.....	134
Figura 4.15	Modelo probit sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.	135
Figura 4.16	Matriz de confusión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC.....	136

Figura 4.17	Modelo probit sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.	137
Figura 4.18	Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC	138
Figura 4.19	Modelo probit sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 5%.	139
Figura 4.20	Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC	140
Figura 4.21	Modelo probit sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 25%.	141
Figura 4.22	Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC	142
Figura 4.23	Modelo probit ordenado sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 5%.	142
Figura 4.24	Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC	143
Figura 4.25	Modelo probit ordenado sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 25%.	144
Figura 4.26	Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC	145

Figura 4.27	Modelo probit ordenado sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.....	146
Figura 4.28	Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC	147
Figura 4.29	Modelo probit ordenado sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.....	147
Figura 4.30	Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC	148
Figura 4.31	Modelo probit ordenado sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 5%.....	149
Figura 4.32	Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC	150
Figura 4.33	Modelo probit ordenado sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 25%.....	150
Figura 4.34	Estimación por kernel (suavizado) de la densidad empírica del logaritmo natural de la respuesta (log-ICTPC) e histograma de probabilidad. Se omitieron tres observaciones asociadas a ingresos pequeños (52.66, 187.71 y 367.30 pesos mexicanos). Fuente: elaboración propia.....	151
Figura 4.35	154
Figura 4.36	Mapas con la estimación del porcentaje de la población bajo la LPI y LPEI.....	157
Figura 4.37	Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal)	159

Figura 4.38	Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal).	160
Figura 4.39	Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal).	161
Figura 4.40	165
Figura 4.41	Matriz de confusión entre la respuesta binaria discretizada observada (columnas) y los valores ajustados (renglones).	165
Figura 4.42	Estimación del porcentaje de la población bajo la LPI y LPEI, en el primer renglón se muestran las estimaciones a partir del método BV y en el segundo renglón a partir del método HMC.	169
Figura 4.43	Matriz de confusión entre la respuesta ordinal observada (columnas) y las categorías ajustadas (renglones)	170

LISTA DE CUADROS

Cuadro 2.3	Resumen de los métodos de inferencia Bayesiana variacional.	55
Cuadro 3.1	Resumen sobre los tres modelos de regresión a nivel unidad.	69
Cuadro 3.2	Bloques de un programa Stan.	97
Cuadro 3.3	Cantidades fijadas para los experimentos de simulación.	108
Cuadro 3.4	Métricas básicas del ajuste.	110
Cuadro 3.5	Valores iniciales para la simulación.	112
Cuadro 3.6	Tamaño del hogar ajustado.	117
Cuadro 4.1	Modelo log-normal con único intercepto en cada área pequeña. Porcentaje de muestreo: 25%.	124
Cuadro 4.2	Modelo probit sesgado con único intercepto en cada área pequeña. Porcentaje de muestreo: 25%.	124
Cuadro 4.3	Modelo probit ordenado sesgado con único intercepto en cada área pequeña. Porcentaje de muestreo: 25%.	125
Cuadro 4.4	Modelo log-normal sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.	126
Cuadro 4.5	Métricas del ajuste del modelo log-normal sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.	127
Cuadro 4.6	Modelo log-normal sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.	128
Cuadro 4.7	Métricas del ajuste del modelo log-normal sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.	129
Cuadro 4.8	Modelo probit sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 5%.	130
Cuadro 4.9	Métricas del ajuste del modelo probit sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 5%.	131
Cuadro 4.10	Modelo probit sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 25%.	131

Cuadro 4.11	Métricas del ajuste del modelo probit sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 25%....	132
Cuadro 4.12	Modelo probit sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.....	133
Cuadro 4.13	Métricas del ajuste del modelo probit sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 5%....	134
Cuadro 4.14	Modelo probit sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.....	135
Cuadro 4.15	Métricas del ajuste del modelo probit sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 25%....	136
Cuadro 4.16	Modelo probit sesgado latente con único intercepto para todas las áreas pequeñas. Porcentaje de muestreo: 5%.....	137
Cuadro 4.17	Métricas del ajuste del modelo probit sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 5%.138	
Cuadro 4.18	Modelo probit sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 25%.....	139
Cuadro 4.19	Métricas del ajuste del modelo probit sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 25%.....	140
Cuadro 4.20	Modelo probit ordenado sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 5%.....	141
Cuadro 4.21	Métricas del ajuste del modelo probit ordenado sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 5%....	142
Cuadro 4.22	Modelo probit ordenado sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 25%.....	143
Cuadro 4.23	Métricas del ajuste del modelo probit ordenado sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 25%....	144
Cuadro 4.24	Modelo probit ordenado sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.....	145

Cuadro 4.25	Métricas del ajuste del modelo probit ordenado sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.	146
Cuadro 4.26	Modelo probit ordenado sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.	146
Cuadro 4.27	Métricas del ajuste del modelo probit ordenado sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.	147
Cuadro 4.28	Modelo probit ordenado sesgado latente con único intercepto para todas las áreas pequeñas. Porcentaje de muestreo: 5%.	148
Cuadro 4.29	Métricas del ajuste del modelo probit ordenado sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 5%.	149
Cuadro 4.30	Modelo probit ordenado sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 25%.	149
Cuadro 4.31	Métricas del ajuste del modelo probit ordenado sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 25%.	150
Cuadro 4.32	Estadísticos resumen del ingreso corriente total per cápita por Alcaldía y ámbito.	152
Cuadro 4.32	Estadísticos resumen del ingreso corriente total per cápita por Alcaldía y ámbito.	153
Cuadro 4.33	Estadísticos resumen del ingreso corriente total per cápita por Alcaldía y ámbito. Fuente: elaboración propia con información de la ENIGH 2024.	153
Cuadro 4.34	Ajuste del Modelo log-normal sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible.	154
Cuadro 4.34	Ajuste del Modelo log-normal sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible.	155
Cuadro 4.35	Los porcentajes de población bajo la LPI y LPEI son de 18.81% y 2.58% para el método VB, y de 31.14% y 5.60% para el método HMC. La medición estatal oficial es de 25.70% y 4.45%.	156

Cuadro 4.36	Sesgo promedio calculado a partir de la muestra <i>a posteriori</i>	158
Cuadro 4.37	Métricas del ajuste entre los valores observados y pronósticos en escala original.....	159
Cuadro 4.38	Ajuste del Modelo log-normal sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible.	160
Cuadro 4.39	Métricas del ajuste entre los valores observados y pronósticos en escala original.....	161
Cuadro 4.40	Ajuste del Modelo log-normal sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible.	161
Cuadro 4.41	Métricas del ajuste entre los valores observados y pronósticos en escala original.....	162
Cuadro 4.42	Ajuste del modelo probit sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible.	162
Cuadro 4.42	Ajuste del modelo probit sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible.	163
Cuadro 4.43	Los porcentajes de población bajo la LPI son de 31.70% para el método VB y 27.25% para el método HMC.	163
Cuadro 4.43	Los porcentajes de población bajo la LPI son de 31.70% para el método VB y 27.25% para el método HMC.	164
Cuadro 4.44	Métricas del ajuste entre los valores binarios observados y pronósticos.	166
Cuadro 4.45	Ajuste del modelo probit ordenado sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible.	166
Cuadro 4.45	Ajuste del modelo probit ordenado sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible.	167
Cuadro 4.46	Ajuste del modelo probit ordenado sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible. Los porcentajes obtenidos con BV son 19.26% y 7.59%, con HMC son 22.61% y 4.88%. Los totales estales oficiales son 25.7% y 4.5%.	
	Fuente: elaboración propia.....	168

Cuadro 4.47	Métricas del ajuste entre los valores ordinales observados y pronósticos	170
Cuadro 4.48	Estimaciones de los coeficientes de regresión en el modelo log-normal sesgado.	171
Cuadro 4.49	Estimaciones de los coeficientes de regresión en el modelo probit sesgado.	172
Cuadro 4.50	Estimaciones de los coeficientes de regresión en el modelo probit ordenado sesgado.	172

CAPÍTULO 1. INTRODUCCIÓN

Quizás uno de los retos más importantes dentro del paradigma de inferencia Bayesiana, consiste en determinar o tomar muestras de la distribución *a posteriori* para algún modelo estadístico de interés, ya que pueden existir una gran cantidad de parámetros o variables latentes involucrados en este. La promesa de la inferencia Bayesiana variacional (BV), consiste en transformar el problema de encontrar o muestrear de la distribución *a posteriori* a un problema de optimización: ganar tiempo de cómputo a expensas de perder un poco de precisión. El objetivo de esta optimización se mide en términos de la divergencia Kullback-Leibler (KL) entre la verdadera *a posteriori* y la aproximación variacional propuesta.

Por otro lado, el planteamiento central de los modelos de áreas pequeñas es que no se dispone del registro completo de las observaciones en cada región, lo que puede conducir a realizar inferencia, y en general, estimaciones sesgadas de la población, especialmente subestimando la incertidumbre. Desde la perspectiva de inferencia Bayesiana, es posible modelar las observaciones faltantes como variables latentes, no obstante, conforme el tamaño de cada subpoblación aumenta, los métodos usuales para realizar esta clase de inferencia (muestreo de la distribución *a posteriori* basado en cadenas de Márkov Monte Carlo, *Markov chain Monte Carlo* (MCMC) como *Gibbs Sampler*, *Metropolis-Hastings* y *Hamiltonian Monte Carlo*) se vuelven prohibitivos, de ahí el potencial de la propuesta variacional.

El tercer elemento principal de investigación, consiste en el uso de la distribución normal asimétrica como el objeto básico empleado para construir modelos de regresión, esta densidad generaliza la distribución normal y la hace más flexible para capturar relaciones donde la asimetría es inherente y relevante en los datos: para datos continuos se ajustó el modelo log-normal sesgado, mientras que para datos binarios y ordinales se ajustó el modelo probit, ambos con el enfoque de variable latente cuya función de distribución es la normal sesgada.

La propuesta generada por estos tres elementos motiva la hipótesis de investigación que se

plantea enseguida. Posteriormente, se muestra el objetivo general del proyecto y se exponen los objetivos particulares.

El resto del documento se estructura de la siguiente manera. En el [Capítulo 2](#) se presenta la revisión de literatura con la teoría necesaria para estimar los modelos de regresión propuestos. En el [Capítulo 3](#), se describe la metodología para hilar los conceptos y elementos descritos previamente y dar lugar a los modelos de regresión propuestos bajo el régimen de áreas pequeñas, particularmente, se hace uso del teorema de Bayes y el proceso de truncamiento oculto; posteriormente se muestra como programar estos modelos en el lenguaje de programación probabilística Stan. En el [Capítulo 4](#), se presentan los resultados obtenidos con el ajuste de cada uno de los tres modelos en áreas pequeñas propuestos, tanto para un estudio de simulación como para un conjunto de datos reales. En este último, se realizó la estimación del ingreso corriente total per cápita (ICTPC) en los hogares de la Ciudad de México, México, en el año 2025. Finalmente, en el [Capítulo 5](#) se discuten los resultados más relevantes que se obtuvieron, también se emiten algunas recomendaciones concretas.

1.1 Objetivos

El objetivo general de este proyecto es mostrar la teoría y metodología para estimar dos tipos de modelos de regresión mediante inferencia Bayesiana variacional en áreas pequeñas: un modelo log-normal y un modelo probit ordenado con variable latente, ambos basados en el uso de la distribución normal asimétrica.

Los objetivos específicos son:

- Caracterizar la distribución normal sesgada, listando sus propiedades, parametrizaciones y describir su conexión con el proceso de truncamiento oculto.
- Describir de forma breve los métodos Bayesianos basados en MCMC.
- Presentar una introducción al método de inferencia Bayesiana variacional y algunas de sus variantes.
- Describir el modelo de regresión Bayesiana log-normal sesgado en el contexto de áreas

pequeñas.

- Describir el modelo de regresión Bayesiana probit sesgado ordenado variable latente en el contexto de áreas pequeñas.
- Implementar los modelos de regresión log normal sesgado y probit ordenado sesgado con variable latente en el lenguaje de programación probabilística Stan, a través de la interfaz `cmdstanr` del lenguaje de programación R.
- Estimar las dos clases de modelos propuestos para analizar el ICTPC a nivel municipal, de acuerdo a fuentes de información oficiales y los criterios e indicadores establecidos por el Consejo Nacional de Evaluación de la Política de Desarrollo Social (Coneval).
- Seleccionar los conjuntos de covariables más relevantes para el estudio del ICTPC a partir de la técnica de búsqueda estocástica de variables.

1.2 Hipótesis

La hipótesis de esta investigación es que el método de inferencia Bayesiana variacional es una alternativa viable, en términos de precisión y tiempo de cómputo, a los métodos Bayesianos usuales basados en muestreo de la distribución *a posteriori*, específicamente, al método Hamiltoniano Monte Carlo; para estimar los modelos de regresión log-normal y probit ordenado sesgado con variable latente, ambos basados en el uso de la distribución normal asimétrica.

CAPÍTULO 2. REVISIÓN DE LITERATURA

Este capítulo consta de tres apartados donde se abordan los conceptos principales de la investigación. En la [Sección 2.1](#) se revisa la distribución normal sesgada en una y varias dimensiones, la parametrización centrada y dos representaciones estocásticas útiles. Luego, en la [Sección 2.2](#) se hace un breve resumen sobre el paradigma Bayesiano de la estadística y sobre la técnica de muestreo con cadenas de Márkov Monte Carlo (MCMC) conocida como Hamiltoniano Monte Carlo (HMC). Finalmente, la [Sección 2.3](#) muestra algunos enfoques del método Bayesiano variacional, así como el algoritmo de optimización que se emplea más adelante.

2.1 Distribución normal asimétrica y truncamiento oculto

En esta sección presentamos la distribución normal asimétrica, así como su extensión multivariada y su génesis debido al proceso de *truncamiento oculto*. Esta densidad fue formulada como una extensión de la distribución normal por Azzalini ([1985](#)) y posteriormente extendida al caso multivariado por Azzalini y Valle ([1996](#)), sin embargo, es posible trazar su aparición a varias décadas atrás, no como una extensión de la distribución normal pero sí como resultado de manipular estas densidades.¹ No obstante, otros autores también han contribuido al estudio y desarrollo de esta clase de densidades, por ejemplo Arnold, Beaver et al. ([2002](#)), Domínguez-Molina et al. ([2007](#)) y Arellano-Valle y Azzalini ([2008](#)), entre otros autores. En la [Sección 2.1.1](#) presentamos la distribución normal asimétrica o normal sesgada junto a varias de su propiedades, similares a la densidad normal. Luego, en la [Sección 2.1.2](#) presentamos la extensión multivariada que llamamos distribución normal asimétrica multivariada o normal sesgada multivariada y también listamos algunas propiedades y la relación con la densidad normal multivariada. En la Sección [Sección 2.1.3](#) y [Sección 2.1.4](#) se presenta una parametrización alternativa para

¹Una revisión más extensa sobre su historia se encuentra en <http://azzalini.stat.unipd.it/SN/faq-h.html>.

ambas distribuciones, llamada parametrización centrada o con parámetros centrados, la cuál es conveniente desde el punto de vista práctico para realizar inferencia sobre los parámetros de estas densidades. Posterior a esto, en la [Sección 2.1.5](#) presentamos el mecanismo de *truncamiento oculto* que bajo ciertas especificaciones da lugar a la densidad normal sesgada en una o varias dimensiones. Este fenómeno también es relevante desde el enfoque de inferencia, ya que, como se verá más adelante, la densidad normal sesgada también puede representarse mediante un mecanismo similar a la aumentación de datos. Finalmente, pero no menos importante, en la [Sección 2.1.6](#) se muestra otro mecanismo estocástico que da origen a la densidad normal sesgada, en una o varias dimensiones.

2.1.1 Distribución normal asimétrica

Se dice que la variable aleatoria X tiene distribución (ley) normal asimétrica o normal sesgada si su función de densidad está dada por

$$f_X(x | \mu, \sigma^2, \lambda) = \frac{2}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right) \Phi\left(\lambda \frac{x - \mu}{\sigma}\right) I_{(-\infty, \infty)}(x), \quad (2.1)$$

donde $(\mu, \sigma^2, \lambda) \in \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$ son los parámetros de localidad, varianza y forma. Las funciones ϕ y Φ denotan la función de densidad y distribución de una variable aleatoria normal estándar. A continuación describimos algunas de sus propiedades. Sea $X \sim SN(\mu, \sigma^2, \lambda)$ entonces (Azzalini [2005](#)):

1. La familia de distribuciones normales asimétricas es cerrada bajo la transformación de cambio de localidad y escala, para $(a, b) \in \mathbb{R} \times \mathbb{R}^+$

$$a + bX \sim SN(a + b^2\mu, b^2\sigma^2, \lambda). \quad (2.2)$$

2. Si $\lambda = 0$, entonces $X \sim N(\mu, \sigma^2)$. Además, $-X \sim SN(\mu, \sigma^2, -\lambda)$.
3. $\lim_{\lambda \rightarrow \infty} f_X(x | \mu, \sigma^2, \lambda) \rightarrow NT(x | \mu, \sigma^2; 0, \infty)$, es decir, haciendo que el parámetro de forma crezca (disminuya) sin límite, la densidad normal asimétrica converge a una densidad normal truncada a la izquierda (derecha) en cero.

4. La función generadora de momentos está dada por

$$m_X(t) = 2 \exp \left\{ \mu t + \frac{\sigma^2 t^2}{2} \right\} \Phi(\sigma \rho t), \quad (2.3)$$

donde $\rho \triangleq \lambda / \sqrt{1 + \lambda^2}$. De ahí que:

5. $\mathbb{E}[X] = \mu + \sigma \rho \sqrt{2/\pi}$. $\text{Var}[X] = \sigma^2 (1 - 2\rho^2/\pi)$.

6. El coeficiente de asimetría está dado por

$$\frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{(\text{Var}[X])^{3/2}} = \sqrt{\frac{2}{\pi}} (4 - \pi) \lambda^3 / \pi (1 - \frac{2}{\pi} \lambda^2)^{3/2}.$$

Usualmente, esta cantidad se denota con γ_1 . Bayes y Branco (2007) sugieren que como $\gamma_1 \in (-0.9953, 0.9953)$, esta densidad es adecuada solo para modelar sesgo leve o moderado.

En la Figura 2.1 se dibuja la densidad $SN(x | 0, 1, \lambda)$ para los valores $\lambda \in \{0, \pm 2, \pm 5\}$. En este caso, cuando $\mu = 0$ y $\sigma^2 = 1$, decimos que X tiene densidad normal sesgada estándar, y de igual modo al caso normal, es común denotar esta variable aleatoria con Z .

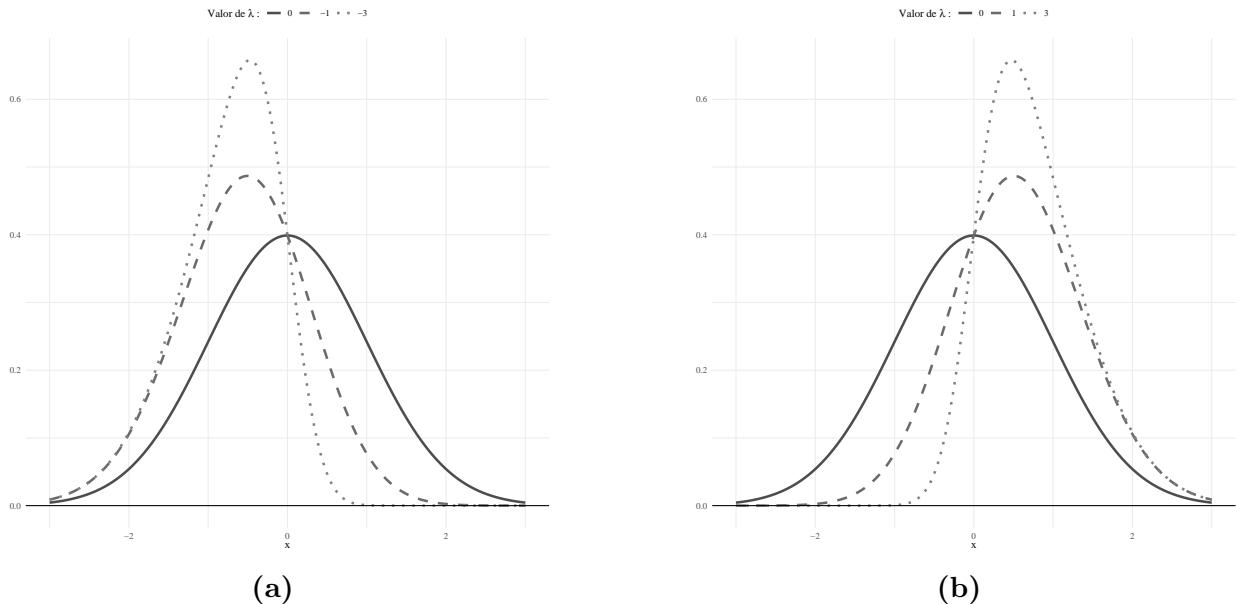


Figura 2.1: Densidad normal sesgada estándar para varios valores de λ . Fuente: elaboración propia.

2.1.2 Distribución normal asimétrica multivariada

Sea $\mathbf{X}_n = (X_1, X_2, \dots, X_n)^T$ un vector aleatorio. De acuerdo con Arnold, Beaver et al. (2002) y Azzalini y Valle (1996), entre otros autores, se dice que \mathbf{X}_n tiene distribución (ley) normal asimétrica o normal sesgada multivariada si su función de densidad está dada por

$$f_{\mathbf{X}_n}(\mathbf{x}_n | \boldsymbol{\mu}_n, \Sigma, \boldsymbol{\lambda}_n) = 2\phi_n(\mathbf{x}_n | \boldsymbol{\mu}_n, \Sigma) \Phi(\boldsymbol{\lambda}_n^T \boldsymbol{\sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_n)), \quad (2.4)$$

donde $(\boldsymbol{\mu}_n, \Sigma, \boldsymbol{\lambda}_n) \in \mathbb{R}^n \times \mathbb{M}_n(\mathbb{R}) \times \mathbb{R}^n$ son los parámetros de localidad, covarianza y forma.² ϕ_n denota la función de densidad normal multivariada de dimensión n . Definimos $\boldsymbol{\sigma} \equiv \text{diag}(\Sigma)$ como una matriz diagonal cuyas entradas son la raíz cuadrada de los elementos diagonales de Σ , de igual modo, definimos $\bar{\Sigma} = \boldsymbol{\sigma}^{-1} \bar{\Sigma} \boldsymbol{\sigma}^{-1}$ como una matriz de correlación obtenida a partir de la matriz de covarianza Σ .

A continuación describimos algunas de las propiedades básicas de esta distribución, las cuáles podemos ver que son analógas al caso univariado. Sea $\mathbf{X}_n \sim SN_n(\boldsymbol{\mu}_n, \Sigma, \boldsymbol{\lambda}_n)$, entonces (Azzalini 2005):

1. La familia de distribuciones normales asimétricas multivariadas es cerrada bajo la transformación de cambio de localidad y escala.
2. Si $\boldsymbol{\lambda}_n = \mathbf{0}_n$, entonces $\mathbf{X}_n \sim N_n(\boldsymbol{\mu}_n, \Sigma)$. Además, $-\mathbf{X}_n \sim SN_n(\boldsymbol{\mu}_n, \Sigma, -\boldsymbol{\lambda}_n)$
3. La función generadora de momentos está dada por

$$m_{\mathbf{X}}(\mathbf{t}) = 2 \exp \left(\boldsymbol{\mu}_n^T \mathbf{t} + \frac{1}{2} \mathbf{t}^T \Sigma \mathbf{t} \right) \Phi \left((2/\pi)^{-1/2} \boldsymbol{\mu}_Z^T \boldsymbol{\sigma} \mathbf{t} \right), \quad (2.5)$$

donde $\boldsymbol{\sigma} \equiv \text{diag}(\Sigma)$, $\boldsymbol{\mu}_Z \equiv (1 + \boldsymbol{\lambda}_n^T \bar{\Sigma} \boldsymbol{\lambda}_n)^{-1/2} \bar{\Sigma} \boldsymbol{\lambda}_n$. De ahí que:

4. $\mathbb{E}[\mathbf{X}_n] = \boldsymbol{\mu}_n + \boldsymbol{\sigma} \boldsymbol{\mu}_Z \sqrt{\frac{2}{\pi}}$. $\text{Var}[\mathbf{X}_n] = \Sigma - \boldsymbol{\sigma} \boldsymbol{\mu}_Z^T \boldsymbol{\mu}_Z \boldsymbol{\sigma}$.
5. Suponga que partimos el vector \mathbf{X}_n y sus parámetros como $\mathbf{X}_n = (\mathbf{X}_1, \mathbf{X}_2)^T$, $\boldsymbol{\mu}_n = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)^T$, $\boldsymbol{\lambda}_n = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)^T$ y $\Sigma = \{\Sigma_{ij}\}_{(i,j) \in \{1,2\}}$ (una matriz en bloques). Si \mathbf{X}_1 es de

²Denotamos como $\mathbb{M}_n(\mathbb{R})$ al conjunto de matrices cuadradas de dimensión n con entradas en los reales que son simétricas y definidas positivas: en otras palabras, es el conjunto de matrices de covarianza.

dimensión n_1 , entonces su distribución marginal es $\mathbf{X}_1 \sim SN_{n_1}(\boldsymbol{\mu}_1, \Sigma_{11}, \boldsymbol{\lambda}_{1|2})$, donde

$$\boldsymbol{\lambda}_{1|2} \equiv \frac{\lambda_1 + \bar{\Sigma}_{11}^{-1} \bar{\Sigma}_{12} \boldsymbol{\lambda}_2}{\sqrt{1 + \boldsymbol{\lambda}_2^T \bar{\Sigma}_{22|1} \boldsymbol{\lambda}_2}} \quad (2.6)$$

$$\bar{\Sigma}_{22|1} \equiv \bar{\Sigma}_{22} - \bar{\Sigma}_{21} \bar{\Sigma}_{11}^{-1} \bar{\Sigma}_{12}.$$

El estado del arte sobre la distribución normal asimétrica multivariada consiste en la clase de densidades llamada normal asimétrica unificada, *skew unified normal* (SUN), la cuál extiende la clase descrita arriba modificando el término de asimetría en dos aspectos (Durante 2019; Arellano-Valle y Azzalini 2022):

- Añade un parámetro adicional dentro de la función Φ , lo que permite una mayor regulación de la forma, lo que conduce a la distribución normal sesgada extendida multivariada, *extended skew-normal* (ESN). Esto es análogo al desarrollo de Arnold, Beaver et al. (2002).
- Permite que el mecanismo que induce la asimetría sea multivariado, es decir, generaliza al vector de forma $\boldsymbol{\lambda}_n$ a una matriz Λ . Esta modificación genera la familia normal sesgada cerrada, *closed skew-normal* (CSN) de Domínguez-Molina et al. (2007).

Además de incrementar la flexibilidad, estas extensiones confieren propiedades de cerradura para marginales, condicionales y distribuciones conjuntas, produciendo así una clase general (Durante 2019). Concretamente, la densidad está dada por

$$\phi_p(\mathbf{x}_n | \boldsymbol{\mu}_n, \Sigma) \frac{\Phi_p(\boldsymbol{\gamma} + \Lambda^T \bar{\Sigma}^{-1} \boldsymbol{\sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_n); \mathbf{0}_n, \Gamma - \Lambda^T \bar{\Sigma}^{-1} \Lambda)}{\Phi_n(\boldsymbol{\gamma}; \mathbf{0}_n, \Gamma)}, \quad (2.7)$$

y escribimos $\mathbf{X}_n \sim SUN_{n,p}(\boldsymbol{\mu}_n, \Sigma_{n \times n}, \Lambda_{n \times p}, \boldsymbol{\gamma}_n, \Gamma_{p \times p})$. Note que con la elección adecuada de estos parámetros es posible recuperar la densidad normal asimétrica que describimos previamente e incluso la densidad normal multivariada.

Durante (2019) muestra que la elección de la familia SUN como *a priori* para los parámetros β en un modelo de regresión logística resulta ser conjugada, más aún, Onorati y Liseo (2025) propone la familia *perturbed unified skew-normal* (pSUN), cuya utilidad es ser la distribución

conjugada para los coeficientes β en cualquier modelo de regresión binaria, siempre que la función liga pueda ser expresada como una mezcla de escala de funciones de distribución gaussianas.

No obstante, en la discusión posterior no es necesario emplear directamente la forma de la densidad normal asimétrica, y por tanto, ninguna de estas generalizaciones, por lo que únicamente consideramos el caso más simple, es decir, aquel que describimos al inicio de esta sección. En la [Figura 2.2](#) se dibujan los contornos de la densidad $SN_2(\mathbf{x}|\mathbf{0}, I, \boldsymbol{\lambda})$ para los valores $\boldsymbol{\lambda} = \pm[2, 2]^T$.

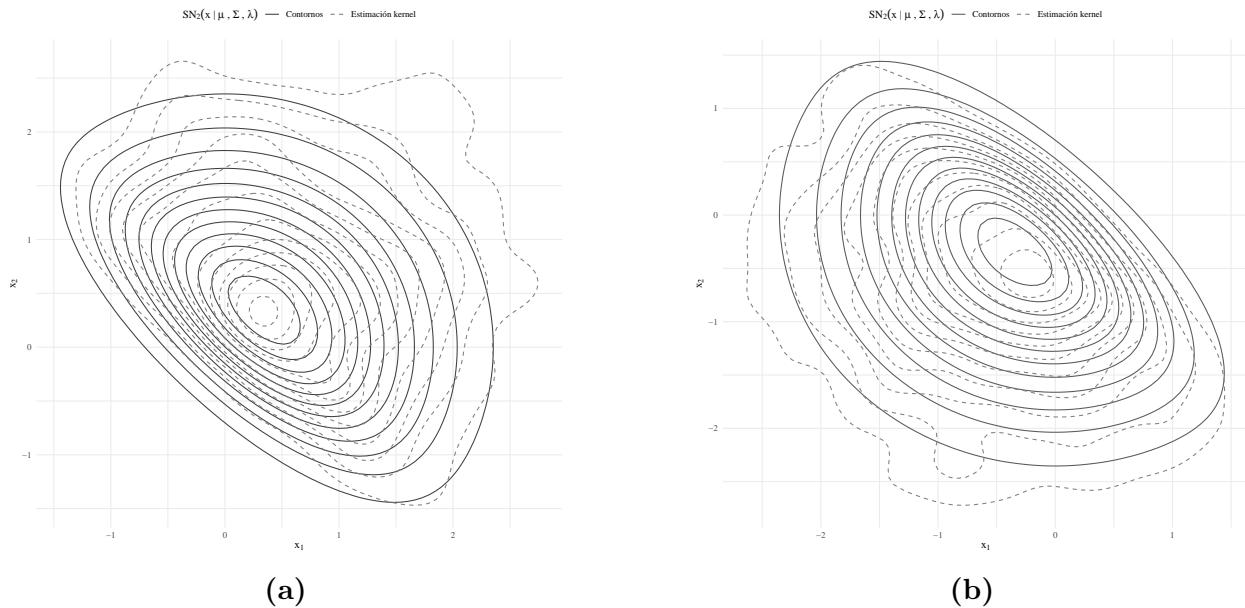


Figura 2.2: Las líneas continuas representan los contornos analíticos de la densidad, mientras que las líneas discontinuas muestran los contornos estimados a partir de la generación de muestras aleatorias de esta densidad en dos dimensiones. Fuente: elaboración propia.

2.1.3 Distribución normal asimétrica: parametrización centrada

Vale la pena notar que, a diferencia del caso normal, la media y varianza no coinciden con los parámetros de localidad y escala, específicamente, la media está desplazada hacia el signo de ρ o λ y la varianza está re-escalada en función del parámetro de correlación o forma. Por otro lado, la inferencia sobre estos parámetros presenta dificultades prácticas, por ejemplo (Arellano-Valle y Azzalini [2008](#); Bayes y Branco [2007](#); Azzalini [1985](#)):

-
- El estimador de máxima verosimilitud para el parámetro de asimetría puede ser infinito.
 - La matriz de información de Fisher se vuelve singular a medida que $\lambda \rightarrow 0$, esto implica que los estimadores de máxima verosimilitud no sean asintóticamente normales.
 - Existen máximos locales en la función de (log-)verosimilitud.

De manera adicional, Arellano-Valle y Azzalini (2008) señala que estos problemas se deben completamente a que la parametrización directa no adecuada para la estimación, ya que los parámetros son identificables. En el Anexo 1 se cubre brevemente este concepto estadístico.

Motivado por estas razones, Azzalini (1985) propone una parametrización que remedia el problema de singularidad en la matriz de información³ y la llama distribución normal asimétrica con parámetros centrados, la idea general es simple: remover tanto el desplazamiento en el parámetro de localidad como el encogimiento en el parámetro de escala. Así, sea la variable aleatoria X definida como

$$X = \mu + \sigma (\text{Var}[Z]^{-1/2} (Z - \mathbb{E}[Z])), \quad (2.8)$$

si además se reparametriza el parámetro de forma λ en términos del coeficiente de asimetría γ_1 , entonces se dice que X tiene distribución normal asimétrica con parámetros centrados, y lo denotamos como $X \sim SN^C(\mu, \sigma^2, \gamma_1)$, donde γ_1 es igual a como se define previamente. La densidad de probabilidad de X está dada por la siguiente expresión (Pérez-Rodríguez, Acosta-Pech et al. 2018; Azevedo, Bolfarine y Andrade 2011), donde es posible notar la similitud con la densidad en la Ecuación 2.1, sin embargo, tiene una expresión más extensa

$$f_X(x \mid \mu, \sigma^2, \gamma_1) = \frac{2}{\sigma^*} \phi\left(\frac{x - \mu^*}{\sigma^*}\right) \Phi\left(\lambda^*\left(\frac{x - \mu^*}{\sigma^*}\right)\right) I_{(-\infty, \infty)}(x), \quad (2.9)$$

donde $\mu^* = \mu - s\gamma_1^{1/3}$, $\sigma^{2*} = \sigma^2 \times (1 + s^2\gamma_1^{2/3})$, $\lambda^* = s\gamma_1^{1/3}/\sqrt{r^2 + s^2\gamma_1^{2/3}(r^2 - 1)}$, $r = \sqrt{2/\pi}$ y $s = (2/(4 - \pi))^{1/3}$. Así mismo, es posible invertir ambas parametrizaciones para obtener los parámetros directos o los parámetros centrados. En la discusión posterior, únicamente centramos la media y varianza, dejando el parámetro de forma sin modificación. De este

³Azzalini y Capitanio (1999, Sección 5.2) comenta que mediante cálculos analíticos detallados, la parametrización centrada remueve la singularidad en la matriz de información de Fisher cuando $\lambda = 0$.

modo, es posible mantener sólo un tipo de notación. Por ejemplo, a partir de las propiedades previas, la variable aleatoria X con distribución

$$X \sim SN \left(\mu - \sigma \sqrt{\frac{2}{\pi}} \rho^2, \sigma^2 / \sqrt{1 - \frac{2}{\pi} \rho^2}, \lambda \right), \quad (2.10)$$

cumple que $\mathbb{E}[X] = \mu$ y $\text{Var}[X] = \sigma^2$. En la [Figura 2.3](#) se muestra la densidad normal sesgada con parámetros centrados, es decir $SN^c(x | 0, 1, \lambda)$, para los valores $\lambda \in \{0, \pm 2, \pm 5\}$.

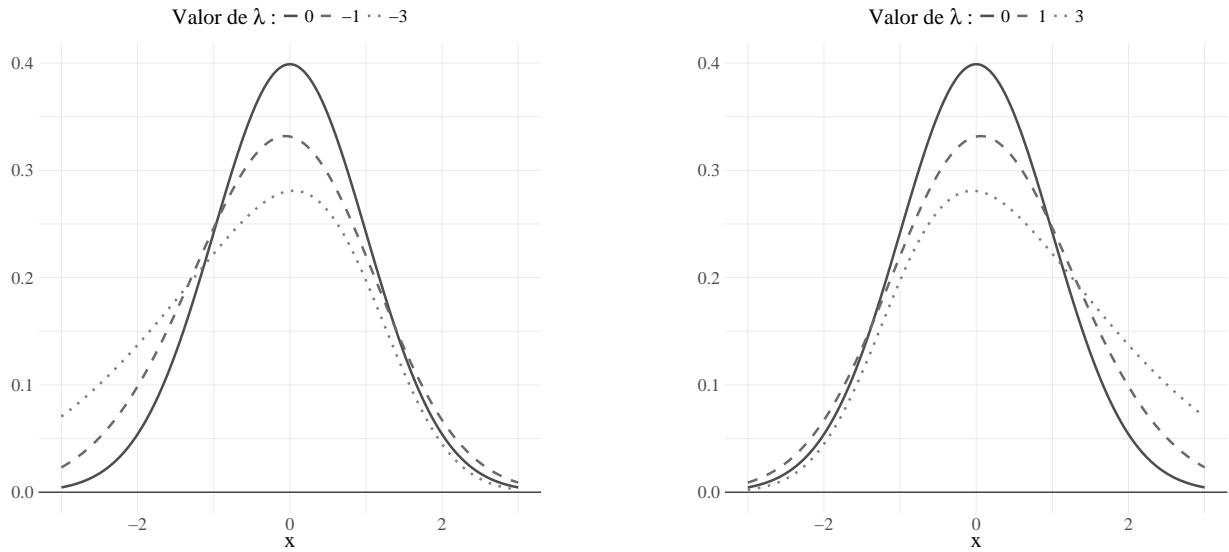


Figura 2.3: Densidad normal sesgada estándar con parámetros centrados para varios valores de λ . Fuente: elaboración propia.

2.1.4 Distribución normal asimétrica multivariada: parametrización centrada

Por otro lado, Arellano-Valle y Azzalini ([2008](#)) presentan una propuesta sobre como aplicar la parametrización centrada cuando se trabaja con la densidad normal asimétrica multivariada: la invitación es centrar cada entrada del vector \mathbf{X}_n y multiplicar por una matriz diagonal a fin de escalar cada entrada. En la [Figura 2.4](#) se muestran los contornos de la densidad $SN_2^c(\mathbf{x}|\mathbf{0}, I, \boldsymbol{\lambda})$, con $\boldsymbol{\lambda} = \pm[2, 2]^T$: es decir, la misma densidad que se muestra en la [Figura 2.2](#) pero con parámetros centrados: podemos notar que ahora la densidad tiene un comportamiento más ‘regular’.

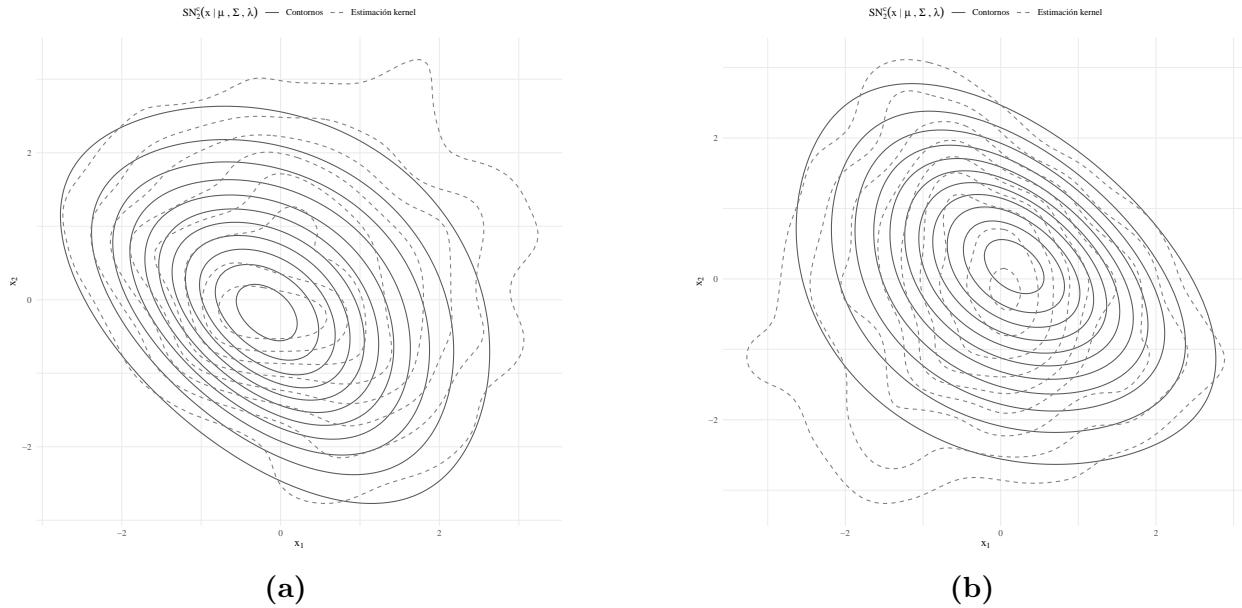


Figura 2.4: Superficie de la densidad normal sesgada con parámetros centrados en dos dimensiones. Fuente: elaboración propia.

2.1.5 Proceso de truncamiento oculto

Una representación estocástica de la densidad normal asimétrica que es útil para la exposición siguiente es mediante el proceso de *truncamiento oculto*, ya que ofrece una forma simple de generar variables aleatorias de esta distribución. Este proceso puede surgir en diversos contextos, quizás incluso de forma inconsciente: suponga que se observan dos variables aleatorias normales correlacionadas, digamos V_1 y V_2 . Ahora, si registramos alguna de estas variables siempre que la otra exceda cierto umbral v (es decir $U = V_1$ siempre que $V_2 > v$), entonces inducimos sesgo en las observaciones retenidas (U), y este sesgo es proporcional a su correlación ρ , de este modo, las observaciones que filtramos (U) de acuerdo a la otra variable (V_2) tendrán distribución normal asimétrica. Este mecanismo puede extenderse a más dimensiones, si $V_1, V_2, \dots, V_n, V_{n+1}$ tiene distribución normal multivariada, entonces al filtrar V_1, V_2, \dots, V_n condicionado a que V_{n+1} exceda cierto umbral, obtenemos una densidad normal asimétrica en n dimensiones. (Azzalini 2013; Arnold y Beaver 2000).

Aunque existen varios enfoques para esta aplicación en particular, nos concentramos en el caso que nos ocupa, así, el caso más sencillo es considerando una distribución normal

bivariada:

$$\begin{bmatrix} V \\ W \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

si definimos $U = V$ siempre que $W > 0$, entonces $U \sim SN(\mu, \sigma^2, \lambda)$, con $\lambda = \rho/\sqrt{1 - \rho^2}$. Vale la pena notar que $\rho \in (-1, 1)$, mientras que $\lambda \in (-\infty, \infty)$, así, conforme $\rho \rightarrow \pm 1$ ocurre que $\lambda \rightarrow \pm\infty$, es decir que cerca del borde de este rango de valores, λ se comporta de forma asintótica. Note que podemos invertir la expresión que define ρ y escribirla en términos de λ . En la [Figura 2.5](#) se dibujan ambas funciones. Podemos notar que la relación de λ para valores de ρ pequeños, digamos menores a 0.5, es similar a un comportamiento lineal, no obstante, el comportamiento de λ cerca de ± 1 puede resultar contraintuitivo: en el escenario descrito arriba, podemos pensar que no existe gran diferencia si tomamos $\rho_1 = \text{Cov}[V, W] = 0.995$ comparado con $\rho_2 = \text{Cov}[V, W] = 0.999$, no obstante, estos valores de correlación tienen asociado los valores de forma $\lambda_1 = 9.96$ y $\lambda_2 = 22.34$.

Un resultado principal que es relevante para esta exposición, es escribir la densidad conjunta de (U, W) como

$$f_{U,V}(u, w) \equiv f_{U|W}(u | w) f_W(w) = N(u | \mu + w\rho, \sigma^2(1 - \rho^2)) N(w | 0, \sigma^2) I_{(0, \infty)}(w), \quad (2.11)$$

donde $f_W(w)$ es la densidad de una normal truncada a la derecha en cero. En el [Anexo 1](#) se prueba como obtener estas densidades. La utilidad práctica de esta representación es evitar usar explícitamente la densidad normal asimétrica, ya que se expresa como el producto de una densidad normal por una densidad normal truncada, así, el proceso de truncamiento oculto junto a la parametrización centrada descrita en la [Sección 2.1.3](#) hacen la inferencia de sobre los parámetros de la distribución normal sesgada más estable desde el punto de vista analítico.

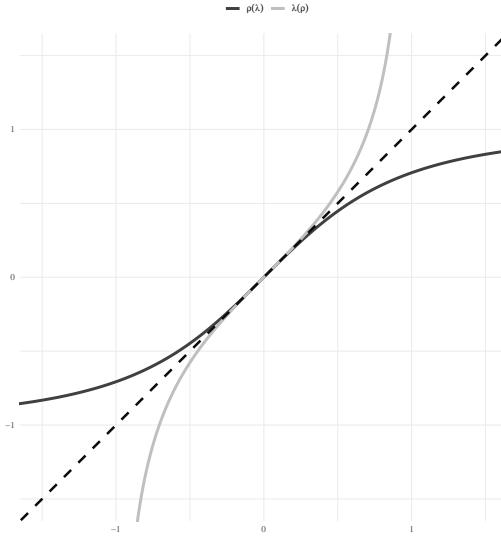


Figura 2.5: Observe la naturaleza dual entre el parámetro de correlación en la densidad normal original y el parámetro de sesgo en la densidad normal asimétrica resultante, además, el mapeo $\lambda(\rho)$ es invertible. Se agregó la función identidad identidad en el intervalo $(-1/2, 1/2)$, cuando ambos parámetros se acercan a cero, estos se comportan de manera lineal. Fuente: elaboración propia.

De igual forma, también es posible generalizar esta situación para el caso donde V tiene una distribución normal multivariada:

$$\begin{bmatrix} \mathbf{V} \\ W \end{bmatrix} \sim N_{n+1} \left(\begin{bmatrix} \boldsymbol{\mu} \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} \bar{\Sigma}_V & \mathbf{1}_n \rho \\ \mathbf{1}_n^T \rho & 1 \end{bmatrix} \right), \quad (2.12)$$

donde

$$\bar{\Sigma}_V \equiv (1 - \rho_i^2) I_n + \rho^2 J_n = (1(i=j) + \rho^2 1(i \neq j))_{ij}, \quad (2.13)$$

esta elección de matriz de covarianza induce correlación entre las observaciones $V_1, V_2 \dots, V_n$ que son filtradas por W , además de que permite la generalización correcta del proceso de truncamiento oculto a más de una dimensión. En el [Anexo 1](#) se discuten las implicaciones de esta matriz de covarianzas. Ahora bien, si definimos $\mathbf{U} = \mathbf{V}$ siempre que $W > 0$, entonces, usando un argumento análogo al caso univariado, no es difícil ver que

$$f_{\mathbf{U}, W}(\mathbf{u}, w) = \prod_{i=1}^n 2N(u_i | \mu_i + w\rho, \sigma^2(1 - \rho^2)) N(w | 0, \sigma^2) I_{(0, \infty)}(w), \quad (2.14)$$

además, la densidad marginal de \mathbf{U} corresponde a una normal asimétrica multivariada, como la definida por Azzalini y Capitanio (1999). Este enunciado se demuestra en el [Anexo 1](#); no obstante, para nuestra exposición no es necesario emplear este hecho, y como mencionamos previamente, el objetivo principal es eludir el uso explícito de la densidad normal asimétrica.

2.1.6 Representación alternativa

Existen otras representaciones estocásticas de la densidad normal sesgada, por ejemplo, una combinación lineal convexa entre una variable aleatoria normal y una normal truncada en $(0, \infty)$: sea $V \sim N(0, 1)$ y $W \sim NT(0, 1; 0, \infty)$, además, si U y W son independientes, entonces

$$X = \mu + \sigma(\sqrt{1 - \rho^2}V + \rho W) \sim SN(\mu, \sigma^2, \lambda). \quad (2.15)$$

Nuevamente, se define $\rho \triangleq \lambda/\sqrt{1 + \lambda^2}$. Si \mathbf{V}_n es normal multivariada, al aplicar la [Ecuación 2.15](#) a cada entrada V_i y asignarlo a X_i , se obtiene la generalización para más dimensiones, es decir \mathbf{X}_n tendrá distribución normal sesgada multivariada. Esta representación se emplea en el [Capítulo 3](#) para obtener la estructura de covarianzas de los modelos propuestos. Además, también es posible centrar los parámetros de localidad y escala en cada entrada del vector. Bayes y Branco (2007), Arnold, Beaver et al. (2002), Azzalini y Capitanio (1999) y Azzalini (1985) discuten esta representación alternativa.

2.2 Inferencia Bayesiana

El paradigma de inferencia Bayesiana considera que las cantidades de interés que desconocemos, por ejemplo parámetros en un modelo, son variables aleatorias y tienen asociada una función de densidad de probabilidad. El planteamiento general puede resumirse como sigue: sean \mathbf{y} los datos observados, sea $p(\mathbf{y} | \boldsymbol{\theta})$ la función de verosimilitud del modelo propuesto, sea $\boldsymbol{\theta} \in \Theta$ el vector de parámetros del modelo y sea $\pi(\boldsymbol{\theta})$ la distribución *a-priori* sobre $\boldsymbol{\theta}$. Actualizamos nuestras creencias sobre $\boldsymbol{\theta}$ -es decir $\pi(\boldsymbol{\theta})$ - de

acuerdo al teorema de Bayes:

$$\pi(\boldsymbol{\theta} | \mathbf{y}) = \frac{p(\mathbf{y} | \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})}{p(\mathbf{y})}, \quad (2.16)$$

donde $p(\mathbf{y}) = \int p(\mathbf{y} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ es llamada verosimilitud marginal o distribución predictiva *a priori*⁴. Si disponemos de información previa sobre estas cantidades desconocidas, por ejemplo, de conocimiento experto o experimentos anteriores, podemos emplear un enfoque *Bayesiano subjetivo*, donde incorporamos toda la información disponible al modelo por medio de la densidad $\pi(\boldsymbol{\theta})$.

Así mismo, con la distribución *a-priori* también es posible reflejar nuestro estado de ignorancia sobre $\boldsymbol{\theta}$: si no disponemos de información relevante, podemos emplear una densidad $\pi(\boldsymbol{\theta})$ con información vaga o bastante general, por ejemplo, el soporte de los parámetros. Esta alternativa constituye el enfoque *Bayesino objetivo*: las densidades $\pi(\boldsymbol{\theta})$ en este enfoque reciben los nombres *objetivas, no informativas, o de referencia*⁵.

Uno de los retos que presenta la inferencia Bayesiana es determinar la distribución *a posteriori* $\pi(\boldsymbol{\theta} | \mathbf{y})$. Usualmente, no es posible determinarla (i.e., identificar la densidad de probabilidad) ya que no conocemos $p(\mathbf{y})$, por lo que nos limitamos únicamente a generar muestras aleatorias de esta densidad: son varios, útiles y populares los esquemas de muestreo basados en cadenas de Márkov Monte Carlo, *Markov Chain Monte Carlo* (MCMC), por ejemplo *Metropolis-Hastings*, *Gibbs sampler* y *Hamiltonian Monte Carlo*.

Usualmente, en inferencia Bayesiana usamos $\pi(\boldsymbol{\theta})$ para referirnos a la distribución *a priori* sobre los parámetros $\boldsymbol{\theta}$, y en un estilo similar, $\pi(\boldsymbol{\theta} | \mathbf{y})$ representa la distribución *a posteriori* de $\boldsymbol{\theta}$; no obstante, sin riesgo de ambigüedad, también podemos usar p en lugar de π para denotar también a estas dos densidades, y en la discusión posterior usaremos casi exclusivamente p , teniendo en mente que podemos intercambiar esta notación.

Es relevante señalar que es posible efectuar inferencia Bayesiana sin emplear métodos

⁴También recibe el nombre de evidencia, especialmente en inferencia Bayesiana variacional, como se verá en el capítulo siguiente (Rohde y Wand 2016).

⁵Sin embargo, reservamos este último para una clase de densidades *a-priori* desarrolladas por Berger y Bernardo (1992).

MCMC, por ejemplo, empleando distribuciones *a priori* conjugadas. De igual modo, existen otros métodos para realizar inferencia Bayesiana de forma aproximada, es decir, se busca aproximar la densidad *a posteriori* $p(\boldsymbol{\theta} | \mathbf{y})$ con alguna densidad manejable o común, mediante la optimización de alguna función objetivo o criterio, tal es el caso de la inferencia Bayesiana variacional (BV) que se describe en el capítulo siguiente.

Iniciamos este capítulo presentando los estimadores de Bayes en la [Sección 2.2.1](#), posteriormente, en la [Sección 2.2.2](#) y [2.2.3](#) presentamos de forma concisa los dos temas centrales que dan lugar a los métodos MCMC: los métodos Monte Carlo y las cadenas de Markov con espacio de estados continuo. Posteriormente, en la [Sección 2.2.4](#) describimos brevemente los tres métodos MCMC mencionados previamente, aunque puedan parecer distintos en su planteamiento, en las secciones subsecuentes se bosqueja su relación.

2.2.1 Estimadores Bayesianos

Para caracterizar completamente un problema desde la perspectiva de inferencia Bayesiana, es necesario utilizar elementos de teoría de la decisión. De acuerdo con Ghosh, Delampady y Samanta ([2006](#)), es posible abordar problemas de inferencia de una manera matemáticamente más formal a través de la teoría de la decisión, ya que proporciona un marco conceptual unificado para abordar problemas muy diversos.

Típicamente, un problema de decisión involucra un espacio de acciones $a \in \mathcal{A}$, estados desconocidos de la naturaleza $\theta \in \Theta$, una función $u : \mathcal{A} \times \Theta \rightarrow \mathbb{R}$ que asigna utilidades a cada consecuencia (a, θ) , o alternativamente, una función de ‘pérdida’ no negativa $\ell(a, \theta)$ que representa la discrepancia o error incurrido cuando el parámetro es θ y la acción es a , y una especificación $\pi(\theta)$ en forma de una distribución de probabilidad con las creencias actuales sobre los posibles valores de los estados de la naturaleza.

- La respuesta óptima al problema de inferencia es una $a \in \mathcal{A}$ que maximiza la utilidad esperada:

$$\int_{\Theta} u(a, \theta) \pi(\theta) d\theta. \quad (2.17)$$

-
- Por otro lado, si se trabaja con la función de pérdida, la respuesta óptima al problema de inferencia es una $a \in \mathbb{A}$ que minimiza la pérdida esperada:

$$\int_{\Theta} \ell(a, \theta) \pi(\theta) d\theta. \quad (2.18)$$

También, definimos a un procedimiento de decisión δ como una forma sistemática de elegir acciones a basadas en observaciones $x_{1:n}$, usualmente, es una función $a = \delta(x_{1:n})$. Así, un procedimiento o estimador de Bayes es un procedimiento de decisión δ que escoge una acción a que minimiza la pérdida esperada $a posteriori$ $\rho(a, x_{1:n})$ para cada $x_{1:n}$:

$$\rho(a, x_{1:n}) \triangleq \mathbb{E}[\ell(a, \theta) | x_{1:n}], \quad (2.19)$$

esta cantidad también recibe el nombre de riesgo *a posteriori*. Note que los estimadores de Bayes dependen de la función de pérdida y la distribución *a priori*. De este modo, cada función de riesgo tiene asociado un estimador de Bayes. En el siguiente cuadro se resumen los estimadores Bayesianos para algunas funciones de pérdida usuales. En general, en el contexto de modelos de regresión, la pérdida error cuadrado medio se emplea con respuestas continuas, mientras que la perdida 0-1 se emplea con respuestas discretas, como en clasificación.

Cuadro 2.1: Algunas funciones de pérdida univariadas comunes.

Función de pérdida	Definición	Estimador Bayesiano
Error cuadrado medio	$\ell(a, \theta) = (a - \theta)^2$	$\mathbb{E}[\pi(\boldsymbol{\theta} x_{1:n})]$
Error absoluto	$\ell(a, \theta) = a - \theta $	$m_\theta : \int_{-\infty}^{m_\theta} \pi(\boldsymbol{\theta} x_{1:n}) d\theta = 1/2$
0-1	$1(a \neq \theta)$	$\arg \max_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta} x_{1:n})$

2.2.2 Métodos de Monte Carlo

Los métodos de Monte Carlo (MC) son una técnica basada en la simulación de variables aleatorias que es útil en varias áreas de la estadística, su origen se suele señalar al final de la década de 1940 (Brooks et al. 2011). A grandes rasgos, su actividad principal consiste en calcular integrales de forma numérica -ya sea en una o varias dimensiones-. Con frecuencia

en probabilidad y estadística, estas aparecen al momento de evaluar esperanzas⁶ o momentos en general, de ahí que sean útiles tanto en el paradigma Bayesiano como frecuentista.

La idea básica en los métodos MC es que estamos interesados en calcular

$$\mathbb{E}_f[h(X)] \triangleq \int_{\mathcal{X}} h(x) f_X(x) dx,$$

si $x_{1:m} = (x_1, \dots, x_m)$ es una muestra aleatoria de tamaño m que proviene de $f_X(x)$ entonces la cantidad

$$\bar{h}_m \triangleq \frac{1}{m} \sum_{i=1}^m h(x_i)$$

converge casi seguramente a $\mathbb{E}_f[h(X)]$ por la ley de los grandes números (Robert P. y Casella 2004). Además, si h^2 tiene esperanza finita bajo f , la velocidad de convergencia de \bar{h}_m puede ser medida ya que la varianza también puede ser estimada de la muestra,

$$\text{Var}[\bar{h}_m] \triangleq \frac{1}{m} \int_{\mathcal{X}} [h(x) - \mathbb{E}_f[h(X)]] f(x), dx,$$

y aproximamos la varianza con

$$v_m \triangleq \frac{1}{m^2} \sum_{i=1}^m [h(x_i) - \bar{h}_m]^2,$$

por tanto, la velocidad de convergencia es $\mathcal{O}(\sqrt{1/n})$.

En el contexto de los métodos MCMC, asumimos que podemos generar una muestra aleatoria de la distribución *a posteriori*, y a partir de estas realizaciones, calculamos cantidades de interés.

Otra técnica de simulación que es útil en el quehacer estadístico es el método *Bootstrap*; este es bastante similar a los métodos MC, salvo que este calcula cantidades de interés efectuando re-muestreo de forma intensiva: esto es, dada una muestra aleatoria $\mathbf{x} = (x_1, \dots, x_m)$, se

⁶Y como un caso particular, aproximar probabilidades: sea A un evento de interés, definimos la variable aleatoria X como $1_A(x)$ -es decir, $X = 1$ si $x \in A$ y $X = 0$ de otro modo-, entonces $\mathbb{E}[X] = \mathbb{P}[A]$.

obtienen B muestras de tamaño m realizando muestreo con remplazo de \boldsymbol{x} . Después, con estas B muestras de tamaño m es posible cuantificar la precisión o incertidumbre de las cantidades de interés.

2.2.3 Cadenas de Markov en un espacio de estados general

Una cadena de Márkov es un proceso estocástico, es decir, una colección de variables aleatorias $\{X_t\}_{t \geq 0}$ que poseen ciertas propiedades especiales. De manera simplista, es posible clasificar a los procesos estocásticos de acuerdo a: (1) el índice o tiempo que parametriza esta colección y (2) el conjunto donde las variables aleatorias toma valores, llamado espacio de estados. Generalmente, cada uno de estos criterios puede ser discreto o contable, por ejemplo: colecciones ordenadas en el tiempo $\{X_0, X_1, \dots\}$, espacio de estados finito: $\{x_0, x_1, \dots, x_S\}$; o bien, continuo o no contable $\{X_t : t \in \mathbb{R}^+\}$, espacio de estados continuo $A \subseteq \mathbb{R}$. En la discusión posterior, centramos nuestra atención únicamente a procesos indexados por el tiempo. Así, decimos que el proceso estocástico a tiempo discreto $\{X_n\}_{n \geq 0}$ es una cadena de Markov si para todo $A \subseteq \mathcal{S}$, donde \mathcal{S} es el espacio de estados, se cumple que (Spade 2020)

$$\mathbb{P}[X_k \in A \mid X_0 = x_0, \dots, X_{k-1} = x_{k-1}] = \mathbb{P}[X_k \in A \mid X_{k-1} = x_{k-1}]. \quad (2.20)$$

Cuando el espacio de estados es discreto, es decir $\mathcal{S} = \{x_0, x_1, \dots, \}$, la probabilidad de transición en un paso p_{ij} nos dice la probabilidad de transitar del estado j al estado i en el tiempo k al $k + 1$. Simbólicamente, escribimos

$$p_{ij} \equiv \mathbb{P}[X_k = i \mid X_{k-1} = j]. \quad (2.21)$$

Por otra parte, cuando nos trasladamos al caso de espacio de estados continuo, en lugar de tener la probabilidad de transición en un paso p_{ij} , hablamos de un *kernel de transición* $K(x, A)$, el cuál es una función de densidad de probabilidad, este kernel se define como

(Spade 2020; Gilks, Richardson y Spiegelhalter 1995; Greenberg 2012).

$$K(x, A) \triangleq \mathbb{P}[X_k \in A \mid X_{k-1} = x],$$

para $x \in \mathcal{S}$ y $A \in \mathcal{B}(\mathcal{S})$ ⁷. Los métodos MCMC funcionan construyendo una cadena de Márkov -en un espacio de estados general, i.e., ya sea continuo o discreto- que es ergódica⁸ y tiene como distribución estacionaria o invariante la densidad *a-posteriori* (o en general, alguna densidad objetivo). La condición de ergodicidad implica convergencia a la distribución estacionaria y convergencia a los promedios de la trayectoria, esto último dice que a partir de una muestra de la cadena de Márkov, podemos estimar características de la densidad objetivo (Gilks, Richardson y Spiegelhalter 1995).

Cuando se considerara un espacio de estados general es posible desarrollar la teoría de MCMC, ya que usualmente trabaja con conjuntos de la forma $\mathbb{R}^p \subseteq \mathcal{S}$; de este modo, podemos explorar la densidad objetivo π posiblemente intratable.

2.2.4 Métodos Markov Chain Monte Carlo

Después de mostrar el panorama general, definimos que un método MCMC para la simulación de una distribución π es cualquier técnica que genera una cadena de Markov $\{X_t\}_{t \geq 0}$ ergódica cuya distribución estacionaria es π (Robert P. y Casella 2004). Esto es, genera realizaciones o trayectorias de la cadena de Márkov ergódica

$$\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots, \boldsymbol{\theta}^{(k)}, \dots,$$

de tal modo que se explore el espacio de parámetros completamente y que converga a la distribución estacionaria; en la práctica, consideramos cierto número de iteraciones como un periodo de calentamiento o *warmup* y a partir de este punto, los números simulados

⁷Aquí $\mathcal{B}(\mathcal{S})$ denota el conjunto de Borel con respecto a \mathcal{S} .

⁸Una cadena de Markov es ergódica si satisface tres propiedades, (1) es irreducible: dado cualquier valor inicial de la cadena, con probabilidad positiva es posible alcanzar cualquier conjunto A , (2) es aperiódica: esta propiedad garantiza que la cadena no queda atrapada en un comportamiento cíclico, por ejemplo, saltar entre dos estados por siempre y (3) es recurrente positiva: con probabilidad uno, la cadena revisita cada región del espacio de estados. (Vea Gilks, Richardson y Spiegelhalter 1995, Sección 3.2)

constituyen una muestra (pseudo)aleatoria que es la base para calcular cantidades de interés con el método de Monte Carlo.

Vale la pena señalar que, a pesar de que la teoría detrás de los métodos MCMC garantiza la convergencia a la densidad objetivo, en la práctica no siempre se obtienen resultados satisfactorios. Por esta razón, se han construido varios criterios que permiten evaluar la calidad de la simulación MCMC, las herramientas más comunes para evaluar la simulación son (1) el *tamaño efectivo de muestra*, este compara las realizaciones de la cadena de Markov contra simulaciones independientes e identicamente distribuidas y (2) el coeficiente \hat{R} o estadístico de Gelman-Rubin, el cuál evalúa la convergencia de las simulaciones MCMC, entre otros.

2.2.5 Algoritmo Hamiltonian Monte Carlo

El método Hamiltoniano Monte Carlo o Monte Carlo Híbrido, *Hybrid Monte Carlo* (HMC), al igual que otros algoritmos populares basados MCMC, tiene su origen en la física. Este algoritmo evita el comportamiento de caminata aleatoria y la sensibilidad a parámetros correlacionados que afectan a muchos de los métodos MCMC mediante una serie de pasos que emplea información del gradiente (Hoffman y Gelman 2014).⁹ Algunas características particulares de este método son (1) únicamente puede muestrear densidades continuas en \mathbb{R}^p y (2) transforma el problema de muestrear de una distribución objetivo a simular dinámicas Hamiltonianas (Brooks et al. 2011; 2025). Estas características permiten que converga a distribuciones objetivo de gran dimensión mucho más rápido que métodos más simples como Metropolis con caminata aleatoria o muestreador de Gibbs (Hoffman y Gelman 2014).

No obstante, estas virtudes no vienen sin costo adicional: como se señaló previamente, se requiere calcular o aproximar las derivadas de primer orden de la función objetivo para generar transiciones eficientes, esto en modelos complejos puede ser costoso e incluso imposible. Así mismo, pese a ser un algoritmo poderoso, su desempeño se ve afectado drásticamente por dos parámetros controlados por el usuario: el tamaño de paso ϵ y el

⁹Otro método MCMC que evita el comportamiento de caminata aleatoria, es el método de caminata t , *t-walk* (Christen y Fox 2010). No obstante, no se revisa en este capítulo.

número de pasos L . En el [Cuadro 2.2](#), mostramos de forma resumida que sucede si no empleamos valores adecuados para estos parámetros.

De acuerdo con Hoffman y Gelman (2014), podemos hacer menos oneroso el requerimiento de calcular el vector gradiente de la distribución objetivo -por ejemplo, la *a posteriori*- a través de diferenciación automática. Así mismo, los autores proponen una modificación al algoritmo HMC llamada el muestreador no vuelta-en-U, *no-U-turn sampler* (NUTS), cuya virtud es que elimina la necesidad de elegir el parámetro L , a la vez que propone un esquema para ajustar automáticamente el parámetro ϵ .

Actualmente, el lenguaje de programación probabilística Stan implementa como único método de muestreo de la distribución *a posteriori* el método HMC con esta variante adaptativa ([2025](#)). Equipados con herramientas de diferenciación automática y el método NUTS, es posible ejecutar el método la variante NUTS del método HMC sin ningún tipo de *tuning* manual (Hoffman y Gelman [2014](#)). En el [Algoritmo 1](#) se muestra el algoritmo HMC básico.

Cuadro 2.2: Algunas consecuencias de elegir malas elecciones para los parámetros L y ϵ en el método HMC. Fuente: elaboración propia basado en Hoffman y Gelman ([2014](#)).

Parámetro	Condición	
	Demasiado pequeño	Demasiado grande
ϵ	Se pierde tiempo de cómputo al realizar muchos pasos pequeños	La simulación será imprecisa Se producen bajas tasas de aceptación
L	Las muestras sucesivas están próximas entre sí: se genera una caminata aleatoria y mezcla lenta	Se generan trayectorias que retroceden y vuelven sobre sus pasos. Mala elección. Es posible que al usar una elección inapropiada de L los parámetros ‘saltan’ de un lado del espacio parámetrico al otro en cada iteración, resultando que la cadena de Márkov incluso no sea ergódica. De igual modo, puede ser que la cadena sea ergódica, pero con movimiento muy lento entre las regiones de baja y alta densidad.

Algoritmo 1: Algoritmo Hamiltonian Monte Carlo

Input: Tamaño de paso ϵ . Número de pasos L . Logaritmo de la función objetivo

$\log p(\boldsymbol{\theta})$. Vector de derivadas parciales $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta})$.

Output: Una muestra (pseudo)aleatoria de tamaño M de la densidad objetivo $p(\boldsymbol{\theta})$.

Initialize: Un valor inicial $\boldsymbol{\theta}^{(0)}$

```
1 for  $t = 0, 1, 2, \dots, M$  do
2   Generar  $\mathbf{r}^{(0)} \sim N(\mathbf{r}, I)$            // Similar al muestreador de Gibbs
3
4   Asignar  $\boldsymbol{\theta}^{(m)} \leftarrow \boldsymbol{\theta}^{(m-1)}$ ,  $\tilde{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}^{(m-1)}$ ,  $\tilde{\mathbf{r}}^m \leftarrow \mathbf{r}^{(0)}$ 
5   for  $i = 1, 2, \dots, L$  do
6     | Asignar  $(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{r}}) \leftarrow \text{Leapfrog}(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{r}}, \epsilon)$ 
7   end
8   Con probabilidad  $\alpha$ , asignar  $\boldsymbol{\theta}^{(m)} \leftarrow \tilde{\boldsymbol{\theta}}$ ,  $\mathbf{r}^{(m)} \leftarrow -\tilde{\mathbf{r}}$ 
9
10  
$$\alpha = \min \left\{ 1, \frac{\exp\{\log p(\tilde{\boldsymbol{\theta}}) - \frac{1}{2}\tilde{\mathbf{r}}^T \tilde{\mathbf{r}}\}}{\exp\{\log p(\boldsymbol{\theta}^{(m-1)}) - \frac{1}{2}\mathbf{r}^{(0)T} \mathbf{r}^{(0)}\}} \right\}$$

11
12
13
14
15 return  $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots, \boldsymbol{\theta}^{(M)}$ 
```

Note que, en el paso 8 del [Algoritmo 1](#), es posible trazar similitudes con el algoritmo Metrópolis-Hastings (MH), específicamente en calcular la probabilidad de aceptación de un nuevo candidato, de este modo, el método HMC también puede ser visto como una instancia del algoritmo MH.

2.2.6 Ejemplo 1: aproximando una distribución normal bivariada

Los métodos de muestreo MCMC son de uso general, es decir, su uso no se limita a generar muestras de la *a posteriori*, si no que también pueden explorar cualquier otro tipo de densidad objetivo. A continuación, se ilustra el método HMC para explorar la densidad

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi \det(\Sigma^{-1})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) 1(\mathbf{x} \in \mathbb{R}^2),$$

$$\boldsymbol{\mu} = (-1, 1)^T$$

$$\Sigma = \begin{pmatrix} 1.0 & 0.7 \\ 0.7 & 2.0 \end{pmatrix},$$
(2.22)

es decir, una normal bivariada donde conocemos el vector de localidad y la matriz de covarianzas. De este modo, el objetivo es explorar la distribución -generar muestras aleatorias- de (x_1, x_2) . A continuación, se describen los pasos generales para este método.

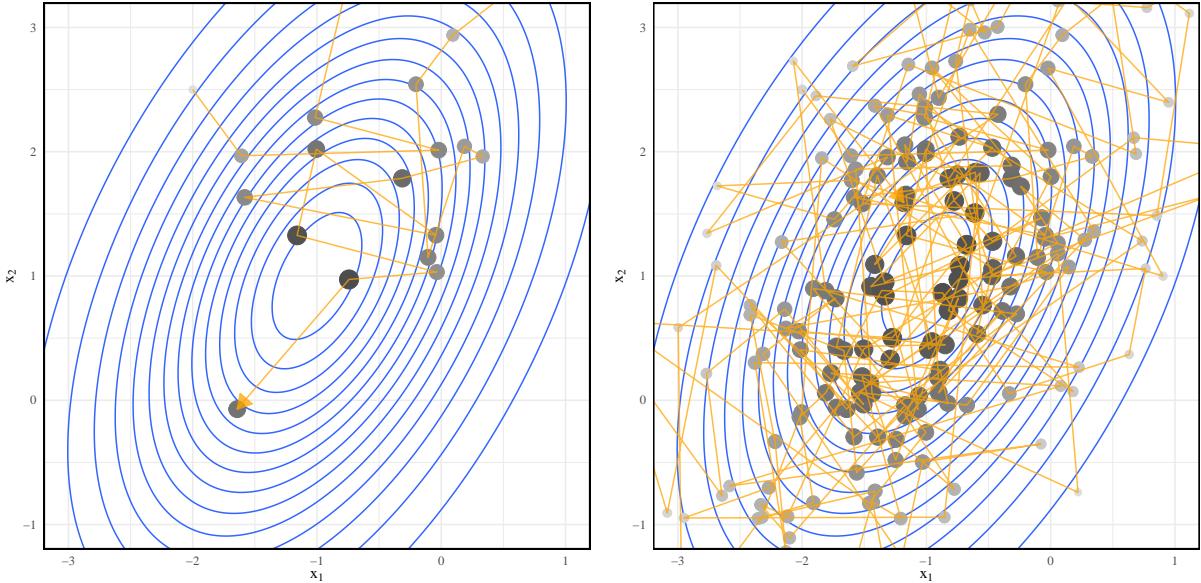
- Método HMC simple, basado en Brooks et al. (2011, Capítulo 5). A partir de esta implementación mostrada en el texto, sólo hace falta calcular la log-densidad y el gradiente de la log-densidad con respecto a cada entrada de (x_1, x_2) . Estas cantidades están dadas por

$$U(q) \propto -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

$$\nabla U(q) = -2\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}),$$
(2.23)

el vector gradiente corresponde a la derivada del vector \mathbf{x} en una forma cuadrática (Petersen y Pedersen 2012).

Todas estas simulaciones se inician en el punto $(-2.0, 2.5)$. En la [Figura 2.6](#) se ilustran esta técnica MCMC, usando 20 y 200 iteraciones en el lado izquierdo y derecho de cada subfigura. Además, se dibujan los contornos de la densidad normal bivariada que se muestrea.



(a) Método HMC, 20 iteraciones.

(b) Método HMC, 200 iteraciones.

Figura 2.6: De izquierda a derecha y de arriba hacia abajo, exploración mediante tres métodos MCMC de una densidad normal en dos dimensiones. El tamaño e intensidad de los puntos corresponde a la densidad en esas zonas, además, se traza la ruta seguida por cada algoritmo.
Fuente: elaboración propia.

2.3 Inferencia Bayesiana variacional

En esta sección mostraremos los elementos conceptuales para desarrollar la teoría de inferencia Bayesiana variacional (BV). El nombre ‘aproximaciones variacionales’ tiene su origen en el cálculo variacional, el cuál se encarga de optimizar un funcional sobre una clase de funciones de las que depende. Las soluciones aproximadas surgen cuando la clase de funciones se restringe de alguna manera, generalmente para mejorar la manejabilidad (Ormerod y Wand 2010). Como se mencionó anteriormente, este enfoque aproxima la densidad *a posteriori* $p(\boldsymbol{\theta} | \mathbf{y})$ -o en general alguna densidad objetivo $p(\boldsymbol{\theta})$ - mediante una densidad de probabilidad $q(\boldsymbol{\theta})$ que pertenece a alguna familia ‘manejable’, es decir, común o sencilla, de distribuciones \mathcal{Q} . La mejor aproximación BV, denotada por $q^* \in \mathcal{Q}$, se encuentra minimizando la divergencia Kullback-Leibler (KL) de $q(\boldsymbol{\theta})$ a $p(\boldsymbol{\theta} | \mathbf{y})$.

Simbólicamente podemos escribir esto como

$$q^*(\boldsymbol{\theta}) = \arg \min_{q \in \mathcal{Q}} \text{KL}(q(\boldsymbol{\theta}) \| p(\mathbf{y} \mid \boldsymbol{\theta})) \quad (2.24)$$

En la literatura podemos identificar dos clases principales de inferencia BV de acuerdo al tipo de restricción impuesta en \mathcal{Q} : los métodos basados en el supuesto de campo medio, *mean field*, que emplean el algoritmo de ascenso de coordenadas, *coordinate ascence algorithm* (CAVI), y los métodos de forma fija que emplean el algoritmo de ascenso del gradiente o alguna otra rutina de optimización.¹⁰ De manera general, el primer método explota la conjugancia de la familia exponencial entre la verosimilitud y las distribuciones *a priori*, de donde se obtienen aproximaciones de forma analítica o cerrada, mientras que el segundo método se extiende a casos fuera de la familia exponencial y no se obtienen aproximaciones analíticas.

Algunas implementaciones populares del método forma fija, son el algoritmo inferencia variacional caja negra, *black box variational inference* (BBVB), *Pathfinder* o inferencia variacional cuasi-Newton paralela, *parallel quasi-Newton variational inference* e inferencia variacional con diferenciación automática, *automatic differentiation variational inference* (ADVI). Más adelante estudiaremos con detenimiento estas dos clases de inferencia BV y en especial a los algoritmos CAVI y ADVI.

De acuerdo con Ormerod y Wand 2010, los métodos basados en campo medio y forma fija también se consideran transformaciones de la densidad. Esto sugiere que existe otro tipo de aproximaciones variacionales que no están basadas en minimizar la divergencia KL, por ejemplo, las aproximaciones por transformación tangente, ya que trabajan con representaciones ‘tangentes’ de funciones cóncavas y convexas. No obstante, en la discusión posterior no se trabaja con este tipo de aproximaciones. Así mismo, las aproximaciones variacionales pueden usarse en contextos frecuentistas, sin embargo, usar inferencia variacional en este escenario es mucho más raro (Ormerod y Wand 2010).

Ahora bien, a grandes rasgos, podemos señalar que los métodos basados en MCMC producen

¹⁰Es posible complementar este último método con ascenso del gradiente estocástico, es decir, tomando muestras o *batches* del conjunto de datos completo) para acelerar la convergencia.

estimaciones más precisas, pero el precio a pagar es en términos del tiempo de cómputo, en cambio, los métodos basados en optimización son más rápidos pero quizás no tan precisos. En este sentido, las características del problema nos pueden ayudar a decidir cuando emplear cada enfoque, por ejemplo: si se tiene una gran cantidad de parámetros o variables latentes, preferimos perder un poco de precisión para mejorar el tiempo de desempeño.

A continuación estudiaremos la función objetivo que el paradigma de inferencia BV busca minimizar.

2.3.1 Divergencia Kullback-Leibler

La divergencia Kullback-Leibler (KL) es una medida de teoría de la información sobre la proximidad entre dos densidades. La divergencia KL entre las densidades p, q se define como

$$\text{KL}(q\|p) \equiv \int q(\theta) \cdot \log \frac{q(\theta)}{p(\theta)} d\theta = \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log p(\theta)], \quad (2.25)$$

donde el valor esperado se toma con respecto a $q(\theta)$. Esta medida es asimétrica, es decir¹¹ $\text{KL}(q\|p) \neq \text{KL}(p\|q)$, además, es no negativa. La divergencia KL es minimizada cuando $q = p$. De forma adicional, puede mostrarse que la divergencia KL no constituye una métrica, ya que carece de las propiedades de simetría y desigualdad del triángulo.

Otro punto a tomar en cuenta, es que la divergencia KL entre p, q , puede ser escrita en términos de sus entropías. Para ver esto, consideramos X, Y variables aleatorias continuas con densidades $p(x)$ y $q(y)$, la entropía de X y la entropía cruzada de X con Y están dadas por

$$\begin{aligned} H(p) &\equiv \mathbb{E}_p[-\log p(x)] = \int_{\mathbb{R}} p(x) \log p(x) dx, \\ H(p, q) &\equiv \mathbb{E}_p[-\log q(y)] = \int_{\mathbb{R}} p(x) \log q(y) dx, \end{aligned} \quad (2.26)$$

¹¹Si $\text{KL}(q\|p)$ es la divergencia KL entre las densidades q y p , la cantidad $\text{KL}(p\|q)$ recibe el nombre de divergencia KL invertida. Note que aunque $\text{KL}(q\|p)$ esté definida, no necesariamente existe la divergencia KL invertida.

por tanto, en la definición de divergencia KL están involucrados estas dos cantidades. Así mismo, es posible desarrollar la expresión que define la divergencia KL entre la aproximación $q(\boldsymbol{\theta})$ propuesta y la verdadera *a posteriori*, en cuyo caso se obtiene

$$\begin{aligned}\text{KL}(q(\boldsymbol{\theta})\|p(\boldsymbol{\theta} \mid \mathbf{y})) &= \mathbb{E}_q[\log q(\boldsymbol{\theta})] - \mathbb{E}_q[\log p(\boldsymbol{\theta} \mid \mathbf{y})] \\ &= \mathbb{E}_q[\log q(\boldsymbol{\theta})] - \mathbb{E}_q[\log p(\mathbf{y}, \boldsymbol{\theta})] + \mathbb{E}_q[\log p(\mathbf{y})] \\ &= \mathbb{E}_q[\log q(\boldsymbol{\theta})] - \mathbb{E}_q[\log p(\mathbf{y}, \boldsymbol{\theta})] + \log p(\mathbf{y}),\end{aligned}\quad (2.27)$$

2.3.2 Ejemplo: divergencia KL entre las densidades normal y normal asimétrica

En este apartado se aproxima la divergencia KL entre dos densidades continuas: la densidad normal y la densidad normal asimétrica, ambas en parametrización estándar, es decir, con localidad cero y varianza unitaria. Ya que la distribución normal es un caso particular de la distribución normal sesgada -con el parámetro de forma $\lambda = 0$ -, se busca contrastar la divergencia producida entre estas dos densidades, fijando $\lambda_p = 0$ en el caso normal y variando $\lambda_q \neq 0$ para el caso asimétrico.

Se emplean métodos Monte Carlo por su implementación sencilla, para ello se usó el lenguaje Python 3.9.0 tanto para la tarea de simulación de números aleatorios y para la visualización gráfica. Los detalles acerca de la implementación se encuentran en el [Anexo 4](#).

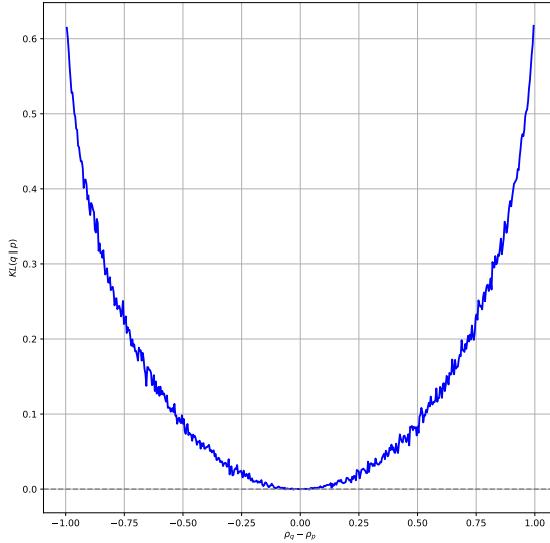
La discusión en la [Sección 2.1](#), proporciona una expresión invertible para el parámetro de forma λ que lo transforma a un parámetro de correlación ρ , es decir, mapea $(-\infty, \infty)$ al intervalo $(-1, 1)$, así, parece más simple considerar el intervalo $(-1, 1)$ en lugar de todo \mathbb{R} . Esta observación permite obtener los dos tipos de gráficos que se muestran en la [Figura 2.7](#). En ambos gráficos, es evidente que la divergencia KL no se comporta de manera lineal, así, en la [Figura 2.7a](#), se hace más evidente observar que cuando la diferencia $\rho_q - \rho_p$ es pequeña -digamos entre $(-0.5, 0.5)$ -, la divergencia KL produce valores pequeños y viceversa.

No obstante, a partir de las propiedades de la distribución normal asimétrica descritas en la [Sección 2.1.1](#), cuando el parámetro de forma $\lambda \rightarrow \pm\infty$, entonces la ley normal sesgada

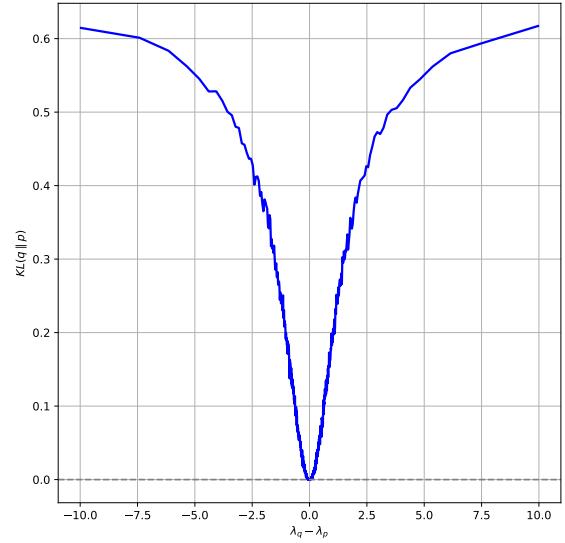
converge a una ley normal truncada a la izquierda (derecha) en cero. De este modo, es posible calcular la divergencia KL entre dos casos extremos: $\lambda = 0$, la densidad normal usual, y $\lambda \rightarrow \pm\infty$, la densidad normal truncada. Así mismo, es posible calcular la divergencia KL entre estas dos densidades, y para ello escribimos¹²

$$\begin{aligned} \text{KL}(NT(x | 0, 1) \| N(x | 0, 1)) &= \int_0^\infty NT(x | 0, 1) \log \frac{NT(x | 0, 1)}{N(x | 0, 1)} dx \\ &= \int_0^\infty \frac{2}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \log \frac{\frac{2}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}}{\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}} dx \\ &= \log(2) \int_0^\infty \frac{2}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx = \log(2), \end{aligned} \quad (2.28)$$

es decir, el valor máximo que podría alcanzar la divergencia KL entre las densidades normal asimétrica y normal, en el caso de localidad cero y escala unitaria, es $\log(2) \simeq 0.6931$. En la Figura 2.7a y la Figura 2.7b se observa que los valores de la divergencia KL están acotados por este valor. Concluimos este ejemplo comentando que, en general, no es claro como determinar cuando un valor de la divergencia KL entre dos densidades es grande o pequeño.



(a) Cálculo de la divergencia KL con ρ .



(b) Cálculo de la divergencia KL con λ .

Figura 2.7: Divergencia KL entre las densidades normal sesgada y normal conforme el valor absoluto del parámetro de correlación/forma crece. Fuente: elaboración propia.

¹²En este caso, no podemos considerar la divergencia invertida, es decir $\text{KL}(N(x | 0, 1) \| NT(x | 0, 1))$, ya que para la densidad normal truncada, en el intervalo $(-\infty, 0)$ evalúa a cero.

2.3.3 Límite inferior de la evidencia

En la [Ecuación 2.27](#), se desarrolla la expresión que define la divergencia KL entre la aproximación variacional propuesta y la verdadera *a posteriori*, no obstante, habitualmente no es posible calcular $p(\mathbf{y})$ -y con ello $\log p(\mathbf{y})$ -, por lo que no es posible minimizar exactamente la divergencia KL. Debido a esto, en la práctica se optimiza una función objetivo que es equivalente salvo por una constante: la función límite inferior de la evidencia, *evidence lower bound* (ELBO).¹³ Esta se define como

$$\text{ELBO}(q) \equiv \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta = \mathbb{E}_q \left[\log \frac{p(\theta, y)}{q(\theta)} \right], \quad (2.29)$$

y al desarrollar esta expresión se tiene que

$$\begin{aligned} \text{ELBO}(q) &= \mathbb{E}_q[\log p(y, \theta)] - \mathbb{E}_q[\log q(\theta)] \\ &= \mathbb{E}_q[\log p(\theta)] + \mathbb{E}_q[\log p(y | \theta)] - \mathbb{E}_q[\log q(\theta)] \\ &= \mathbb{E}_q[\log p(y | \theta)] - \text{KL}(q(\theta) \| p(\theta)). \end{aligned} \quad (2.30)$$

De acuerdo con Blei, Kucukelbir y McAuliffe ([2017](#)), el objetivo variacional en la [Ecuación 2.30](#) refleja el balance usual entre la función de verosimilitud y la *a priori*, ya que el primer término es el valor esperado de una verosimilitud, el cuál favorece densidades que colocan su masa en configuraciones de los parámetros que explican a los datos observados, a la vez que el segundo término es la divergencia KL entre la densidad candidata y la *a priori*, el cuál promueve densidades cercanas a la *a priori*.

Una propiedad de la ELBO es que acota por abajo el logaritmo de la evidencia, es decir, para cualquier $q(\theta)$ se tiene que

$$\log p(y) = \text{KL}(q(\theta) \| p(\theta|y)) + \text{ELBO}(q), \quad (2.31)$$

y dado que $\text{KL}(\cdot \| \cdot) \geq 0$, se tiene que $\log p(y) \geq \text{ELBO}(q)$; de ahí su nombre. Note que

¹³Si se conociera la forma de $p(\mathbf{y})$, entonces puede realizarse inferencia Bayesiana de forma cerrada.

podemos expresar la divergencia KL exacta en la [Ecuación 2.27](#) como

$$\text{KL}(q(\boldsymbol{\theta}) \| p(\boldsymbol{\theta} | \mathbf{y})) = \log p(\mathbf{y}) - \text{ELBO}(q), \quad (2.32)$$

si bien $\log p(\mathbf{y})$ es desconocido, es constante respecto a q . Por lo tanto, el objetivo de minimizar la divergencia KL (un funcional) sobre la aproximación q (una función), es equivalente a maximizar el $\text{ELBO}(q)$, el signo de estas dos cantidades es clave. En consecuencia, esto significa que también podemos obtener las densidades q^* óptimas como

$$q^*(\boldsymbol{\theta}) = \arg \max_{q \in \mathcal{Q}} \text{ELBO}(q). \quad (2.33)$$

Finalmente, recordamos que si no imponemos restricciones en \mathcal{Q} , la mejor aproximación BV, la que minimiza la divergencia KL es $q^* = p(\boldsymbol{\theta} | \mathbf{y})$, pero es intratable. Como se mencionó al inicio de la sección, dependiendo de la restricción impuesta en la clase \mathcal{Q} , los algoritmos de inferencia variacional pueden ser clasificados, en este caso, consideramos exclusivamente a los métodos campo medio y forma fija. En Rohde y Wand ([2016](#)), llaman a estos dos métodos campo medio no paramétrico y campo medio semiparamétrico o paramétrico. En turno, estudiaremos cada una de las dos alternativas.

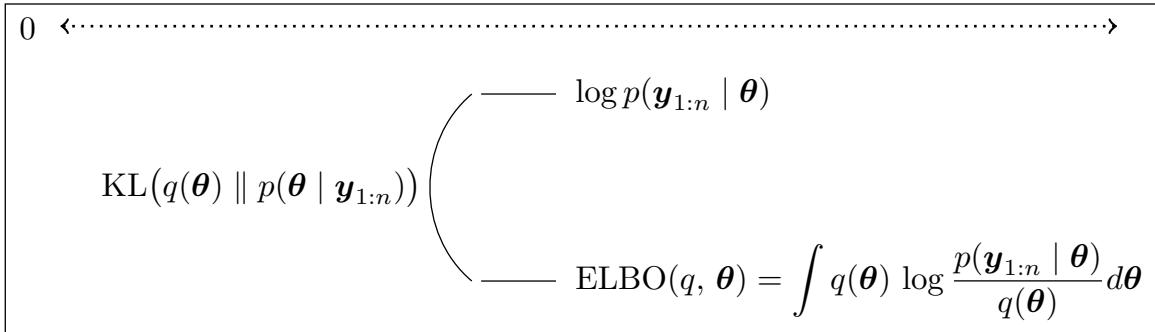


Figura 2.8: Representación del límite inferior de la evidencia en la [Ecuación 2.32](#). La línea punteada representa la recta cero. Por lo general, las cantidades $\text{ELBO}(q)$, y $\log p(\mathbf{y} | \boldsymbol{\theta})$ son negativas. Fuente: elaboración propia basado en Bishop ([2006](#), Figura 9.11).

2.3.4 Restricción Campo Medio

Esta restricción asume que la familia variacional se factoriza como

$$q(\theta_1, \theta_2, \dots, \theta_k) = \prod_{i=1}^k q_i(\theta_i),$$

esto es más general de lo que aparenta: los parámetros pueden ser agrupados y factorizar la distribución de cada grupo (Blei, Kucukelbir y McAuliffe 2017). De acuerdo con Ormerod y Wand (2010), Tran, T.-N. Nguyen y Dao (2021) y Rohde y Wand (2016), esta restricción es no paramétrica, ya que no se especifica la forma de los factores variacionales, así mismo, Blei, Kucukelbir y McAuliffe (2017) señala que en principio, cada factor óptimo puede adoptar cualquier forma paramétrica apropiada para la variable aleatoria que corresponde. Ormerod y Wand (2010) señala que esta restricción tiene sus orígenes en la física estadística.

2.3.5 Algoritmo Inferencia Variacional por Ascenso de Coordenadas

Supongamos que $\boldsymbol{\theta}$ está dividido en k bloques $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)^T$. Se desea aproximar la densidad posterior $p(\theta_1, \theta_2, \dots, \theta_k \mid y)$ con $q(\theta) = q_1(\theta_1)q_2(\theta_2)\cdots q_k(\theta_k)$. De acuerdo con Blei, Kucukelbir y McAuliffe (2017) y Ormerod y Wand (2010), la densidad $q_j(\theta_j)$ que maximiza el límite inferior de la evidencia, es decir, el ELBO(q), cuando $q_1, \dots, q_{j-1}, q_{j+1}, \dots, q_k$ permanecen fijos está dada por

$$q_j(\theta_j) \propto \exp(\mathbb{E}_{-q_j}[\log p(y, \boldsymbol{\theta})]), \quad j = 1, 2, \dots, k, \quad (2.34)$$

y por el supuesto de Campo Medio, es equivalente a la expresión

$$q_j(\theta_j) \propto \exp(\mathbb{E}_{-q_j}[\log p(\theta_j \mid \theta_{-j}, y)]), \quad j = 1, 2, \dots, k, \quad (2.35)$$

estas expresiones son la base del algoritmo inferencia variacional por ascenso de coordenadas (CAVI), el cuál presentamos en el [Algoritmo 2](#). En el [Anexo 2](#) se da una prueba de esto. Las densidades óptima obtenidas mediante el supuesto de campo medio

recuerdan al muestreador de Gibbs: la densidad óptima en la [Ecuación 2.35](#) se obtiene a partir de la densidad condicional completa de θ_j dado el resto de parámetros, así mismo, algunos autores señalan que el primer término del ELBO en la [Ecuación 2.30](#) es la esperanza de la log-verosimilitud completa, la cuál es optimizada por el algoritmo de esperanza maximización (EM).

Blei, Kucukelbir y McAuliffe ([2017](#)) señalan que el algoritmo EM fue diseñado para encontrar las estimaciones de máxima verosimilitud en modelos con variables latentes. Usa el hecho de que el ELBO es igual a $\log p(\mathbf{y})$ (i.e., el logaritmo de la evidencia) cuando $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta} | \mathbf{y})$. En contraste con la inferencia variacional, el algoritmo EM asume que la esperanza sobre $p(\boldsymbol{\theta} | \mathbf{y})$ se puede calcular y la utiliza en problemas de estimación de parámetros que de otro modo serían difíciles.

Algoritmo 2: Inferencia variacional por ascenso de coordenadas (CAVI)

Input: Un modelo $p(y, \boldsymbol{\theta})$
Output: La densidad variacional óptima $q^*(\boldsymbol{\theta}) = q_1^*(\theta_1) \cdots q_k^*(\theta_k)$
Initialize: Factores variacionales $q_j(\theta_j)$

```

1 while la cota inferior de la evidencia (ELBO) no converga do
2   for  $j \in \{1, 2, \dots, k\}$  do
3     Calcular de forma analítica
      
$$q_j^*(\theta_j) \propto \exp(\mathbb{E}_{-q_j}[\log p(\theta_j | \theta_{-j}, y)])$$

4   end
5   Calcular o estimar
      
$$\text{ELBO}(q) = \mathbb{E}[\log p(y, \theta)] - \mathbb{E}[\log q(\boldsymbol{\theta})]$$

6 end
7 return  $q_1^*(\theta_1) \cdots q_k^*(\theta_k)$ 
```

De acuerdo con Tran, T.-N. Nguyen y Dao ([2021](#)), una regla de parada para el [Algoritmo 2](#) es terminar la actualización si el cambio en los parámetros de la *a posteriori* aproximada, es decir $\boldsymbol{\theta}^{(k)}$, entre dos iteraciones consecutivas es menor que cierto umbral ε . De igual modo, si fuera posible calcular el $\text{ELBO}(q)$, podemos parar el algoritmo si el incremento o un porcentaje de este es menor que cierto umbral. Aunado a esto, los autores también

comentan que teóricamente, el ELBO (q) incrementa después de cada iteración del algoritmo CAVI, sin embargo, si en lugar de calcularlo de forma analítica este se aproxima de forma numérica, esta estrategia puede generar iteraciones donde el ELBO es decreciente, por lo que debe elegirse un criterio adecuado, por ejemplo, monitorear un promedio móvil.

Quizás la limitante más importante de este método, es que solamente podemos determinar de forma cerrada o analítica las distribuciones óptimas $q^*(\theta_i)$ de cada parámetro cuando existe conjugancia entre la verosimilitud del modelo y la distribución *a priori* de θ_i . Por otra parte, trabajar con la familia exponencial hace que la tarea de obtener la densidad óptima $q^*(\theta_i)$ sea más sencilla; con respecto a este último punto, Blei, Kucukelbir y McAuliffe (2017) muestra la forma de estas actualizaciones.

2.3.6 Ejemplo 1: aproximando una distribución normal bivariada (revisitado)

A continuación, se ilustra el algoritmo de inferencia variacional campo medio, aquí el interés bastante general, es decir, buscamos aproximar una densidad de probabilidad bivariada, que en este caso no corresponde a una densidad *a posteriori*. Para ilustrar el método, nuevamente se empleada la densidad Gausiana en dos dimensiones de la Sección 2.2.6 dada por

$$\begin{aligned} p(x_1, x_2 \mid \boldsymbol{\mu}, \Sigma) &= N_2(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma), \\ \boldsymbol{\mu} &= (-1, 1)^T \\ \Sigma &= \begin{bmatrix} 1.0 & 0.7 \\ 0.7 & 2.0 \end{bmatrix}, \end{aligned} \tag{2.36}$$

este escenario también se ilustra en D. Nguyen (2013) y Bishop (2006). Empleando el supuesto campo medio, podemos escribir la aproximación q como

$$q(x_1, x_2) = q(x_1)q(x_2), \tag{2.37}$$

después, es necesario determinar las densidades óptimas $q^*(x_i)$, por lo que empleamos el resultado en la [Ecuación 2.35](#), de ahí que

$$q^*(x_1) \propto \exp \left\{ \mathbb{E}_{-x_1} [\log p(x_1 | x_2, \boldsymbol{\mu}, \Sigma)] \right\}, \quad (2.38)$$

aquí podemos notar que $p(x_1 | x_2, \boldsymbol{\mu}, \Sigma) = N(x_1 | \mu_{1|2}, \sigma_{1|2}^2)$ por un resultado conocido, además $\mu_{1|2} = \mu_1 + \sigma_{12}\sigma_{22}^{-1}(z_2 - \mu_2)$ y $\sigma_{1|2}^2 = \sigma_{11} - \sigma_{12}\sigma_{22}^{-1}\sigma_{21}$, aquí $\sigma_{ii} = \sigma_i^2$. Por tanto, escribimos

$$q^*(x_1) = \exp \left\{ \mathbb{E}_{q(x_2)} \left[\log \frac{1}{\sqrt{2\pi\sigma_{1|2}^2}} \exp \left(-\frac{(x_1 - \mu_{1|2})^2}{2\sigma_{1|2}^2} \right) \right] \right\},$$

desarrollando la expresión y removiendo constantes que no contengan x_1

$$q^*(x_1) \propto \exp \left\{ \mathbb{E}_{q(x_2)} \left[-\frac{x_1^2 - 2x_1\mu_{1|2}}{2\sigma_{1|2}^2} \right] \right\},$$

note que $\mu_{1|2}$ depende de z_2 , así que desarrollamos y tomamos la esperanza con respecto a $q(x_2)$

$$\begin{aligned} q^*(x_1) &= \exp \left\{ -\frac{1}{2\sigma_{1|2}^2} [x_1^2 - 2x_1 \mathbb{E}_{q(x_2)}(\mu_1 + \sigma_{12}\sigma_{22}^{-1}[x_2 - \mu_2])] \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma_{1|2}^2} [x_1^2 - 2x_1 (\mu_1 + \sigma_{12}\sigma_{22}^{-1} [\mathbb{E}_{q(x_2)}[x_2] - \mu_2])] \right\}, \end{aligned}$$

de este modo, identificamos que esta densidad variacional óptima está dada por

$$\begin{aligned} q^*(x_1) &= N(x_1 | \mu_1^*, \sigma_1^{2*}), \\ \mu_1^* &= \mu_1 + \sigma_{12}\sigma_{22}^{-1}(\mathbb{E}_{q(x_2)}[x_2] - \mu_2), \\ \sigma_1^{2*} &= \sigma_{1|2} \end{aligned} \quad (2.39)$$

la densidad óptima $q^*(x_2)$ se obtiene inmediatamente por simetría, de este modo hemos aproximado $p(x_1, x_2) \approx q^*(x_1)q^*(x_2)$. En la [Figura 2.9](#) se comparan los contornos de la densidad real contra la aproximación Campo Medio. Notamos que la aproximación

variacional está centrada en la densidad real, no obstante, su capacidad para reensamblar la estructura de correlación es limitada.

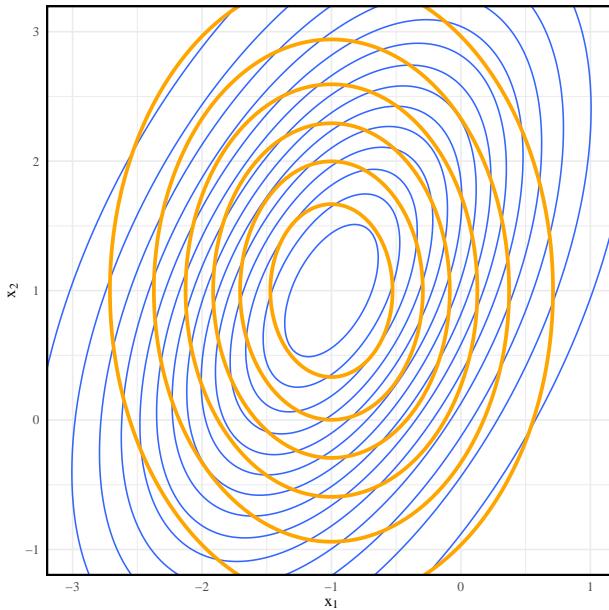


Figura 2.9

2.3.7 Ejemplo 2: inferencia para la distribución normal

Ahora, se muestra un ejemplo de inferencia Bayesiana variacional donde se aproxima la densidad *a posteriori* de los parámetros en una distribución normal. Para tal propósito, se emplea el supuesto campo medio de la [Sección 2.3.4](#). En este escenario, la distribución *a priori* conjugada de (μ, σ^2) es la densidad normal gamma-inversa, lo que permite realizar inferencia de forma cerrada y sirve como referencia para comparar la aproximación que realiza el método BV.

Sea $\mathbf{y} = (y_1, y_2, \dots, y_n)$ una muestra aleatoria de tamaño n de $f_Y(y | \mu, \sigma^2)$, se asume que (μ, σ^2) son ambos desconocidos y por ello les asignamos una *a priori*, como se mencionó

previamente, la distribución normal gamma-inversa es conjugada para este modelo:

$$\begin{aligned}
p(\mu, \sigma^2) &\equiv p(\mu | \sigma^2) \times p(\sigma^2) \\
&= N(\mu | \mu_0, \lambda_0 \sigma^2) \times IG(\sigma^2 | a_0, b_0) \\
&\equiv NIG(\mu_0, \lambda_0, a_0, b_0).
\end{aligned} \tag{2.40}$$

No es difícil mostrar que la distribución *a posteriori* $p(\mu, \sigma^2 | \mathbf{x})$ también es $NIG(\mu, \sigma^2 | \mu_n, \lambda_n, a_n, b_n)$, donde $\mu_n = \lambda_n(\lambda_0^{-1}\mu_0 + n\bar{y})$, $\lambda_n = n + \lambda_0^{-1}$, $a_n = a_0 + n/2$ y $b_n = b_0 + \frac{1}{2} \left(\mu_0^2 \lambda_0^{-1} + \sum_{i=1}^n x_i^2 - \mu_n^2 \lambda_n^{-1} \right)$ (Murphy 2007). Ahora bien, para la aproximación BV empleamos el supuesto de campo medio, es decir $p(\mu, \sigma^2 | \mathbf{x}) = q(\mu)q(\sigma^2)$, luego, obtenemos las densidades óptimas $q^*(\mu)$ y $q^*(\sigma^2)$ de forma analítica, por comodidad consideramos el logaritmo de cada densidad:

$$\begin{aligned}
\log q^*(\mu) &\propto \mathbb{E}_{-\mu} [\log p(\mathbf{y}, \mu, \sigma^2)] \\
&\propto \mathbb{E}_{q(\sigma^2)} [\log p(\mathbf{y} | \mu, \sigma^2) + \log p(\mu | \sigma^2)] \\
&= \mathbb{E}_{q(\sigma^2)} \left[\log \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp\left(-\frac{\sum(y_i - \mu)^2}{2\sigma^2}\right) + \log \frac{\lambda_0}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\lambda_0}{2\sigma^2}(\mu - \mu_0)^2\right) \right] \\
&\propto \mathbb{E}_{q(\sigma^2)} \left[-\frac{\sum(y_i - \mu)^2}{2\sigma^2} - \frac{\lambda_0}{2\sigma^2}(\mu - \mu_0)^2 \right] \\
&\propto \mathbb{E}_{q(\sigma^2)} \left[-\frac{n\mu^2 - 2n\mu\bar{y}}{2\sigma^2} - \frac{\lambda_0}{2\sigma^2}(\mu^2 - 2\mu\mu_0) \right] \\
&= \mathbb{E}_{q(\sigma^2)} \left[-\frac{\lambda_0}{2\sigma^2} \left(\left(\frac{n}{\lambda_0} + 1 \right) \mu^2 - 2 \left(\frac{n\bar{y}}{\lambda_0} + \mu_0 \right) \mu \right) \right] \\
&= \mathbb{E}_{q(\sigma^2)} \left[-\frac{\lambda_0}{2\sigma^2} \left(\frac{n + \lambda_0}{\lambda_0} \mu^2 - 2 \left(\frac{n\bar{y} + \lambda_0\mu_0}{\lambda_0} \right) \mu \right) \right] \\
&= \mathbb{E}_{q(\sigma^2)} \left[-\frac{\cancel{\lambda_0}(n + \lambda_0)}{2\sigma^2 \cancel{\lambda_0}} \left(\mu^2 - 2 \left(\frac{n\bar{y} + \lambda_0\mu_0}{\cancel{\lambda_0}} \right) \mu \right) \right] \\
&= \mathbb{E}_{q(\sigma^2)} \left[-\frac{n + \lambda_0}{2\sigma^2} \left(\mu^2 - 2 \left(\frac{n\bar{y} + \lambda_0\mu_0}{n + \lambda_0} \right) \mu \right) \right]
\end{aligned} \tag{2.41}$$

ahora, tomando la esperanza con respecto a $q(\sigma^2)$,

$$= -\frac{n + \lambda_0}{2} \mathbb{E}_{q(\sigma^2)} \left[\frac{1}{\sigma^2} \right] \left(\mu^2 - 2 \left(\frac{n\bar{y} + \lambda_0\mu_0}{n + \lambda_0} \right) \mu \right), \tag{2.42}$$

completando el término cuadrado en μ y elevando a e ambos lados, podemos identificamos el kernel gaussiano, es decir que

$$\begin{aligned} q^*(\mu) &= N(\mu \mid \mu^*, \sigma^{2*}), \\ \mu^* &= (n\bar{y} + \lambda_0\mu_0)/(n + \lambda_0), \\ \sigma^{2*} &= [\mathbb{E}_{q(\sigma^2)}[1/\sigma^2](n + \lambda_0)]^{-1}, \end{aligned} \tag{2.43}$$

procedemos análogamente para obtener $q^*(\sigma^2)$

$$\begin{aligned} \log q^*(\sigma^2) &\propto \mathbb{E}_{-\sigma^2} [\log p(\mathbf{y}, \mu, \sigma^2)] \\ \log q^*(\sigma^2) &\propto \mathbb{E}_{q(\mu)} [\log p(\mathbf{y} \mid \mu, \sigma^2) + \log p(\mu \mid \sigma^2) + \log p(\sigma^2)] \\ &= \mathbb{E}_{q(\mu)} \left[\log \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp(-\frac{\sum(y_i - \mu)^2}{2\sigma^2}) + \log \frac{\lambda_0}{\sqrt{2\pi\sigma^2}} \exp(-\frac{\lambda_0}{2\sigma^2}(\mu - \mu_0)^2) \right. \\ &\quad \left. + \frac{b_0^{a_0}}{\Gamma(a_0)} (\sigma^2)^{-(a_0-1)} \exp(-b_0/\sigma^2) \right] \\ &\propto \mathbb{E}_{q(\mu)} \left[-\frac{n}{2} \log(\sigma^2) - \frac{\sum(y_i - \mu)^2}{2\sigma^2} - \frac{1}{2} \log(\sigma^2) - \frac{\lambda_0}{2\sigma^2}(\mu - \mu_0)^2 - (a_0 - 1) \log(\sigma^2) - \frac{b_0}{\sigma^2} \right] \\ &= \mathbb{E}_{q(\mu)} \left[-\left(a_0 + \frac{n+1}{2} \right) \log(\sigma^2) + \left(b_0 + \frac{1}{2} \sum(y_i - \mu)^2 + \frac{\lambda_0}{2}(\mu - \mu_0)^2 \right) / \sigma^2 \right], \end{aligned} \tag{2.44}$$

ahora, tomando la esperanza con respecto a $q(\mu)$

$$= -\left(a_0 + \frac{n+1}{2} \right) \log(\sigma^2) + \left(b_0 + \frac{1}{2} \mathbb{E}_{q(\mu)} \left[\sum(y_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2 \right] \right) / \sigma^2, \tag{2.45}$$

elevando a e ambos lados, podemos identificar el kernel Gamma-Inversa, es decir que

$$\begin{aligned} q^*(\sigma^2) &= IG(\sigma^2 \mid a^*, b^*), \\ a^* &= a_0 + (n+1)/2, \\ b^* &= b_0 + \frac{1}{2} \mathbb{E}_{q(\mu)} \left[\sum(y_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2 \right], \end{aligned} \tag{2.46}$$

notamos que ambas densidades optimas dependen de los momentos de la otra densidad, esto sugiere emplear un esquema iterativo como en el [Algoritmo 2](#), para ello, desarrollamos las

expresiones de los parámetros óptimos. Para $\sigma^{2\star}$ de $q^\star(\mu)$ se tiene que

$$\mathbb{E}_{q(\sigma^2)}[1/\sigma^2] = \frac{a^\star}{b^\star}, \quad (2.47)$$

como $q^\star(\sigma^2) = IG(\sigma^2 \mid a^\star, b^\star)$, entonces $(\sigma^2)^{-1} \sim \text{Gamma}(a^\star, b^\star)$, de ahí que podemos evaluar inmediatamente ese momento. Ahora, para b^\star de $q^\star(\sigma^2)$ se calcula

$$\frac{1}{2}\mathbb{E}_{q(\mu)} \left[\sum (y_i - \mu^2) + \lambda_0(\mu - \mu_0)^2 \right] = \frac{1}{2} \sum \mathbb{E}_{q(\mu)} [y_i^2 + \mu^2 - 2y_i\mu] + \frac{\lambda_0}{2} \mathbb{E}_{q(\mu)} [\mu^2 + \mu_0^2 - 2\mu\mu_0], \quad (2.48)$$

agrupando términos y tomando esperanza con respecto a $q(\mu)$

$$= \frac{n + \lambda_0}{2} \mathbb{E}_{q(\mu)}[\mu^2] - (n\bar{y} + \lambda_0\mu_0) \mathbb{E}_{q(\mu)}[\mu] + \frac{\sum y_i^2 + \lambda_0\mu_0^2}{2}, \quad (2.49)$$

como $q^\star(\mu)$ es una densidad Gaussiana, nuevamente podemos evaluar los momentos de forma inmediata:

$$\mathbb{E}_{q(\mu)}[\mu] = (n\bar{y} + \lambda_0\mu_0)/(n + \lambda_0), \quad (2.50)$$

ahora, usando la relación $\text{Var}[\mu] = \mathbb{E}[\mu^2] - (\mathbb{E}[\mu])^2$ despejamos el segundo momento poblacional y escribimos

$$\mathbb{E}_{q(\mu)}[\mu^2] = [\mathbb{E}_{q(\sigma^2)}[1/\sigma^2](n + \lambda_0)]^{-1} + (\mathbb{E}_{q(\mu)}[\mu])^2, \quad (2.51)$$

equipados con estas expresiones, sólo necesitamos especificar (1) los valores para los hiperparámetros $\mu_0, \lambda_0, a_0, b_0$ y (2) valores iniciales para los momentos poblacionales $\mathbb{E}_{q(\sigma^2)}[\sigma^2]$, $\mathbb{E}_{q(\mu)}[\mu]$ y $\mathbb{E}_{q(\mu)}[\mu^2]$. En la [Figura 2.10](#) se muestra una implementación de este algoritmo. Notamos que con un par de iteraciones, el algoritmo parece reproducir de forma exitosa la forma de la densidad posterior exacta, la cuál es normal gamma-inversa. La programación de este método se realizó en Python 3.9.0, y los códigos empleados pueden consultarse en el [Anexo 4](#).

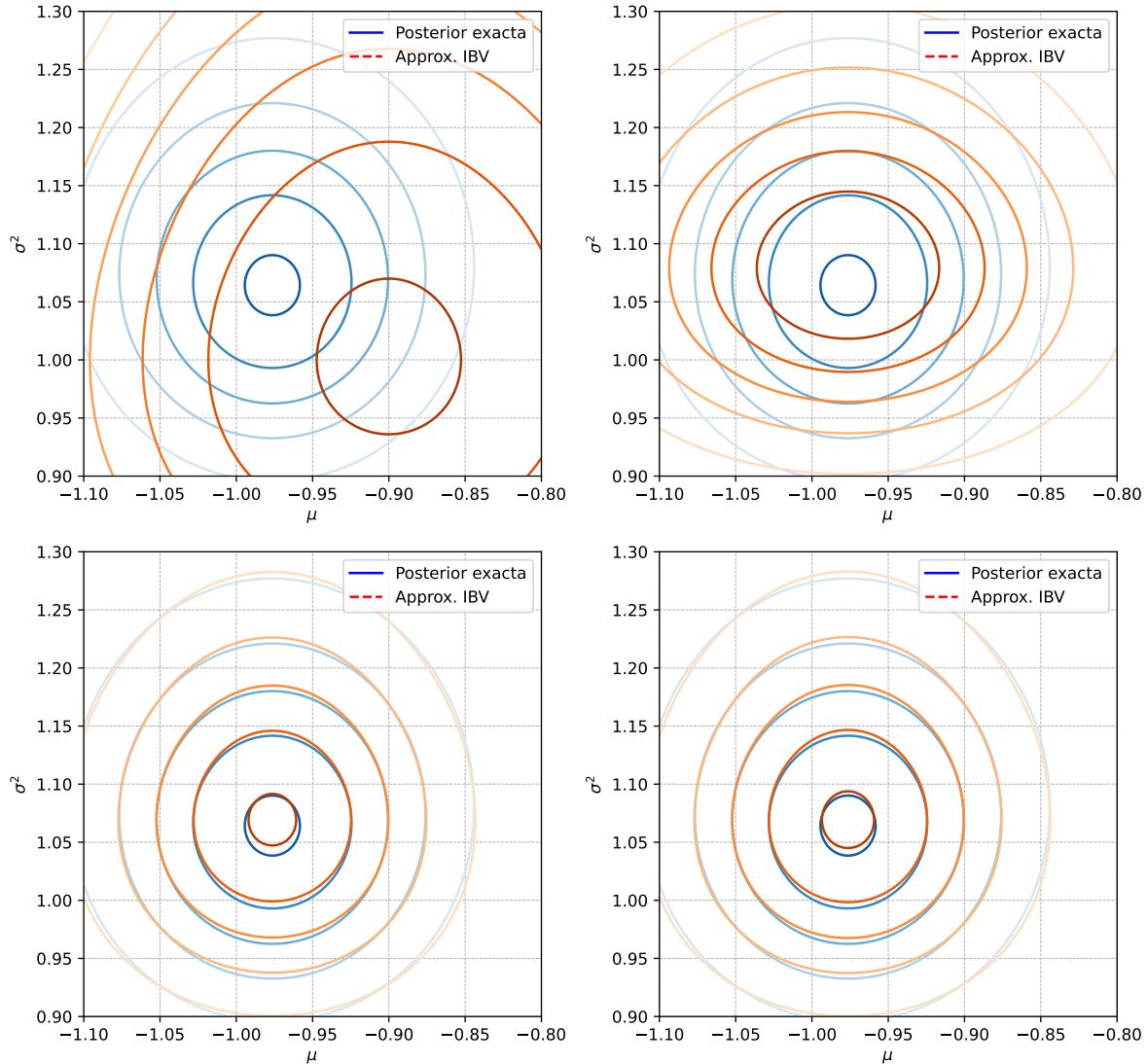


Figura 2.10: De izq. a der. y de arriba hacia abajo: inicialización, iteraciones 1, 20 y 200.
Fuente: elaboración propia basada en Bishop (2006).

2.3.8 Ejemplo 3: inferencia para la distribución normal asimétrica

Este ejemplo ilustra el método de inferencia BV para el modelo de regresión con respuesta continua que se plantea más adelante en la [Sección 3.2](#), sin embargo, aquí se sigue un enfoque analítico. Iniciamos con el proceso de truncamiento oculto descrito en la [Sección 2.1.5](#), sea

$$\begin{bmatrix} \mathbf{V}_n \\ W \end{bmatrix} \sim N_{n+1} \left(\begin{bmatrix} X\boldsymbol{\beta} \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} [(1 - \rho^2)I_n + \rho^2 J_n] & \rho \mathbf{1}_n \\ \rho \mathbf{1}_n^T & 1 \end{bmatrix} \right), \quad (2.52)$$

y si definimos $\mathbf{Y}_n = \mathbf{V}_n$ siempre que $W > 0$, entonces por los resultados de la [Sección 2.1.1](#) y [Sección 2.1.2](#), tanto Y_j como \mathbf{Y}_n tienen ley normal asimétrica univariada y multivariada, además, ya se mostró que

$$p(\mathbf{y}_n \mid w) = \prod_{i=1}^n N(y_i \mid \mathbf{x}_i^T \boldsymbol{\beta} + w\rho, \sigma^2(1 - \rho^2)) 2N(w \mid 0, \sigma^2) I_{(0, \infty)}(w). \quad (2.53)$$

Para hacer un tratamiento Bayesiano completo, debemos asignar una distribución *a priori* a los parámetros $\boldsymbol{\theta} = (\rho, \sigma^2, \boldsymbol{\beta})^T$, la cuál establecemos como

$$p(\boldsymbol{\theta}) \equiv p(\rho, \sigma^2, \boldsymbol{\beta}) = p(\sigma^2) p(\rho) \propto \frac{1}{\sigma^2} \frac{\sqrt{1 + \rho^2}}{1 - \rho^2}, \quad (2.54)$$

como se verá en la [Sección 3.5](#), esta distribución *a priori* corresponde a la distribución de referencia. Ahora, para determinar las densidades variacionales óptimas, empleamos la restricción campo medio, esto es, aproximamos la densidad *a posteriori* como

$$p(\boldsymbol{\theta}, w \mid \mathbf{y}) \equiv p(\boldsymbol{\beta}, \sigma^2, \rho, w \mid \mathbf{y}) = q(\boldsymbol{\beta}) q(w) q(\sigma^2) q(\rho). \quad (2.55)$$

Iniciamos determinando la densidad óptima $q^*(\boldsymbol{\beta})$, para ello escribimos

$$\begin{aligned} \log q^*(\boldsymbol{\beta}) &\propto \mathbb{E}_{-\boldsymbol{\beta}} [\log p(\mathbf{y}, w, \rho, \sigma^2, \boldsymbol{\beta})] \\ &= \mathbb{E}_{q(\rho, \sigma^2, w)} [\log p(\mathbf{y} \mid \boldsymbol{\beta}, w, \sigma^2, \rho)] \end{aligned}$$

$$= \mathbb{E}_q \left[\log \left(\frac{1}{2\pi\sigma^2(1-\rho^2)} \right)^{n/2} \exp \left(-\frac{1}{2} \frac{\sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - w\rho)^2}{\sigma^2(1-\rho^2)} \right) \right],$$

note que $\sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \mathbf{1}_n w\rho)^2 = (\mathbf{y} - X\boldsymbol{\beta} - w\rho\mathbf{1}_n)^T(\mathbf{y} - X\boldsymbol{\beta} - w\rho\mathbf{1}_n)$, ahora, aplicando logaritmo en ambos lados e ignorando constantes

$$\propto \mathbb{E}_q \left[-\frac{1}{2\sigma^2(1-\rho^2)} (\mathbf{y} - X\boldsymbol{\beta} - w\rho\mathbf{1}_n)^T(\mathbf{y} - X\boldsymbol{\beta} - w\rho\mathbf{1}_n) \right],$$

desarrollando la forma cuadrática en $\boldsymbol{\beta}$ e ignorando constantes

$$\propto \mathbb{E}_q \left[-\frac{1}{2\sigma^2(1-\rho^2)} (\boldsymbol{\beta}^T X^T X \boldsymbol{\beta} - 2\boldsymbol{\beta} [X^T \mathbf{y} - X^T w\rho\mathbf{1}_n]) \right],$$

luego, evaluamos y sepáramos con cuidado las esperanzas¹⁴

$$= -\frac{1}{2} \mathbb{E}_q \left[\frac{1}{\sigma^2} \right] \mathbb{E}_q \left[\frac{1}{1-\rho^2} \right] \left(\boldsymbol{\beta}^T X^T X \boldsymbol{\beta} - 2\boldsymbol{\beta} \left(X^T \mathbf{y} - X^T \mathbf{1}_n \mathbb{E}_q[w] \frac{\mathbb{E}_q[\rho/(1-\rho^2)]}{\mathbb{E}_q[1/(1-\rho^2)]} \right) \right),$$

completando la forma cuadrática en $\boldsymbol{\beta}$,

$$= -\frac{1}{2} \mathbb{E}_q \left[\frac{1}{\sigma^2} \right] \mathbb{E}_q \left[\frac{1}{1-\rho^2} \right] (\boldsymbol{\beta} - \boldsymbol{\mu}_{\beta^*})^T (X^T X) (\boldsymbol{\beta} - \boldsymbol{\mu}_{\beta^*}),$$

es decir que $q^*(\boldsymbol{\beta}) = N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_{\beta^*}, \Sigma_{\beta^*})$, donde

$$\boldsymbol{\mu}_{\beta^*} = (X^T X)^{-1} X^T \left(\mathbf{y} - \mathbf{1}_n \mathbb{E}_q[w] \frac{\mathbb{E}_q[\rho/(1-\rho^2)]}{\mathbb{E}_q[1/(1-\rho^2)]} \right)$$

$$\Sigma_{\beta^*} = \mathbb{E}_q \left[\frac{1}{\sigma^2} \right] \mathbb{E}_q \left[\frac{1}{1-\rho^2} \right] (X^T X)^{-1}.$$

Continuamos determinando la densidad óptima de σ^2 , por tanto escribimos

$$\begin{aligned} \log q^*(\sigma^2) &\propto \mathbb{E}_{-\sigma^2} [\log p(\mathbf{y}, w, \boldsymbol{\beta}, \sigma^2, \rho)] \\ &\propto \mathbb{E}_{q(w, \boldsymbol{\beta}, \rho)} [\log p(\mathbf{y} \mid w, \boldsymbol{\beta}, \sigma^2, \rho) + \log p(\sigma^2)], \end{aligned}$$

¹⁴Ya que, por ejemplo, en general no es posible separar las expresiones de la forma $\mathbb{E}_q[a(\rho)b(\rho)]$ como $\mathbb{E}_q[a(\rho)]\mathbb{E}_q[b(\rho))$. Así mismo, momentos como $\mathbb{E}_q[w a(\rho)]$ se factorizan como $\mathbb{E}_q[w]\mathbb{E}_q[a(\rho)]$ debido a la restricción Campo Medio que se propuso previamente.

ignorando constantes escribimos

$$\begin{aligned} & \propto \mathbb{E}_q \left[-\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \left(\frac{\sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} + \rho w)^2}{1 - \rho^2} \right) - \frac{1}{2} \log(\sigma^2) - \frac{w^2}{2\sigma^2} - \log \sigma^2 \right] \\ & = \mathbb{E}_q \left[-\left(\frac{n+1}{2} - 1 \right) \log(\sigma^2) - \frac{1}{2\sigma^2} \left(\frac{\sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} + \rho w)^2}{1 - \rho^2} + w^2 \right) \right], \end{aligned}$$

tomando la esperanza con respecto a la densidad variacional q

$$= -\left(\frac{n+1}{2} - 1 \right) \log(\sigma^2) - \frac{1}{\sigma^2} \mathbb{E}_q \left[\frac{\sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} + \rho w)^2}{1 - \rho^2} + w^2 \right] / 2,$$

elevando a e ambos lados, podemos identificar el kernel Gamma-Inversa, es decir que

$$\begin{aligned} q^*(\sigma^2) &= IG(\sigma^2 | a_{\sigma^2}, b_{\sigma^2}), \\ a_{\sigma^2} &= (n+1)/2, \\ b_{\sigma^2} &= \frac{1}{2} \mathbb{E}_q \left[\frac{\sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} + \rho w)^2}{1 - \rho^2} + w^2 \right]. \end{aligned}$$

Luego, obtenemos la densidad óptima de w , por tanto escribimos

$$\begin{aligned} \log q^*(w) &\propto \mathbb{E}_{-w} [\log p(\mathbf{y}, w, \rho, \sigma^2, \boldsymbol{\beta})] \\ &= \mathbb{E}_q [\log p(\mathbf{y} | \rho, \sigma^2, w, \boldsymbol{\beta}) + \log p(w | \sigma^2)] \end{aligned}$$

aplicando logaritmo e ignorando constantes

$$\propto \mathbb{E}_q \left[-\frac{1}{2\sigma^2(1 - \rho^2)} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - w\rho)^2 - \frac{1}{2\sigma^2} w^2 \right],$$

desarrollando el cuadrado en w , tomando la suma e ignorando constantes

$$\propto \mathbb{E}_q \left[-\frac{1}{2\sigma^2(1-\rho^2)} \left(nw^2\rho^2 - 2w\rho \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta}) \right) - \frac{w^2}{2\sigma^2} \right],$$

juntando ambas fracciones y colectando términos con w^2

$$= \mathbb{E}_q \left[-\frac{1}{2\sigma^2(1-\rho^2)} \left(((n-1)\rho^2 + 1)w^2 - 2w\rho \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta}) \right) \right],$$

ahora tomamos esperanzas, note que es necesario expandir esta expresión a fin de evaluarlas de forma correcta

$$= -\frac{1}{2} \mathbb{E}_q \left[-\frac{1}{\sigma^2} \left(\frac{(n-1)\rho^2 - 1}{1-\rho^2} w^2 - 2w \frac{\rho \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})}{1-\rho^2} \right) \right],$$

luego, re-escribimos el kernel en w

$$= -\frac{1}{2} \mathbb{E}_q \left[\frac{1}{\sigma^2} \right] \mathbb{E}_q \left[\frac{(n-1)\rho^2 + 1}{1-\rho^2} \right] \left(w^2 - 2w \frac{\mathbb{E}_q \left[\frac{\rho}{1-\rho^2} \right] \sum_{i=1}^n (y_i - \mathbf{x}_i^T \mathbb{E}_q[\boldsymbol{\beta}])}{\mathbb{E}_q[(n-1)\rho^2 + 1]/(1-\rho^2)} \right),$$

es decir que $q^*(w) = N(w \mid \mu_w, \sigma_w^2)$, truncada a la izquierda en cero, donde

$$\begin{aligned} \mu_w &= \frac{\mathbb{E}_q \left[\frac{\rho}{1-\rho^2} \right] \sum_{i=1}^n (y_i - \mathbf{x}_i^T \mathbb{E}_q[\boldsymbol{\beta}])}{\mathbb{E}_q \left[\frac{(n-1)\rho^2 + 1}{1-\rho^2} \right]} \\ \sigma_w^2 &= \left(\mathbb{E}_q \left[\frac{1}{\sigma^2} \right] \left[\frac{(n-1)\rho^2 + 1}{1-\rho^2} \right] \right)^{-1}. \end{aligned}$$

Para finalizar, es necesario determinar la densidad óptima $q^*(\rho)$, por tanto escribimos

$$\begin{aligned} \log q^*(\rho) &\propto \mathbb{E}_{-\rho} [\log p(\mathbf{y}, w, \sigma^2, \boldsymbol{\beta})] \\ &\propto \mathbb{E}_q [\log p(\mathbf{y} \mid \rho, \sigma^2, \boldsymbol{\beta}) + \log p(\rho)] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}_q \left[\log \left(\frac{1}{2\pi\sigma^2(1-\rho^2)} \right)^{n/2} \exp \left(-\frac{1}{2} \frac{\sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - w\rho)^2}{\sigma^2(1-\rho^2)} \right) + \log \frac{\sqrt{1+\rho^2}}{1-\rho^2} \right] \\
&\propto \mathbb{E}_q \left[-\frac{n-2}{2} \log(1-\rho^2) - \frac{\sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - w\rho)^2}{2\sigma^2(1-\rho^2)} + \frac{1}{2} \log(1+\rho^2) \right],
\end{aligned}$$

juntando términos semejantes y tomando esperanzas:

$$= -\frac{n-2}{2} \log(1-\rho^2) - \frac{1}{2} \mathbb{E}_q \left[\frac{1}{\sigma^2} \right] \mathbb{E}_q \left[\frac{\sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - w\rho)^2}{1-\rho^2} \right] + \frac{1}{2} \log(1+\rho^2),$$

note que aquí no es posible ignorar más constantes o simplificar la expresión, por tanto, no es sencillo identificar el kernel de la densidad óptima $q^*(\rho)$ como una densidad ‘común’, en otras palabras, no es viable emplear únicamente el método campo medio para realizar inferencia BV sobre los parámetros $(\rho, \sigma^2, \boldsymbol{\beta})^T$. En la siguiente sección mostraremos como emplear un método híbrido, basado en los resultados que se acaban de obtener, para dar solución al problema de inferencia.

Vale la pena notar que, este caso de estimación no es el mismo que partir de una muestra independiente e idénticamente distribuida $y_{1:n} = (y_1, y_2, \dots, y_n)$ proveniente de $SN(y_i \mid \mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2, \lambda)$: para usar el truncamiento oculto, en este escenario se plantea

$$\begin{bmatrix} V_1 \\ W_1 \end{bmatrix}, \begin{bmatrix} V_2 \\ W_2 \end{bmatrix}, \dots, \begin{bmatrix} V_n \\ W_n \end{bmatrix} \sim N_2 \left(\begin{bmatrix} V_i \\ W_i \end{bmatrix} \mid \begin{bmatrix} \mathbf{x}_i^T \boldsymbol{\beta} \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \quad (2.56)$$

y se define $y_i = V_i \mid (W_i > 0)$. La diferencia es sutil pero relevante: el caso descrito previamente asume que $y_i \not\perp y_j$ de forma marginal, mientras que este plantea $y_i \perp y_j$. Para nuestra exposición, resulta más realista no asumir independencia marginal entre observaciones de un mismo dominio. Por otra parte, en algunos casos se prefiere trabajar

con esta otra versión. Note que la densidad de $p(y_{1:n}, w_{1:n})$ está dada por

$$p(y_{1:n}, w_{1:n}) \equiv p(y_{1:n} | w_{1:n})p(w_{1:n}) = \prod_{i=1}^n N(y_i | \mathbf{x}_{ij}^T \boldsymbol{\beta} + w_i \rho, \sigma^2(1 - \rho_i^2)) N(w_i | 0, \sigma^2) 1(w_i > 0), \quad (2.57)$$

en este caso, en las densidades óptimas $q^*(\boldsymbol{\beta})$ y $q^*(\sigma)$, se debe tomar en cuenta que $p(w_{1:n}) \propto -\sum_{i=1}^n w_i^2 / 2\sigma^2$. Si se considera una densidad óptima para cada entrada del vector $w_{1:n}$, entonces $q^*(w_i) \equiv q^*(w)$ por independencia entre W_i y W_j , por otro lado, la densidad óptima para $w_{1:n}$ de forma conjunta debe reensamblar una densidad normal multivariada donde cada entrada ha sido truncada a \mathbb{R}^+ .

2.3.9 Restricción Forma Fija

En este método de inferencia Bayesiana variaciona, se asume una forma paramétrica fija de la densidad aproximada, es decir, aquí el usuario es quien asigna una densidad de probabilidad para cada parámetro o grupo de parámetros de interés, de tal manera que los soportes coincidan, y a partir de esta elección se busca la configuración de parámetros que, nuevamente, como en el caso de la restricción campo medio, maximice el límite inferior de la evidencia. Por ejemplo, para un parámetro cuyo dominio sean los reales positivos, se podría aproximar la densidad óptima con una distribución gamma, o bien, mapear a los reales y aproximar con una distribución normal. Ahora, la familia de densidades \mathcal{Q} está dada por

$$\mathcal{Q} = \{q(\boldsymbol{\theta}; \boldsymbol{\xi}) : \boldsymbol{\xi} \in \Xi\}, \quad (2.58)$$

esto es, se fija la familia de densidades y únicamente varía $\boldsymbol{\xi}$ sobre el espacio paramétrico Ξ , de este modo, las densidades óptimas q^* están dadas por

$$q^*(\boldsymbol{\theta}) = \arg \max_{\boldsymbol{\xi} \in \Xi} \text{ELBO}(q; \boldsymbol{\xi}). \quad (2.59)$$

Este enfoque se aborda por Tran, T.-N. Nguyen y Dao (2021), Rohde y Wand (2016) y Ormerod y Wand (2010). Rohde y Wand (2016) proponen emplear dos nomenclaturas para

cada tipo de método de inferencia variacional: no paramétrica, que coindice con la restricción campo medio y paramétrica, que coincide con el método forma fija.

No obstante, dado que el objetivo del paradigma variacional es realizar aproximaciones, también es posible construir un método híbrido: debido a que la utilidad de la restricción campo medio depende de la propiedad de conjugación, para aquellas densidades que no tienen este atributo, es posible aproximar su aportación a la distribución *a posteriori* mediante una distribución tratable. En este caso, la familia variacional \mathcal{Q} se define como

$$\mathcal{Q} = \{q(\boldsymbol{\theta}, \boldsymbol{\phi}) : q(\boldsymbol{\theta}, \boldsymbol{\phi}) = q(\boldsymbol{\theta}_1) \dots, q(\boldsymbol{\theta}_p) q(\boldsymbol{\phi}, \xi), \xi \in \Xi\}, \quad (2.60)$$

de acuerdo con Rohde y Wand (2016), $q(\boldsymbol{\phi}, \xi)$ es alguna familia paramétrica preespecificada de densidades en las variables $\boldsymbol{\psi}$ (la parte fija o paramétrica), mientras que no se insiste en alguna forma paramétrica de $q(\boldsymbol{\theta}_i)$ (la parte campo medio o no paramétrica). En este caso, el objetivo variacional del método híbrido de inferencia BV está dado por

$$q^*(\boldsymbol{\theta}) = \arg \max_{q \in \mathcal{Q}, \xi \in \Xi} \text{ELBO}(q; \boldsymbol{\xi}). \quad (2.61)$$

2.3.10 Ejemplo 3: inferencia para la distribución normal asimétrica (revisitado)

Siguiendo el supuesto de campo medio de la inferencia BV, no fue posible identificar la densidad óptima $q^*(\rho)$, por lo tanto, se propone emplear el supuesto de forma fija para aproximar la densidad *a posteriori* de esta cantidad. Aunque es posible usar este método para todos los parámetros del modelo, probablemente no es la estrategia adecuada, ya que se puede explotar el hecho de que las densidades óptimas $q^*(\beta)$, $q^*(w)$ y $q^*(\sigma^2)$ son tratables. Vale la pena resaltar que la asignación de una densidad fija para ρ , no significa asignar una *a priori* para este parámetro, sin embargo, el soporte de la densidad asignada debe estar contenido en el soporte de la *a priori*.

Ahora bien, una elección natural para la densidad óptima $q(\rho)$ es emplear una densidad con soporte en el intervalo $(-1, 1)$, no obstante, parece que no hay muchas densidades

comunes que satisfagan esta condición, por ejemplo, Pérez-Rodríguez, Villaseñor et al. (2017) emplearon una versión análoga al kernel de la siguiente transformación para generar muestras con el algoritmo Metropolis-Hastings, reemplazando ρ por el parámetro de asimetría γ_1 ,

$$q(\rho) = 2\text{Be}(\rho | a, b) - 1 \quad (2.62)$$

No obstante, dado que trabajamos con un problema de optimización, otra alternativa es mapear ρ a la recta real, siendo como candidato la transformación que define la relación entre los parámetros de forma/correlación, es decir $\rho = \frac{\lambda}{\sqrt{1-\lambda^2}}$, además, dado que también transformamos la distribución *a priori*, la densidad resultante, ya ajustada por el Jacobiano de la transformación, está dada por (López 2024)

$$p(\lambda, \sigma) = p(\sigma^2)p(\lambda) = \frac{1}{\sigma^2} \frac{\sqrt{1+2\lambda^2}}{(1+\lambda^2)^2}, \quad (2.63)$$

así, podemos escribir el modelo en términos de λ como sigue

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2, \lambda, w | \mathbf{y}) &\propto p(\mathbf{y} | w, \boldsymbol{\beta}, \sigma^2, \lambda) \pi(\boldsymbol{\theta}) \\ &= \left(\frac{1+\lambda^2}{2\pi\sigma^2} \right)^{n/2} \exp \left(-\frac{1+\lambda^2}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \frac{\lambda}{\sqrt{1+\lambda^2}} w)^2 \right) \\ &\times \frac{2}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{w^2}{2\sigma^2} \right) I_{(0, \infty)}(w) \quad (\text{verosimilitud}) \\ &\times \frac{1}{\sigma^2} \frac{\sqrt{1+2\lambda^2}}{(1+\lambda^2)^2}. \quad (\text{distribución } a \text{ priori}) \end{aligned} \quad (2.64)$$

Dado que la densidad $q(\lambda)$ es fija, los parámetros variacionales ahora consisten de $(\mu_\lambda, \sigma_\lambda^2)^T$, además de los parámetros de las densidades óptimas q^* . Ahora bien, el procedimiento para determinar los valores óptimos desde el punto de vista variacional es análogo al método campo medio: es necesario optimizar el límite inferior de la evidencia con respecto a estos parámetros. Una forma simple de hacer esto es empleando el método de ascenso del gradiente, aunque se podría emplear algún otro metodo de optimización. Ahora bien, el objetivo es

maximizar el ELBO con respecto a $(\mu_\lambda, \sigma_\lambda^2)^T$, simbólicamente escribimos

$$\begin{aligned}
(\mu_\lambda, \sigma_\lambda^2)^\star &\equiv \arg \max_{\mu_\lambda, \sigma_\lambda^2} \text{ELBO}(q) \\
&\equiv \arg \max_{\mu_\lambda, \sigma_\lambda^2} \mathbb{E}_q [\log p(\mathbf{y}, w, \boldsymbol{\beta}, \sigma^2, \lambda) + \log J_{T^{-1}}(\lambda)] \\
&\quad + \mathbb{H}[q(\lambda | \mu_\lambda, \sigma_\lambda^2)] + \mathbb{H}[q^\star(\sigma^2 | a_{\sigma^2}, b_{\sigma^2})] + \mathbb{H}[q^\star(w | \mu_w, \sigma_w^2)] + \mathbb{H}[q^\star(\boldsymbol{\beta} | \boldsymbol{\mu}_\beta, \Sigma_\beta)],
\end{aligned} \tag{2.65}$$

en general, el planteamiento es simple pero la implementación requiere calcular varias cantidades: específicamente, esperanzas condicionales y a partir de estas, (productos de) derivadas. El método de ascenso del gradiente requiere calcular $\frac{\partial}{\partial \mu_\lambda} \text{ELBO}(q)$ y $\frac{\partial}{\partial \sigma_\lambda^2} \text{ELBO}(q)$, sin embargo, es posible facilitar esta tarea usando el *truco de reparametrización* (Jens 2023; Kucukelbir et al. 2017; Tran, T.-N. Nguyen y Dao 2021), es decir, representar a λ como una parte estocástica y otra determinista, específicamente empleamos $\lambda = \mu_\lambda + \sigma_\lambda \tilde{\lambda}$, de este modo, podemos derivar a μ_λ y σ_λ y posteriormente, dada una realización de $\tilde{\lambda} \sim N(0, 1)$, podemos estimar con integración Monte Carlo el vector gradiente. A pesar de ser simple, este principio permite simplificar cálculos. Ahora bien, expandiendo la expresión en la Ecuación 2.65 escribimos¹⁵

$$\begin{aligned}
\text{ELBO}(q) &= \mathbb{E}_q \left[\frac{n}{2} \left(\log \frac{1}{2\pi} + \log \frac{1}{\sigma^2} + \log(1 + \lambda^2) \right) - \frac{1 + \lambda^2}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - w \frac{\lambda}{\sqrt{1 + \lambda^2}})^2 \right. \\
&\quad + \log 2 + \frac{1}{2} \left(\log \frac{1}{2\pi} + \log \frac{1}{\sigma^2} \right) - \frac{w^2}{2\sigma^2} \quad (\text{log-verosimilitud}) \\
&\quad \left. + \log \frac{1}{\sigma^2} + \frac{1}{2} \log(1 + 2\lambda^2) - 2 \log(1 + \lambda^2) \right] \quad (\text{dist. } a \text{ priori}) \\
&\quad + \frac{1}{2} (1 + \log 2\pi\sigma_\lambda^2) \quad (\text{entropía Gaussiana}) \\
&\quad + \frac{1}{2} (a_{\sigma^2} + \ln(b_{\sigma^2}\Gamma(a_{\sigma^2})) - (1 + a_{\sigma^2})\psi(a_{\sigma^2})) \quad (\text{entropía Gamma-Inv.}) \\
&\quad + \frac{p}{2} (1 + \log(2\pi)) + \frac{1}{2} \log \det(\Sigma_\beta) \quad (\text{entropía Gauss. mult.}) \\
&\quad + \frac{1}{2} (1 + \log(2\pi\sigma_w^2)) + \log \Phi \left(\frac{\mu_w}{\sigma_w} \right) - \frac{\phi(\mu_w/\sigma_w)}{2\Phi(\mu_w/\sigma_w)} \quad (\text{entropía Gauss. trunc.})
\end{aligned}$$

¹⁵En general, podemos omitir términos que no dependen de ρ .

Cuando se derivaron las densidades condicionales óptimas $q^*(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \Sigma_\beta)$, $q^*(\sigma^2 \mid a_{\sigma^2}, b_{\sigma^2})$ y $q^*(w \mid \mu_w, \sigma_w^2)$, se observa que, a pesar de que la restricción Campo Medio asume que cada factor variacional se factoriza, los parámetros de estos factores están acoplados entre sí, y esta naturaleza da lugar al método CAVI de [Algoritmo 2](#). A continuación se calculan las derivadas del ELBO (q) con respecto a μ_λ y σ_λ^2 , tratando el resto de las esperanzas -aunque dependientes de estos parámetros- como fijas: esto se justifica debido al truco de reparametrización. Ahora bien, escribimos

$$\begin{aligned}
\frac{\partial}{\partial \mu_\lambda} \text{ELBO}(q) &= \frac{n-4}{2} \frac{\partial}{\partial \mu_\lambda} \log(1 + [\mu_\lambda + \sigma_\lambda \tilde{\lambda}]^2) \\
&\quad - \frac{1}{2\sigma^2} \frac{\partial}{\partial \mu_\lambda} (1 + [\mu_\lambda + \sigma_\lambda \tilde{\lambda}]^2) \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - w \frac{\lambda}{\sqrt{1+\lambda^2}})^2 \\
&\quad + \frac{1}{2} \frac{\partial}{\partial \mu_\lambda} \log(1 + 2[\mu_\lambda + \sigma_\lambda \tilde{\lambda}]^2) \\
&= \frac{n-4}{2} \left(\frac{2\lambda}{1+\lambda^2} \right) \\
&\quad - \frac{1}{\sigma^2} \left(\lambda \sum_{i=1}^n \mu_i^2 - w [(1+\lambda^2)^{1/2} - \lambda^2(1+\lambda^2)^{-1/2}] \sum_{i=1}^n \mu_i \right) \\
&\quad + \frac{1}{2} \left(\frac{4\lambda}{1+2\lambda^2} \right), \tag{2.66}
\end{aligned}$$

donde $\mu_i \triangleq y_i - \mathbf{x}_i^T \boldsymbol{\beta} - w\lambda/\sqrt{1+\lambda^2}$ es el término dentro de estas sumas. Así mismo, podemos calcular las derivadas con respecto a σ_λ^2 , sin embargo, dado que trabajamos con un problema de optimización, resulta más adecuado mapear este parámetro a los reales, una transformación útil es el tomar el logaritmo y por tal motivo, trabajamos con $\log \sigma_\lambda^2$. Ahora, note que podemos calcular el gradiente del ELBO (q) respecto a $\log \sigma_\lambda^2$ sin mayor dificultad aplicando la regla de la cadena al gradiente del ELBO (q) con respecto a μ_λ : es decir, podemos ver¹⁶ el parámetro $\sigma_\lambda \tilde{\lambda}$ como el parámetro μ_λ multiplicado por una realización de la variable aleatoria $\tilde{\lambda}$, luego, notamos que σ_λ puede ser escrito como función de $\log \sigma_\lambda^2$

¹⁶Esto es, $\frac{\partial}{\partial \mu_\lambda} \mu_\lambda \equiv \frac{\partial}{\partial \sigma_\lambda^2} \sigma_\lambda^2$

mediante $\sqrt{\exp(\log \sigma_\lambda^2)}$, procediendo de esta forma, se tiene que

$$\begin{aligned} \frac{\partial}{\partial \log \sigma_\lambda^2} \text{ELBO}(q) &= \frac{\partial}{\partial \mu_\lambda} \text{ELBO}(q) \frac{\partial}{\partial \log \sigma_\lambda^2} \sqrt{\exp \log(\sigma_\lambda^2)} \\ &= \frac{\tilde{\lambda}}{2} \sigma_\lambda \frac{\partial}{\partial \mu_\lambda} \text{ELBO}(q). \end{aligned} \quad (2.67)$$

Una vez calculadas todas las derivadas parciales, sólo resta tomar esperanza sobre los parámetros $(\lambda, \sigma^2, w, \boldsymbol{\beta})$. En general, las expresiones que involucran a λ son intratables, así que como se mencionó previamente, podemos aproximar el gradiente con integración Monte Carlo. El vector gradiente del $\text{ELBO}(q)$, es decir, con respecto a ambos parámetros variacionales, está dado por

$$\nabla_{(\mu_\lambda, \log \sigma_\lambda^2)} \text{ELBO}(q) = \begin{bmatrix} \frac{n-4}{2} \mathbb{E}_q \left[\frac{2\lambda}{1+\lambda^2} \right] - \mathbb{E}_q \left[\frac{1}{\sigma^2} \left(\lambda \sum_{i=1}^n \mu_i^2 - wh(\lambda) \sum_{i=1}^n \mu_i \right) \right] + \frac{1}{2} \mathbb{E}_q \left[\frac{4\lambda}{1+2\lambda^2} \right] \\ \frac{n-4}{2} \mathbb{E}_q \left[\frac{2\lambda}{1+\lambda^2} \frac{\tilde{\lambda}}{2} \sigma_\lambda \right] - \mathbb{E}_q \left[\frac{1}{\sigma^2} \frac{\tilde{\lambda}}{2} \sigma_\lambda \left(\lambda \sum_{i=1}^n \mu_i^2 - wh(\lambda) \sum_{i=1}^n \mu_i \right) \right] + \frac{1}{2} \mathbb{E}_q \left[\frac{4\lambda}{1+2\lambda^2} \frac{\tilde{\lambda}}{2} \sigma_\lambda \right] \end{bmatrix}, \quad (2.68)$$

con $h(\lambda) \triangleq (1 + \lambda^2)^{1/2} - \lambda^2(1 + \lambda^2)^{-1/2}$. Note que, por ejemplo, desarrollar la esperanza del segundo término en ambas entradas del vector, requiere un poco de trabajo adicional:¹⁷

$$\mathbb{E}_q \left[\frac{1}{\sigma^2} \left(\lambda \sum_{i=1}^n \mu_i^2 - wh(\lambda) \sum_{i=1}^n \mu_i \right) \right] = \mathbb{E}_q \left[\frac{1}{\sigma^2} \right] \mathbb{E}_q \left[\lambda \sum_{i=1}^n \mu_i^2 - wh(\lambda) \sum_{i=1}^n \mu_i \right],$$

la esperanza de $1/\sigma^2$ admite una expresión cerrada, por tanto

$$= \frac{b_{\sigma^2}}{a_{\sigma^2} - 1} \left(\mathbb{E}_q \left[\lambda \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - w \frac{\lambda}{\sqrt{1+\lambda^2}})^2 \right] - \mathbb{E}_q \left[wh(\lambda) \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - w \frac{\lambda}{\sqrt{1+\lambda^2}}) \right] \right),$$

luego, desarrollando la expresión y tomando esperanzas se tiene que

$$= \frac{b_{\sigma^2}}{a_{\sigma^2} - 1} \left(\mathbb{E}_q[\lambda](\mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T X \mathbb{E}_q[\boldsymbol{\beta}] + \text{tr}(X^T X \mathbb{E}_q[\boldsymbol{\beta} \boldsymbol{\beta}^T]) + n \mathbb{E}_q[w^2] \mathbb{E}_q \left[\frac{\lambda^3}{1+\lambda^2} \right] \right)$$

¹⁷Esto se debe a que si X es una variable aleatoria, en general no es posible separar $\mathbb{E}[f(X)g(X)]$ como $\mathbb{E}[f(X)] \mathbb{E}[g(X)]$.

$$\begin{aligned}
& - 2 \mathbb{E}_q[w] \mathbb{E}_q \left[\frac{\lambda^2}{\sqrt{1 + \lambda^2}} \right] \sum_{i=1}^n (y_i - \mathbf{x}_i^T \mathbb{E}_q[\boldsymbol{\beta}]) \\
& - \mathbb{E}_q[w] \mathbb{E}_q(h(\lambda)) \sum_{i=1}^n (y_i - \mathbf{x}_i^T \mathbb{E}_q[\boldsymbol{\beta}]) + n \mathbb{E}_q[w^2] \mathbb{E}_q \left[h(\lambda) \frac{\lambda}{\sqrt{1 + \lambda^2}} \right],
\end{aligned}$$

luego, para el gradiente del ELBO (q) con respecto a $\log \sigma^2$ se procede de forma análoga. De este modo, dados valores iniciales para a_{σ^2} , b_{σ^2} , μ_w , σ_w^2 , $\boldsymbol{\mu}_{\beta}$, Σ_{β} y una muestra aleatoria de tamaño M de $\tilde{\lambda} \sim N(0, 1)$, es posible estimar el vector gradiente del ELBO (q) con integración Monte Carlo. Luego, falta establecer el parámetro de tamaño de paso y el criterio de convergencia, para el primer propósito, se puede utilizar la misma estrategia descrita por Kucukelbir et al. (2017, Algoritmo 1), luego, con respecto a la regla de parada, Tran, T.-N. Nguyen y Dao (2021), comenta que teóricamente el ELBO (q) es una función no decreciente, sin embargo, la estimación Monte Carlo podría no serlo, por tanto, una alternativa viable es considerar un tipo de promedio móvil para monitorear los cambios en el ELBO (q), el cuál también es el tipo de criterio usado por el algoritmo `variational` de Stan (Stan Development Team 2025a).

2.3.11 ADVI: Automatic Differentiation Variational Inference

El algoritmo de inferencia variacional con diferenciación automática, *automatic differentiation variational inference*, (ADVI) fue presentado por Kucukelbir et al. (2017) y propone realizar inferencia variacional de manera simple para el usuario: en palabras de los autores, únicamente es necesario especificar un modelo y los datos.

De acuerdo a la discución previa, el algoritmo ADVI se clasifica dentro de la inferencia BV de forma fija, pero simplifica la implementación en varios aspectos

- Transforma cada parámetro en $\boldsymbol{\theta}$ al espacio sin restricciones \mathbb{R}^p mediante $T : \boldsymbol{\theta} \rightarrow \mathbf{z}$.¹⁸
Luego, asigna una distribución Gaussiana a cada parámetro o variable latente en \mathbf{z} .
- Utiliza el truco de reparametrización para mover el gradiente dentro de la esperanza que define al ELBO, y enseguida calcula estas derivadas con diferenciacion automática.

¹⁸ p denota el número de parámetros o variables latentes.

-
- Posterior a esto, evalúa las derivadas obtenidas en el punto anterior mediante integración Monte Carlo.
 - Aplica un esquema de ascenso del gradiente en los parámetros transformados \mathbf{z} y estima al ELBO. Repite este paso hasta convergencia.
 - Finalmente, aplica la transformación inversa $T^{-1}(\mathbf{z})$ para obtener los parámetros variacionales óptimos $\boldsymbol{\theta}^*$, es decir, los que minimizan la divergencia KL, en la escala original.

Profundizando en el primer aspecto del algoritmo ADVI, este considera dos tipos de aproximaciones Gaussianas: una con densidades independientes -y por tanto, univariadas- y otra empleando una densidad multivariada con covarianza densa. Los autores nombran a estas dos aproximaciones como *Mean Field* y *Full Rank*. Vale la pena detenernos un segundo a analizar esto, ya que existe una mezcla entre la terminología empleada por los autores en Tran, T.-N. Nguyen y Dao (2021), Kucukelbir et al. (2017), Blei, Kucukelbir y McAuliffe (2017) y Ormerod y Wand (2010), entre otros. Una resumen sobre estas metodologías se resume en el [Cuadro 2.3](#).

Cuadro 2.3: Comparación de los métodos de inferencia Bayesiana variacional. Fuente: elaboración propia basada en Tran, T.-N. Nguyen y Dao (2021), Kucukelbir et al. (2017), Blei, Kucukelbir y McAuliffe (2017), Ormerod y Wand (2010) y Rohde y Wand (2016).

Método \ Fuente	Tran, T.-N. Nguyen y Dao (2021), Kucukelbir et al. (2017) Rohde y Wand (2016), Blei, Kucukelbir y McAuliffe (2017) y Ormerod y Wand (2010)
	Descripción
<i>Mean Field,</i>	Distribución óptima: para encontrarla se requiere calcular esperanzas de forma analítica y posteriormente identificar la distribución. Hace uso del supuesto
<i>Non parametric</i>	propone como el producto de distribuciones Gaussianas independientes, para ello se transforma cada parámetro/variable latente a \mathbb{R} ; la tarea es identificar los parámetros óptimos. Su nombre es una reminiscencia del supuesto
<i>Mean Field</i>	Uso: con distribuciones <i>a priori</i> conjugadas.
	Uso: para modelos diferenciables con <i>a priori</i> conjugada o no.

Cuadro 2.3: Comparación de los métodos de inferencia Bayesiana variacional. Fuente: elaboración propia basada en Tran, T.-N. Nguyen y Dao (2021), Kucukelbir et al. (2017), Blei, Kucukelbir y McAuliffe (2017), Ormerod y Wand (2010) y Rohde y Wand (2016). (Continuación)

Método \ Fuente	Tran, T.-N. Nguyen y Dao (2021), Kucukelbir et al. (2017) Rohde y Wand (2016), Blei, Kucukelbir y McAuliffe (2017) y Ormerod y Wand (2010)	
	Descripción	Descripción
<i>Full Rank</i>	Equivalente al método <i>Fixed Form/parametrical Mean Field</i> . Distribución óptima: restringe la familia variacional a $q(\boldsymbol{\theta}) = N(\boldsymbol{\theta} \boldsymbol{\mu}, \Sigma)$, la elección $\Sigma = I$ corresponde al método <i>Mean Field</i> , Σ densa corresponde al método <i>Full Rank</i> , ambos definidos por Kucukelbir et al. (2017).	Distribución óptima: se propone una distribución Gaussiana multivariada con matriz de covarianza densa, nuevamente, se transforma cada parámetro/variable latente a \mathbb{R} ; la tarea es obtener los parámetros óptimos con rutinas de optimización. Uso: para modelos diferenciables con <i>a priori</i> conjugada o no.
<i>Fixed Form</i>		
<i>Par. Mean Field</i>	Distribución óptima: el usuario propone o fija las densidades óptimas de tal modo que el soporte de los parámetros/variables latentes coincida: la tarea es identificar los parámetros óptimos. Uso: general, modelos con <i>a priori</i> conjugada o no.	

Cuadro 2.3: Comparación de los métodos de inferencia Bayesiana variacional. Fuente: elaboración propia basada en Tran, T.-N. Nguyen y Dao (2021), Kucukelbir et al. (2017), Blei, Kucukelbir y McAuliffe (2017), Ormerod y Wand (2010) y Rohde y Wand (2016). (Continuación)

Método \ Fuente	Tran, T.-N. Nguyen y Dao (2021), Kucukelbir et al. (2017) Rohde y Wand (2016), Blei, Kucukelbir y McAuliffe (2017) y Ormerod y Wand (2010)
Descripción	Descripción
<i>Semi parametric</i>	
<i>Mean Field</i>	<p>Familia variacional: se emplea la restricción Campo Medio para los parámetros/variables latentes que sean conjugados y se asigna una distribución apropiada -cuyo soporte coincida- para aquellos que no. Así, la familia está dada por</p> $\mathcal{Q} = \{q(\boldsymbol{\theta}_1) \dots q(\boldsymbol{\theta}_p) q(\boldsymbol{\phi}; \boldsymbol{\xi})\}$ <p>Distribución óptima: Para $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p$, es posible obtener de forma analítica las densidades óptimas q^*. Para el resto de densidades $q(\boldsymbol{\phi}; \boldsymbol{\xi})$, el trabajo es encontrar los parámetros $\boldsymbol{\xi}$ que minimicen la divergencia KL mediante optimización.</p> <p>Uso: modelos con parámetros/variables latentes tanto conjugadas y no conjugadas.</p>

Cuadro 2.3: Comparación de los métodos de inferencia Bayesiana variacional. Fuente: elaboración propia basada en Tran, T.-N. Nguyen y Dao (2021), Kucukelbir et al. (2017), Blei, Kucukelbir y McAuliffe (2017), Ormerod y Wand (2010) y Rohde y Wand (2016). (Continuación)

Método \ Fuente	Tran, T.-N. Nguyen y Dao (2021), Kucukelbir et al. (2017)
Descripción	Descripción
	Rohde y Wand (2016), Blei, Kucukelbir y McAuliffe (2017) y Ormerod y Wand (2010)

Actualmente, el lenguaje Stan implementa el método ADVI para realizar inferencia Bayesiana variacional (Stan Development Team 2025a). Además, es posible trazar una analogía entre los algoritmos NUTS y ADVI, a pesar de que resuelven dos problemas distintos, muestreo y aproximación de la distribución *a posteriori*, ambos eliminan la necesidad de que el usuario intervenga o realice afinamientos al mismo. En otras palabras, NUTS es para HMC lo que ADVI es para VB. En los siguientes apartados, se describen de forma breve los elementos clave que dan lugar al algoritmo ADVI.

Aspectos del algoritmo ADVI

Diferenciación automática. De acuerdo con Baydin et al. (2018) y Margossian (2019), los métodos para el cálculo de derivadas en programas computacionales se pueden clasificar en cuatro categorías: (1) cálculo y programación manual, (2) diferenciación numérica usando aproximaciones por diferencias finitas, (3) diferenciación simbólica usando manipulación de expresiones en sistemas de álgebra computacional y (4) diferenciación automática, *automatic differentiation* (AD), también llamada diferenciación algorítmica.

Baydin et al. (2018) señala que habitualmente se confunde la AD con diferenciación numérica o incluso simbólica, ya que esta técnica proporciona valores numéricos de derivadas, y lo hace utilizando reglas simbólicas de diferenciación, lo que le confiere una naturaleza dual (Griewank 2003). Sin embargo, la AD es un método bien definido y distinto de estos.

La AD se refiere a una colección de técnicas que calculan derivadas a través de la acumulación de valores durante la ejecución del código, lo que genera evaluaciones numéricas en lugar de expresiones derivadas. Esto permite una evaluación precisa de las derivadas con *precisión de máquina*, con solo un pequeño factor constante de sobrecarga con eficiencia asintótica ideal. Así mismo, existen dos modos principales de AD, *Forward* o lineal tangente y *Reverse*, lineal cotangente o adjunto. Este último, constituye la generalización del algoritmo *backpropagation*, un método especializado que ha sido pilar para el entrenamiento de redes neuronales (Baydin et al. 2018).

Modelos diferenciables. Kucukelbir et al. (2017) mencionan que el algoritmo ADVI solo permite aproximar la distribución *a posteriori* de modelos probabilísticos diferenciables. Los miembros que pertenecen a esta clase de modelos tienen variables latentes $\boldsymbol{\theta}$ continuas -esto incluye a los parámetros- y además, el gradiente de la log-densidad conjunta $\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}, \boldsymbol{\theta})$ existe. Como una observación adicional, el gradiente debe ser válido dentro del soporte de la distribución *a priori*, el cuál se denota como

$$\text{supp}(p(\boldsymbol{\theta})) = \{\boldsymbol{\theta} \in \mathbb{R}^p : p(\boldsymbol{\theta}) > 0\} \subseteq \mathbb{R}^p, \quad (2.69)$$

aquí p denota el número de variables latentes y parámetros. Los autores también señalan que es posible superar la restricción de tener densidades diferenciables al marginalizar fuera del modelo a las variables aleatorias discretas involucradas, por ejemplo, como sucede con los modelos de mezclas Gaussianas, o propiamente en los modelos que consideramos en secciones posteriores, específicamente, en la distribución *a priori* para selección estocástica de variables descrita en la Sección 3.6.

Transformación de variables latentes. El algoritmo ADVI propone una transformación T uno a uno definida como

$$T : \text{supp}(p(\boldsymbol{\theta})) \rightarrow \mathbb{R}^p, \quad (2.70)$$

es decir, esta función transforma $\boldsymbol{\theta}$ a $T(\boldsymbol{\theta}) \equiv \mathbf{z}$. Así, la densidad conjunta $p(\mathbf{x}, \mathbf{z})$ puede ser escrita empleando el Jacobiano de la transformación T , es decir

$$p(\mathbf{y}, \mathbf{z}) = p(\mathbf{y}, T^{-1}(\mathbf{z})) |\det J_{T^{-1}}(\mathbf{z})|. \quad (2.71)$$

La idea detrás de esta transformación es sencilla: eliminar las restricciones en el soporte de cada variable latente. Ahora que los parámetros viven en el espacio \mathbb{R}^p sin restricciones, ADVI propone dos densidades Gasussinas fijas, una con covarianza diagonal y otra densa. Así, los autores comentan que se induce una factorización no-Gaussiana en el espacio de coordenadas original, esto significa que, por ejemplo, si $\theta \in (0, 1)$, no se asigna de forma arbitraria $q(\theta) = N(\theta | \mu_\theta, \sigma_\theta^2)$, truncada en $(0, 1)$ por el soporte.

Estandarización elíptica. De acuerdo con Kucukelbir et al. (2017), esta transformación consiste en remover la dependencia de las variables latentes transformadas \mathbf{z} con respecto a los parámetros variacionales $\boldsymbol{\xi}$, es decir, la media y covarianzas de la aproximación gaussiana inducida en el espacio \mathbb{R}^p sin restricciones. Para lograr este objetivo, se recurre a una instancia del truco de reparametrización empleado la [Sección 2.3.10](#): se busca expresar \mathbf{z} como una parte determinista y una parte estocástica, lo cuál es resultado de una propiedad de la clase de densidades elípticas. Los siguientes resultados se presentan en Härdle y Simar (2015)

Se dice que un vector aleatorio \mathbf{Y}_p tiene distribución esférica si su función característica $\psi_Y(t)$ satisface $\psi_Y(t) = \phi(t^T t)$ para alguna función escalar ψ , la cuál recibe el nombre de generador característico de la distribución esférica $S_p(\phi)$. Se denota como $\mathbf{Y}_p \sim S_p(\phi)$. Además, es posible ver las distribuciones esféricas como una extensión de la distribución normal multivariada estándar $N_p(\mathbf{0}, I_p)$.

Por otro lado, se dice que un vector aleatorio X_n tiene distribución elíptica con parámetros $\boldsymbol{\mu}_n$ y $\Sigma_{p \times p}$ si X tiene la misma distribución que $\boldsymbol{\mu}_p + A^T \mathbf{Y}_p$, donde $\mathbf{Y}_k \sim S_k(\varphi)$ y A es una matriz de dimensión $(k \times n)$ tal que $A^T A = \Sigma$ con $\text{rango}(\Sigma) = k$. Se denota como $X \sim EC_n(\boldsymbol{\mu}_n, \Sigma, \phi)$. Así mismo, las distribuciones elípticas pueden ser vistas como una

extensión de la densidad normal multivariada $N_p(\boldsymbol{\mu}_p, \Sigma)$.

De este modo, no es difícil ver que la clase de densidades esféricas es un subconjunto de las densidades elípticas, por lo tanto, el objetivo es mapear \mathbf{z} , cuya densidad es elíptica, a $\Sigma^{-1/2}(\mathbf{z} - \boldsymbol{\mu}) \equiv \boldsymbol{\eta}$, cuya densidad es esférica, consiguiendo de este modo obtener una expresión con términos deterministas y estocásticos.

Algoritmo ADVI

En el [Algoritmo 3](#) se presenta el algoritmo ADVI con la variante *Mean Field*, que es el método base para el ajuste de los modelos de regresión en áreas pequeñas que se plantean en el [Capítulo 3](#). Para implementar este método, se requiere calcular el vector gradiente del ELBO con respecto a los parámetros variacionales $(\boldsymbol{\mu}, \boldsymbol{\omega})$, el cual está dado por

$$\begin{aligned}\nabla_{\boldsymbol{\mu}} \mathcal{L} &= \mathbb{E}_{N(\boldsymbol{\eta})} \left[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{y}, \boldsymbol{\theta}) \nabla_{\mathbf{z}} T^{-1}(\mathbf{z}) + \nabla_{\mathbf{z}} \log |\det J_{T^{-1}}(\mathbf{z})| \right] \\ \nabla_{\boldsymbol{\omega}} \mathcal{L} &= \mathbb{E}_{N(\boldsymbol{\eta})} \left[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{y}, \boldsymbol{\theta}) \nabla_{\mathbf{z}} T^{-1}(\mathbf{z}) + \nabla_{\mathbf{z}} \log |\det J_{T^{-1}}(\mathbf{z}) \boldsymbol{\eta}^T \text{diag}(\exp(\boldsymbol{\omega}))| \right] + \mathbf{1},\end{aligned}\quad (2.72)$$

aquí $T : \boldsymbol{\theta} \rightarrow \mathbf{z}$ es el mapeo del vector de parámetros y variables latentes $\boldsymbol{\theta}$ a \mathbb{R}^p , $J_{T^{-1}}$ es el Jacobiano o matriz de derivadas de la transformación inversa que corresponde a este cambio de variables. $\nabla_{\boldsymbol{\theta}}$ y $\nabla_{\mathbf{z}}$ representan el vector gradiente o vector de derivadas parciales con respecto a cada entrada de estos parámetros.

De acuerdo con Kucukelbir et al. (2017), para establecer la taza de aprendizaje en el esquema de ascenso del gradiente, se utiliza la siguiente estrategia: sea $\mathbf{g}^{(i)}$ el vector gradiente en la iteración i , el tamaño de paso $\boldsymbol{\rho}^{(i)}$ (no debe confundirse con los parámetros $\rho_{1:M}$ del modelo propuesto) está dado por la siguiente relación recurrente

$$\begin{aligned}\rho_k^{(i)} &= \eta i^{-1/2+\varepsilon} (\tau + \sqrt{s_k^{(i)}})^{-1}, \\ s_k^{(i)} &= \alpha(g_k^{(i)})^2 + (1 - \alpha) s_k^{(i-1)}, \\ s_k^{(1)} &= g_k^{2(1)}.\end{aligned}\quad (2.73)$$

Algoritmo 3: Inferencia variacional con diferenciación automática, *automatic differentiation variational inference* (ADVI)

Input: Un conjunto de datos $\mathbf{y} = y_{1:n}$, un modelo $p(\mathbf{y}, \boldsymbol{\theta})$.
Output: parámetros variacionales óptimos $\boldsymbol{\mu}^*$, $\boldsymbol{\omega}$ en el espacio sin restricciones \mathbb{R}^p .
Initialize: $i = 1$. $\boldsymbol{\mu}^{(1)} = \mathbf{0}$. $\boldsymbol{\omega}^{(1)} = \mathbf{0}$ (para la aproximación *Mean-Field*)

```
1 .
2 while el cambio en el límite inferior de la evidencia (ELBO) sea menor a cierto
   umbral: do
3   Tomar  $M$  muestras  $\boldsymbol{\eta}_n \sim N(\mathbf{0}, \mathbf{I})$  de la distribución normal multivariada
   estándar.
4   Aproximar  $\nabla_{\boldsymbol{\mu}} \mathcal{L}$  en la Ecuación 2.72 usando integración Monte Carlo.
5   Aproximar  $\nabla_{\boldsymbol{\omega}} \mathcal{L}$  en la Ecuación 2.72 usando integración Monte Carlo (Mean
   Field).
6   Calcular el tamaño de paso  $\boldsymbol{\rho}^{(i)}$  en la Ecuación 2.73.
7   Actualizar  $\boldsymbol{\mu}^{(i+1)} \leftarrow \boldsymbol{\mu}^{(i)} + \text{diag}(\boldsymbol{\rho}^{(i)}) \nabla_{\boldsymbol{\mu}} \mathcal{L}$ .
8   Actualizar  $\boldsymbol{\omega}^{(i+1)} \leftarrow \boldsymbol{\omega}^{(i)} + \text{diag}(\boldsymbol{\rho}^{(i)}) \nabla_{\boldsymbol{\omega}} \mathcal{L}$ .
9   Actualizar  $i \leftarrow i + 1$ .
10 end
11 return  $\boldsymbol{\mu}^* \leftarrow \boldsymbol{\mu}^{(i)}$ ,  $\boldsymbol{\omega}^* \leftarrow \boldsymbol{\omega}^{(i)}$ 
```

CAPÍTULO 3. METODOLOGÍA

La principal contribución metodológica de esta investigación, es la implementación de un método de inferencia Bayesiana variacional para estimar dos clases de modelos de regresión en áreas pequeñas, ambos basados en el uso de la distribución normal sesgada. Además, con el fin de aliviar el costo computacional, se recurre a la representación estocástica generada por el proceso truncamiento oculto, sumado a esto, se centran los parámetros de localidad y escala de esta densidad.

En la primera sección de este capítulo, se describe el escenario particular del modelo de regresión en áreas pequeñas que estudiamos, el cuál recibe el nombre de *modelo a nivel unidad con error anidado*. Luego, se obtiene el modelo de regresión propuesto para respuesta continua, haciendo uso del proceso de truncamiento oculto descrito en la [Sección 2.1.5](#). Posterior a esto, se detalla como obtener los modelos de regresión para respuesta binaria y respuesta ordinal, podemos adelantar que el tratamiento aplicado es similar al caso previo. Después, se presentan las distribuciones *a priori* que se utilizarán con intención de mantener la objetividad en el análisis Bayesiano, para ello se considera emplear la distribución *a priori* de referencia, así como otra distribución *a priori* que permite seleccionar variables de forma estocástica. Enseguida, se comenta como acomplir la verosimilitud y las distribuciones *a priori* para obtener la distribución *a posteriori* proporcional para cada uno de estos modelos, por medio del teorema de Bayes. Finalizamos el capítulo mostrando como implementar los modelos en el lenguaje de programación probabilística Stan y describiendo el conjunto de datos de estudio.

3.1 Descripción de los modelos en áreas pequeñas

La estimación en áreas pequeñas, *small-area estimation* (SAE), aborda el problema de generar pronósticos y estimaciones confiables de parámetros de interés, con sus medidas de incertidumbre asociadas, para subgrupos (áreas, dominios o regiones) de una población

finita de la cual no hay muestras disponibles o no hay tamaños adecuados (J. Rao y Molina 2015). De forma general, el contexto de áreas pequeñas plantea que tenemos M regiones y en cada región existen N_i objetos o entidades de interés, pero solo se dispone de una muestra de tamaño $n_i < N_i$ de estos objetos en cada una de ellas. Usualmente, se realizan estudios en este tipo de planteamientos, por ejemplo al realizar encuestas estatales, como la Encuesta Nacional de Ingresos y Gastos en los Hogares (ENIGH), por tanto, es atractivo construir modelos de regresión para realizar inferencia, particularmente predicción en los objetos o entidades no muestreadas. En la Figura 3.1 se ilustra una idealización sobre como se ve un conjunto de observaciones $\{y_{ij}\}_{j=1,\dots,N_i}^{i=1,\dots,M}$ en el contexto de áreas pequeñas.

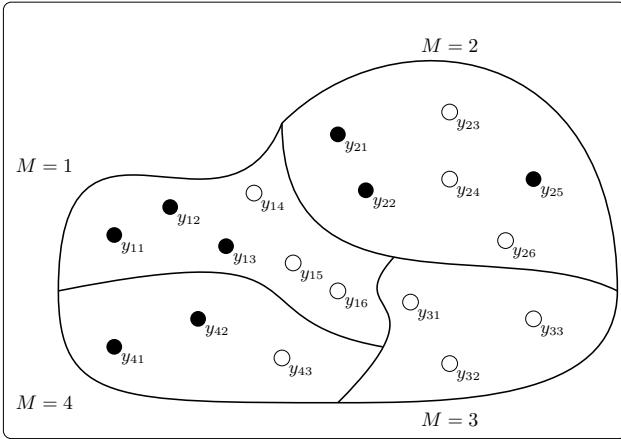


Figura 3.1: Representación de $M = 4$ áreas pequeñas. Las observaciones y_{ij} se indexan de acuerdo a la región y número de observación en esta, los círculos sólidos (vacíos) representan instancias (no) observadas. Fuente: elaboración propia.

De acuerdo con Hedlin (2008), es posible clasificar a los estimadores de un dominio en dos tipos:

1. Estimadores directos: llevan a cabo la inferencia con información exclusiva de cada área pequeña en un periodo de tiempo específico, y casi siempre se construyen a partir de un diseño muestral, por ejemplo, encuestas. López (2024) señala que los estimadores directos se basan en muestreo probabilístico y son caracterizados por las propiedades de insesgamiento y consistencia bajo el diseño muestral en el que se construyen.
2. Estimadores indirectos: utilizan información auxiliar de fuera del dominio o de periodos de tiempo anteriores. Un estimador indirecto toma información o ‘fuerza’ de áreas

vecinas, de tal modo que se pueden hacer inferencias más robustas. En general, los estimadores indirectos emplean modelos estadísticos con el fin de hacer predicciones sobre la(s) variable(s) de estudio.

López (2024) realiza una revisión extensa sobre la estimación en áreas pequeñas, abarcando ambos tipos de estimadores. Ahora bien, vale la pena notar que en la estimación de áreas pequeñas, no es el área la que es ‘pequeña’, sino la muestra tomada del dominio específico: un dominio se considera pequeño si su tamaño de muestra no es lo suficientemente grande como para obtener estimaciones directas con la precisión adecuada. En caso contrario, se dice que el dominio es grande.¹ Por lo tanto, el término área pequeña no se limita a una región en el sentido geográfico, sino que se refiere a cualquier subpoblación para la que no se pueden producir estimaciones directas con una precisión adecuada. Así mismo, Hedlin (2008) también señala que los términos ‘estimador indirecto’ y ‘estimador en áreas pequeñas’ se usan de forma intercambiable.

La aplicación en particular que se propone se ubica dentro del paradigma de estimación indirecta, y se plantean tres modelos de regresión para su estudio: uno para respuesta continua, otro para respuesta binaria y uno más para respuesta ordinal (categorías ordendadas). En el contexto de áreas pequeñas se dispone del conjunto de datos $\{(y_{ij}, \mathbf{x}_{ij})\}_{j=1,\dots,N_i}^{i=1,\dots,M}$, donde \mathbf{x}_{ij}^T es un vector con p covariables asociadas a la respuesta y_{ij} , así, \mathbf{y}_i es el vector de respuestas en la región i , X_i es la matriz de covariables en la región i , \mathbf{y} es el vector de respuestas de las M regiones y X es la matriz de covariables de las M regiones, es decir

$$\begin{aligned}\mathbf{y} &\equiv (\mathbf{y}_1, \dots, \mathbf{y}_M)^T \\ X &\equiv [X_1, \dots, X_M]^T,\end{aligned}\tag{3.1}$$

como se comentó previamente, normalmente sólo se conoce una muestra de tamaño $n_i < N_i$ en cada dominio, mientras que se tiene acceso a toda la matriz de covariables X . En este caso, es posible hacer la distinción entre los elementos observados -o muestreados- y no

¹Hedlin (2008) menciona también que una expresión mejor podría ser ‘estimación de muestra pequeña’.

observados, por tanto, para la región i escribimos:

$$\begin{aligned}\mathbf{y}_i^s &= (y_1, y_2, \dots, y_{n_i})^T \\ \mathbf{y}_i^r &= (y_{n_i+1}, y_{n_i+2}, \dots, y_{N_i})^T\end{aligned}\tag{3.2}$$

y para toda la población:

$$\begin{aligned}\mathbf{y}^s &= (\mathbf{y}_i^s, \mathbf{y}_2^s, \dots, \mathbf{y}_M^s)^T \\ \mathbf{y}^r &= (\mathbf{y}_i^r, \mathbf{y}_2^r, \dots, \mathbf{y}_M^r)^T,\end{aligned}\tag{3.3}$$

los superíndices (r, s) denotan los elementos muestreados y no muestreados, en este orden. Formalmente, para escribir esta representación, se asume que la muestra es intercambiable, sin embargo, en este caso es un supuesto inocente.

Ahora bien, dentro de corriente de SAE, -es decir, la metodología de estimación indirecta o inferencia basada en modelos estadísticos-, J. Rao y Molina (2015) mencionan que es posible clasificar estas técnicas en dos tipos principales:

- Modelos a nivel agregado (o de área), que relacionan los estimadores directos de área pequeña con covariables específicas del área. Estos modelos son necesarios si no se dispone de datos a nivel de unidad (o elemento).
- Modelos a nivel de unidad, los cuales relacionan los valores unitarios de una variable de estudio con covariables específicas de la unidad. Así mismo, un supuesto crítico para esta clase de modelos es que la muestra \mathbf{y}_i^s de cada área obedece al modelo de población asumido, es decir, no existe sesgo de selección. Este punto es particularmente relevante para generar pronósticos acerca de \mathbf{y}^r .

En la siguiente sección, se profundiza acerca del caso de estimación indirecta con el modelo a nivel unidad.

3.1.1 Modelo a nivel unidad con error anidado

Battese, Harter y Fuller (1988) propusieron un modelo a nivel unidad con errores anidados, el cuál identificamos como un modelo de regresión lineal con un efecto aleatorio para cada región de estudio: sea $(y_{ij}, \mathbf{x}_{ij})$ el atributo de interés y la información auxiliar asociada, para las regiones $i = 1, 2, \dots, M$ y observación $j = 1, 2, \dots, N_i$ se asume que

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_i + e_{ij}, \quad (3.4)$$

donde $u_i \sim N(0, \sigma_u^2)$, $e_{ij} \sim N(0, \sigma_e^2)$ y además, u_i es independiente de e_{ij} , para todo par (i, j) .

Además, es posible escribir el modelo en forma matricial como sigue:

$$\begin{bmatrix} \mathbf{y}_i^r \\ \mathbf{y}_i^s \end{bmatrix} = \begin{bmatrix} \mathbf{X}_i^r \\ \mathbf{X}_i^s \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{1}_i^r \\ \mathbf{1}_i^s \end{bmatrix} u_i + \begin{bmatrix} \mathbf{e}_i^r \\ \mathbf{e}_i^s \end{bmatrix}, \quad (3.5)$$

los superíndices (r, s) denotan los elementos no muestreados y muestreados. Adicionalmente, se asume que no existe sesgo, de modo que la inferencia puede realizarse únicamente con los datos observados. Como \mathbf{y}_i^r depende de los efectos aleatorios \mathbf{u}_i^r (siempre no observados), la alternativa frecuentista consiste en obtener la estimaciones de máxima verosimilitud de los parámetros de interés, y la predicción del efecto aleatorio u_i mediante la técnica del mejor predictor lineal insesgado, *best linear unbiased estimator* (BLUP). Concretamente, estos estimadores están dados por (J. Rao y Molina 2015)

$$\begin{aligned} \tilde{\boldsymbol{\beta}}^{\text{BLUE}} &= \left(\sum_{i=1}^M X_i^T V_i^{-1} X_i \right)^{-1} \sum_{i=1}^M X_i^T V_i^{-1} \mathbf{y}_i \\ \hat{u}_i^{\text{BLUE}} &= \sigma_u^2 \mathbf{1}_i^T V_i^{-1} (\mathbf{y}_i - X_i \tilde{\boldsymbol{\beta}}^{\text{BLUE}}), \\ V_i &\equiv \mathbb{V}\text{ar}[\mathbf{y}_i] = \sigma_u^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T + \sigma_e^2 I_{n_i}. \end{aligned} \quad (3.6)$$

Al momento de calcular los estimadores, el modelo toma información de las demás regiones de estudio, de ahí a que se consideren estimadores indirectos. Así mismo, también se puede adoptar un enfoque Bayesiano y tomar muestras de la distribución *a posteriori* para cada uno de estos términos. Ahora bien, siguiendo este sentido, Diallo y J. N. K. Rao (2018)

proponen generalizar este modelo al relajar el supuesto de normalidad: ahora ambos errores pertenecen a la familia normal asimétrica, dotándolo de mayor flexibilidad; específicamente, los autores propusieron

$$\begin{aligned} u_i &\sim SN(-\rho_u \sigma_u \sqrt{2/\pi}, \sigma_u^2, \lambda_u), \\ e_{ij} &\sim SN(-\rho_e \sigma_e \sqrt{2/\pi}, \sigma_e^2, \lambda_e) \end{aligned} \tag{3.7}$$

donde $\rho_u = \lambda_u / \sqrt{1 + \lambda_u^2}$, $\rho_e = \lambda_e / \sqrt{1 + \lambda_e^2}$ y nuevamente u_i es independiente de e_{ij} , para todo par (i, j) . Es posible observar dos cosas:

- Aquí ya han centrado las variables aleatorias a fin de que su valor esperado sea cero, como se comenta en la [Sección 2.1.3](#).
- u_i es un efecto aleatorio dentro de cada área pequeña, sin embargo, los parámetros de forma λ_u y λ_e son los mismos para todas las M regiones.

Luego, con el propósito de obtener la verosimilitud del modelo, los autores recurren a una generalización de la distribución normal sesgada multivariada llamada *Closed Skew Normal* (CSN), la cuál es equivalente a la familia SUN de la [Sección 2.1.2](#) salvo un cambio de parametrización (Arellano-Valle y Azzalini [2022](#); Domínguez-Molina et al. [2007](#)). La familia CSN, como su nombre sugiere, posee las propiedades de cerradura bajo condicionamiento y marginalización, que también posee la distribución normal asimétrica multivariada mostrada previamente, no obstante, esta otra familia de densidades, tiene la propiedad de cerradura bajo conjunción y transformaciones lineales, es decir, las sumas de variables aleatorias CSN se distribuyen CSN. Cabe señalar que esta generalización viene acompañada de mayor complejidad computacional.

El interés principal de esta clase de modelos, es realizar pronósticos y medir la incertidumbre sobre alguna función $h(\mathbf{y}_i^r)$ acerca de los datos faltantes, para ello, la alternativa usual de la estimación en áreas pequeñas, es basar la inferencia a partir de los parámetros estimados con la colección de las variables de estudio observadas y su información auxiliar, esto es

$$\mathbf{y}_i^s = \mathbf{X}_i^s \boldsymbol{\beta} + \mathbf{Z}_i^s \boldsymbol{v} + \boldsymbol{e}_i^s. \tag{3.8}$$

Al inicio de este capítulo, se comentó la propuesta de estimación para tres modelos a nivel unidad, caracterizados por el uso de la distribución normal sesgada y la naturaleza de la variable respuesta: continua, ordinal y binaria. En el [Cuadro 3.1](#) mostramos un resumen de cada uno de estos.

Cuadro 3.1: Resumen sobre los tres modelos de regresión a nivel unidad. k denota el número de categorías ordenadas y δ_c el umbral que separa a las clases $c - 1$ y c . Aquí $\eta_{ij} \triangleq \mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}$.

Fuente: elaboración propia.

Modelo de regresión	Respuesta	Definición
Log-normal sesgado	Continua en \mathbb{R}	$p(y_{ij}) = SN(y_{ij} \eta_{ij}, \sigma^2, \lambda_i)$, $y_{ij} \equiv \log \tilde{y}_{ij}$
Probit sesgado con variable latente	Binaria: $\{0, 1\}$	$\mathbb{P}(y_{ij} = 1) = 1 - \Phi_{SN}(-\eta_{ij}; \lambda_i)$
Probit ordinal sesgado con variable latente	Ordinal: $\{0, 1, \dots, k\}$	$\mathbb{P}(y_{ij} \leq c) = \Phi_{SN}(\delta_{c-1} - \eta_{ij}; \lambda_i) - \Phi_{SN}(\delta_c - \eta_{ij}; \lambda_i)$

En cada uno de estos modelos, no se incluyen explícitamente efectos aleatorios como parte del componente lineal, ya que como se verá en los apartados siguientes, se obtiene una aportación equivalente por medio de los parámetros de forma/correlación ρ_i . De igual manera, en los tres casos propuestos se asume que tanto σ^2 como $\boldsymbol{\beta}$ gobiernan las M regiones, mientras que los parámetros exclusivos de cada área pequeña son ρ_i , así mismo, también puede emplearse un intercepto μ_i asociado a cada área pequeña.

En general, los modelos de regresión que se presentan a continuación, son análogos al procedimiento de López (2024), y estos se deducen a partir de la representación estocástica descrita en la [Sección 2.1.6](#). A continuación, se muestra su derivación. Note que es posible bosquejar similitudes entre este planteamiento y un modelo de regresión con efectos mixtos. Sea

$$\mathbf{V}_i = \mu_i \mathbf{1}_{N_i} + \mathbf{X}_i^T \boldsymbol{\beta} + \rho_i W_i + e_{ij} \sqrt{1 - \rho_i^2}, \quad (3.9)$$

donde $e_i \sim N_{N_i}(\mathbf{0}, \sigma^2 \mathbf{I}_{N_i})$, $W_i \sim N(0, \sigma^2)$ y también $e_{ij} \perp W_i$. Por propiedades de la distribución normal se sigue que $e_{ij} \perp e_{ij'}$. Note que fijamos $\text{Var}[\mathbf{e}_i] = \sigma^2(1 - \rho^2)\mathbf{I}_{N_i}$, aunque la elección de esta covarianza pueda parecer extraña, conduce a que la covarianza de \mathbf{V}_i reproduzca apropiadamente una generalización del proceso de truncamiento oculto a más

de dos dimensiones, como se muestra en la Sección 2.1.5, así mismo, su estructura induce dependencias entre los elementos dentro de un mismo dominio, es decir, entre V_{ij} y $V_{ij'}$. Luego, con el fin de mantener simple la notación, en este momento no insistimos en hacer un tratamiento Bayesiano completo, es decir, omitimos la dependencia de \mathbf{V}_i en $\rho_i \cdot \sigma^2$ y $\boldsymbol{\beta}$. El propósito ahora es describir la distribución de \mathbf{V}_i y de $[\mathbf{V}_i, W_i]^T$, note que podemos escribir \mathbf{V}_i como una combinación lineal de \mathbf{e}_i y W_i como se muestra a continuación:

$$\mathbf{V}_i = \mu_i \mathbf{1}_{N_i} + X_i \boldsymbol{\beta} + \underbrace{\begin{bmatrix} \sqrt{1 - \rho_i^2} I_{N_i} & \rho_i \mathbf{1}_{N_i} \end{bmatrix}}_A \begin{bmatrix} \mathbf{e}_i \\ W_i \end{bmatrix},$$

por propiedades de la densidad normal multivariada, se sigue que

$$\begin{aligned} \mathbf{V}_i &\sim N_{N_i+1} \left(\mu_i \mathbf{1}_{N_i} + X_i \boldsymbol{\beta}, A \begin{bmatrix} \sigma^2 I_{N_i} & 0 \\ 0 & \sigma^2 \end{bmatrix} A^T \right) \\ \Sigma_{V_i} &= \sigma^2 (1 - \rho_i^2) I_{N_i} + \sigma^2 \rho_i^2 \mathbf{1}_{N_i} \mathbf{1}_{N_i}^T \\ &= \sigma^2 \left[(1 - \rho_i^2) I_{N_i} + \rho_i^2 J_{N_i} \right], \end{aligned}$$

a partir del planteamiento, $\text{Cov}[e_{ij}, W_i] = 0$. Ahora, dado que \mathbf{e}_i tiene ley Gaussiana multivariada, nuevamente, por propiedades de esta distribución, la densidad conjunta de $[\mathbf{V}_i, W_i]^T$ tiene ley Gaussiana, la nuestra única tarea restante es determinar $\text{Cov}[\mathbf{V}_i, W_i]$, para ello escribimos

$$\text{Cov}[\mathbf{V}_i, W_i] = \text{Cov}[\sqrt{1 - \rho_i^2} \mathbf{e}_i + \rho_i W_i + X_i \boldsymbol{\beta} + \mu_i \mathbf{1}_{N_i}, W_i],$$

por linealidad de la covarianza podemos separar términos y sacar constantes, por tanto

$$\begin{aligned} &= \sqrt{1 - \rho_i^2} \text{Cov}[\mathbf{e}_i, W_i] + \rho_i \text{Cov}[W_i, W_i] \mathbf{1}_{N_i} + \text{Cov}[X_i \boldsymbol{\beta} + \mu_i \mathbf{1}_{N_i}, W_i] \xrightarrow{0} \\ &= \rho_i \text{Var}[W_i] \mathbf{1}_{N_i} = \rho_i \sigma^2 \mathbf{1}_{N_i}, \end{aligned}$$

de este modo, se obtiene que

$$\begin{bmatrix} \mathbf{V}_i \\ W_i \end{bmatrix} \sim N_{N_i+1} \left(\begin{bmatrix} \mu_i \mathbf{1}_{N_i} + X_i \boldsymbol{\beta} \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{U_i} & \rho_i \sigma^2 \mathbf{1}_{N_i} \\ \rho_i \sigma^2 \mathbf{1}_{N_i}^T & \sigma^2 \end{bmatrix} \right), \quad (3.10)$$

este es el modelo básico a partir del cuál podemos generar variables aleatorias normales asimétricas, es decir, definimos $\mathbf{U}_i = \mathbf{V}_i$ si $W_i > 0$, o de forma equivalente, $U_{ij} = V_{ij}$ si $W_i > 0$, para $j = 1, 2, \dots, N_i$, y por los resultados de la [Sección 2.1.1](#) y [Sección 2.1.2](#), U_{ij} y \mathbf{U}_i tienen ley normal asimétrica univariada y multivariada.

3.2 Modelo log-normal sesgado

Antes de obtener la densidad log-normal asimétrica, recordamos como generar la densidad log-normal usual, esto es: si $X \sim N(\mu, \sigma^2)$, entonces se dice que $Y = \exp(X)$ tiene distribución log-normal con densidad

$$f_Y(y|\mu, \sigma^2) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(\ln y - \mu)^2}{\sigma^2}\right\} I_{(0, \infty)}(y), \quad (3.11)$$

y escribimos $Y \sim \text{LogN}(\mu, \sigma^2)$. Ahora, hacemos que X tenga distribución normal asimétrica y se define $Y = \exp(X)$, entonces, se dice que Y tiene distribución log-normal sesgada con densidad

$$f_Y(y|\mu, \sigma^2, \lambda) = \frac{2}{y\sigma} \phi\left(\frac{\ln y - \mu}{\sigma}\right) \Phi\left(\lambda \frac{\ln y - \mu}{\sigma}\right) I_{(0, \infty)}(y), \quad (3.12)$$

y escribimos $Y \sim \text{LogSN}(\mu, \sigma^2, \lambda)$. Usando la técnica de la función de distribución, no es difícil obtener esta expresión. Note que, por propiedades de estas distribuciones, podemos expresar la variable aleatoria Y en ambos casos como $Y = e^{\mu + \sigma Z}$, donde Z es normal (sesgada) estándar. Vale la pena notar algunos aspectos concretos

- El nombre de estas densidades podría resultar confuso y es conveniente interpretarse como sigue: la variable aleatoria X es normal (sesgada) después de tomar logaritmo a Y .

-
- Similar al caso anterior, en ambas densidades, la interpretación de los parámetros es antes de tomar el logaritmo, es decir, (μ, σ^2) se refieren al parámetro de localidad y escala de X , antes de aplicar esta transformación.

Por otra parte, en un modelo de regresión lineal, usualmente se modela la media o esperanza de la variable respuesta y en el caso del modelo log-normal usual, se tiene que

$$\mathbb{E}[y_i | \mu, \sigma^2] = \mathbb{E}[\exp(x_i) | \mu, \sigma^2] = m_{x_i}(1), \quad (3.13)$$

donde $x_i \sim N(\mu, \sigma^2)$, se identifica esta esperanza como la función generadora de momentos Gaussiana evaluada en $t = 1$, en cuyo caso es igual a $\exp(\mu + \sigma^2/2)$. Es decir, el pronóstico de y_i no sólo consiste en tomar exponencial al pronóstico de x_i , si no que además se multiplica por un factor de $\exp(\sigma^2/2)$.² Aunque es posible emplear una procedimiento similar para el modelo log-normal sesgado, dada la naturaleza del método, los pronósticos de y_i se obtienen mediante integración Monte Carlo.

Algunas aplicaciones que emplean la distribución Log-normal sesgada se muestran en Morán-Vásquez, Giraldo-Melo y Mazo-Lopera (2023), Gómez-Déniz y Calderín-Ojeda (2020) y Martínez-Flórez, Vergara-Cardozo y González (2013).

Ahora bien, el modelo propuesto pertenece a la misma corriente de estimación indirecta a nivel unidad: se supone que, condicional a los parámetros, el logaritmo la variable objetivo sigue una distribución log-normal sesgada, lo cual intenta mimetizar dos aspectos inherentes al conjunto de datos de ingresos: siempre son positivos y su distribución empírica es asimétrica. Para la observación (i, j) se plantea:

$$\tilde{y}_{ij} | \lambda_i, \mu_i, w_i, \sigma^2, \boldsymbol{\beta} \sim \text{LogSN}(\mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}, \sigma^2, \lambda_i), \quad (3.14)$$

es decir que $\tilde{y}_{ij} \equiv \exp(\mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \tilde{e}_{ij})$, donde $\tilde{e}_{ij} \sim \text{LogSN}(0, 1, \lambda_i)$, por tanto podemos definir $y_{ij} \triangleq \log(\tilde{y}_{ij}) = \mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta} + e_{ij}$, donde $e_{ij} \sim SN(0, 1, \lambda_i)$. Esto significa que \tilde{y}_{ij} es la variable aleatoria que modela el fenómeno de interés, cuyo soporte son los reales positivos

²Si únicamente se toma exponencial, se obtiene un pronóstico para la mediana de y_i

(por ejemplo, el conjunto de datos del ingreso corriente total pér capita) y además admite un parámetro de asimetría, mientras que la transformación $y_{ij} \triangleq \log(\tilde{y}_{ij})$ tiene soporte en todos los reales y también admite un parámetro de asimetría: en la práctica, \tilde{y}_{ij} son los valores observados y por simplicidad modelamos $\log(\tilde{y}_{ij})$. Ahora, al centrar los parámetros de localidad y escala, para y_{ij} , el logaritmo de la observación (i, j) , se plantea:

$$y_{ij} \mid \lambda_i, \sigma^2, \mu_i, \boldsymbol{\beta} \sim SN(\mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta} - \sigma \rho_i \sqrt{2/\pi}, \sigma^2/(1 - 2\rho_i^2/\pi), \lambda_i), \quad (3.15)$$

y de forma equivalente:

$$y_{ij} = \mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \sigma \frac{e_{ij} - \mathbb{E}[e_{ij}]}{\sqrt{\text{Var}[e_{ij}]}} , \quad (3.16)$$

donde $e_{ij} \sim SN(0, \sigma^2, \lambda_i)$. También es posible escribir el modelo en forma matricial como sigue

$$\begin{bmatrix} \mathbf{y}_i^r \\ \mathbf{y}_i^s \end{bmatrix} = \begin{bmatrix} \mathbf{1}_i^r \\ \mathbf{1}_i^s \end{bmatrix} \mu_i + \begin{bmatrix} \mathbf{X}_i^r \\ \mathbf{X}_i^s \end{bmatrix} \boldsymbol{\beta} - \frac{\sigma \sqrt{\frac{2}{\pi} \rho_i^2}}{\sqrt{1 - \frac{2}{\pi} \rho_i^2}} \begin{bmatrix} \mathbf{1}_i^r \\ \mathbf{1}_i^s \end{bmatrix} + \frac{1}{\sqrt{1 - \frac{2}{\pi} \rho_i^2}} \begin{bmatrix} \mathbf{e}_i^r \\ \mathbf{e}_i^s \end{bmatrix} , \quad (3.17)$$

como ya se centraron los parámetros de localidad y escala, dada la discusión previa, se garantiza que $\mathbb{E}[y_{ij}] = \mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}$ y $\text{Var}[y_{ij}] = \sigma^2$; de igual modo, los superíndices (r, s) denotan nuevamente a los elementos no muestreados y muestreados de cada región. Note que a diferencia del planteamiento en la [Ecuación 3.7](#), aquí se estima un parámetro de forma para cada área pequeña, de este modo se obtiene un resultado similar a incorporar un efecto aleatorio para cada región, sin embargo, dada la naturaleza del caso de estudio, se prioriza la estimación del parámetro de asimetría en cada dominio; μ_i es el parámetro de intercepto de área pequeña. En general, esta representación es más simple, ya que no incorporamos la estructura de errores anidados (es decir, efectos aleatorios) y no es necesario recurrir a la familia CSN en virtud que solo requiere el proceso de truncamiento oculto de la [Sección 2.1.5](#).³ De forma concreta, se emplea el siguiente mecanismo de selección para la

³Esto debido a que el proceso de truncamiento oculto es menos costoso en términos computacionales que escribir directamente las densidades normales sesgadas.

área i , sea

$$\begin{bmatrix} V_{i1} \\ \vdots \\ V_{iN_i} \\ W_i \end{bmatrix} \sim N_{N_i+1} \left(\begin{bmatrix} \mu_i \mathbf{1}_{N_i} + X_i \boldsymbol{\beta} - \sigma \frac{\sqrt{\frac{2}{\pi}} \rho_i^2}{\sqrt{1-\frac{2}{\pi} \rho_i^2}} \mathbf{1}_{N_i} \\ 0 \end{bmatrix}, \frac{\sigma^2}{1-\frac{2}{\pi} \rho_i^2} \begin{bmatrix} 1 & \cdots & \rho_i^2 & \rho_i \\ \vdots & \ddots & \vdots & \vdots \\ \rho_i^2 & \cdots & \rho_i^2 & \rho_i \\ \rho_i & \cdots & \rho_i & 1 \end{bmatrix} \right), \quad (3.18)$$

y definimos $y_{ij} \triangleq V_{ij}$ o $\mathbf{y}_i \triangleq \mathbf{V}_i$ siempre que $W_i > 0$. En este caso, ya se centraron los parámetros de localidad y escala para aliviar en parte la dificultad sobre la inferencia de este modelo. Es útil recordar que al definir y_{ij} como arriba, entonces por los resultados de la Sección 2.1.5, la distribución marginal de y_{ij} es como se muestra en la Ecuación 3.15. A continuación escribimos la verosimilitud del modelo asociada a los M dominios, haciendo uso de esta representación estocástica, sea $\boldsymbol{\theta} = (\rho_{1:M}, \sigma^2, \mu_{1:M}, \boldsymbol{\beta})$,

$$\begin{aligned} p(\mathbf{y}_1, \dots, \mathbf{y}_M, w_{1:M} \mid \boldsymbol{\theta}) &= \prod_{i=1}^M \left[\prod_{j=1}^{N_i} N(y_{ij} \mid \eta_{ij}, \sigma_i^2) \right] 2N(w_i \mid 0, \sigma_i^{2*}) I_{(0, \infty)}(w_i) \\ \eta_{ij} &\triangleq \mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \rho_i w_i - \sigma \rho_i \sqrt{2/\pi} \\ \sigma_i^2 &\triangleq \sigma^2 (1 - \rho_i^2) / (1 - 2\rho_i^2/\pi)) \\ \sigma_i^{2*} &\triangleq \sigma^2 / (1 - 2\rho_i^2/\pi)). \end{aligned} \quad (3.19)$$

Note que $w_{1:M}$ son variables no observadas, generadas durante el proceso de truncamiento oculto. De forma técnica, no es preciso denominar ‘verosimilitud’ a densidad en la Ecuación 3.19, ya que no observamos a $w_{1:M}$.

3.3 Modelo probit sesgado latente

Esta sección inicia mostrando el planteamiento general del modelo probit a partir de la técnica de variable latente. Luego, al igual que en la sección previa, se escribe la verosimilitud del modelo. Podemos mencionar que el procedimiento para obtener este modelo es bastante similar a obtener el modelo log-normal sesgado, salvo dos diferencias principales:

1. Se fija $\sigma^2 \equiv 1$, de tal modo que se obtiene cierta simplificación en las expresiones

involucradas en la verosimilitud y distribución *a priori*, como se verá más adelante.

2. Se emplea una variable latente que desempeña el rol de ‘intermediario’ entre el proceso truncamiento oculto y la respuesta binaria observada.

Sea $e_{ij} \sim SN(0, 1, \lambda_i)$, de modo que $Z_{ij} \triangleq \mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta} + e_{ij} \sim SN(\mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}, 1, \lambda_i)$, en este momento no se insiste en centrar los parámetros de localidad y escala. Definimos la variable aleatoria Y_{ij} como

$$Y_{ij} = \begin{cases} 1 & \text{si } Z_{ij} > \delta_0 \\ 0 & \text{si } Z_{ij} < \delta_0 \end{cases}, \quad (3.20)$$

con $\delta_0 \in \mathbb{R}$, entonces $Y_{ij} \mid p_{ij} \sim \text{Bern}(p_{ij})$, donde

$$\begin{aligned} p_{ij} &\triangleq \mathbb{P}(Z_{ij} > \delta_0) \\ &= \mathbb{P}(\mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta} + e_{ij} > \delta_0) \\ &= 1 - \mathbb{P}(e_{ij} < \delta_0 - \mu_i - \mathbf{x}_{ij}^T \boldsymbol{\beta}) \\ &= 1 - \Phi_{SN}(\delta_0 - \mu_i - \mathbf{x}_{ij}^T \boldsymbol{\beta}; \lambda_i), \end{aligned} \quad (3.21)$$

de esta forma, se define el modelo probit cuya función liga es la distribución normal sesgada, δ_0 es el umbral que permite obtener $y_{ij} \in \{0, 1\}$. Se observa que, si fijamos $\delta_0 = 0$ y $\lambda_i = 0$, obtenemos el modelo probit usual, ya que

$$p_{ij} = 1 - \Phi_{SN}(-\mu_i - \mathbf{x}_{ij}^T \boldsymbol{\beta}; 0) = \Phi(\mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}),$$

por simetría de Φ . Así mismo, con respecto a la identificación del modelo, es posible notar dos aspectos:

1. El problema de dejar que δ_0 tome cualquier valor en \mathbb{R} , es que su efecto se confunde con el intercepto μ_i , esto es:

$$1 - \Phi_{SN}(\delta_0 - \mu_i - \mathbf{x}_{ij}^T \boldsymbol{\beta}; \lambda_i) = 1 - \Phi_{SN}((\delta_0 + c) - (\mu_i + c) - \mathbf{x}_{ij}^T \boldsymbol{\beta}; \lambda_i), \quad (3.22)$$

es decir, para cualquier real c , las cantidades $(\mu_i, \boldsymbol{\beta}, \delta_0)$ y $(\mu_i + c, \boldsymbol{\beta}, \delta_0 + c)$ son indistinguibles. Por tanto, a partir de ahora, fijamos $\delta_0 = 0$ por identificación de estos parámetros.

2. En el modelo probit usual, se fija $\sigma^2 \equiv 1$, ya que de otra forma se estimaría $(\mu_i/\sigma, \boldsymbol{\beta}/\sigma)$ en lugar de $(\mu_i, \boldsymbol{\beta})$:

$$\Phi\left(\frac{1}{\sigma}(\mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta})\right) = \Phi\left(\frac{\mu_i}{\sigma} + \mathbf{x}_{ij}^T \frac{\boldsymbol{\beta}}{\sigma}\right). \quad (3.23)$$

Similar al caso previo, si consideramos centrar el parámetro de escala en la distribución normal sesgada -esto es, $e_{ij} \sim (0, 1/\sqrt{\text{Var}[e_{ij}]}, \lambda_i)$ -, no es suficiente especificar $\sigma^2 \equiv 1$, ya que aún así se estimaría $(\sqrt{\text{Var}[e_{ij}]}\mu_i, \sqrt{\text{Var}[e_{ij}]}\boldsymbol{\beta})$ en lugar de $(\mu_i, \boldsymbol{\beta})$:

$$1 - \Phi_{SN}(\text{Var}[e_{ij}](-\mu_i - \mathbf{x}_{ij}^T \boldsymbol{\beta})) = 1 - \Phi_{SN}(-\sqrt{\text{Var}[e_{ij}]}\mu_i - \mathbf{x}_{ij}^T \sqrt{\text{Var}[e_{ij}]}\boldsymbol{\beta}), \quad (3.24)$$

donde $\text{Var}[e_{ij}] = 1 - 2\rho_i^2/\pi$. En el [Anexo 1](#) se cubre brevemente el concepto de identificabilidad en un modelo estadístico.

A continuación, se presenta el enfoque de variable latente, el cuál sortea los posibles problemas de identificación del modelo al emplear una estrategia similar a la técnica de aumentación de datos, así mismo, el método es análogo al desarrollado por Albert y Chib ([1993](#)). Sea $\boldsymbol{\theta} = (\rho_i, \sigma^2, \mu_i, \boldsymbol{\beta})$; ya se observó que $Y_{ij} | p_{ij} \sim \text{Bern}(p_{ij})$ y además $p(z_{ij} | \boldsymbol{\theta}) = SN(z_{ij} | \mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}, 1, \lambda_i)$, obtengamos la distribución conjunta $p(y_{ij}, z_{ij} | \boldsymbol{\theta})$, para ello, escribimos $p(y_{ij}, z_{ij} | \boldsymbol{\theta}) = p(y_{ij} | z_{ij}, \boldsymbol{\theta})p(z_{ij} | \boldsymbol{\theta})$, por el planteamiento, ya conocemos a esta última densidad y sólo hace falta determinar $p(y_{ij} | z_{ij}, \boldsymbol{\theta})$, para tal propósito, note que $Y_{ij} \in \{0, 1\}$, entonces, para $Y_{ij} = 1$:

$$\begin{aligned} \mathbb{P}(Y_{ij} = 1 | Z_{ij}, \boldsymbol{\theta}) &= P(Y_{ij} = 1 | Z_{ij}) \\ &= \mathbb{P}(Y_{ij} = 1 | Z_{ij} > 0) + \overbrace{P(Y_{ij} = 1 | Z_{ij} < 0)}^0 \\ &= \mathbb{P}(Y_{ij} = 1 | Z_{ij} > 0) = 1, \end{aligned} \quad (3.25)$$

empleando un argumento análogo obtenemos $P(Y_{ij} = 0 \mid Z_{ij}, \boldsymbol{\theta}) = 1$. Con estas dos expresiones, podemos notar que $Y_{ij} \mid Z_{ij}, \boldsymbol{\theta}$ es una variable aleatoria degenerada, ya que está completamente determinada por el signo de Z_{ij} , así, esta función de densidad condicional está dada por

$$p(y_{ij} \mid z_{ij}, \boldsymbol{\theta}) = 1(y_{ij} = 1)1(z_{ij} > 0) + 1(y_{ij} = 0)1(z_{ij} < 0), \quad (3.26)$$

desde un punto de vista numérico es evidente que esta expresión siempre evalúa a uno para todos todos los valores de su soporte. Ahora bien, juntando estas dos densidades, podemos expresar la densidad conjunta $p(y_{ij}, z_{ij} \mid \boldsymbol{\theta})$ como

$$\begin{aligned} p(y_{ij}, z_{ij} \mid \boldsymbol{\theta}) &= p(y_{ij} \mid z_{ij}, \boldsymbol{\theta}) p(z_{ij} \mid \boldsymbol{\theta}) \\ &= [1(y_{ij} = 1)1(z_{ij} > 0) + 1(y_{ij} = 0)1(z_{ij} < 0)] SN(z_{ij} \mid \mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}, 1, \lambda_i), \end{aligned} \quad (3.27)$$

este enfoque nos permite realizar inferencia en la variable latente z_{ij} como si se tratara de un modelo de regresión lineal usual (Albert y Chib 1993). Una vez planteado el modelo, podemos usar el mismo proceso de truncamiento oculto de la Ecuación 3.18 -fijando $\sigma^2 \equiv 1$ - para expresar la densidad normal asimétrica $p(z_{ij} \mid \boldsymbol{\theta})$ en la Ecuación 3.27. De forma análoga, si definimos $Z_{ij} = V_{ij}$ siempre que $W_i > 0$, para la observación (i, j) se tiene que

$$\begin{aligned} p(y_{ij}, z_{ij}, w_i \mid \boldsymbol{\theta}) &\equiv p(y_{ij} \mid z_{ij}, w_i, \boldsymbol{\theta}) p(z_{ij} \mid w_i, \boldsymbol{\theta}) p(w_i \mid \boldsymbol{\theta}) \\ &= p(y_{ij} \mid z_{ij}, \boldsymbol{\theta}) p(z_{ij} \mid \boldsymbol{\theta}) p(w_i, \boldsymbol{\theta}) \\ &= 1(y_{ij}, z_{ij}) N(z_{ij} \mid \eta_{ij}, \sigma_i^2) 2N(w_i \mid 0, \sigma_i^{2\star}) I(w_i > 0), \end{aligned}$$

con $1(y_{ij}, z_{ij}) \triangleq [1(y_{ij} = 1)1(z_{ij} > 0) + 1(y_{ij} = 0)1(z_{ij} < 0)]$, de manera adicional, η_{ij} , σ_i^2 y $\sigma_i^{2\star}$ se definen en la Ecuación 3.19, salvo que $\sigma^2 \equiv 1$. Notamos que Y_{ij} depende sólo del signo de Z_{ij} -esta es la parte latente-, mientras que Z_{ij} se genera a partir de truncamiento oculto con W_i . Procediendo de forma análoga a la sección previa, ahora escribimos la verosimilitud

del modelo asociada a los M dominios y a los N_i objetos en cada uno de ellos,

$$\begin{aligned}
& p(\mathbf{y}_1, \dots, \mathbf{y}_M, \mathbf{z}_1, \dots, \mathbf{z}_M, w_{1:M} \mid \boldsymbol{\theta}) \\
& \equiv \prod_{i=1}^M \left[\prod_{j=1}^{N_i} p(y_{ij} \mid z_{ij}, \boldsymbol{\theta}) p(z_{ij} \mid \boldsymbol{\theta}) \right] p(w_i \mid \boldsymbol{\theta}) \\
& = \prod_{i=1}^M \left[\prod_{j=1}^{N_i} 1(y_{ij}, z_{ij}) N(z_{ij} \mid \eta_{ij}, \sigma_i^2) \right] 2N(w_i \mid 0, \sigma^{2\star}) 1_{(0, \infty)}(w_i).
\end{aligned} \tag{3.28}$$

3.4 Modelo probit ordenado sesgado latente

En este apartado consideramos la extensión del modelo probit sesgado latente cuando tenemos más de dos categorías y es posible ordenarlas, en otras palabras, la respuesta es ordinal. La deducción del modelo es análoga al caso del modelo binario, pero con algunas diferencias: quizás las características más relevantes son el uso de la distribución categórica en la respuesta -que generaliza la distribución Bernoulli- y la restricciones que se impone en los puntos de corte o umbrales, además de que fijamos el primero de ellos para poder identificar, y por lo tanto, estimar el resto de estos.

Nuevamente, sea $e_{ij} \sim SN(0, 1, \lambda_i)$, de modo que $Z_{ij} \triangleq \mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta} + e_{ij} \sim SN(\mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}, 1, \lambda_i)$.

Definimos la variable aleatoria Y_{ij} como

$$Y_{ij} = \begin{cases} 0 & \text{si } -\infty < Z_i \leq \delta_0 \\ 1 & \text{si } \delta_0 < Z_i \leq \delta_1 \\ \dots & \\ k-1 & \text{si } \delta_{k-1} < Z_i \leq \delta_k \\ k & \text{si } \delta_k < Z_i < \infty \end{cases},$$

entonces $Y_i \mid p_{0:k} \sim \text{Cat}(p_0, \dots, p_k)$, esto es, multinomial con $n = 1$. Ahora, definimos

$\delta_{-1} = -\infty$ y $\delta_{k+1} = \infty$, entonces

$$\begin{aligned}
p_{ij} &= \mathbb{P}(\delta_{i-1} < Z_{ij} \leq \delta_i) \\
&= \mathbb{P}(\delta_{i-1} < \mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta} + e_{ij} \leq \delta_i) \\
&= \mathbb{P}(\delta_{i-1} - \mu_i - \mathbf{x}_{ij}^T \boldsymbol{\beta} < e_{ij} \leq \delta_i - \mu_i - \mathbf{x}_{ij}^T \boldsymbol{\beta}) \\
&= \Phi_{SN}(\delta_{i-1} - \mu_i - \mathbf{x}_{ij}^T \boldsymbol{\beta}) - \Phi_{SN}(\delta_i - \mu_i - \mathbf{x}_{ij}^T \boldsymbol{\beta}),
\end{aligned}$$

de forma análoga a la sección anterior, así definimos el modelo probit ordenado sesgado, los δ_i son los umbrales que permiten obtener $y_{ij} \in \{0, 1, \dots, k\}$. Note que para obtener k categorías, es necesario definir $k - 1$ umbrales δ_i , los cuales están ordenados, es decir

$$-\infty \equiv \delta_{-1} < \delta_0 < \delta_1 < \dots < \delta_k < \delta_{k+1} \equiv \infty,$$

sin embargo, para evitar el problema de identificación en los puntos de corte δ_i , es necesario considerar algún tipo de restricción, por ejemplo, fijar uno de estos; la alternativa usual es asignar δ_0 a cero, lo cuál es análogo a la restricción en la sección previa. Así mismo, como se comentó en el apartado previo, también establecemos $\sigma^2 \equiv 1$.

Luego, se procede de forma análoga al modelo probit, el siguiente paso es emplear el método de variable latente: ya observamos que $Y_{ij} \mid p_{0:k} \sim \text{Cat}(p_0, p_1, \dots, p_k)$ y $Z_{ij} \sim SN(\mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}, 1, \lambda_i)$, además, sea $\boldsymbol{\theta} = (\rho_i, \delta_{1:k}, \mu_i, \boldsymbol{\beta})$, note que se desea hacer inferencia sobre los puntos de corte δ_1 al δ_k . Para continuar el desarrollo, se obtiene la distribución conjunta $p(y_{ij}, z_{ij} \mid \boldsymbol{\theta})$, para ello, escribimos $p(y_{ij}, z_{ij} \mid \boldsymbol{\theta}) = p(y_{ij} \mid z_{ij}, \boldsymbol{\theta}) p(z_{ij} \mid \boldsymbol{\theta})$, por el planteamiento, ya conocemos a esta última densidad y sólo hace falta determinar $p(y_{ij} \mid z_{ij}, \boldsymbol{\theta})$, y para tal propósito, note que $Y_{ij} \in \{0, 1, \dots, k\}$, entonces, para $Y_{ij} = c$:

$$\begin{aligned}
\mathbb{P}(Y_{ij} = c \mid Z_{ij}, \boldsymbol{\theta}) &= P(Y_{ij} = c \mid Z_{ij}) \\
&= P(Y_{ij} = c \mid Z_{ij} \in (\delta_{c-1}, \delta_c]) + P(Y_{ij} = c \mid Z_{ij} \notin (\delta_{c-1}, \delta_c]) \xrightarrow{0} \\
&= P(Y_{ij} = c \mid Z_{ij} \in (\delta_{c-1}, \delta_c)) = 1,
\end{aligned}$$

dado que $c \in \{0, 1, \dots, k\}$, podemos notar que $Y_{ij} \mid Z_{ij}, \boldsymbol{\theta}$ es una variable aleatoria degenerada, ya que está completamente determinada por el intervalo al que pertenece Z_{ij} , así, su función de densidad está dada por⁴

$$p(y_{ij} \mid z_{ij}, \boldsymbol{\theta}) = \sum_{c=0}^k 1(y_{ij} = c) 1(z_{ij} \in (\delta_{c-1}, \delta_c])$$

desde un punto de vista numérico, es evidente que esta densidad siempre evalúa a uno. Ahora bien, juntando estas dos densidades, podemos expresar la densidad conjunta $p(y_{ij}, z_{ij} \mid \boldsymbol{\theta})$ como

$$p(y_{ij}, z_{ij} \mid \boldsymbol{\theta}) = \left[\sum_{c=0}^k 1(y_{ij} = c) 1(z_{ij} \in (\delta_{c-1}, \delta_c]) \right] SN(z_{ij} \mid \mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}, 1, \lambda_i). \quad (3.29)$$

Empleando un argumento similar a las dos secciones previas, empleamos el mismo proceso el truncamiento oculto de la [Ecuación 3.18](#) para expresar la densidad normal asimétrica $p(z_{ij} \mid \boldsymbol{\theta})$ en la [Ecuación 3.29](#). De forma análoga, si definimos definimos $Z_{ij} = V_{ij}$ siempre que $W_i > 0$, para la observación (i, j) se tiene que

$$\begin{aligned} p(y_{ij}, z_{ij}, w_i, \boldsymbol{\theta}) &\equiv p(y_{ij} \mid z_{ij}, w_i, \boldsymbol{\theta}) p(z_{ij} \mid w_i, \boldsymbol{\theta}) p(w_i \mid \boldsymbol{\theta}), \\ &= p(y_{ij} \mid z_{ij}, \boldsymbol{\theta}) p(z_{ij} \mid \boldsymbol{\theta}) p(w_i, \boldsymbol{\theta}) \\ &= 1(y_{ij}, z_{ij}, \delta_{0:k}) \times \\ &\quad N(z_{ij} \mid \eta_{ij}, \sigma_i^2) N(w_i \mid 0, \sigma_i^{2*}) I(w_i > 0), \end{aligned} \quad (3.30)$$

con $1(y_{ij}, z_{ij}, \delta_{0:k}) \triangleq \sum_{c=1}^k 1(y_{ij} = c) 1(z_{ij} \in (\delta_{c-1}, \delta_c])$, de manera adicional, η_{ij} , σ_i^2 y σ_i^{2*} se definen en la [Ecuación 3.19](#), salvo que $\sigma^2 \equiv 1$. De forma análoga, notamos que Y_{ij} depende sólo del intervalo al que Z_{ij} pertenece -esta es la parte latente-, mientras que Z_{ij} se genera a partir de truncamiento oculto con W_i . Finalmente, escribimos la verosimilitud del modelo

⁴En el último umbral se abusa de la notación ya que $z_{ij} \in (\delta_{k-1}, \infty]!$.

asociada a los M dominios y a los N_i objetos en cada uno de ellos,

$$\begin{aligned}
 & p(\mathbf{y}_1, \dots, \mathbf{y}_M, \mathbf{z}_1, \dots, \mathbf{z}_M, w_{1:M} \mid \boldsymbol{\theta}) \\
 & \equiv \prod_{i=1}^M \left[\prod_{j=1}^{N_i} p(y_{ij} \mid z_{ij}, \boldsymbol{\theta}) p(z_{ij} \mid \boldsymbol{\theta}) \right] p(w_i \mid \boldsymbol{\theta}) \\
 & = \prod_{i=1}^M \left[\prod_{j=1}^{N_i} 1(y_{ij}, z_{ij}, \delta_{0:k}) N(z_{ij} \mid \eta_{ij}, \sigma_i^2) \right] 2N(w_i \mid 0, \sigma^{2\star}) 1_{(0, \infty)}(w_i),
 \end{aligned} \tag{3.31}$$

3.5 Distribución *a priori* de referencia

En la [Sección 2.2](#) se inició la discusión introduciendo el paradigma Bayesiano de la estadística y el matiz derivado de los enfoques subjetivo y objetivo, ambos determinados a través de la distribución *a priori*. Algunas ventajas del paradigma Bayesiano es que permite generar resultados más ricos en contenido que el tratamiento frecuentista o clásico, ya que entre otras cosas, es posible estimar fácilmente funciones de interés a partir de la muestra *a posteriori*, de igual modo, los intervalos de credibilidad miden la certeza con que cierto parámetro esté contenido en algún subconjunto de \mathbb{R} . Sin embargo, usualmente el costo de oportunidad es la complejidad computacional, en contraste con los métodos frecuentistas que son más simples en su aplicación, y la necesidad de especificar distribuciones *a priori*.⁵

Dentro del contexto de modelos en áreas pequeñas, se busca hacer más robusta la inferencia sobre estimaciones y pronósticos al tomar información prestada de otros dominios, en este sentido, el enfoque Bayesiano subjetivo brinda una método formal para inyectar en el modelo información previa o relevante, especialmente cuando los datos disponibles son escasos, y de este modo, contribuir al estudio. Sin embargo, también es posible que la influencia de la distribución *a priori* domine a la verosimilitud, es decir, a la muestra observada del modelo paramétrico asumido, provocando que no sea posible aprender acerca de los datos. En los tres modelos propuestos, se plantea seguir un enfoque objetivo para generar resultados reproducibles e imparciales, además de que no se cuenta con el conocimiento experto para

⁵No obstante, no siempre es el caso, por ejemplo, con modelos conjugados o simples en los que la inferencia se realiza de forma cerrada.

introducir información previa.

A pesar de que la teoría señala que no existe densidad *a priori* que no aporte ninguna información al modelo, Kass y Wasserman (1996) señalan que la elección de esta a partir de reglas formales o principios que reflejan nulo estado de conocimiento (como el de Laplace o razón insuficiente) y no en información previa o creencias, es quién las considera objetivas. De igual modo, estos autores describen varios métodos para obtener distribuciones *a priori* objetivas, algunos de estos son:

- El método de Jeffrey, el cuál propone usar:

$$p(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))},$$
$$I(\boldsymbol{\theta}) = \mathbb{E} \left\{ \frac{\partial \theta_i}{\partial \theta_j \partial \theta_i} \log p(\boldsymbol{\theta} | \mathbf{y}) \right\}, \quad (3.32)$$

donde $I(\boldsymbol{\theta})$ es la matriz de información esperada o de Fisher. Cuando $\boldsymbol{\theta}$ es multivariado, se hace un ajuste a este método.

- El método de Laplace o principio de razón insuficiente:

$$p(\boldsymbol{\theta}) \propto 1, \quad (3.33)$$

es decir, se asume que cada entrada θ_i puede asumir, con igual probabilidad, cualquiera de los valores dentro de su soporte.

- El método de Bernardo, cuya distribución se define como la *a priori* mínimamente informativa, en el sentido que maximiza la información faltante al calcular la divergencia Kullback-Leibler entre los datos observados y el modelo paramétrico asumido.

Continuamos este apartado explorando esta última alternativa, la cuál llamaremos distribución *a priori* de referencia⁶. De acuerdo con Kass y Wasserman (1996), el método de Bernardo presenta dos innovaciones: (1) define el concepto de información faltante y (2) desarrolla un procedimiento secuencial para abordar los parámetros de ruido. De forma

⁶Como se mencionó en la Sección 2.2, el adjetivo ‘de referencia’ puede emplearse como sinónimo de objetivo o no informativo, pero aquí lo reservamos para nombrar a este método

adicional, cuando no hay tales parámetros de ruido y se satisfacen ciertas condiciones de regularidad, la *a priori* de referencia coincide con la *a priori* de Jeffrey. No obstante, en general se generan densidades distintas a las de Jeffrey.

A grandes rasgos, la distribución de referencia considera el escenario donde se desea realizar inferencia sobre el vector de parámetros $\boldsymbol{\theta} = (\psi, \omega)$, donde ψ puede ser considerado como los parámetros de interés y ω como parámetros de ruido, es decir, no constituyen el interés principal. Por lo tanto, la *a priori* resultante no solo depende del modelo muestral -es decir la verosimilitud del modelo-, si no que también del problema en cuestión (Robert 2007).

En términos simples, la *a priori* de referencia $\pi(\boldsymbol{\theta})$ será aquella que maximice la información sobre $\boldsymbol{\theta}$ con respecto a la densidad *a posteriori*, la cuál se puede calcular a través de la divergencia Kullback-Leibler. En general, para obtener la distribución *a priori* de referencia involucra trabajo analítico, además de que la notación que emplea es bastante compleja (Berger y Bernardo 1992).

Ahora, el vector de parámetros de interés en el modelo log-normal sesgado está dado por

$$\boldsymbol{\theta} = (\rho_1, \dots, \rho_M, \sigma^2, \mu_1, \dots, \mu_M, \boldsymbol{\beta}), \quad (3.34)$$

y es de dimensión $2M + p + 1$. Dado que la novedad principal del modelo planteado radica en la estimación de los parámetros de forma/correlación ρ_i de cada región, podemos ordenar los elementos del vector $\boldsymbol{\theta}$ de acuerdo a su relevancia de estudio, por ejemplo

$$\rho_1 \succeq \rho_2 \succeq \dots \succeq \rho_M \succ \sigma^2 \succ \mu_1 \succeq \mu_2 \succeq \dots \succeq \mu_M \succeq \boldsymbol{\beta}, \quad (3.35)$$

así, esto motiva aplicar el algoritmo de referencia para un grupo de parámetros ordenados (Berger y Bernardo 1992). López (2024) obtuvo la distribución *a priori* de referencia para este modelo, y encontró que es proporcional a

$$p(\rho_1, \dots, \rho_M, \sigma^2, \mu_1, \dots, \mu_M, \boldsymbol{\beta}) \propto \frac{1}{\sigma^2} \prod_{i=1}^M \frac{\sqrt{1 + \rho_i^2}}{1 - \rho_i^2}, \quad (3.36)$$

alternativamente, para la transformación $\lambda_i = \rho/\sqrt{1+\rho_i^2}$ la distribución *a priori* está dada por

$$p(\boldsymbol{\theta}) = \frac{1}{\sigma^2} \left(\frac{1}{\sqrt{2} + \cosh^{-1}(\sqrt{2})} \right)^M \prod_{i=1}^M \frac{\sqrt{1-2\lambda_i^2}}{(1-\lambda_i^2)^2}, \quad (3.37)$$

es decir, $\rho_{1:M}$, σ^2 , $\boldsymbol{\mu}_{1:M}$ y $\boldsymbol{\beta}$ son independientes *a priori*, además de que la distribución es uniforme o plana en $\boldsymbol{\mu}_{1:M}$ y $\boldsymbol{\beta}$, análogo al principio de razón insuficiente. Motivados por esto último, es posible emplear una densidad en estos parámetros que nos resulte de mayor utilidad. En la siguiente sección, se presenta una *a priori* que permite realizar búsqueda estocástica de variables, es decir, seleccionar a los elementos *a posteriori* más relevantes de $\boldsymbol{\beta}$.

Por otro lado, como una alternativa a la *a priori* de referencia, Bayes y Branco (2007) propone una aproximación a de la regla de Jeffrey para el parámetro de forma λ en el contexto del modelo normal sesgado, la cuál está dada por

$$p(\lambda) \simeq \sqrt{\frac{2}{\pi} \left(1 + \frac{2\lambda^2}{\pi^2/4} \right)^{-3/4}}, \quad (3.38)$$

la expresión dentro de la raíz cuadrada corresponde a la densidad de una variable aleatoria $t(a, b, d)$, donde $a = 0$ es el parámetro de no centralidad, $b = \pi^2/4$ es el parámetro de escala y $d = 1/2$ son los grados de libertad. En este mismo sentido, si empleamos como distribución *a priori* $\pi(\rho) = I_{(-1, 1)}(\rho)$, resulta que la densidad *a priori* para el parámetro de forma, es decir la transformación $\lambda = \rho/\sqrt{1-\rho^2}$, $\pi(\lambda)$ tiene distribución $t(0, 1/2, 2)$.⁷ Sin embargo, surge un aspecto interesante a partir de la representación de la distribución $t(a, b, d)$ como una mezcla de escala normal gamma, es decir, si escribimos

$$p(\lambda) = \int N\left(\lambda \mid 0, \frac{b}{w}\right) \text{Gamma}\left(w \mid \frac{b}{2}, \frac{b}{2}\right) dw, \quad (3.39)$$

y además, si se considera la reparametrización $\gamma = \sigma\rho$, $\alpha = \sigma\sqrt{1-\rho^2}$ y usamos como

⁷Como señalan los autores, esto ilustra como una distribución uniforme no es invariante ante transformaciones invertibles del parámetro de interés.

densidad *a priori* para (μ, σ^2) la densidad usual de Jeffrey, es decir $p(\mu, \sigma^2) \propto 1/\sigma^2$, entonces juntando estas partes se tiene que

$$p(\mu, \sigma^2, \lambda) \propto \frac{1}{\alpha^2} \exp\left(-\frac{1}{2} \frac{\gamma^2 w}{\alpha^2 b}\right) w^{(d+1)/2} \exp\left(-\frac{d}{2} w\right), \quad (3.40)$$

lo cuál resulta en que el modelo sea conjugado y sea posible obtener las densidades condicionales completas para aplicar el método del muestreador de Gibbs, por tanto, en teoría se dispone de una alternativa para implementar un método Bayesiano variacional basado en la restricción campo medio, ya que de acuerdo con la discusión en la Sección 2.3.4, las densidades óptimas q^* pueden deducirse a partir de las densidades condicionales completas.

Además de esto, Wand et al. (2011) proponen un algoritmo para realizar inferencia sobre los parámetros de la normal asimétrica univariada empleando el método campo medio para ‘modelos elaborados’; de acuerdo con Tran, T.-N. Nguyen y Dao (2021) y Wand et al. (2011), se dice que un modelo es elaborado si admite una expresión jerárquica como

$$\begin{aligned} \mathbf{y} | \boldsymbol{\theta}, \eta &\sim p(\mathbf{y} | \boldsymbol{\theta}, \eta) \\ \eta | \boldsymbol{\theta} &\sim p(\eta | \boldsymbol{\theta}) \\ \theta &\sim p(\boldsymbol{\theta}), \end{aligned} \quad (3.41)$$

no obstante, la propuesta de Wand et al. (2011) no genera una solución analítica o cerrada, sino que depende aún de aproximaciones numéricas, específicamente integrales.

3.6 Distribución *a priori* para la búsqueda estocástica de variables

George y McCulloch (1993) presentaron un método para realizar selección de variables en el contexto de regresión lineal Bayesiana, empleando el muestreador de Gibbs y se nombró selección de variables con búsqueda estocástica, *stochastic search variable selection*, (SSVS). Como señalan los autores, esta técnica es similar al enfoque de aplicar distribuciones *a priori* del tipo *spike and slab* propuestas originalmente por Mitchell y Beauchamp (1988).

De acuerdo con Gilks, Richardson y Spiegelhalter (1995, Capítulo 12), a diferencia de otros métodos usuales para la selección de modelos con herramientas como el AIC, C_p de Mallow y BIC a través de los 2^p modelos posibles, la técnica SSVS busca estocásticamente aquellos subconjuntos ‘prometedores’ de predictores. Este procedimiento asigna una densidad de probabilidad al conjunto de todos los modelos de regresión posibles, de forma que los modelos prominentes reciben la probabilidad más alta, y luego utiliza el muestreador de Gibbs para simular una muestra correlacionada a partir de esta distribución. De esta manera, los modelos más probables aparecen con mayor frecuencia en la muestra.

De forma concreta, la metodología SSVS considera la siguiente distribución *a priori* para los parámetros de regresión β

$$\beta_k \mid \gamma_k \sim (1 - \gamma_k)N(0, \tau_k^2) + \gamma_k N(0, c_k^2 \tau_k^2), \quad (3.42)$$

donde $\gamma_i \sim \text{Bernoulli}(p_i)$, c_k y τ_k son hiperparámetros. Este método genera realizaciones de β_k de acuerdo a las densidades en la Figura 3.2. En una realización de la distribución *a posteriori*, los parámetros se estiman como cero si son generados por la densidad más concentrada, y se estiman como uno en caso contrario. En general, el uso de distribuciones *a priori* en β tiene un efecto de encogimiento, por ejemplo, una distribución normal conduce a una regularización Ridge, una mezcla de escala normal exponencial, conduce a una regularización tipo lasso, una mezcla de escala normal gamma-inversa, al método de determinación automática de relevancia, *automatic relevance determination* (ARD) entre otros.

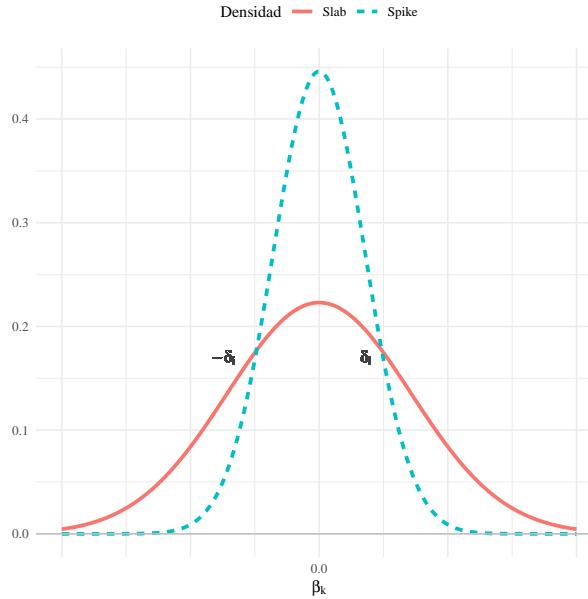


Figura 3.2: Distribución *a priori* SSVS. Fuente: elaboración propia.

Ahora, aunque es posible aprender τ_i^2 y c_i^2 por medio de los datos, se opta por fijarlos para no agregar otro nivel o jerarquía en el análisis; por ejemplo, si consideramos que las covariables están estandarizadas, entonces están en escalas similares y podemos tomar $\tau_i^2 = \tau^2$ y $c_i^2 = c^2$. Por su parte, en Gilks, Richardson y Spiegelhalter (1995) mencionan que el enfoque SSVS permite seleccionar variables de acuerdo con su ‘significancia práctica’ en lugar de significancia estadística, regresando a la idea anterior, en la Figura 3.2 se grafican las densidades $p(\beta_i | \gamma_i = 0) = N(\beta_i | 0, \tau_i^2)$ y $p(\beta_i | \gamma_i = 1) = N(\beta_i | 0, \tau_i^2 c_i^2)$, además, no es difícil ver que estas distribuciones se interceptan en el punto $\delta_i = \tau_i \sqrt{2c_i^2 \log(c_i)/(c_i^2 - 1)}$. Así, $|\beta_i| \leq \delta_i$ corresponde a la región donde $N(\beta_i | 0, \tau_i^2)$ cubre a $N(\beta_i | 0, \tau_i^2 c_i^2)$ y viceversa. De este modo, τ_i y c_i deberían ser seleccionados de tal manera que si $|\beta_i| \leq \delta_i$, entonces preferimos establecer $\beta_i = 0$, es decir, excluir la covariable i . Procediendo de esta manera, podemos fijar el valor de significancia práctica δ_i y proponer otro valor para c_i^2 , que sabemos debe ser chico -para emular un pico como en *spike and slab*-, de ahí podemos despejar τ_i . En resumen, esto se interpreta como seleccionar aquellos coeficientes cuya magnitud sea mayor al umbral δ_i .

Más adelante se implementará esta distribución *a priori* en el lenguaje Stan, no obstante,

dado que el programa no soporta variables aleatorias discretas, no es posible incluir directamente a los parámetros γ_i cuya distribución es Bernoulli. La alternativa usual es integrar la densidad $p(\beta_k | \gamma_k)$ con respecto a γ_i , en este caso, reemplazamos la integral por una suma, procediendo así, se obtiene que

$$\begin{aligned}
p(\beta_k) &= \sum_{\gamma_k \in \{0, 1\}} p(\beta_k, \gamma_k) \\
&= \sum_{\gamma_k \in \{0, 1\}} p(\beta_k | \gamma_k) p(\gamma_k) \\
&= \sum_{\gamma_k \in \{0, 1\}} [(1 - \gamma_k)N(\beta_k | 0, \tau^2) + \gamma_k N(\beta_k | 0, c^2\tau^2)] p_k^{\gamma_k} (1 - p_k)^{\gamma_k} \\
&= N(\beta_k | 0, \tau^2) \sum_{\gamma_k \in \{0, 1\}} (1 - \gamma_k) p_k^{\gamma_k} (1 - p_k)^{\gamma_k} \\
&\quad + N(\beta_k | 0, c^2\tau^2) \sum_{\gamma_k \in \{0, 1\}} \gamma_k p_k^{\gamma_k} (1 - p_k)^{\gamma_k} \\
&= (1 - p_k)N(\beta_k | 0, \tau^2) + p_k N(\beta_k | 0, c^2\tau^2) \\
&= \text{MixN}(\beta_k | 0, \tau^2, 0, \tau^2 c^2, p_k),
\end{aligned} \tag{3.43}$$

donde $\text{MixN}(x | \mu_1, \sigma_1^2, \mu_2, \sigma_2^2, p) = pN(x | \mu_1, \sigma_1^2) + (1-p)N(x | \mu_2, \sigma_2^2)$. Podemos identificar la última expresión como la función de densidad asociada a un modelo de mezclas gaussiano con dos componentes. Note que cada término corresponde a la mezcla de normales que se muestra en la [Figura 3.3](#). Este método es análogo al uso de una *a priori* del tipo *spike-slab*, donde se ha reemplazado el pico o la delta de Dirac por una densidad continua.

Dado que hemos integrado sobre γ_k , la incertidumbre de esta variable está contenida en el parámetro p_k . Para completar la especificación de la *a priori* SSVS, es necesario asignar una densidad para las probabilidades de inclusión $p_{1:p}$. Con el propósito de mantener la objetividad en el análisis, asignamos la siguiente distribución no informativa

$$\pi(p_{1:p}) \equiv \prod_{k=1}^p \text{Be}(p_k | 1/2, 1/2), \tag{3.44}$$

por simplicidad se asume que p_k y $p_{k'}$ son independientes *a priori*, de ahí que la densidad conjunta sea igual al producto de las densidades marginales. Aunque en la práctica la

inclusión de β_k pueda tener influencia en la inclusión de $\beta_{k'}$, por simplicidad se omite este hecho. Tal elección de distribución *a priori* corresponde al método de Jeffrey para el parámetro de probabilidad en la distribución Bernoulli y en general para la distribución Binomial. A partir de la muestra *a posteriori* de p_k , se genera una muestra de las covariables seleccionadas, digamos, mediante la simulación de una variable aleatoria indicadora o Bernoulli. Así, los modelos ‘más prominentes’ tendrán mayor frecuencia de aparición.

En resumen, se asigna la siguiente *a priori* para $\boldsymbol{\beta}$ y $p_{1:p}$

$$\begin{aligned}\pi(\boldsymbol{\beta}, p_{1:p} | c_{1:p}, \tau_{1:p}) &\equiv \pi(\boldsymbol{\beta} | p_{1:p}, c_{1:p}, \tau_{1:p}) \pi(p_{1:p}) \\ &= \prod_{k=1}^p \text{MixN}(\beta_k | 0, \tau_k^2, 0, \tau_k^2 c_k^2, p_k) \text{Be}(p_k | 1/2, 1/2) \\ &= \prod_{k=1}^p [p_k N(\beta_k | c_k^2 \tau_k^2) + (1 - p_k) N(\beta_k | \tau_k^2)] \frac{p_k^{-1/2} (1 - p_k)^{-1/2}}{\text{Beta}(1/2, 1/2)}.\end{aligned}\tag{3.45}$$

Como ya se mencionó, la *a priori* SSVS también depende de los hiperparámetros c_k y τ_k que de acuerdo con George y McCulloch (1993), pueden obtenerse a partir de significancia práctica, a continuación se comenta otra alternativa para especificarlos: la estimación $\beta_k = 0$ tiene alta probabilidad de estar en el intervalo $(-3\tau_k, 3\tau_k)$, por lo que τ_k puede fijarse estableciendo el valor de τ_k en el cuál se decide considerar la estimación β_k como diferente de cero; en el caso del modelo continuo, dado que la variable objetivo se encuentra en escala logaritmo, la interpretación de β_k es en términos de porcentajes, así, desde un punto de vista práctico, fijamos este umbral como $3\tau_k = 1\%$, por lo que $\tau_k = 1/300$. Se estandarizan las covariables para que estén en escalas similares, de modo que únicamente se requiere una tupla (τ, c) , la cuál se fija como $(1/300, \sqrt{10 \times 300^2})$, es decir, las estimaciones diferentes de cero serán generadas por la densidad $N(0, \tau^2 c^2 = 10)$. En los modelos binario y ordinal, se asignan los mismos valores para esta tupla, pero se toma el valor de significancia práctica como 0.1. Finalmente, es posible observar que no se asigna este tipo de *a priori* en los parámetros $\mu_{1:M}$ ya que siempre se desea incluirnos en el modelo.

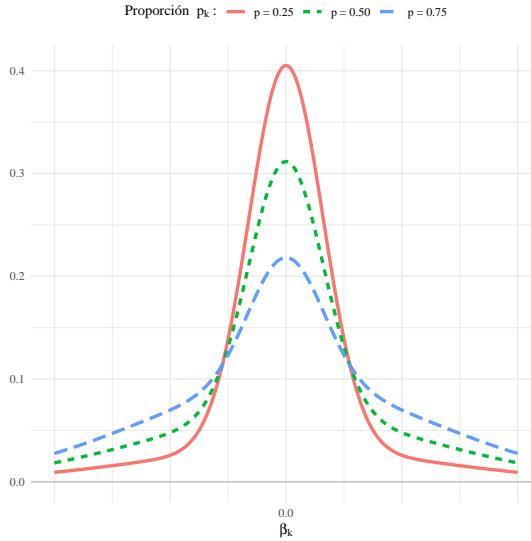


Figura 3.3: Distribución *a priori* SSVS con distintas proporciones de mezcla. Fuente: elaboración propia.

3.7 Distribución *a posteriori*

Usando el teorema de Bayes con la verosimilitud de los modelos en la Ecuación 3.19, Ecuación 3.19 y Ecuación 3.19, junto a la distribución *a priori* compuesta por la densidad de referencia y la densidad SVSS, se obtiene el kernel de la distribución *a posteriori* -es decir, salvo una constante-. El tratamiento para obtener la *a posteriori* es identico en los tres modelos, y podemos señalar que cada uno de estos casos no admite una forma estándar, así que la alternativa usual es emplear algún método de cadenas de Márkov Monte Carlo, *Markov chain Monte Carlo* (MCMC).

Sea $\boldsymbol{\theta}$ el vector de parámetros de interés en las M regiones, dada la discusión previa, se tiene que

$$\begin{aligned} \pi(\boldsymbol{\theta}) &\equiv p(\rho_{1:M}, \sigma^2) \times p(\boldsymbol{\beta} | p_{1:p}) \\ &= \frac{1}{\sigma^2} \prod_{i=1}^M \frac{\sqrt{1 + \rho_i^2}}{1 - \rho_i^2} \times [p_k N(\beta_k | c_k^2 \tau_k^2) + (1 - p_k) N(\beta_k | \tau_k^2)] \frac{p_k^{-1/2} (1 - p_k)^{-1/2}}{\text{Beta}(1/2, 1/2)}, \end{aligned} \quad (3.46)$$

a continuación escribimos las distribuciones *a posteriori*, salvo una constante, para cada modelo

-
- Modelo log-normal sesgado: sea $w_{1:M}$ la colección de variables latentes que generan a cada y_{ij} , originadas a partir del proceso de truncamiento oculto, esta densidad tiene dimensión $2M + 2p + 1$, los parámetros son $\boldsymbol{\theta} = (\rho_{1:M}, \sigma^2, \mu_{1:M}, \boldsymbol{\beta}, p_{1:M})$:

$$\begin{aligned}\pi(\boldsymbol{\theta}, w_{1:M} \mid \mathbf{y}) &\propto p(\mathbf{y}, w_{1:M} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \equiv p(\mathbf{y} \mid \boldsymbol{\theta}, w_{1:M}) p(w_{1:M} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &= \prod_{i=1}^M \left[\prod_{j=1}^{N_i} N(y_{ij} \mid \eta_{ij}, \sigma_i^2) \right] 2N(w_i \mid 0, \sigma_i^{2*}) 1_{(0,\infty)}(w_i) \times \pi(\boldsymbol{\theta}).\end{aligned}\quad (3.47)$$

- Modelo probit sesgado latente: sea z_{ij} las variables latentes que permiten generar a los y_{ij} binarios y sea $w_{1:M}$ la colección de variables latentes que generan a cada z_{ij} , originadas a partir del proceso de truncamiento oculto, esta densidad tiene dimensión $2M + 2p$, los parámetros son $\boldsymbol{\theta} = (\rho_{1:M}, \mu_{1:M}, \boldsymbol{\beta}, p_{1:M})$:

$$\begin{aligned}\pi(\boldsymbol{\theta}, w_{1:M}, \mathbf{z} \mid \mathbf{y}) &\propto p(\mathbf{y}, \mathbf{z}, w_{1:M} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \equiv p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z} \mid \boldsymbol{\theta}, w_{1:M}) p(w_{1:M} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &= \prod_{i=1}^M \left[\prod_{j=1}^{N_i} 1(y_{ij}, z_{ij}) N(z_{ij} \mid \eta_{ij}, \sigma_i^2) \right] 2N(w_i \mid 0, \sigma_i^{2*}) 1_{(0,\infty)}(w_i) \times \pi(\boldsymbol{\theta}).\end{aligned}\quad (3.48)$$

- Modelo probit ordenado sesgado latente: sea z_{ij} las variables latentes que permiten generar a los y_{ij} binarios y sea $w_{1:M}$ la colección de variables latentes que generan a cada z_{ij} , originadas a partir del proceso de truncamiento oculto, esta densidad tiene dimensión $2M + 2p + k - 1$, los parámetros son $\boldsymbol{\theta} = (\rho_{1:M}, \mu_{1:M}, \delta_{1:k}, \boldsymbol{\beta}, p_{1:M})$:

$$\begin{aligned}\pi(\boldsymbol{\theta}, w_{1:M}, \mathbf{z} \mid \mathbf{y}) &\propto p(\mathbf{y}, \mathbf{z}, w_{1:M} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \equiv p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z} \mid \boldsymbol{\theta}, w_{1:M}) p(w_{1:M} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &= \prod_{i=1}^M \left[\prod_{j=1}^{N_i} 1(y_{ij}, z_{ij}, \delta_{1:k}) N(y_{ij} \mid \eta_{ij}, \sigma_i^2) \right] 2N(w_i \mid 0, \sigma_i^{2*}) 1_{(0,\infty)}(w_i) \times \pi(\boldsymbol{\theta}).\end{aligned}\quad (3.49)$$

En este escenario, es relevante notar que los parámetros $\delta_{1:k}$ están ordenados, es decir, el soporte de δ_c está acotado entre $(\delta_{c-1}, \delta_{c+1}]$. El lenguaje Stan puede manejar este tipo de restricciones de forma automática. Por otra parte, para realizar un tratamiento Bayesiano completo, se asigna una *a priori* para los puntos de corte $\delta_{1:k}$, y con el

objetivo de continuar dentro del enfoque objetivo, se asigna la siguiente distribución no informativa basada en el principio de razón insuficiente

$$p(\delta_c) \propto 1(\delta_c \in (\delta_{c-1}, \delta_{c+1})) \quad (3.50)$$

es decir, uniforme para cada punto dentro de su soporte. Otra alternativa práctica, es asignar una densidad plana, por ejemplo $p(\delta_c) = NT(\delta_c | 0, 100^2; \delta_{c-1}, \delta_{c+1})$.

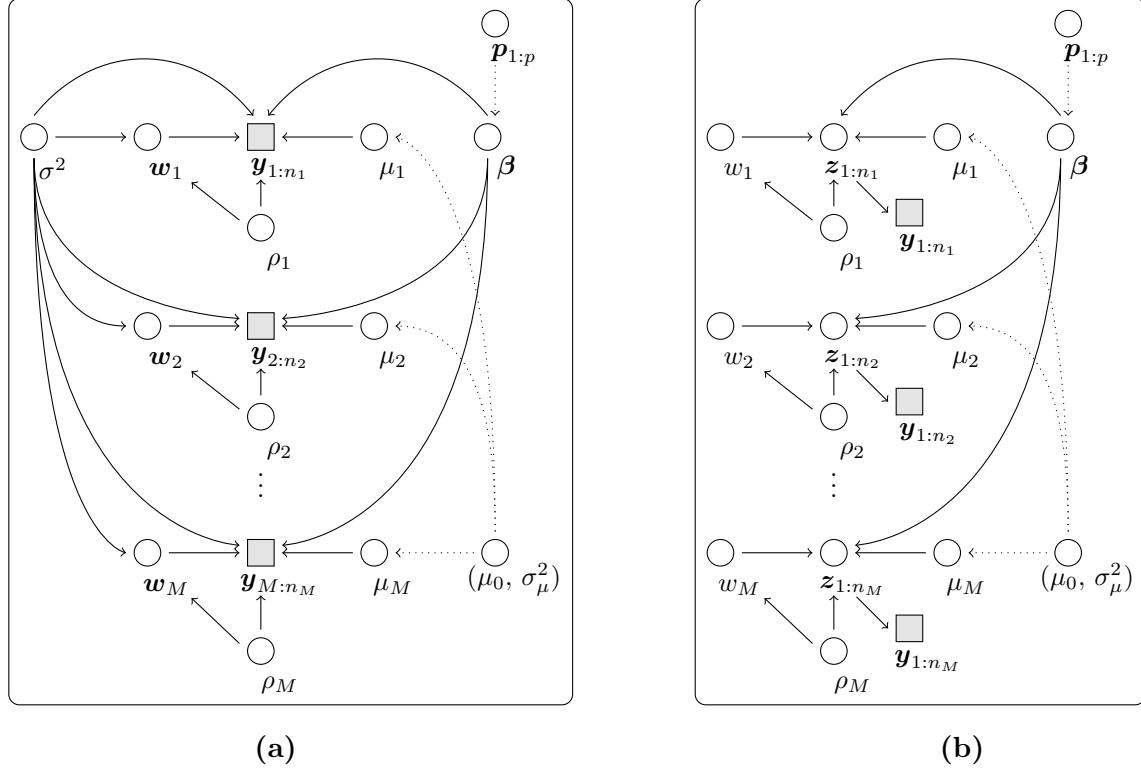
Por otro lado, la [Ecuación 3.51](#) muestra el planteamiento jerárquico del modelo log-normal sesgado, y en la [Figura 3.4](#) se muestran los grafos dirigidos acíclicos, *directed acyclic graph* (DAG), que corresponden tanto a esta representación jerárquica como a la del modelo probit sesgado con variable latente. Las líneas punteadas relacionan parámetros y sus hiperparámetros. Los símbolos redondos indican parámetros y variables latentes, mientras que los símbolos cuadrados representan a los datos observados.

$$\begin{aligned} y_{ij} \mid \rho_i, \sigma^2, w_i, \mu_i, \boldsymbol{\beta} &\sim N\left(\mu_i - \mathbf{x}_{ij}^T \boldsymbol{\beta} - w\rho_i + \frac{\sigma \sqrt{\frac{2}{\pi} \rho_i^2}}{\sqrt{1 - \frac{2}{\pi} \rho_i^2}}, \frac{\sigma^2(1 - \rho_i^2)}{\sqrt{1 - \frac{2}{\pi} \rho_i^2}}\right) \\ w_i \mid \rho_i, \sigma^2 &\sim NT\left(0, \frac{\sigma^2}{\sqrt{1 - \frac{2}{\pi} \rho_i^2}}\right) \\ p(\sigma^2, \rho_1, \dots, \rho_M) &\propto \frac{1}{\sigma^2} \prod_{i=1}^M \frac{\sqrt{1 + \rho_i^2}}{1 - \rho_i^2} \\ p(\mu_1, \dots, \mu_M) &= N_M(\mu_0 \mathbf{1}_M, \sigma_\mu^2 I_M) \\ \mu_0 &\sim N(0, 100^2) \\ \sigma_\mu^2 &\sim \text{InvGamma}(0.001, 0.001) \\ p(\boldsymbol{\beta} \mid p_{1:p}) &\propto \prod_{k=1}^p \text{MixNorm}(\beta_k \mid 0, \tau^2 c^2, 0, c^2, \gamma_k) \\ p(p_{1:p}) &= \prod_{k=1}^p \text{Be}(\gamma_k \mid 1/2, 1/2), \end{aligned} \quad (3.51)$$

En la siguiente sección, se sugiere usar esta *a priori* jerárquica en $\mu_{1:M}$ que permite que sus elementos compartan información entre sí, lo cuál es útil cuando los tamaños de muestra son

pequeños o incluso cero. A menos que se indique lo contrario, se fija $\mu_0 \equiv 0$ y $\sigma_\mu^2 = 100^2$.

Figura 3.4: Grafo dirigido acíclico del que corresponde la representación jerárquica del modelo log-normal sesgado con parámetros centrados. Fuente: elaboración propia



3.8 Predicción de nuevas observaciones ($n_i > 0$ y $n_i = 0$)

En las secciones previas, para mantener simple la notación de los modelos, se sugiere que se conocen todos los pares (i, j) , es decir, $i = 1, 2, \dots, M$ y $j = 1, 2, \dots, N_i$. No obstante, de forma general en el contexto de áreas pequeñas no se cumple esto. Por tanto, es posible hacer la distinción entre la colección de \mathbf{y}^s (observados) y \mathbf{y}^r (faltantes), considerando a este último como variables latentes, así, se escribe

$$\pi(\boldsymbol{\theta}, z_{1:M}, \mathbf{y}^r | \mathbf{y}^s) \propto p(\mathbf{y}^s | z_{1:M}, \boldsymbol{\theta}, \mathbf{y}^r) \times p(z_{1:M}, \mathbf{y}^r | \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}). \quad (3.52)$$

Dada la independencia condicional $y_{ij} \mid z_i \perp y_{ij'} \mid z_u$, entonces $\mathbf{y}^s \mid z_{1:M} \perp \mathbf{y}^r \mid z_{1:M}$, lo cuál permite remover la dependencia de \mathbf{y}^s en \mathbf{y}^r , y además, desarrollando

$$p(z_{1:M}, \mathbf{y}^r \mid \boldsymbol{\theta}) = p(z_{1:M} \mid y^r, \boldsymbol{\theta}) \times p(\mathbf{y}^r \mid \boldsymbol{\theta}) = p(z_{1:M} \mid \boldsymbol{\theta}) \times p(\mathbf{y}^r \mid \boldsymbol{\theta}), \quad (3.53)$$

es posible escribir la expresión previa como

$$\pi(\boldsymbol{\theta}, z_{1:M}, \mathbf{y}^r \mid \mathbf{y}^s) \propto p(\mathbf{y}^s \mid z_{1:M}, \boldsymbol{\theta}) \times p(z_{1:M} \mid \boldsymbol{\theta}) \times p(y^r \mid \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}). \quad (3.54)$$

No obstante, de aquí parece que es posible integrar \mathbf{y}^r de forma sencilla:

$$\begin{aligned} \int \pi(\boldsymbol{\theta}, z_{1:M}, \mathbf{y}^r \mid \mathbf{y}^s) d\mathbf{y}^r &\propto \int p(\mathbf{y}^s \mid z_{1:M}, \boldsymbol{\theta}) \times p(z_{1:M} \mid \boldsymbol{\theta}) \times p(\mathbf{y}^r \mid \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \\ &= p(\mathbf{y}^s \mid z_{1:M}, \boldsymbol{\theta}) \times p(z_{1:M} \mid \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \underbrace{\int p(\mathbf{y}^r \mid \boldsymbol{\theta}) d\mathbf{y}^r}_1 \\ &= p(\mathbf{y}^s \mid z_{1:M}, \boldsymbol{\theta}) \times p(z_{1:M} \mid \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}), \end{aligned} \quad (3.55)$$

ya que $p(\mathbf{y}^r \mid \boldsymbol{\theta})$ es una densidad propia. Esto significa que si $\mathbf{y}^s \mid z_{1:M} \perp \mathbf{y}^r \mid z_{1:M}$, entonces \mathbf{y}^r no aporta información al modelo. Ahora bien, en el escenario cuando $0 < n_i < N_i$, esto es, existen observaciones en la región i , el pronóstico de y_{ij}^* se genera a partir de la muestra *a posteriori* e integración Monte Carlo:

$$p(y_{ij}^* \mid \mathbf{y}^s) = \int p(y_{ij}^* \mid \boldsymbol{\theta}, z_{1:M}) p(\boldsymbol{\theta}, z_{1:M} \mid \mathbf{y}^s) d\boldsymbol{\theta} \approx \frac{1}{S} \sum_{k=1}^S p(y_{ij}^* \mid \boldsymbol{\theta}^{(k)}, z_{1:M}^{(k)}), \quad (3.56)$$

donde $p(y_{ij}^* \mid \boldsymbol{\theta}, z_{1:M})$ es el modelo muestral, es decir, el mismo mecanismo estocástico propuesto para generar a los datos \mathbf{y}^s , y S denota el número de muestras *a posteriori*.

Con frecuencia, los modelos de áreas pequeñas necesitan generar pronósticos aún en los dominios donde $n_i = 0$. Este caso presenta un desafío mayor, ya que si bien σ^2 , y $\boldsymbol{\beta}$ toman información directamente de otras áreas pequeñas, los efectos a nivel de área ρ_i y μ_i únicamente son informados de manera indirecta, lo cuál puede ocasionar que los parámetros se encojan o bien colapsen hacia los límites de su soporte. En este escenario, la alternativa

usual es realizar agrupamiento o *pooling* por medio de una estructura jerárquica en los parámetros, específicamente aquellos que son propios de cada región. Por ejemplo, puede proponerse

$$\begin{aligned}\lambda_{1:M} &\sim N_M(\mu_0, \sigma_\lambda^2) & \mu_{1:M} &\sim N_M(\mu_0 \mathbf{1}_M, \sigma_\mu^2 I_M) \\ \mu_0 &\equiv 0 & \mu_0 &\sim N(0, 100^2) \\ \sigma_\lambda^2 &\sim \text{InvGamma}(0.001, 0.001) & \sigma_\mu^2 &\sim \text{InvGamma}(0.001, 0.001).\end{aligned}\tag{3.57}$$

Esta estructura jerárquica *a priori* intenta ser objetiva al asignar distribuciones planas de acuerdo a la regla de Laplace. Se fija la media del parámetro de forma en cero para no incrementar otro nivel adicional la complejidad, es decir, otra capa jerárquica, en el modelo. De igual modo, fuera del panorama Bayesiano objetivo, puede incluirse información de áreas vecinas en ambos parámetros.

3.9 Implementación con métodos *Markov Chain Monte Carlo*

De acuerdo a la discusión en la [Sección 2.2](#), podemos tratar de identificar las distribuciones condicionales completas en cada una de las densidades *a posteriori* en la sección previa para aplicar el algoritmo muestreador de Gibbs, y para aquellas variables que no se pueda identificar una densidad condicional completa, podemos usar el algoritmo Metrópolis-Hastings e implementar un método híbrido. Para el modelo log-normal sesgado, este es precisamente el enfoque de MCMC que fue aplicado por López ([2024](#)). No obstante, el interés principal es sobre el estudio de los métodos Bayesianos variacionales, y con el fin de contrastarlos contra los métodos de MCMC usuales, se usará el algoritmo Hamiltoniano Monte Carlo (HMC) con la variante automática *No-U-Turn sampler* (NUTS), esto significa que no se requiere trabajo adicional de programación o desarrollo analítico, por tal motivo, no se replica el análisis y tampoco es necesario obtener las densidades condicionales completas. Para los modelos probit sesgado y probit ordenado sesgado, ambos siguiendo el enfoque de variable latente, el tratamiento es análogo pero novedoso.

3.10 Implementación en Stan

Stan es un lenguaje de programación probabilística (PP) en el sentido de que una variable aleatoria es genuinamente un objeto básico (Stan Development Team [2025a](#)). La PP es un paradigma de programación que se sitúa en la intersección de los campos del aprendizaje automático o *machine learning* (ML), la estadística y los lenguajes de programación en general. Este método aprovecha la semántica formal, los compiladores y otras herramientas de los lenguajes de programación para crear evaluadores de inferencia eficientes para modelos y aplicaciones de aprendizaje automático, utilizando los algoritmos de inferencia y la teoría de la estadística. (Meent et al. [2021](#)).

De acuerdo con Meent et al. ([2021](#)) y Pichler, Jewson y Avalos-Pacheco ([2025](#)), una enfoque de la PP consiste en automatizar la inferencia Bayesiana mediante herramientas de las ciencias de la computación, permitiendo al usuario escribir modelos Bayesiano en un formato simple. Por ejemplo, los lenguajes BUGS, *Bayesian inference using Gibbs sampling* y JAGS, *just another Gibbs sampling* fueron pioneros en esta tarea.

Stan es un lenguaje imperativo, es decir, declara una secuencia de instrucciones que especifican como debe ejecutarse algo. Además, está basado en asignaciones, *loops*, condicionales, variables locales, funciones y estructuras de datos tipo *array* (Stan Development Team [2025a](#)). Para su ejecución, un código Stan se traduce a un programa C++ mediante el compilador stanc de Stan (Carpenter et al. [2017](#)).

Un programa de Stan se compone de seis bloques, de los cuáles, ninguno es obligatorio. En el [Cuadro 3.2](#) describimos cada uno de estos. A su vez, cada bloque de Stan hace distinción entre *declarations* o declaraciones y *statements* o instrucciones. En general, primero declaramos las variables y después indicamos la secuencia de instrucciones que la definen.

Cuadro 3.2: Bloques de un programa Stan. Fuente: elaboración propia, basado en Stan Development Team (2025a) y Carpenter et al. (2017).

Bloque	Descripción
<code>data {...}</code>	Se declaran los datos requeridos para ajustar el modelo.
<code>transformed data{...}</code>	Aquí se declaran y definen nuevas variables calculadas a partir del bloque anterior. También se definen constates.
<code>parameters{...}</code>	Se ejecuta cada vez que la log probabilidad es evaluada (lo que puede ocurrir varias veces por iteración). Todas las variables declaradas con soporte restringido se transforman a un espacio sin restricciones. ⁸ Este bloque <code>parameters</code> define los parámetros y variables latentes involucrados en el modelo. Es en este apartado donde se establecen las restricciones sobre los parámetros, por ejemplo, para que sean no negativas, o estén acotados en algún intervalo de \mathbb{R} .

Cuadro 3.2: Bloques de un programa Stan. Fuente: elaboración propia, basado en (Stan Development Team 2025b; Stan Development Team 2025a; Carpenter et al. 2017). (Continuación)

Bloque	Descripción
<code>transformed parameters{...}</code>	Como su nombre sugiere, permite transformar parámetros dentro de un modelo. Cualquier variable declarada aquí forma parte de la salida generada. Cualquier variable definida totalmente en términos de los datos o de alguna transformación de estos, por eficiencia debe declararse y definirse en el bloque previo. Es relevante mencionar que dentro de este bloque es posible realizar tanto transformación en los parámetros como cambios de variable. La distinción clave es que una transformación muestrea u optimiza un parámetro y después lo transforma para emplearlo donde se requiera, mientras que un cambio de variables primero transforma el parámetro y después genera muestras de él. Sólo el cambio de variables requiere añadir el ajuste del log Jacobiano de forma manual en el bloque siguiente. Una regla de pulgar es la siguiente: si se asigna una distribución a los parámetros transformados, es necesario agregar el log Jacobiano adecuado.

Cuadro 3.2: Bloques de un programa Stan. Fuente: elaboración propia, basado en (Stan Development Team 2025b; Stan Development Team 2025a; Carpenter et al. 2017). (Continuación)

Bloque	Descripción
<code>model{...}</code>	<p>El propósito de este bloque es definir la función de log-probabilidad, esto es, la verosimilitud y <i>a priori</i>, en el espacio de parámetros restringido. Sólo aquí se permiten las declaraciones de probabilidad, ya sea por medio de la notación de distribución <code>~</code> o incrementando la log-verosimilitud <code>target +=</code>.</p> <p>Este bloque se ejecuta después del bloque <code>transformed parameters</code> cada vez que se evalúa la función de log-probabilidad.</p> <p>En caso de no especificarse una <i>a priori</i> para algún grupo de parámetros, se les asigna por defecto una distribución uniforme en su soporte. Si está acotada, entonces resulta ser una distribución propia.</p>
<code>generated quantities{...}</code>	<p>Este bloque permite definir de manera eficiente valores que dependen de parámetros y datos, pero que no afectan a la estimación. Nada declarado aquí afecta a los valores de los parámetros generados.</p> <p>Este bloque solo se llama una vez por muestra o iteración, no por cada evaluación de la función de log-probabilidad. Se puede utilizar para realizar inferencia <i>a posteriori</i>. Además, los valores de todas las demás variables declaradas en bloques previos (excepto variables locales), pueden ser utilizadas.</p>

⁸Se pretende que la densidad de probabilidad definida por un programa de Stan tenga soporte sin restricciones, es decir, sin regiones de probabilidad cero, lo que simplifica enormemente la tarea de desarrollar muestreadores u optimizadores.

Desde el punto de vista práctico, el lenguaje Stan es atractivo en el sentido de que que permite implementar una amplia variedad de modelos Bayesianos de forma sencilla, sin embargo, existen dos restricciones principales al momento de definirlos y construirlos:

1. No se pueden optimizar o incluir parámetros discretos. Con respecto a este punto, después de marginalizar la variable aleatoria Bernoulli en la distribución SSVS de la [Sección 3.6](#), podemos ignorar esta restricción, ya que además, el enfoque de variable latente en los modelos de clasificación remueve la presencia de las variables binarias y_{ij} , más allá de las funciones indicadoras $1(y_{ij}, z_{ij})$ y $1(y_{ij}, z_{ij}, \delta_{1:c})$ que numéricamente evalúan uno.
2. El modelo debe ser diferenciable. Para esta observación, una extensa gama de modelos en ML son diferenciables, entre ellos, los modelos de regresión que consideramos aquí.

A continuación, se describen los bloques de código Stan empleados para realizar inferencia sobre los modelos de regresión propuestos, se inicia describiendo al modelo log-normal sesgado, posteriormente, al modelo probit ordenado con variable latente, ya que al simplificar este último, se obtiene el código Stan para el modelo restante.

3.10.1 Modelo log-normal sesgado centrado

A continuación se describe la implementación del modelo de regresión normal-sesgado. En virtud de que este se obtiene al tomar el logaritmo de la variable objetivo, no es necesario que el programa distinga entre cuando la respuesta está en escala original y cuando está transformada.

- En las líneas 1-16 se define el bloque `data`, el cuál contiene información sobre la identificación del modelo: número de observaciones, valores faltantes, covariables, región de las observaciones; así como el vector con la variable respuesta (`y_obs`) y la matriz diseño asociada a los datos observados y no observados (`X_obs`, `X_mis`). Además, se pase información acerca de los hiperparámetros c , τ y el umbral ε para la técnica SSVS.

```

1  data {
2      int<lower=1> K;
3      int<lower=0> N_obs;
4      int<lower=0> N_mis;
5      int<lower=1> p;
6      vector[N_obs] y_obs;
7      matrix[N_obs, p] X_obs;
8      matrix[N_mis, p] X_mis;
9      matrix[N_obs + N_mis, p] X_all;
10     array[N_obs] int<lower=1, upper=K> group;
11     array[N_mis] int<lower=1, upper=K> group_mis;
12     array[N_obs+N_mis] int<lower=1, upper=K> group_all;
13     real<lower=0> c;
14     real<lower=0> tau;
15     real epsilon;
16 }

```

- Aquí se programan los parámetros del modelo y sus respectivas restricciones. Por ejemplo, σ y \tilde{z} se declaran entre $(0, \infty)$, además, se restringe la correlación de cada área pequeña al intervalo $(0, 1)$, al igual que las probabilidades de inclusión.

```

17 parameters {
18     real<lower=0> sigma;
19     vector<lower=0, upper=1>[K] rho;
20     vector[K] mu;
21     vector[p] b;
22     vector<lower=0>[K] z_tilde;
23     vector<lower=0, upper=1>[p] pr;
24 }
25
26 transformed parameters{
27     vector<lower=0>[K] z = z_tilde .* sigma ./ sqrt(1 - 2*square(rho)/pi());
28 }

```

-
- Luego, en el bloque `model` se declarada la log-verosimilitud del modelo paramétrico asumido, junto a sus distribuciones *a priori*. Para hacer esto, se dispone de dos alternativas, la primera emplea el símbolo reservado `~`, que relaciona un parámetro en el lado izquierdo con una densidad de probabilidad en el lado derecho, la segunda alternativa consiste en aumentar la log-verosimilitud a través de la asignación `target+=`, donde el kernel de una densidad de probabilidad en el lado derecho. Por ejemplo, si `y` es un parámetro cuya distribución es normal estándar, entonces ambas declaraciones tienen el mismo efecto:

```
y ~ normal(0, 1);
target+=normal_lupdf(y|0, 1)
```

En general, Stan trabaja con el kernel de las densidades, por lo que la segunda declaración es ligeramente más eficiente. Así mismo, en caso de no especificar una *a priori* para algún parámetro, Stan le asignará una distribución uniforme, y por tanto, impropia.

De igual modo, se emplea una parametrización no centrada en `z` y `z_tilde`, es decir, si $Z \sim NT(0, \sigma^2, (0, \infty))$, entonces $Z = \sigma \tilde{Z}$, donde $\tilde{Z} \sim NT(0, 1, (0, \infty))$. Esto permite mejorar la geometría *a posteriori*, y con ello la eficiencia del muestreo (Stan Development Team 2025b).

```
29 model {
30   // a priori de rho
31   target += sum(0.5 * log1p(square(rho)) - log1m(square(rho)));
32   // a priori de sigma
33   // target += -sigma;
34   sigma ~ inv_gamma(0.001, 0.001);
35   // distribución de z_tilde
36   z_tilde ~ normal(0, 1);
37   // a priori de mu
38   mu ~ normal(0, 100);
39   // a priori de beta (SSVS)
```

```

40     pr ~ beta(0.5, 0.5);
41     for (i in 1:p) {
42         target += log_mix(pr[i],
43                             normal_lpdf(b[i] | 0, tau * c),
44                             normal_lpdf(b[i] | 0, tau));
45     }
46     // Verosimilitud  $p(y^s / \theta, z_{1:K})$ 
47     vector[N_obs] sig = sigma * sqrt(1 - square(rho[group])) ./ sqrt(1 -
48                                     → 2/pi() * square(rho[group]));
49     vector[N_obs] intercept = rho[group] .* z[group] - sigma*rho[group] .*
50                                     → sqrt(2/pi()) ./ sqrt(1 - 2/pi() * square(rho[group]));
51     target += normal_id_glm_lupdf(y_obs | X_obs, mu[group] + intercept, b,
52                                     → sig);
53 }
```

- Finalmente, el bloque `generated quantities` se ejecuta después de cada iteración si se usa el método HMC/NUTS, y después de terminar el proceso de optimización en el caso BV/ADVI. En general, este fragmento de código permite realizar inferencia *a posteriori*. De acuerdo con la guía del usuario de Stan (Stan Development Team 2025b), es más eficiente generar una variable en este bloque en lugar del bloque `transformed parameters`, por lo tanto, si alguna cantidad no desempeña ningún rol en el modelo, debe ser definida en este apartado. Para generar valores ajustados y pronósticos se sigue el mecanismo de truncamiento oculto: dado que $p(y_{ij}, z_i | \boldsymbol{\theta}) = p(y_{ij} | \boldsymbol{\theta}, z_i) p(z_i | \boldsymbol{\theta})$, dada una realización de la *a posteriori* $\boldsymbol{\theta}^{(t)}$ y $z_i^{(t)}$, se genera $y_{ij} \sim p(y_{ij} | \boldsymbol{\theta}^{(t)}, z_i^{(t)})$.

```

51 generated quantities {
52     // SSVS
53     vector<lower=0, upper=1>[p] m_ind;
54     for (j in 1:p) {
55         if (abs(b[j]) > epsilon) {
56             m_ind[j] = 1;
57         } else {
```

```

58         m_ind[j] = 0;
59     }
60 }
61 // Valores ajustados
62 int<lower=1, upper=N_obs+N_mis> j;
63 vector[N_obs+N_mis] y_all;
64 real eta_ij;
65 real<lower=0> sigma_i;
66 for(i in 1:(N_obs+N_mis)){
67     j = group_all[i];
68     eta_ij = mu[j] + X_all[i]*b + rho[j] .* z[j]- sigma .* rho[j] .*
69             ↵ sqrt(2/pi()) ./ sqrt(1 - 2*square(rho[j])/pi());
70     sigma_i = sigma * sqrt(1 - square(rho[j]))./ sqrt(1 - 2/pi() *
71             ↵ square(rho[j]));
72     y_all[i] = normal_rng(eta_ij, sigma_i);
73 }
74 }
```

3.10.2 Modelo probit ordenado sesgado latente centrado

La diferencia principal para desarrollar las implementaciones tipo probit (es decir, binarias u ordinales), es la incorporación de restricciones en las variables aleatorias normales. Esta implementación se basa en la guía del usuario de Stan (Stan Development Team 2025b), la cuál está basada en el enfoque de variable latente desarrollado por Albert y Chib (1993). En esta misma guía, se muestran otras formas de definir el modelo probit ordenado, sin embargo, este enfoque aprovecha el uso de variables latentes. Para la ejecución de este modelo, es posible utilizar de nuevo partes del código previo incorporando estas adecuaciones. Enseguida se describen las modificaciones realizadas a la implementación anterior, el símbolo `...` significa que utilizamos las mismas instrucciones dentro de este bloque que en el modelo log-normal sesgado.

- En el bloque `data` se pasa información sobre el número de valores en cada categoría

`Ni`, digamos, el número de valores cero, uno y dos, y las posiciones `ni` que ocupan en el vector respuesta observado \mathbf{y}^s . Por ejemplo, si $\mathbf{y}^s = (0, 1, 2, 2, 1, 0)$, entonces $N0 = N1 = N2 = 2$ y $n0 = (1, 6)$, $n1 = (2, 5)$ y $n2 = (3, 4)$.

```

1  data {
2      ...
3      // Restrictions
4      int<lower=1, upper=N_obs-2> N0;
5      int<lower=1, upper=N_obs-2> N1;
6      int<lower=1, upper=N_obs-2> N2;
7      array[N0] int<lower=1, upper=N_obs> n0;
8      array[N1] int<lower=1, upper=N_obs> n1;
9      array[N2] int<lower=1, upper=N_obs> n2;
10 }
```

- En el bloque `transformed data` se fijan las cantidades $\sigma^2 \equiv 1$ y $\delta_0 \equiv 0$ por identificación del modelo. Al hacer esto, no entran a la parte de muestreo u optimización. Aunque es posible sustituir directamente estas cantidades por estos valores en los bloques siguientes, mantenerlo así permite reutilizar fragmentos de código.

```

11 transformed data {
12     real<lower=0> sigma=1.0;
13     real delta0=0.0;
14 }
```

- En la definición del umbral δ_1 se acota su soporte al intervalo (δ_0, ∞) . Dado el uso de variables latentes, los parámetros `z_tilde` y `z`, que representan a la densidad normal truncada a la izquierda en cero, se reemplazan por `w_tilde` y `w`. La parte latente son los parámetros `zi`, para las cuáles se establecen restricciones de acuerdo con los puntos de corte δ_0 y δ_1 .

```

15 parameters {
16     ...

```

```

17   real<lower=0> w_tilde;
18   real<lower=delta0> delta1;
19   vector<upper=delta0>[N0] z0;
20   vector<lower=delta0, upper=delta1>[N1] z1;
21   vector<lower=delta1>[N2] z2;
22 }
```

- En el bloque `transformed parameters` se crea el vector `z_obs`, en cuyas posiciones `Ni` se les asigna el parámetro `zi` con las restricciones adecuadas

```

23 transformed parameters {
24   vector[N_obs] z_obs;
25   for (n in 1:N0) {
26     z_obs[n0[n]] = z0[n];
27   }
28   for (n in 1:N1) {
29     z_obs[n1[n]] = z1[n];
30   }
31   for (n in 1:N2) {
32     z_obs[n2[n]] = z2[n];
33   }
34   vector<lower=0>[K] w = w_tilde .* sigma ./ sqrt(1 - 2*square(rho)./pi());
35 }
```

- Se asigna una *a priori* plana dentro de su soporte al umbral δ_1 . Aunado a esto, se agrega el mismo término de log-verosimilitud en la variable `z_obs`, es decir, $\log p(z^s | \theta, w_{1:M})$. El resto de términos del modelo permanece igual.

```

36 model {
37   ...
38   // a priori para delta1
39   delta1 ~ normal(0, 100);
40   // distribución de w_tilde
41   target += normal_id_glm_lupdf(z_obs | X_obs, intercept, b, sig);
```

```
42 }
```

- Finalmente, se generan valores ajustados y predicciones para los z_{ij} mediante el mismo proceso de truncamiento oculto, sólo que ahora se generan un vector `z_all`. De manera análoga, la selección de variables con SSVS permanece igual.

```
43 generated quantities {
44     ...
45     // Ajustados
46     int<lower=1, upper=N_obs+N_mis> j;
47     vector[N_obs+N_mis] z_all;
48     real eta_ij;
49     real<lower=0> sigma_i;
50     for(i in 1:(N_obs+N_mis)){
51         j = group_all[i];
52         eta_ij = mu[j] + X_all[i]*b + (rho[j] .* w[j]) - sigma .* rho[j] .*
53             ↳ sqrt(2/pi()) ./ sqrt(1 - 2/pi() * square(rho[j]));
54         sigma_i = sigma * sqrt(1 - square(rho[j]))./ sqrt(1 - 2/pi() *
55             ↳ square(rho[j]));
56         z_all[i] = normal_rng(eta_ij, sigma_i);
57     }
58 }
```

3.10.3 Modelo probit sesgado latente centrado

La implementación de este modelo es totalmente análoga al modelo probit ordenado, salvo que se omite las restricciones inducidas por el segundo umbral δ_1 , así como los datos `N2` y `n2`.

3.11 Simulación y cantidades de interés

A continuación mostramos las cantidades que consideramos los experimentos siguientes, y que permanecen constantes a lo largo de todas las simulaciones a menos que se indique lo

contrario.

Cuadro 3.3: Cantidades fijadas para los experimentos de simulación. La implementación se realizó en el lenguaje R.

Cantidades	Definición
$M \leftarrow 4$	Número de áreas pequeñas.
$N_{1:M} \leftarrow (1500, 2000, 1500, 2000)$	Tamaño de cada subregión.
$\rho_{1:M} \leftarrow (0.5, 0.75, 0.85, 0.95)$	Parámetro de forma en cada región.
$\sigma^2 \leftarrow 1.5$ (log-normal sesgado)	Parámetro de varianza de todas las regiones.
$\sigma^2 \leftarrow 1.0$ (probit ordenado sesgado)	
$\beta \leftarrow (1.9, -0.5, -1.4, 0.0, 0.0)$	Coeficientes de regresión de todas las regiones.
$\mu_{1:4} \leftarrow (7.0, 6.5, 6.5, 5.0)$ (log-normal sesgado, varios y único int.)	Término constante en cada región.
$\mu_{1:4} \leftarrow (0.6, 0.7, 0.65, 0.95)$ (probit ordenado sesgado, varios y único int.)	
$\mu_{1:4} \leftarrow (0.0, 0.0, 0.0, 0.0)$ (sin int.)	
Porcentaje $\in \{5, 25\}$	Porcentaje de muestreo en todas las regiones.
Iteraciones $\leftarrow 75,000$ (para el método BV)	Número de interacciones para los algoritmos
Iteraciones $\leftarrow 6,000$ (para el método HMC)	ADVI de BV y NUTS de HMC. Para este último, las primeras 5,000 son iteraciones <i>burn-in</i> y las 1,000 restantes son de muestreo, se toma <i>thin</i> de uno.

Cuando se menciona probit ordenado sesgado, también se hace referencia al modelo binario, ya que este se obtiene al considerar cualesquiera dos categorías. Es importante ahondar sobre varios aspectos de las cantidades que se acaban de declarar

- El número de iteraciones no son equivalentes para ambos métodos estudiados, ya que una iteración de HMC realiza exploración o muestreo de la distribución *a posteriori*,

mientras que una iteración VB se optimiza el límite inferior de la evidencia y se actualiza los parámetros con ascenso del gradiente. En general, este último procedimiento es más rápido que generar una muestra de HMC, por lo que podríamos sesgar el tiempo y desempeño de la ejecución al fijar el mismo número de muestras.

- La calidad y velocidad de la implementación VB está influida por tres cantidades, el factor de escala del tamaño de paso η , el número de muestras Monte Carlo M para estimar el gradiente y el número de muestras Monte Carlo para estimar el ELBO. Aunque Kucukelbir et al. (2017) y Stan Development Team (2025b) sugieren valores plausibles para estas cantidades, `grad_samples = 1`, `elbo_samples = 100` y una búsqueda automática de $\eta \in \{0.01, 0.1, 1, 10, 100\}$ que obtiene la convergencia más rápida, cada aplicación puede tener potencialmente un mejor desempeño al calibrar estos hiperparámetros, no obstante, sólo donde se indica lo contrario se modificaron.
- Otro aspecto relevante es que solo se consideran dos porcentajes de muestreo: 5% y 25%, ya que en un escenario real, como en el conjunto de datos estudiado, los tamaños de muestra son usualmente pequeños. Esta decisión también permite hacer más breves y concisos los resultados reportados que enseguida se describen.

Para cada método de estimación, se ajustaron los tres tipos de modelos de regresión propuestos. También, se consideró un escenario de validación, es decir, se dividió la muestra de forma aleatoria en un conjunto de entrenamiento y prueba, con la proporción de 80%-20% respectivamente, esto con el fin de evaluar los pronósticos cuando $n_i > 0$. En el Cuadro 3.4 se muestran las métricas básicas que se calcularon a partir de este ajuste. Las primeras cuatro métricas se calculan para el modelo log-normal sesgado, las métricas cinco a ocho se calculan para el modelo probit sesgado con variable latente, ya que se basan en la construcción de una tabla de valores observados contra ajustados (matriz de confusión), finalmente, para el modelo probit ordenado sesgado con variable latente, se calculan las métricas cinco y nueve.

Enseguida se describe cada uno de los resultados básicos obtenidos mediante estimación.

- Cuando se usa toda la información disponible, se reportaron las medias *a posteriori* de cada parámetro, así como los intervalos de credibilidad empíricos construidos con

los cuantiles 2.5% y 97.5%. De acuerdo con la Sección 2.2, implicitamente se asume la función de pérdida error cuadrado medio, lo cuál es adecuado con respuesta continua. Para los modelos de clasificación, la elección usual es la pérdida 0-1, lo que conduce a que el estimador de Bayes sea la moda *a posteriori*, sin embargo, por simplicidad aún empleamos las medias *a posteriori*.

- De manera adicional, se reportan los modelos seleccionados con el método SSVS y la probabilidad asociada a tal realización. Con el fin de facilitar la comparación, estos se acomodan en gráficos de mosaico.
- Se realiza nuevamente este ajuste pero considerando el caso de entrenamiento-prueba. A partir de aquí, se genera un cuadro comparativo con las métricas descritas previamente, donde se evalúan las predicciones de las observaciones y_{ij}^* en el conjunto de prueba.
- A partir de los resultados obtenidos con la validación, se dibujan gráficos o tablas de aciertos y errores según el modelo que corresponda, a fin de comparar los ajustes.

Cuadro 3.4: Métricas básicas del ajuste. Fuente: elaboración propia. Aquí, las cantidades VP , VN , FP y FN denotan verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

Métrica	Definición	Interpretación
Corr.	$\frac{\sum_{i,j} (y_{ij} - \bar{y}_{ij})(\tilde{y}_{ij} - \bar{\tilde{y}}_{ij})}{\sqrt{\sum_{i,j} (y_{ij} - \bar{y}_{ij})^2 \sum_{i,j} (\tilde{y}_{ij} - \bar{\tilde{y}}_{ij})^2}}$	Mide la asociación entre la colección de pares (y_{ij}, \tilde{y}_{ij}) . Util para datos continuos.
MAE	$\sum_{i=1}^M \sum_{j=1}^{N_i} y_{ij} - \tilde{y}_{ij} $	Mide el promedio de las desviaciones absolutas, está en las mismas unidades que los datos y es más robusto a <i>outliers</i> que la raíz del error cuadrado medio (RMSE), ya que trata a los errores de forma lineal.

Cuadro 3.4: Métricas básicas del ajuste. Fuente: elaboración propia. Aquí, las cantidades VP , VN , FP y FN denotan verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. (Continuación)

Métrica	Definición	Interpretación
RMSE	$\left(\sum_{i=1}^M \sum_{j=1}^{N_i} (y_{ij} - \tilde{y}_{ij})^2 \right)^{1/2}$	Mide la magnitud promedio del error de predicción, está en las mismas unidades que los datos. Además, penaliza más a los errores grandes, misma razón por la cuál es sensible a <i>outliers</i> .
MAPE	$\frac{1}{\sum_i N_i} \sum_{i=1}^M \sum_{j=1}^{N_i} \frac{ y_{ij} - \tilde{y}_{ij} }{y_{ij}}$	Mide el promedio de las desviaciones absolutas expresadas como porcentaje del valor real, por tal motivo no tiene escala.
Exactitud	$\frac{VP + VN}{VP + VN + FP + FN}$	El número de predicciones correctas. Puede resultar inapropiado si las clases están desbalanceadas.
Verdaderos positivos (Recuperación)	$\frac{VP}{VP + FN}$	Proporción de verdaderos positivos detectados del total de positivos.
Verdaderos negativos (Especificidad)	$\frac{VN}{VN + FP}$	Análogo a la recuperación.

Cuadro 3.4: Métricas básicas del ajuste. Fuente: elaboración propia. Aquí, las cantidades VP , VN , FP y FN denotan verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. (Continuación)

Métrica	Definición	Interpretación
F1-Score	$\frac{2TP}{2TP + FP + FN}$	Media armónica entre la precisión y la recuperación. ⁹ Apropiado para clases desbalanceadas.
τ de Kendall ¹⁰	$\frac{\sum_{(i,j) \preceq (i',j')} \text{sgn}(y_{ij} - y_{i'j'})\text{sgn}(\tilde{y}_{ij} - \tilde{y}_{i'j'})}{\binom{n}{2}}$	Mide la asociación entre la colección de pares (y_{ij}, \tilde{y}_{ij}) . Útil para datos ordinales.

Para el modelo de regresión con respuesta continua, se consideraron los siguientes valores iniciales para el muestreador HMC y el proceso de optimización BV.

Cuadro 3.5: Valores iniciales para la simulación.

Parámetro (s)	Valor inicial	Descripción
μ_i, σ^2, λ	$\hat{\lambda}\mathbf{1}_M, \hat{\mu}\mathbf{1}_M, \sigma^2$	$\hat{\lambda}, \hat{\sigma}^2$ y $\hat{\mu}$ se obtuvieron mediante máxima verosimilitud al ajustar el modelo $y_{ij} = \mu + e_{ij}$, con $e_{ij} \sim SN(0, \sigma^2, \lambda)$. Se empleó la función <code>selm</code> de la librería <code>sn</code> de R.
β	$\hat{\beta}$	Se obtiene a partir del ajuste del modelo de regresión $y_{ij} = \mu + \mathbf{x}_{ij}^T \boldsymbol{\beta} + e_{ij}$, con $e_{ij} \sim N(0, \sigma^2)$.

Por razones que se verán más adelante, para los dos modelos de clasificación propuestos, se consideran tres submodelos: uno que incluye interceptos o términos independientes para cada región, otro con un único intercepto en todas la regiones y uno más que no incluye interceptos. En el modelo log-normal sesgado se hace una excepción y únicamente se considera el caso con interceptos en cada área. De igual modo, es relevante mencionar que se trabaja con covariables estandarizadas en todos los modelos.

⁹Se define la relación de orden $(i, j) \preceq (i', j')$ si y solo si $i < i'$ o bien $i = i'$ y $j < j'$. Por ejemplo, $(1, 2) \preceq (1, 3), (2, 1) \preceq (3, 1)$.

¹⁰Se define a la precisión como $VP/(VP + FP)$.

3.12 Caso de estudio: medición de la pobreza

En este apartado, se describe una aplicación real para los tres modelos de regresión en áreas pequeñas propuestos. En síntesis, se estima el ingreso corriente total per cápita (ICTPC) en los hogares de la Ciudad de México, a nivel municipal o alcaldía. Este estudio forma parte de uno de los rubros considerados para construir la medición multidimensional de la pobreza (MMP) en México.

La medición de la pobreza adquiere relevancia porque permite identificar con mayor precisión las carencias que enfrentan los grupos afectados y orientar intervenciones públicas más efectivas. En México, esta medición estuvo durante años a cargo del Consejo Nacional de Evaluación de la Política de Desarrollo Social (Coneval), cuya metodología retomaba elementos del enfoque propuesto por Alkire y Foster; actualmente, la responsabilidad recae en el Instituto Nacional de Estadística y Geografía (INEGI). Desde una perspectiva integral, la pobreza se relaciona con condiciones de vida que vulneran la dignidad humana, restringen el ejercicio de derechos y obstaculizan la satisfacción de necesidades básicas, dificultando así la plena participación social de las personas Coneval ([2023](#)).

Durante décadas, el producto interno bruto per cápita fue utilizado como indicador principal del bienestar económico de la población; sin embargo, su capacidad explicativa resulta limitada para capturar la complejidad de las condiciones de vida contemporáneas. La evaluación del bienestar requiere enfoques más amplios que reconozcan que la pobreza no puede reducirse únicamente al nivel de ingresos, pues intervienen dimensiones que no se reflejan de manera directa en la capacidad monetaria de los hogares. Desde esta perspectiva multidimensional, surge la necesidad de integrar indicadores que permitan comprender de manera más completa las privaciones que experimentan los individuos. Contar con mediciones adecuadas no solo mejora el diagnóstico social, sino que constituye un elemento clave para diseñar políticas públicas que respondan de manera efectiva a las distintas manifestaciones de la pobreza Sáenz Vela ([2020](#)).

3.12.1 Antecedentes

El Consejo Nacional de Evaluación de la Política de Desarrollo Social (Coneval) fue un organismo autónomo y descentralizado cuya tarea principal consistía en la generación e implementación de diversas técnicas estadísticas para cuantificar la pobreza desde varios ámbitos, abordando así la medición multidimensional de la pobreza (MMP) en México. La última edición sobre la MMP que realizó el Coneval fue en 2022, siendo ahora el Instituto Nacional de Estadística y Geografía (INEGI) el organismo encargado de darle continuidad a esta medición. En 2024 el INEGI publicó su primera edición de la MMP tomando como base los mismos lineamientos propuestos por el Coneval. Estos lineamientos tienen sustento en los artículos 36 y 37 de la Ley General de Desarrollo Social (LGDS) y determinan que la MMP se construye a partir de dos fuentes de datos oficiales y productos generados de estas: (1) el Censo de Población y Vivienda (CPV) y (2) la Encuesta Nacional de Ingresos y Gastos en los Hogares (ENIGH).

Los lineamientos establecen que la MMP debe tener una periodicidad mínima bienal en los niveles estatal y nacional, y quinquenal a nivel municipal. Para este último caso resulta complejo realizar la MMP dado que la ENIGH solo es representativa a nivel estatal y nacional, y por tanto, se hace uso de otros métodos de estimación. En 2020, Coneval construyó la MMP a nivel municipal a partir del ajuste de modelos en áreas pequeñas para realizar predicciones sobre atributos de interés, entre ellos, el ingreso corriente total per cápita, mediante la técnica del mejor predictor heterocedástico empírico, *empirical best predictor heterokedastic* (EBPH), la cual está basada en la teoría de modelos lineales mixtos y es equivalente al planteamiento en la [Ecuación 3.4](#), pero permitiendo que la varianza de los hogares sea distinta entre sí, de ahí su nombre.

El objetivo principal de la MMP es estimar el número y porcentaje de la población en situación de pobreza. Para realizar esta estimación, se establece que la pobreza se compone de dos espacios analíticos principales: la dimensión de derechos sociales o carencias, que consta de seis indicadores dicotómicos: rezago educativo, acceso a los servicios de salud, acceso a la seguridad social, calidad y espacios de la vivienda, acceso a los servicios básicos

en la vivienda y acceso a la alimentación; y el espacio de bienestar económico, cuyo indicador es el ICTPC. Al incorporar estos dos rubros, se clasifica a la población en uno de cuatro cuadrantes de pobreza (INEGI 2025; Coneval 2023):

- I. Población en situación de pobreza multidimensional: población con ingreso inferior al valor monetario de las líneas de pobreza por ingresos¹¹ (LPI) respectivas y con al menos una carencia social, en este cuadrante, la población se desagrega en dos grupos:
 - I'. Población en situación de pobreza extrema: su ingreso es inferior al valor monetario de las líneas de pobreza extrema por ingresos¹² (LPEI) y presenta tres o más carencias sociales.
 - I''. Población en situación de pobreza moderada: percibe un ingreso inferior a las LPI y presenta entre una y dos carencias.
- II. Población vulnerable por carencias sociales: población con una o más carencias sociales, pero cuyo ingreso es igual o superior a las LPI respectivas,
- III. Población vulnerable por ingresos: población sin carencias sociales y con ingreso inferior a las LPI respectivas,
- IV. Población no pobre multidimensional y no vulnerable: población con ingreso igual o superior a las LPI respectivas y sin ninguna carencia social.

Es relevante señalar que los indicadores de seguridad social, alimentación e ingreso no pueden obtenerse directamente a partir de información del CPV, por lo que necesitan ser estimados a través de modelos de regresión en áreas pequeñas (Coneval 2021). En este sentido, el modelo propuesto permite abordar la dimensión de la pobreza que corresponde a la vulnerabilidad por ingresos. Así, el objetivo es realizar predicciones del ICTPC para todos los hogares no muestreados de los cuales se tiene información auxiliar proveniente del CPV 2020, y para tal propósito, se emplean y procesan los datos siguiendo los criterios establecidos por el Coneval. De este modo, se explorar una alternativa para desarrollar esta parte de la MMP municipal para el año 2025, el siguiente periodo quinquenal con respecto a la medición previa.

¹¹Valor monetario de la canasta alimentaria más el valor monetario de la canasta no alimentaria: 3296.92 y 4564.97 pesos mexicanos para los ámbitos rural y urbano.

¹²Valor de la canasta alimentaria: 1800.55 y 2354.65 pesos mexicanos para los ámbitos rural y urbano.

3.12.2 Fuentes de información y procesamiento

Las estimaciones con modelos de regresión en áreas pequeñas se realizaron empleando datos oficiales: el CPV 2020 y la ENIGH 2025. Esta última consta de diecisiete productos o tablas: viviendas, hogares, población, gastos monetarios y no monetarios en los hogares, entre otras. Así mismo, la desagregación más pequeña en estas tablas es a nivel vivienda, seguido por los hogares y finalmente se encuentra la población. Por su parte, se dispone de datos del CPV hasta nivel de hogar.

Para actualizar el conjunto de datos de hogares con la información proveniente de la ENIGH 2024, fue necesario realizar ligeros ajustes al código Stata que general las tablas para realizar la MMP municipal de la pobreza¹³¹⁴¹⁵. De este modo, el conjunto de datos conformado por la variable objetivo e información de covariables auxiliares (conteos, indicadores, categorías) fue generado de acuerdo a los mismos criterios que el Coneval usó en el año 2020 para la medición municipal de la pobreza.

Tras integrar las fuentes de información, el conjunto de datos para la Ciudad de México se compone de 79,881 observaciones a nivel hogar, de las cuáles únicamente se tiene registro del ICTPC en 2,329 hogares adicionales. A su vez, la base de datos consta de 52 covariables continuas (conteos, porcentajes, categorías ordenadas), 81 covariables binarias (indicadores de carencias) y 7 covariables categóricas (más de dos niveles no ordenados). Un análisis de componentes principales sobre las 52 covariables continuas -no reportado aquí- señala que es posible reducir la dimensión a 26 componentes recuperando hasta el 95.27% de la estructura de covarianzas. Se empleará esta técnica para reducir la dimensionalidad, no obstante, esto implica que se pierde la interpretación de las covariables continuas, en cambio, aún se dispone de los indicadores binarios. En el [Anexo 3](#) se muestra una lista con las covariables incluidas. Aunque el planteamiento del modelo propuesto en la [Ecuación 3.51](#) no requiere que la matriz diseño sea de rango completo, se optó por remover los indicadores binarios que

¹³Pobreza a nivel municipal: <https://www.coneval.org.mx/Medicion/Paginas/Pobreza-municipio-2010-2020.aspx>

¹⁴MMP estatal 2020 (Coneval): https://www.coneval.org.mx/Medicion/MP/Paginas/Pobreza_2022.aspx

¹⁵MMP 2024 estatal (INEGI): <https://www.inegi.org.mx/desarrollosocial/pm/>

generan dependencias lineales perfectas.

El ICTPC se obtiene a partir de la medición multidimensional de la pobreza 2024 elaborada por el INEGI, para su construcción, se considera el cociente entre ingreso corriente total (ICT) del hogar y el tamaño del hogar ajustado. El ingreso corriente se compone del ingreso monetario: salarios, transferencias, rentas, entre otros; y el ingreso no monetario: pagos y regalos en especie. La ENIGH levanta registro de los ingresos percibidos por los individuos encuestados hasta seis meses previos a la entrevista. El ICT se obtiene sumando los ingresos monetarios y no monetarios promedio percibidos durante este periodo de tiempo por cada integrante del hogar, analizados a precios constantes del 2018. Luego, para obtener el ICTPC, se re-escala el ICT de acuerdo al tamaño del hogar ajustado, esto es, a cada integrante del hogar se le asigna un peso entre (0, 1) de acuerdo al grupo etario al que pertenece, estos pesos están distribuidos como se indica en el Cuadro 3.6

Cuadro 3.6: Tamaño del hogar ajustado. Fuente: elaboración propia con información del programa de cómputo de la MMP 2024.

Tamaño ajustado	Condición
0.7031	Menor de seis años
0.7382	Mayor de seis y menor de trece años
0.7057	Mayor de trece años y menor de diecinueve años
0.9945	Mayor de veinte años y datos perdidos
1.000	Único integrante del hogar

De igual modo, cuando se trabaja con los modelos probit y probit ordenado, se discretiza la variable ICTPC de acuerdo a la LPI y LPEI que corresponde al tipo de entorno: rural o urbano.

- En el caso binario, definimos la presencia de carencias como el caso de interés, es decir, $y_{ij} = 1$; por tanto, se define

$$y_{ij} = \begin{cases} 1 & \text{si el contexto } (i, j) \text{ es urbano y } \tilde{y}_{ij} < \text{LPI urbana} \\ 1 & \text{si el contexto } (i, j) \text{ es rural y } \tilde{y}_{ij} < \text{LPI rural} \\ 0 & \text{de otro modo} \end{cases} \quad (3.58)$$

-
- En el caso ordinal, de forma natural se obtienen tres categorías de acuerdo a los umbrales de ingresos: por debajo de la línea de pobreza extrema, por debajo de la línea de pobreza y sin esta carencia. De forma análoga al caso binario, se define

$$y_{ij} = \begin{cases} 0 & \text{si el contexto } (i, j) \text{ es urbano y } \tilde{y}_{ij} < \text{LPEI urbana} \\ 0 & \text{si el contexto } (i, j) \text{ es rural y } \tilde{y}_{ij} < \text{LPEI rural} \\ 1 & \text{si el contexto } (i, j) \text{ es urbano y } \tilde{y}_{ij} < \text{LPI urbana} \\ 1 & \text{si el contexto } (i, j) \text{ es rural y } \tilde{y}_{ij} < \text{LPI rural} \\ 2 & \text{de otro modo} \end{cases}, \quad (3.59)$$

donde \tilde{y}_{ij} es el ICTPC para la observación (i, j) . Como nota general, en el modelo probit ordenado, es posible distinguir cuando una respuesta es mayor o menor que otra, en este caso, la categoría cero se considera más pequeña, y conforme la respuesta latente crece, se obtiene la categoría uno y dos, por su lado, el modelo binario no tiene orden en la respuesta, sino, una cualidad de interés.

3.12.3 Ruta de trabajo

Como se mencionó previamente, el objetivo de los modelos en áreas pequeñas es generar pronósticos para aquellas regiones con tamaños de muestra insuficientes para realizar estimación directa, e incluso donde no se dispone de observaciones sobre la variable objetivo ($n_i = 0$). No obstante, en este caso se tienen observaciones para todos los $M = 16$ dominios. Por tanto, en este escenario se propuso la siguiente ruta de trabajo:

- Entrenar al modelo log-normal sesgado con información de catorce regiones y predecir las dos restantes. De forma concreta, se propone predecir el log-ICTPC para las alcaldías Miguel Hidalgo y Milpa Alta, ya que en estas dos demarcaciones se registra el ICTPC promedio más grande y pequeño, de esta manera, es posible comparar los pronósticos con valores reales bien diferenciados a fin de evaluar su desempeño.
- Repetir el paso anterior con el método Hamiltoniano Monte Carlo (HMC) a fin de

comparar las métricas de ajuste y el tiempo de ejecución promedio.

- Ajustar los tres modelos propuestos con todas las observaciones disponibles y empleando ambos métodos de inferencia, a fin de calcular y reportar estimaciones *a posteriori*; además, con esta muestra aproximada es posible estimar del porcentaje de la población con ICTPC por debajo de la línea de pobreza. En este caso, como se dispone de información en todas las regiones ($n_i > 0$), fijamos $\mu_0 = 0$ y $\sigma_\mu^2 = 100^2$ en la Ecuación 3.51, es decir, no se emplea la estrategia de agrupamiento.
- Repetir el paso anterior, pero únicamente realizar el ajuste con la muestra observada dividida aleatoriamente en un conjunto de 80% entrenamiento y 20% prueba. En este contexto, únicamente se consideran las métricas de ajuste, el tiempo de ejecución y los gráficos de dispersión o tablas de aciertos y errores entre los valores observados y los ajustados.

Para el ajuste HMC se emplea una única cadena, en este caso, Stan calcula una versión del estadístico de Gelman-Rubin \hat{R} que divide m cadenas paralelas en $2m$ cadenas para evaluar la convergencia. De este modo, es posible estimar esta cantidad incluso cuando se usa una única trayectoria de muestreo HMC. Con respecto al tercer punto, para calcular los porcentajes de la población, primero se identifica a los hogares con ingresos inferiores a las líneas de pobreza, de acuerdo al ámbito, urbano o rural, y a continuación se multiplica el tamaño del hogar por su factor de expansión asociado, finalmente, se agrupa el número de personas con estas carencias. De este modo, se obtiene una estimación del total poblacional con vulnerabilidad por ingresos, con la particularidad de que esta medición está a nivel municipal. Para calcular estos porcentajes de la población, sólo se tomó en cuenta los pronósticos generados a partir de la información del CPV 2020, es decir que se excluyó a los valores del ICTPC que brinda la MMP 2024 ya que los factores de expansión de ambas fuentes de información representan a toda la ciudad: en otras palabras, no se duplicó la información.

Cuando no se tiene información en todas las regiones de estudio, es recomendable calibrar o reponerderar los factores de expansión para mejorar la precisión estadística, sin embargo, en este caso se contó con información completa de las $M = 16$ regiones y esta técnica no se llevó a cabo. Por otro lado, el Coneval realizaba ajustes a los totales poblacionales de cada región

a fin de que la MMP a nivel municipal coincidiera con la medición estatal, así mismo, dado que en el CPV 2020 se omitieron algunos registros con datos nulos, la suma de los factores de expansión en esta fuente de datos no suma al total poblacional de la entidad en el año 2020. En este mismo sentido, quizás sería posible calibrar los factores de expansión a fin de incluir la información de la MMP 2024 para calcular el porcentaje de la población bajo alguna línea de pobreza, no obstante, no se exploró esta alternativa.

CAPÍTULO 4. RESULTADOS

En este capítulo mostramos los principales resultados del ajuste de los dos modelos de regresión Bayesiana en áreas pequeñas descritos en el [Capítulo 3](#): un modelo para datos continuos y un modelo para datos ordinales, a los cuales denominamos log-normal sesgado y probit ordenado sesgado latente. Para realizar inferencia sobre los parámetros y variables latentes, se consideran los métodos Hamiltoniano Monte Carlo y variacional Bayesiano de forma fija, ambos descritos en el [Capítulo 2](#).

En las secciones [4.1](#) y [4.2](#), se exhiben dos estudios de simulación preliminares sobre la generación de variables respuesta y_{ij} binarias y el ajuste del modelo log-normal sesgado sin restricciones en los ρ_i . Luego, en las secciones [4.3](#) a [4.5](#) se muestran los resultados para los experimentos de simulación descritos en la [Sección 3.11](#), donde se proponen $M = 4$ regiones y pocos predictores para cada clase de modelo. En estos apartados, se reportan y comparan entre sí las estimaciones de los parámetros para ambos métodos de inferencia, con los valores reales, así como su tiempo de cómputo.

Posterior a esto, en la [Sección 4.6](#) se inicia presentando estadísticos descriptivos acerca del conjunto de datos del ingreso corriente total per cápita (ICTPC) de la Ciudad de México en 2025. Después, se ajustan los tres modelos de regresión propuestos al logaritmo de los ingresos y a dos conjuntos discretizados de este. En las secciones [4.7](#) a [4.8](#), se reportan estimaciones, tiempos de ejecución y los porcentajes de la población bajo alguna línea de pobreza. Se concluye el estudio sobre los ingresos en la Ciudad revelando los parámetros β que fueron seleccionados más del 75% de las veces por cada modelos de regresión y por ambos métodos.

El uso de negritas sobre los distintos resultados, indica que la estimación obtenida con ese método -indicado en la columna- está más cerca del valor real, así mismo, las negritas indican que cierta métrica es mejor en un método que en otro, por ejemplo, correlación con la respuesta, tiempo de ejecución, etcétera.

4.1 Efecto de ρ_i con μ_i

Con el propósito de evaluar el efecto o la interacción entre estos dos parámetros, se desarrolla el siguiente experimento sencillo. Dada la malla de valores generada a partir de $\rho_i \in \pm\{0.01, 0.02, \dots, 0.99\}$ y $\mu_i \in \pm\{0.0, 0.1, \dots, 2.5\}$. En cada punto de este rástre se generan $n = 500$ variables aleatorias binarias y_i a través de dos métodos equivalentes y siguiendo el enfoque de variable latente:

1. Por medio del signo de z_i : $y_i = 1(z_{ij} > 0)$, generando $z_i \sim SN(\mu_i, 1, \lambda_i)$.
2. Con una variable indicadora Bernoulli: $y_i \sim Bern(p_i)$, cuya probabilidad de éxito está dada por

$$p_i = \mathbb{P}(Y_i = 1) = \mathbb{P}(Z_i > 0) = 1 - \Phi_{SN}(0 - \mu_i; \lambda_i) \quad (4.1)$$

En la [Figura 4.2](#) se muestra la proporción de valores $\frac{1}{n} \sum_{i=1}^n y_i$, es decir, aquellos iguales a uno, para cada punto de la cuadrícula. No consideramos el uso de parámetros centrados.

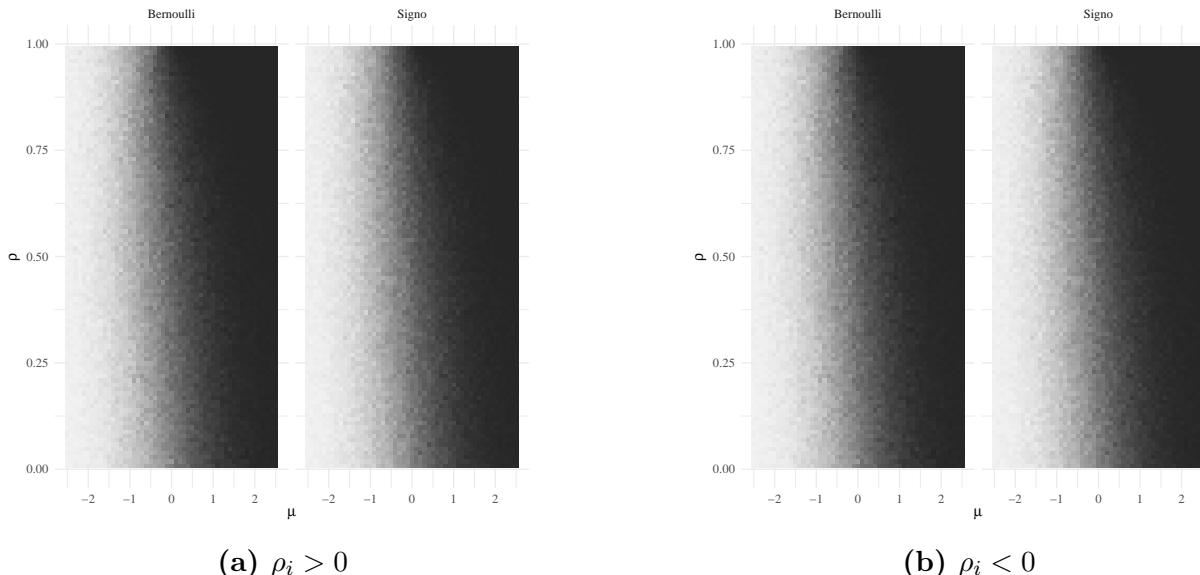


Figura 4.1: Proporción de valores $y_{ij} = 1$ simulados.

De forma adicional, se dibuja la función liga para el modelo probit sesgado latente para los valores $\lambda_i \in \{0, 1, 3\}$. En la [Sección 3.3](#) se obtiene su expresión, por comodidad, aquí se

repite, además, $\eta_{ij} = \mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}$.

$$\mathbb{P}(y_{ij} = 1 \mid p_{ij}) = 1 - \Phi_{SN}(-\eta_{ij}, \lambda_i).$$

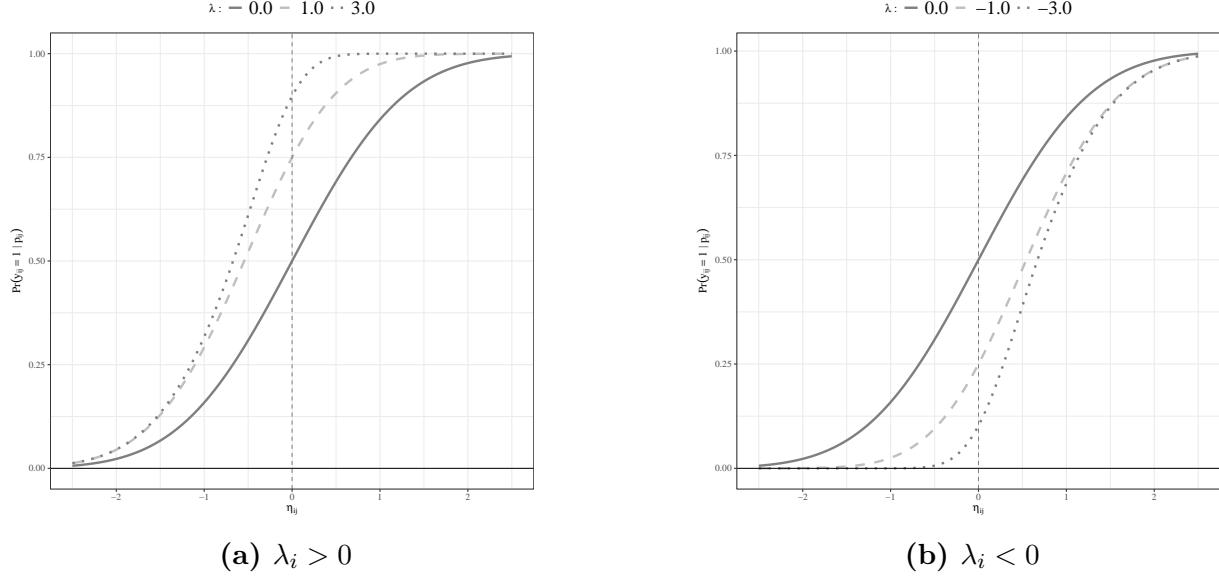


Figura 4.2: Función liga del modelo probit ordenado sesgado.

4.2 Ajuste de los modelos de regresión sin restricción en ρ_i

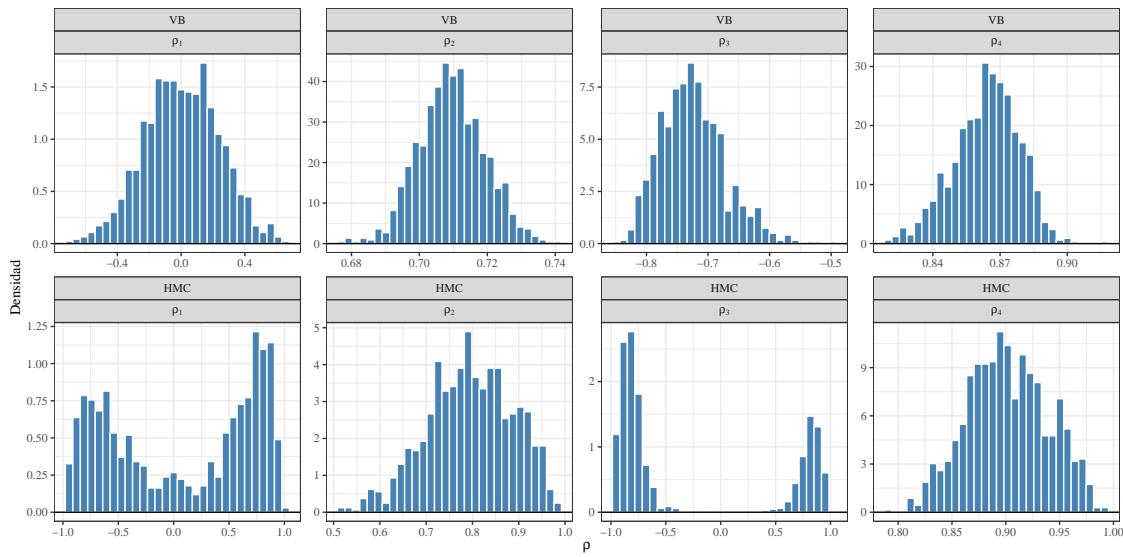
A continuación, se presenta un caso especial de estimación donde se permite que $\rho_i \in (-1, 1)$, es decir, el parámetro de forma λ_i toma valores en \mathbb{R} . Se considera un único intercepto, y además, se omite la salida de los parámetros μ_i y $\boldsymbol{\beta}$ para centrarnos en la inferencia sobre ρ_i . El porcentaje de muestreo fue del 25%.

4.2.1 Modelo log-normal sesgado

Cuadro 4.1: Modelo log-normal con único intercepto en cada área pequeña. Porcentaje de muestreo: 25%.

Par.	Valor real	Bayes Variacional			Hamiltoniano MC		
		Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%
ρ_1	-0.500	0.005	-0.455	0.449	0.082	-0.898	0.925
ρ_2	0.750	0.709	0.692	0.728	0.795	0.597	0.951
ρ_3	-0.850	-0.725	-0.812	-0.611	-0.261	-0.948	0.938
ρ_4	0.950	0.864	0.833	0.889	0.902	0.830	0.971

Figura 4.3

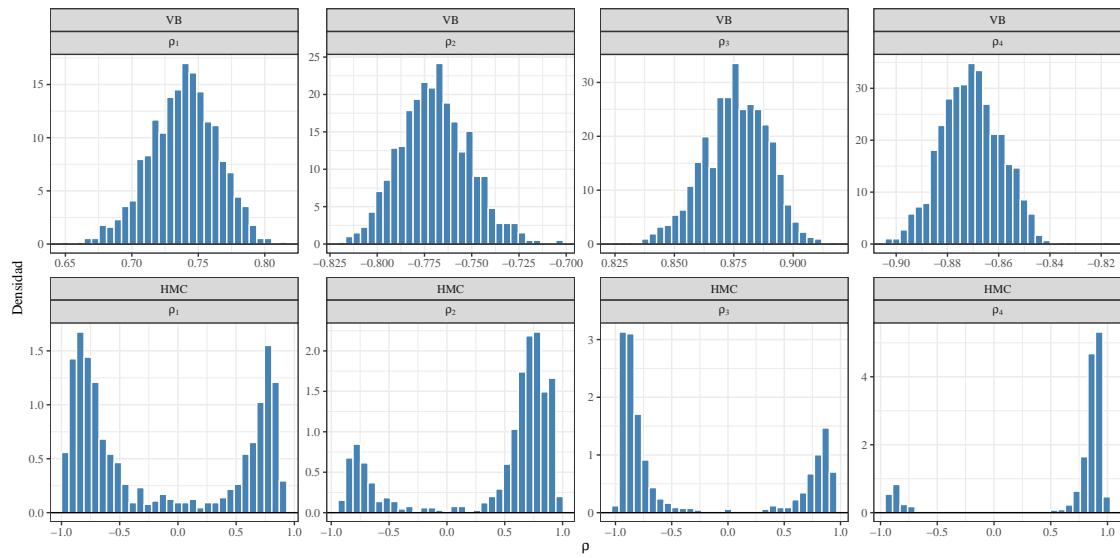


4.2.2 Modelo probit sesgado con variable latente

Cuadro 4.2: Modelo probit sesgado con único intercepto en cada área pequeña. Porcentaje de muestreo: 25%.

Par.	Valor real	Bayes Variacional			Hamiltoniano MC		
		Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%
ρ_1	-0.500	0.739	0.686	0.786	-0.141	-0.943	0.859
ρ_2	0.750	-0.770	-0.804	-0.731	0.404	-0.844	0.933
ρ_3	-0.850	0.876	0.847	0.900	-0.310	-0.960	0.919
ρ_4	0.950	-0.871	-0.893	-0.848	0.652	-0.907	0.962

Figura 4.4

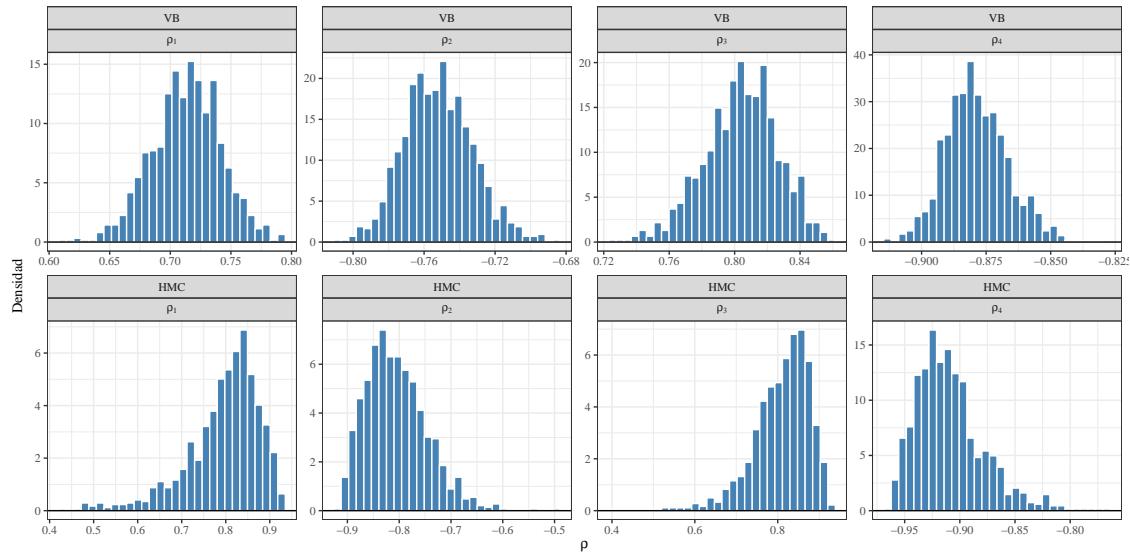


4.2.3 Modelo probit ordenado sesgado con variable latente

Cuadro 4.3: Modelo probit ordenado sesgado con único intercepto en cada área pequeña.
Porcentaje de muestreo: 25%.

Par.	Bayes Variacional				Hamiltoniano MC			
	Valor real	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	
ρ_1	-0.500	0.713	0.653	0.768	0.795	0.575	0.912	
ρ_2	0.750	-0.752	-0.788	-0.711	-0.805	-0.896	-0.677	
ρ_3	-0.850	0.804	0.757	0.843	0.810	0.646	0.908	
ρ_4	0.950	-0.879	-0.900	-0.853	-0.909	-0.954	-0.831	

Figura 4.5



4.3 Ajuste del Modelo log-normal sesgado, datos simulados

4.3.1 Modelos con interceptos en cada área pequeña

Cuadro 4.4: Modelo log-normal sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.

Par.	Bayes Variacional				Hamiltoniano MC			
	Valor real	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	
ρ_1	0.500	0.169	0.021	0.509	0.615	0.038	0.978	
ρ_2	0.750	0.306	0.048	0.713	0.782	0.151	0.984	
ρ_3	0.850	0.877	0.792	0.938	0.938	0.834	0.993	
ρ_4	0.950	0.888	0.835	0.929	0.942	0.867	0.993	
σ^2	1.500	1.375	1.266	1.496	1.992	1.300	4.300	
μ_1	6.000	6.651	6.344	6.942	6.393	3.823	8.831	
μ_2	7.000	7.732	7.458	7.995	7.542	3.966	10.236	
μ_3	6.500	7.902	7.639	8.169	7.484	3.850	10.683	
μ_4	5.000	6.513	6.311	6.719	6.352	3.749	9.001	
β_1	1.900	1.925	1.802	2.046	1.915	1.799	2.029	
β_2	-0.500	-0.487	-0.616	-0.366	-0.478	-0.616	-0.361	
β_3	-1.400	-1.349	-1.471	-1.222	-1.362	-1.466	-1.244	
β_4	0.000	0.013	-0.048	0.072	0.017	-0.035	0.073	
β_5	0.000	-0.001	-0.057	0.060	0.001	-0.060	0.060	

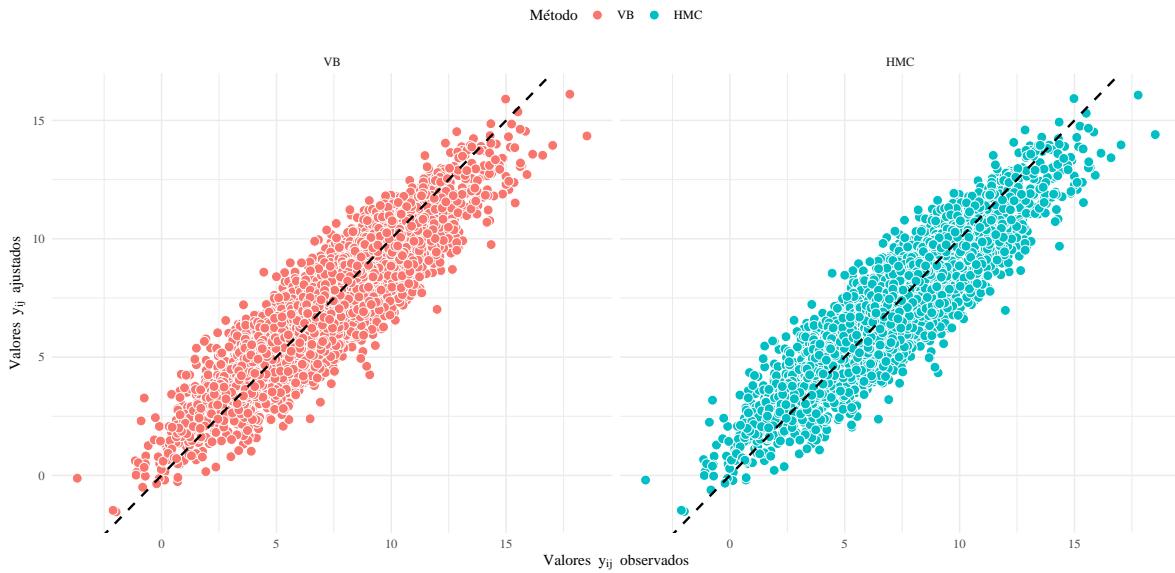


Figura 4.6: Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC

Cuadro 4.5: Métricas del ajuste del modelo log-normal sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.

Métrica	Bayes Variacional	Hamiltoniano MC
Corr.	0.905	0.905
MAE	0.939	0.940
RMSE	1.183	1.185
MAPE	20.039	20.126
Tiempo (s)	11.390	58.530

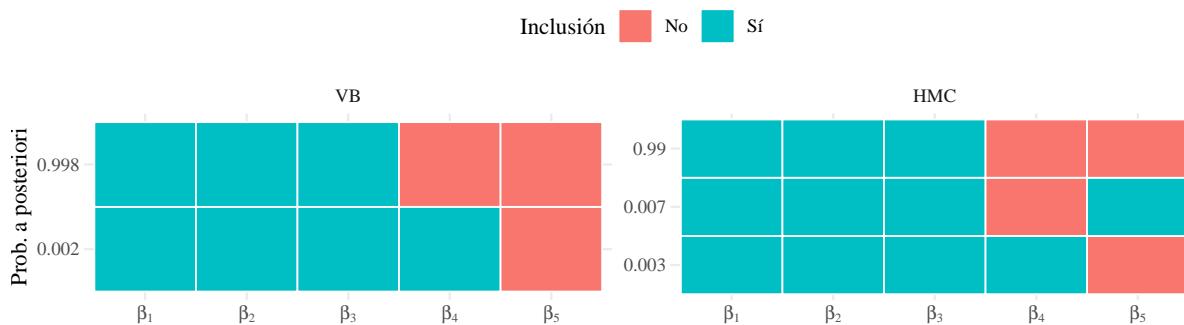


Figura 4.7: Modelo log-normal sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.

Cuadro 4.6: Modelo log-normal sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.

Par.	Valor real	Bayes Variacional			Hamiltoniano MC		
		Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%
ρ_1	0.500	0.111	0.013	0.374	0.552	0.033	0.972
ρ_2	0.750	0.578	0.411	0.726	0.756	0.386	0.979
ρ_3	0.850	0.724	0.606	0.818	0.832	0.633	0.982
ρ_4	0.950	0.845	0.808	0.878	0.898	0.809	0.987
σ^2	1.500	1.321	1.275	1.368	1.705	1.253	3.635
μ_1	6.000	6.616	6.481	6.757	6.499	2.964	8.508
μ_2	7.000	7.944	7.833	8.049	7.849	4.710	10.642
μ_3	6.500	7.542	7.403	7.688	7.385	4.408	9.568
μ_4	5.000	6.370	6.275	6.477	6.143	3.014	8.714
β_1	1.900	1.889	1.828	1.947	1.886	1.833	1.940
β_2	-0.500	-0.539	-0.601	-0.481	-0.534	-0.592	-0.475
β_3	-1.400	-1.388	-1.454	-1.330	-1.389	-1.446	-1.334
β_4	0.000	-0.020	-0.067	0.028	-0.015	-0.060	0.028
β_5	0.000	0.030	-0.020	0.077	0.028	-0.014	0.071

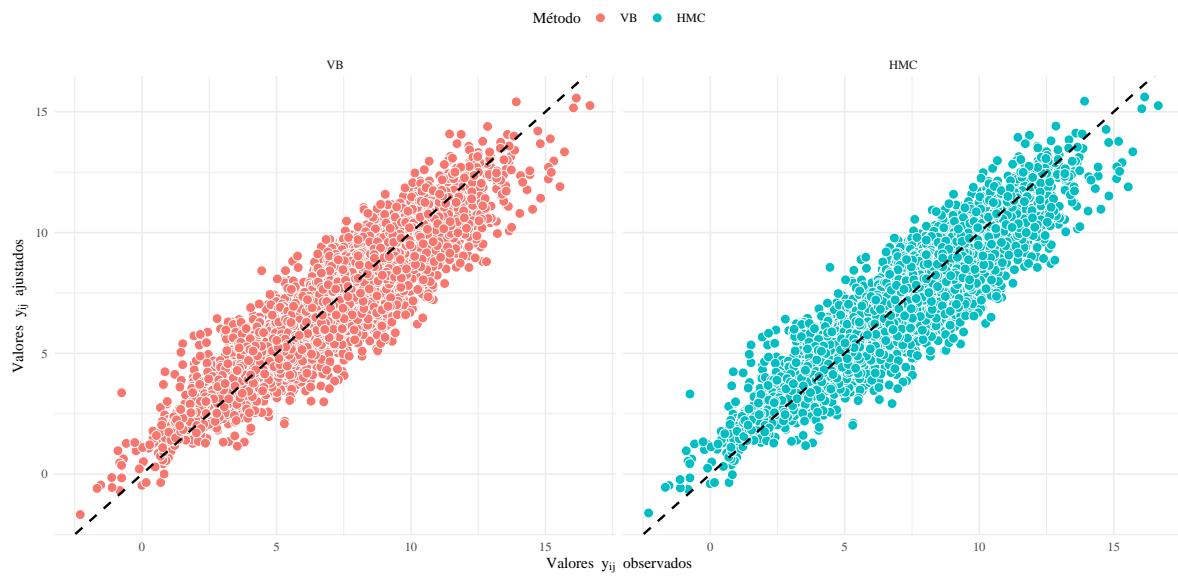


Figura 4.8: Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC

Cuadro 4.7: Métricas del ajuste del modelo log-normal sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.

Métrica	Bayes Variacional	Hamiltoniano MC
Corr.	0.906	0.907
MAE	0.915	0.913
RMSE	1.158	1.156
MAPE	21.579	21.164
Tiempo (s)	31.110	442.910

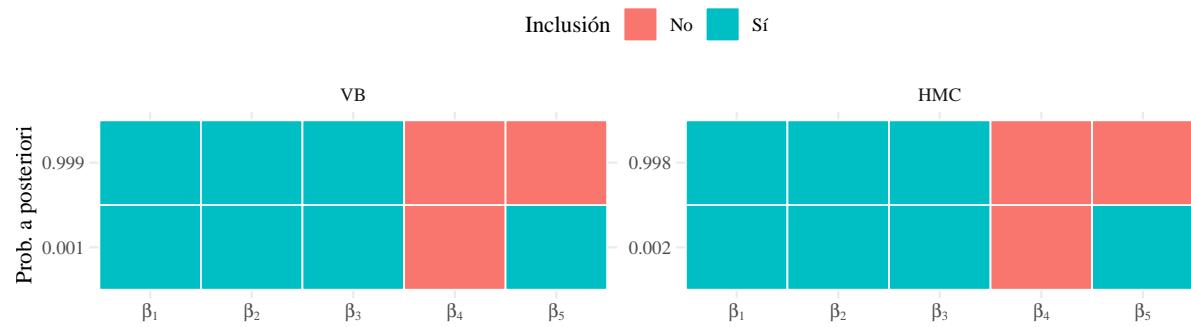


Figura 4.9: Modelo log-normal sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.

4.4 Ajuste del Modelo probit sesgado con variable latente, datos simulados

4.4.1 Modelos sin interceptos en cada área pequeña

Cuadro 4.8: Modelo probit sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 5%.

Par.	Valor real	Bayes Variacional			Hamiltoniano MC		
		Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%
ρ_1	0.500	0.208	0.034	0.565	0.612	0.078	0.933
ρ_2	0.750	0.152	0.019	0.458	0.462	0.031	0.890
ρ_3	0.850	0.470	0.211	0.743	0.904	0.456	0.996
ρ_4	0.950	0.482	0.318	0.648	0.741	0.343	0.937
β_1	1.900	2.741	2.628	2.845	1.859	1.339	2.292
β_2	-0.500	-0.038	-0.099	0.027	-0.355	-0.577	-0.019
β_3	-1.400	-2.038	-2.143	-1.930	-1.444	-1.804	-1.056
β_4	0.000	0.011	-0.053	0.073	0.035	-0.048	0.221
β_5	0.000	0.012	-0.049	0.072	0.021	-0.040	0.127

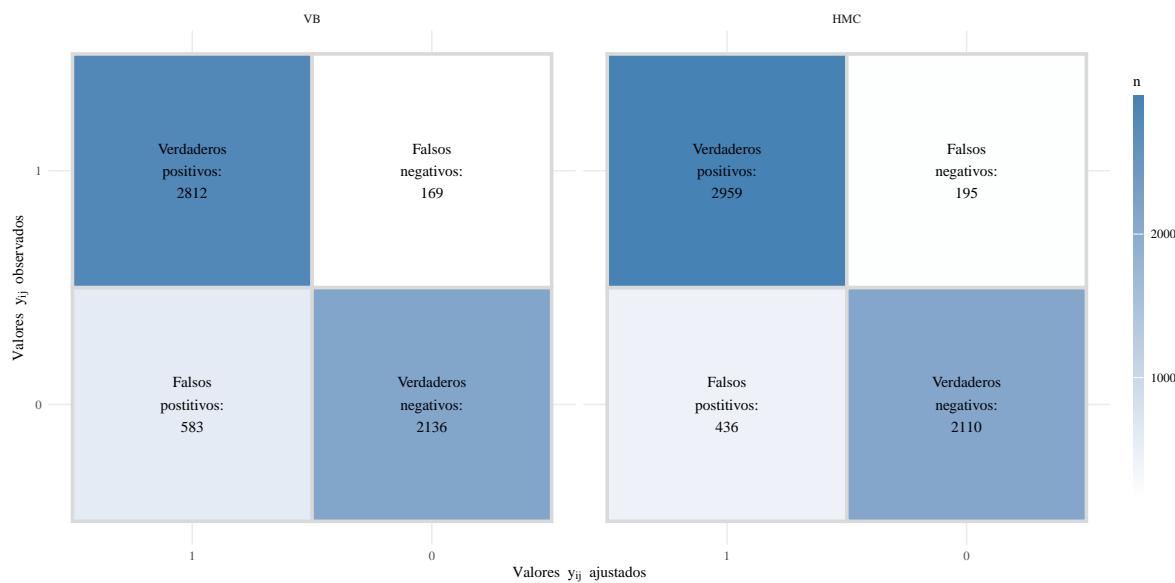


Figura 4.10: Matriz de confusión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC

Cuadro 4.9: Métricas del ajuste del modelo probit sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 5%.

Métrica	Bayes Variacional	Hamiltoniano MC
Exactitud	0.868	0.889
Verdaderos positivos	0.943	0.938
Verdaderos negativos	0.943	0.938
F1-Score	0.882	0.904
Tiempo (s)	17.720	28.580

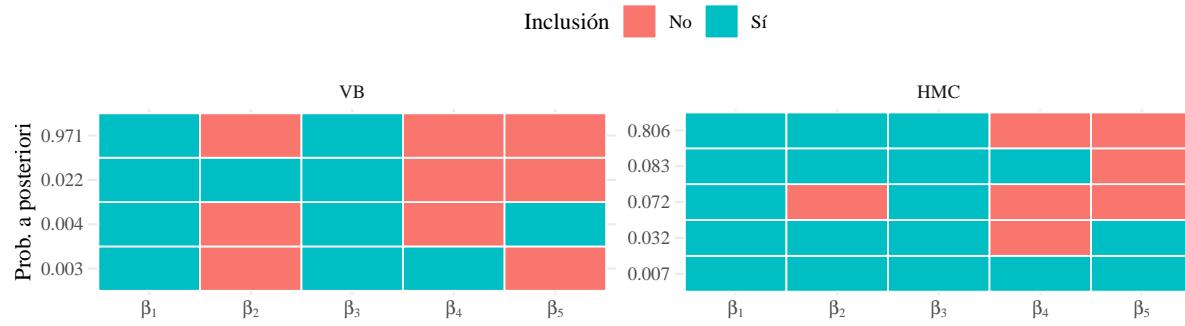


Figura 4.11: Modelo probit sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 5%.

Cuadro 4.10: Modelo probit sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 25%.

Par.	Valor real	Bayes Variacional			Hamiltoniano MC		
		Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%
ρ_1	0.500	0.299	0.246	0.356	0.582	0.258	0.863
ρ_2	0.750	0.365	0.331	0.401	0.731	0.373	0.904
ρ_3	0.850	0.427	0.367	0.493	0.725	0.389	0.925
ρ_4	0.950	0.423	0.379	0.466	0.864	0.640	0.957
β_1	1.900	3.329	3.279	3.379	2.016	1.602	2.320
β_2	-0.500	-0.814	-0.865	-0.765	-0.502	-0.615	-0.385
β_3	-1.400	-2.547	-2.596	-2.495	-1.524	-1.786	-1.175
β_4	0.000	-0.005	-0.045	0.038	-0.010	-0.067	0.040
β_5	0.000	0.011	-0.031	0.055	0.015	-0.035	0.064

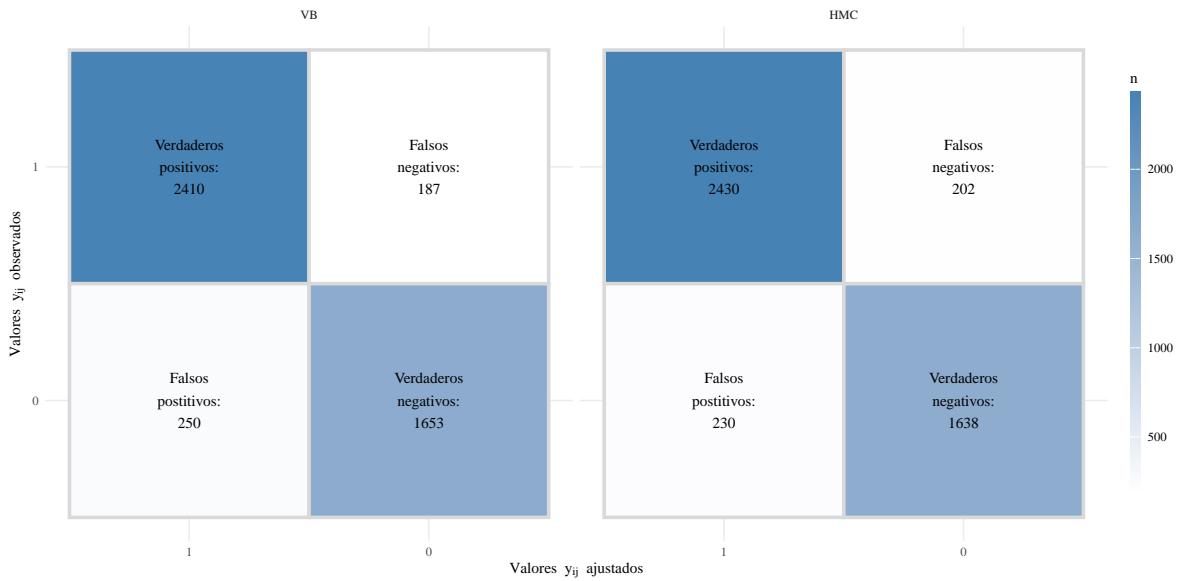


Figura 4.12: Matriz de confusión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC

Cuadro 4.11: Métricas del ajuste del modelo probit sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 25%.

Métrica	Bayes Variacional	Hamiltoniano MC
Exactitud	0.903	0.904
Verdaderos positivos	0.928	0.923
Verdaderos negativos	0.928	0.923
F1-Score	0.917	0.918
Tiempo (s)	56.690	100.760

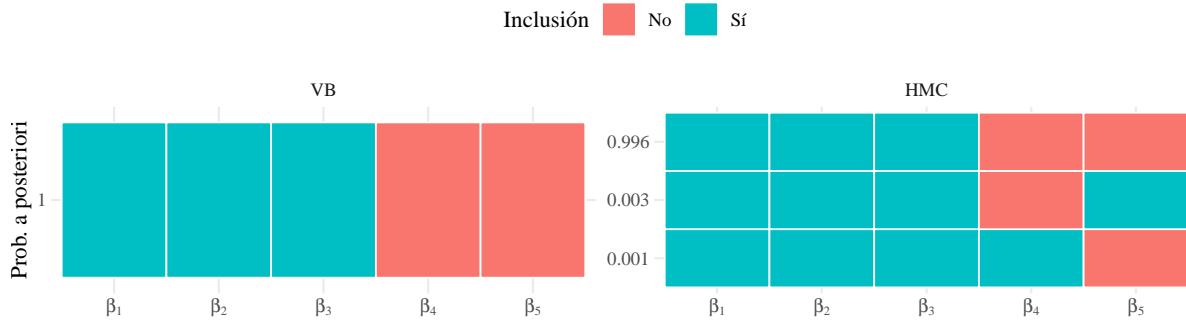


Figura 4.13: Modelo probit sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 25%.

4.4.2 Modelos con interceptos en cada área pequeña

Cuadro 4.12: Modelo probit sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.

Par.	Bayes Variacional				Hamiltoniano MC			
	Valor real	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	
ρ_1	0.500	0.150	0.017	0.474	0.447	0.025	0.850	
ρ_2	0.750	0.140	0.018	0.463	0.499	0.034	0.925	
ρ_3	0.850	0.214	0.022	0.640	0.861	0.131	0.992	
ρ_4	0.950	0.181	0.017	0.573	0.961	0.830	0.998	
μ_1	0.600	1.420	1.189	1.654	0.970	-0.056	1.643	
μ_2	0.700	1.688	1.480	1.891	1.134	0.142	1.863	
μ_3	0.600	2.677	2.400	2.975	1.870	0.272	3.077	
μ_4	0.500	2.519	2.296	2.759	1.889	0.353	2.947	
β_1	1.900	3.361	3.235	3.470	2.160	1.570	2.776	
β_2	-0.500	-0.575	-0.702	-0.453	-0.292	-0.504	-0.014	
β_3	-1.400	-2.455	-2.566	-2.341	-1.580	-2.086	-1.122	
β_4	0.000	-0.001	-0.058	0.056	-0.009	-0.116	0.058	
β_5	0.000	0.017	-0.043	0.079	0.037	-0.030	0.191	

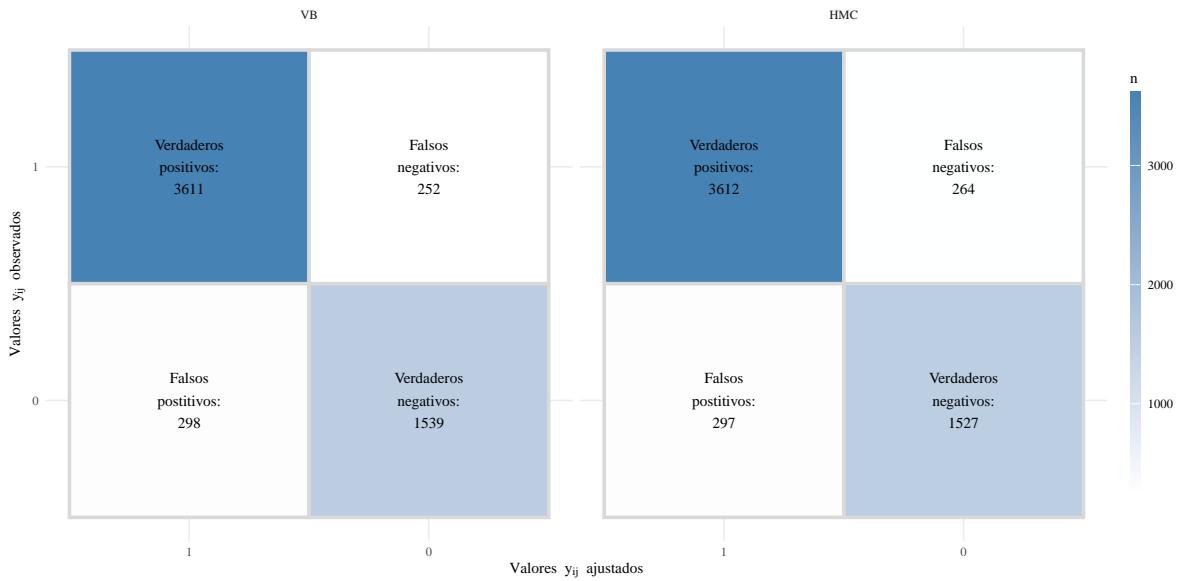


Figura 4.14: Matriz de confusión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC

Cuadro 4.13: Métricas del ajuste del modelo probit sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.

Métrica	Bayes Variacional	Hamiltoniano MC
Exactitud	0.904	0.902
Verdaderos positivos	0.935	0.932
Verdaderos negativos	0.935	0.932
F1-Score	0.929	0.928
Tiempo (s)	20.920	102.380

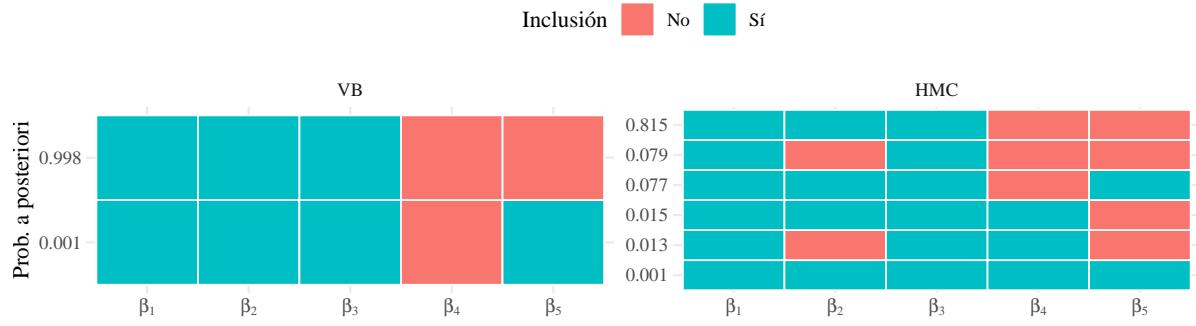


Figura 4.15: Modelo probit sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.

Cuadro 4.14: Modelo probit sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.

Par.	Bayes Variacional				Hamiltoniano MC			
	Valor real	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	
ρ_1	0.500	0.074	0.008	0.279	0.425	0.018	0.877	
ρ_2	0.750	0.070	0.006	0.252	0.562	0.062	0.922	
ρ_3	0.850	0.108	0.011	0.379	0.932	0.822	0.983	
ρ_4	0.950	0.083	0.011	0.299	0.814	0.378	0.956	
μ_1	0.600	1.847	1.740	1.958	1.144	-0.002	1.722	
μ_2	0.700	2.405	2.312	2.494	1.545	0.393	2.194	
μ_3	0.600	2.434	2.307	2.561	1.538	-0.158	2.696	
μ_4	0.500	2.397	2.293	2.507	1.444	0.094	2.352	
β_1	1.900	3.473	3.422	3.526	2.115	1.467	2.521	
β_2	-0.500	-0.865	-0.918	-0.809	-0.518	-0.647	-0.359	
β_3	-1.400	-2.595	-2.650	-2.541	-1.560	-1.881	-1.089	
β_4	0.000	-0.010	-0.051	0.030	-0.018	-0.075	0.037	
β_5	0.000	0.014	-0.033	0.058	0.011	-0.040	0.063	

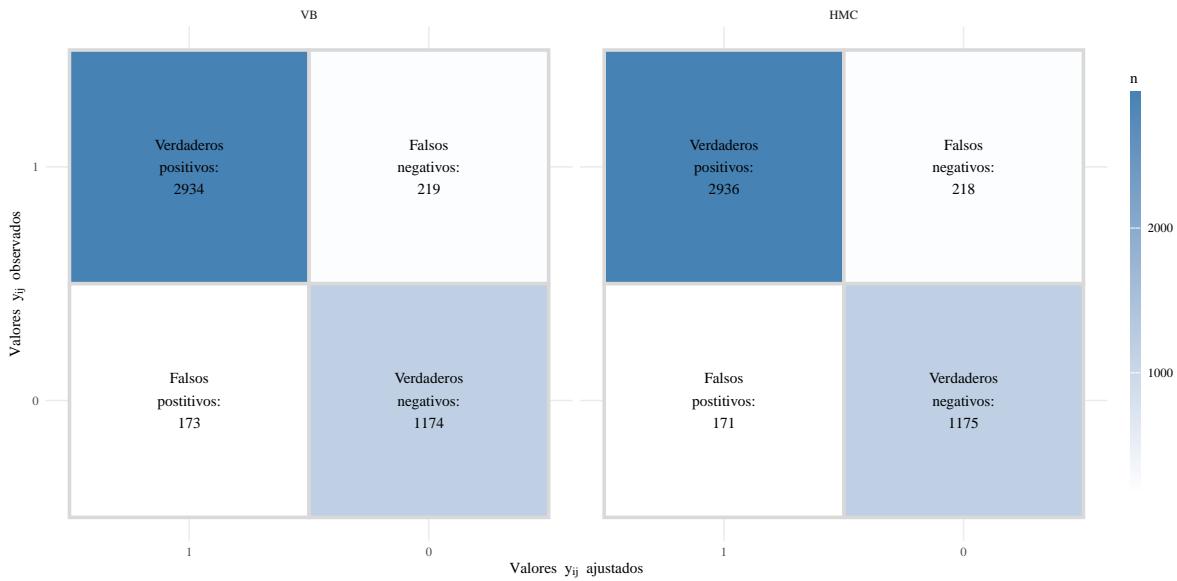


Figura 4.16: Matriz de confusión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC

Cuadro 4.15: Métricas del ajuste del modelo probit sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.

Métrica	Bayes Variacional	Hamiltoniano MC
Exactitud	0.913	0.914
Verdaderos positivos	0.931	0.931
Verdaderos negativos	0.931	0.931
F1-Score	0.937	0.938
Tiempo (s)	67.060	404.410

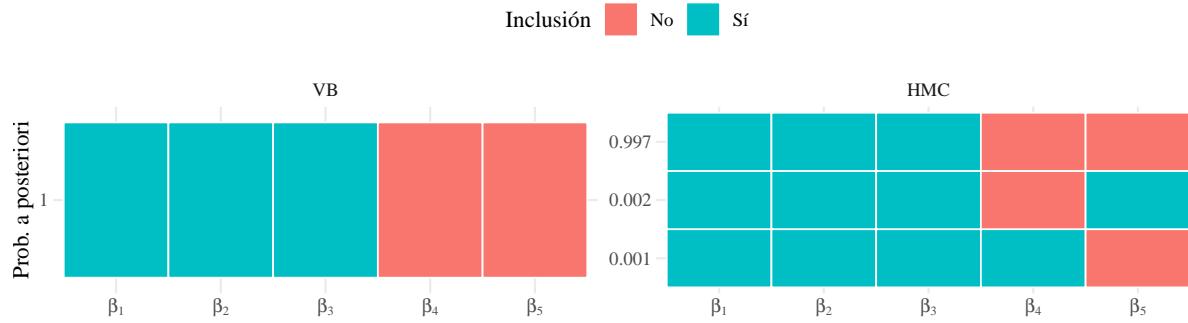


Figura 4.17: Modelo probit sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.

4.4.3 Modelos con único intercepto en todas las áreas pequeñas

Cuadro 4.16: Modelo probit sesgado latente con único intercepto para todas las áreas pequeñas. Porcentaje de muestreo: 5%.

Par.	Bayes Variacional				Hamiltoniano MC			
	Valor real	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	
ρ_1	0.500	0.191	0.024	0.560	0.562	0.050	0.893	
ρ_2	0.750	0.156	0.019	0.488	0.528	0.019	0.918	
ρ_3	0.850	0.266	0.042	0.656	0.900	0.240	0.998	
ρ_4	0.950	0.204	0.031	0.583	0.918	0.386	0.991	
μ	0.600	1.855	1.744	1.975	1.171	0.680	1.687	
β_1	1.900	3.159	3.040	3.276	1.825	1.360	2.364	
β_2	-0.500	-0.625	-0.740	-0.508	-0.271	-0.555	-0.015	
β_3	-1.400	-2.276	-2.388	-2.163	-1.351	-1.776	-0.958	
β_4	0.000	-0.001	-0.059	0.066	0.001	-0.078	0.073	
β_5	0.000	0.016	-0.045	0.074	0.051	-0.032	0.134	

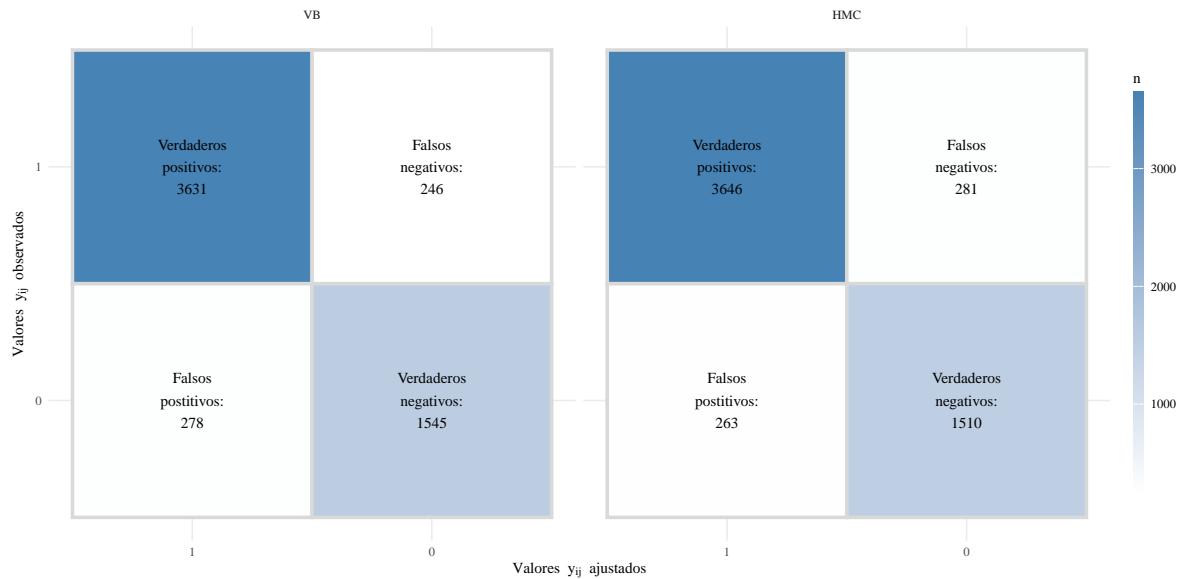


Figura 4.18: Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC

Cuadro 4.17: Métricas del ajuste del modelo probit sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 5%.

Métrica	Bayes Variacional	Hamiltoniano MC
Exactitud	0.908	0.905
Verdaderos positivos	0.937	0.928
Verdaderos negativos	0.937	0.928
F1-Score	0.933	0.931
Tiempo (s)	19.270	37.190

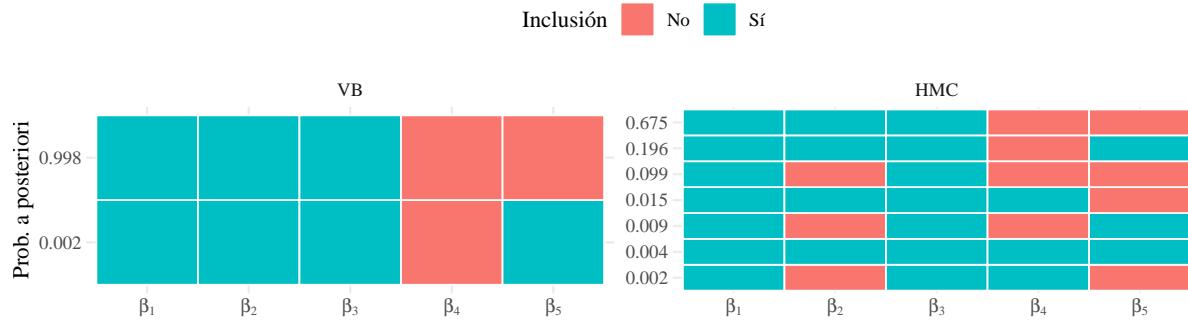


Figura 4.19: Modelo probit sesgado latente con único intercepto en todas las áreas pequeñas.
Porcentaje de muestreo: 5%.

Cuadro 4.18: Modelo probit sesgado latente con único intercepto en todas las áreas pequeñas.
Porcentaje de muestreo: 25%.

Par.	Bayes Variacional				Hamiltoniano MC			
	Valor real	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	
ρ_1	0.500	0.149	0.046	0.321	0.386	0.028	0.758	
ρ_2	0.750	0.069	0.009	0.235	0.385	0.017	0.779	
ρ_3	0.850	0.110	0.012	0.428	0.885	0.485	0.972	
ρ_4	0.950	0.083	0.009	0.292	0.736	0.218	0.921	
μ	0.600	2.270	2.218	2.326	1.497	1.058	1.880	
β_1	1.900	3.475	3.426	3.529	2.282	1.913	2.652	
β_2	-0.500	-0.868	-0.919	-0.815	-0.554	-0.685	-0.431	
β_3	-1.400	-2.577	-2.631	-2.526	-1.685	-1.930	-1.415	
β_4	0.000	-0.010	-0.053	0.033	-0.013	-0.072	0.043	
β_5	0.000	0.018	-0.024	0.062	0.015	-0.043	0.068	

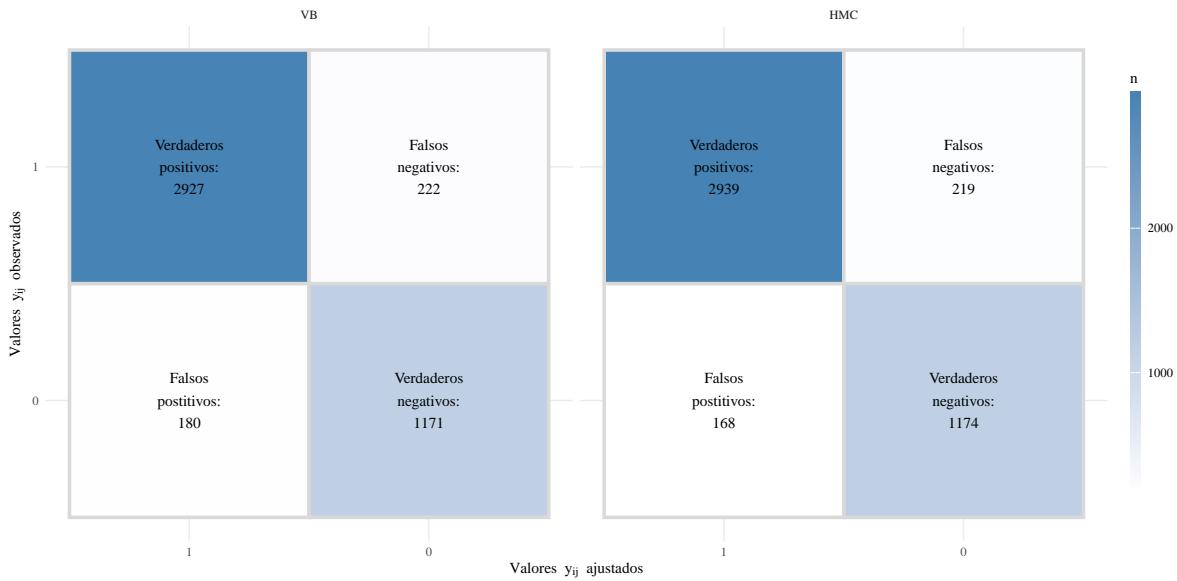


Figura 4.20: Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC

Cuadro 4.19: Métricas del ajuste del modelo probit sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 25%.

Métrica	Bayes Variacional	Hamiltoniano MC
Exactitud	0.911	0.914
Verdaderos positivos	0.930	0.931
Verdaderos negativos	0.930	0.931
F1-Score	0.936	0.938
Tiempo (s)	62.480	148.340

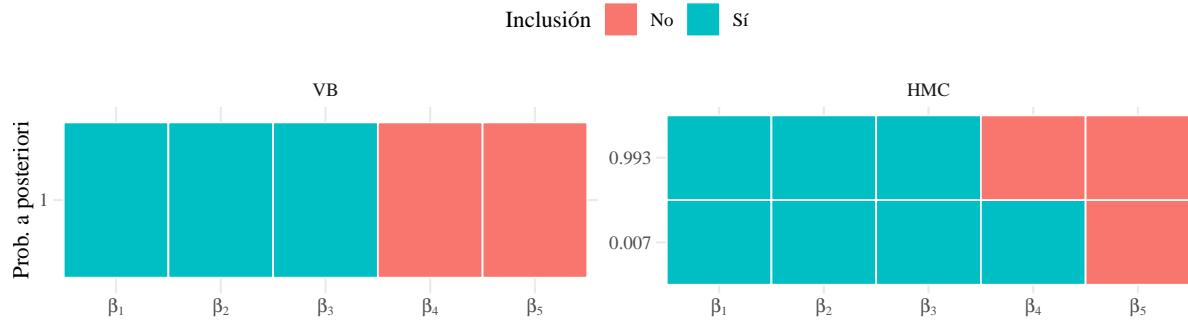


Figura 4.21: Modelo probit sesgado latente con único intercepto en todas las áreas pequeñas.
Porcentaje de muestreo: 25%.

4.5 Ajuste del Modelo probit ordenado sesgado con variable latente, datos simulados

4.5.1 Modelos sin interceptos en cada área pequeña

Cuadro 4.20: Modelo probit ordenado sesgado latente sin interceptos en cada área pequeña.
Porcentaje de muestreo: 5%.

Par.	Bayes Variacional				Hamiltoniano MC			
	Valor real	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	
ρ_1	0.500	0.270	0.060	0.601	0.559	0.093	0.885	
ρ_2	0.750	0.162	0.023	0.504	0.504	0.059	0.889	
ρ_3	0.850	0.440	0.195	0.701	0.899	0.488	0.989	
ρ_4	0.950	0.478	0.319	0.643	0.733	0.307	0.949	
δ_1	1.200	1.013	0.866	1.196	1.011	0.706	1.317	
β_1	1.900	3.165	3.053	3.271	2.071	1.486	2.629	
β_2	-0.500	-0.712	-0.839	-0.595	-0.487	-0.722	-0.282	
β_3	-1.400	-2.381	-2.505	-2.263	-1.619	-2.050	-1.122	
β_4	0.000	0.013	-0.044	0.079	0.053	-0.037	0.293	
β_5	0.000	0.004	-0.049	0.061	0.011	-0.058	0.084	

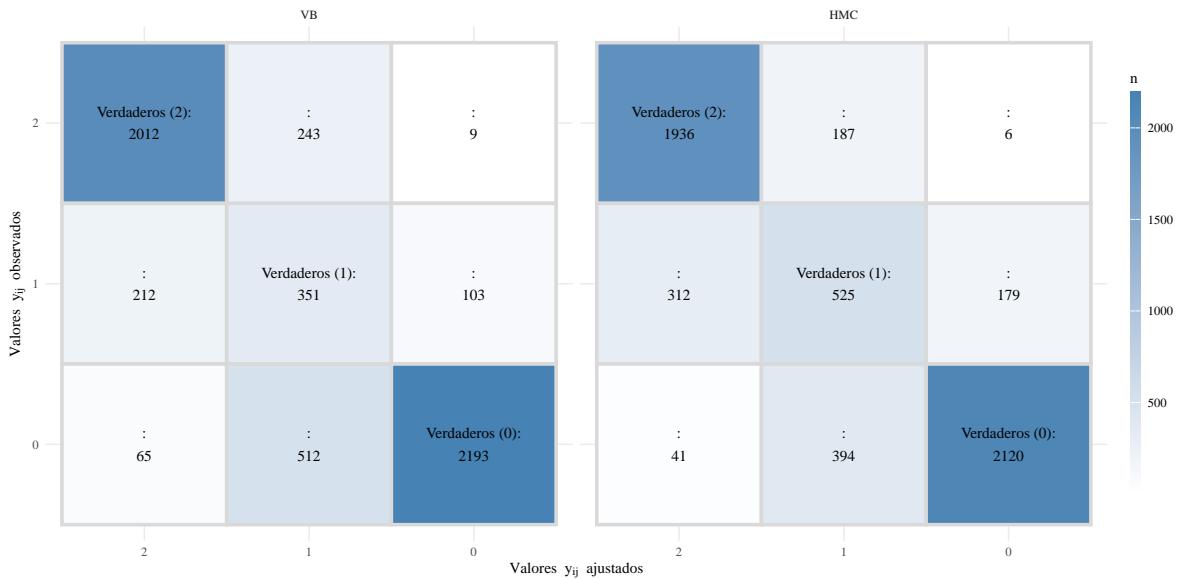


Figura 4.22: Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC

Cuadro 4.21: Métricas del ajuste del modelo probit ordenado sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 5%.

Métrica	Bayes Variacional	Hamiltoniano MC
Exactitud	0.799	0.804
τ de Kendall	0.811	0.816
Tiempo (s)	16.410	29.510

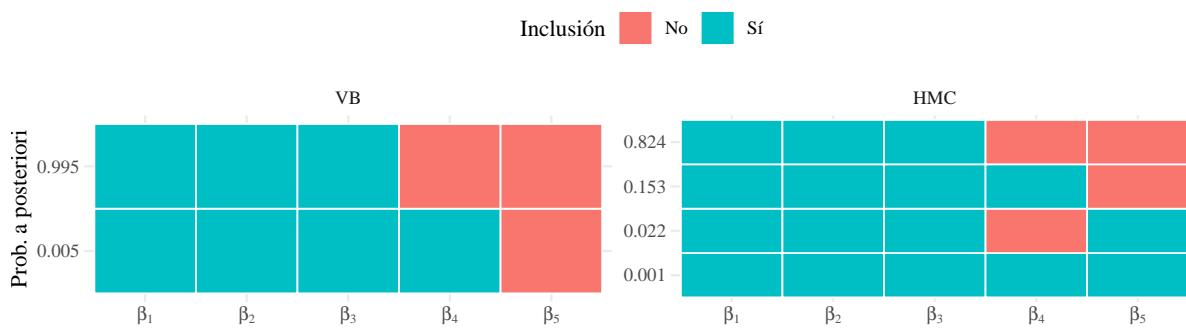


Figura 4.23: Modelo probit ordenado sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 5%.

Cuadro 4.22: Modelo probit ordenado sesgado latente sin interceptos en cada área pequeña.
Porcentaje de muestreo: 25%.

Par.	Bayes Variacional				Hamiltoniano MC			
	Valor real	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	
ρ_1	0.500	0.264	0.208	0.322	0.506	0.232	0.765	
ρ_2	0.750	0.307	0.267	0.352	0.738	0.499	0.876	
ρ_3	0.850	0.401	0.336	0.466	0.739	0.422	0.904	
ρ_4	0.950	0.431	0.380	0.484	0.903	0.839	0.945	
δ_1	1.200	1.499	1.424	1.575	1.257	1.062	1.451	
β_1	1.900	3.089	3.038	3.138	2.090	1.797	2.359	
β_2	-0.500	-0.818	-0.868	-0.770	-0.560	-0.654	-0.463	
β_3	-1.400	-2.291	-2.343	-2.238	-1.533	-1.739	-1.324	
β_4	0.000	-0.011	-0.051	0.031	-0.015	-0.066	0.031	
β_5	0.000	0.023	-0.017	0.065	0.019	-0.029	0.066	

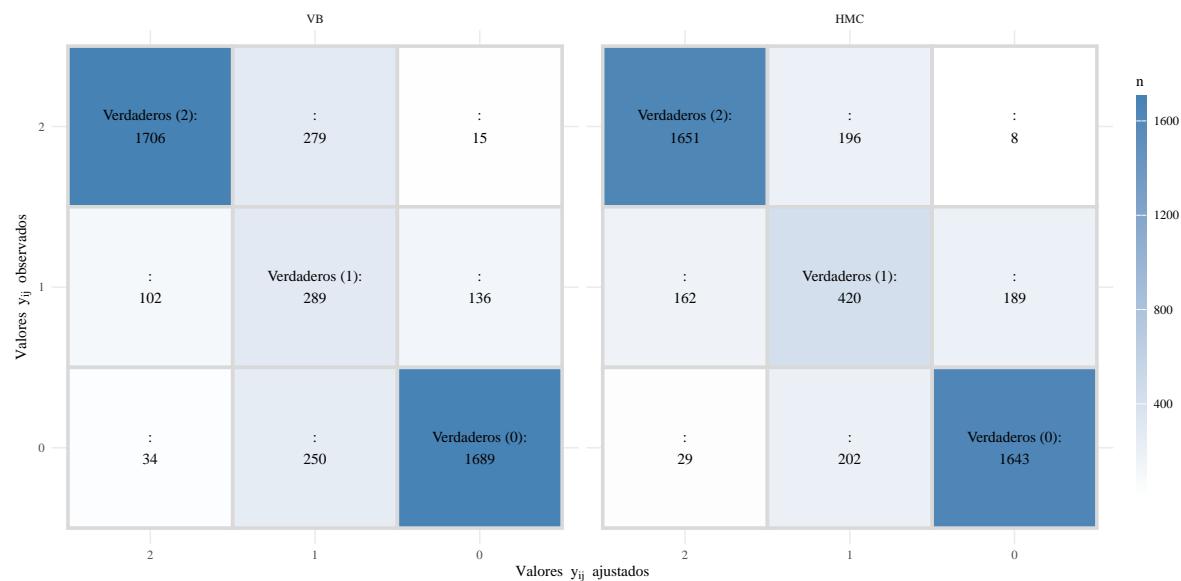


Figura 4.24: Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC

Cuadro 4.23: Métricas del ajuste del modelo probit ordenado sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 25%.

Métrica	Bayes Variacional	Hamiltoniano MC
Exactitud	0.819	0.825
τ de Kendall	0.825	0.831
Tiempo (s)	55.400	127.740

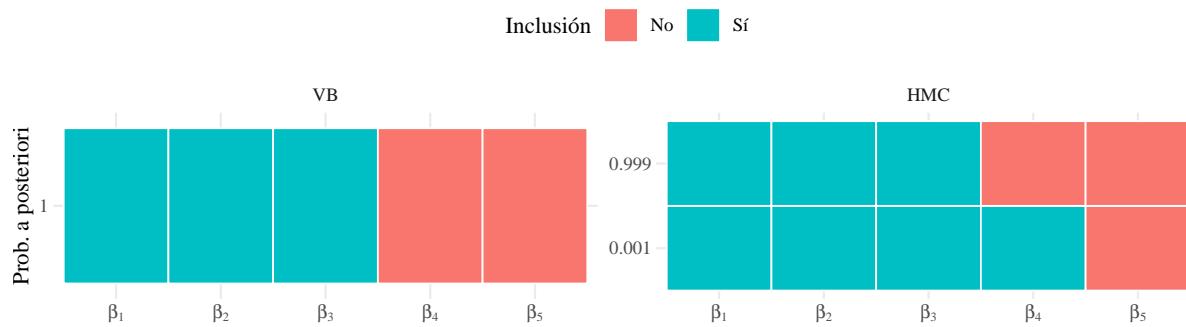


Figura 4.25: Modelo probit ordenado sesgado latente sin interceptos en cada área pequeña. Porcentaje de muestreo: 25%.

4.5.2 Modelos con interceptos en cada área pequeña

Cuadro 4.24: Modelo probit ordenado sesgado latente con interceptos en cada área pequeña.
Porcentaje de muestreo: 5%.

Par.	Bayes Variacional				Hamiltoniano MC			
	Valor real	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	
ρ_1	0.500	0.144	0.014	0.451	0.419	0.026	0.825	
ρ_2	0.750	0.142	0.015	0.484	0.490	0.041	0.849	
ρ_3	0.850	0.227	0.024	0.637	0.893	0.525	0.983	
ρ_4	0.950	0.196	0.027	0.558	0.865	0.353	0.975	
δ_1	1.200	1.537	1.381	1.711	1.305	0.946	1.677	
μ_1	0.600	1.526	1.295	1.756	1.188	0.012	1.829	
μ_2	0.700	1.561	1.362	1.754	1.167	0.133	1.851	
μ_3	0.600	2.530	2.239	2.816	1.977	0.263	3.106	
μ_4	0.500	2.035	1.786	2.278	1.572	-0.257	2.685	
β_1	1.900	3.341	3.233	3.454	2.274	1.707	2.814	
β_2	-0.500	-0.669	-0.787	-0.553	-0.438	-0.632	-0.248	
β_3	-1.400	-2.439	-2.544	-2.324	-1.725	-2.151	-1.345	
β_4	0.000	-0.006	-0.065	0.054	0.010	-0.051	0.114	
β_5	0.000	0.025	-0.033	0.084	0.039	-0.030	0.201	

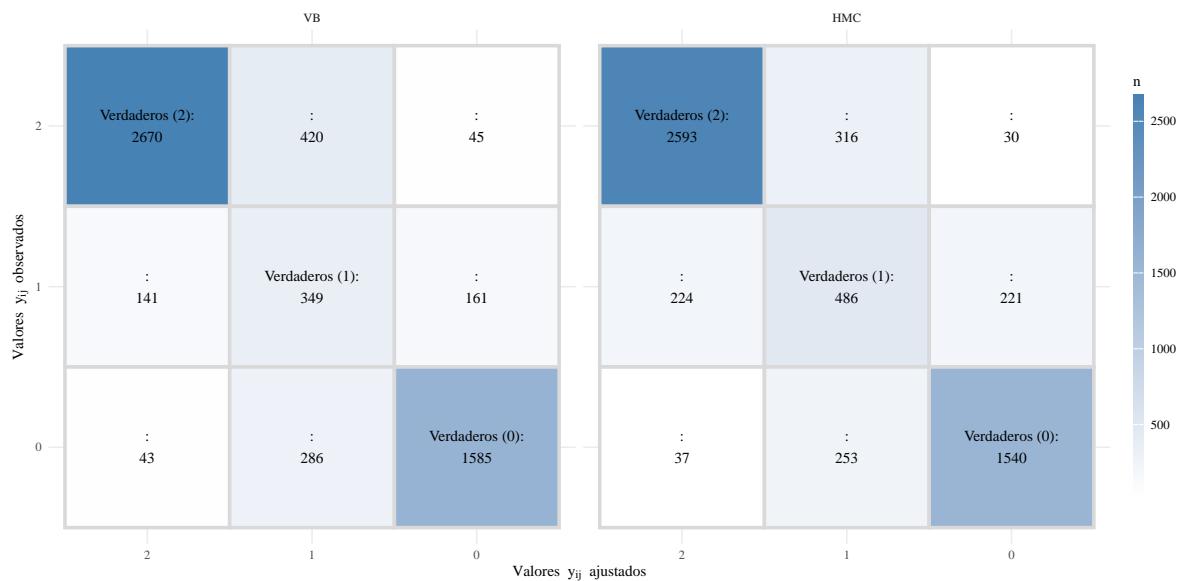


Figura 4.26: Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC

Cuadro 4.25: Métricas del ajuste del modelo probit ordenado sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.

Métrica	Bayes Variacional	Hamiltoniano MC
Exactitud	0.808	0.810
τ de Kendall	0.801	0.806
Tiempo (s)	15.350	64.800

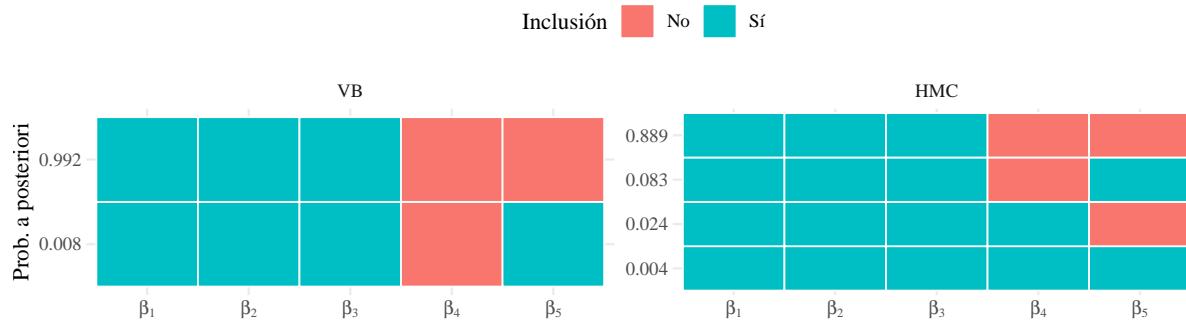


Figura 4.27: Modelo probit ordenado sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 5%.

Cuadro 4.26: Modelo probit ordenado sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.

Par.	Bayes Variacional				Hamiltoniano MC			
	Valor real	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	
ρ_1	0.500	0.069	0.009	0.236	0.300	0.017	0.674	
ρ_2	0.750	0.073	0.008	0.277	0.560	0.046	0.830	
ρ_3	0.850	0.107	0.013	0.360	0.796	0.407	0.915	
ρ_4	0.950	0.087	0.010	0.278	0.877	0.779	0.931	
δ_1	1.200	1.650	1.589	1.717	1.357	1.175	1.561	
μ_1	0.600	1.621	1.517	1.723	1.242	0.666	1.660	
μ_2	0.700	2.024	1.934	2.118	1.466	0.185	2.096	
μ_3	0.600	1.843	1.715	1.969	1.315	-0.328	2.281	
μ_4	0.500	1.893	1.788	1.996	1.342	-0.603	2.360	
β_1	1.900	3.131	3.080	3.180	2.120	1.865	2.393	
β_2	-0.500	-0.778	-0.829	-0.727	-0.548	-0.653	-0.455	
β_3	-1.400	-2.305	-2.361	-2.252	-1.554	-1.783	-1.344	
β_4	0.000	-0.003	-0.044	0.039	-0.005	-0.053	0.041	
β_5	0.000	0.014	-0.030	0.056	0.011	-0.035	0.059	

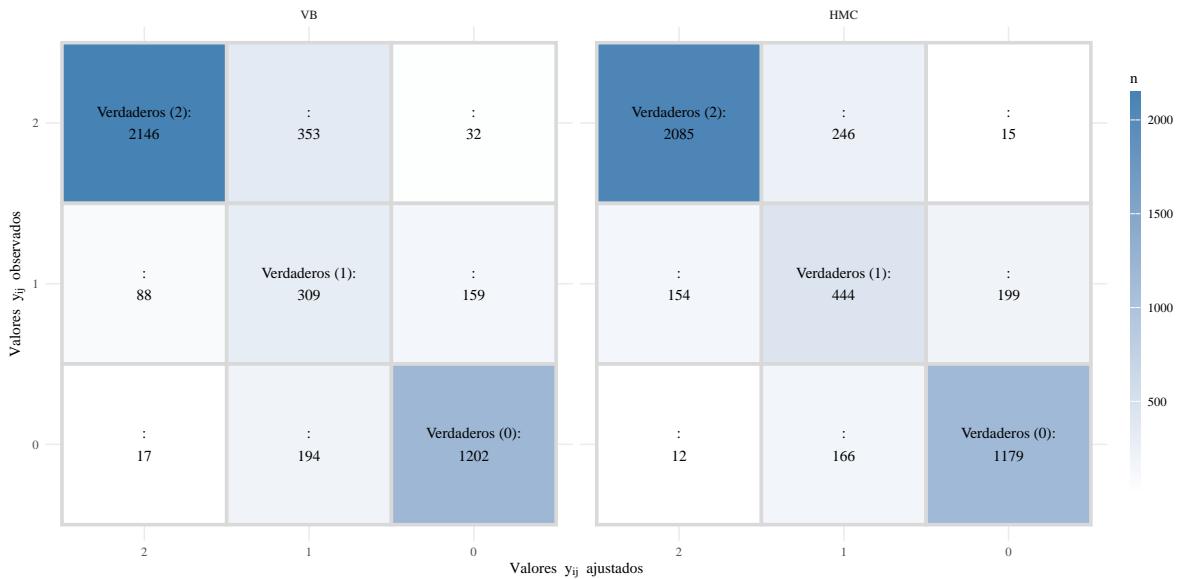


Figura 4.28: Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC

Cuadro 4.27: Métricas del ajuste del modelo probit ordenado sesgado latente con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.

Métrica	Bayes Variacional	Hamiltoniano MC
Exactitud	0.813	0.824
τ de Kendall	0.812	0.827
Tiempo (s)	63.000	483.310

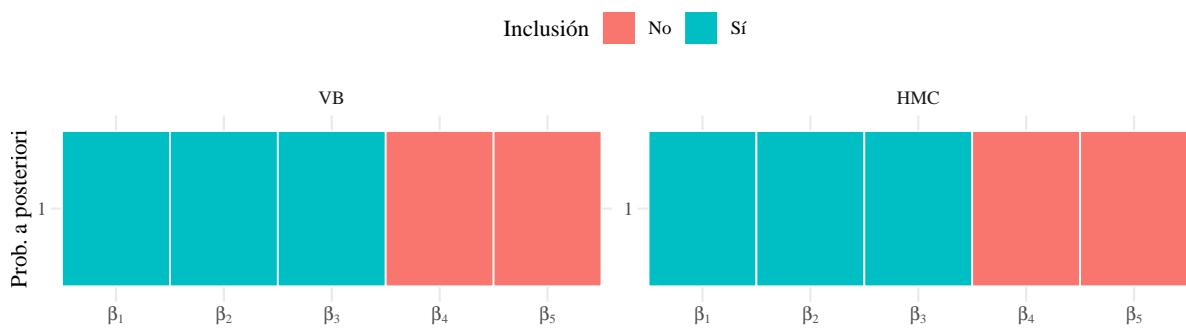


Figura 4.29: Modelo probit ordenado sesgado con interceptos en cada área pequeña. Porcentaje de muestreo: 25%.

4.5.3 Modelos con único intercepto en todas las áreas pequeñas

Cuadro 4.28: Modelo probit ordenado sesgado latente con único intercepto para todas las áreas pequeñas. Porcentaje de muestreo: 5%.

Par.	Bayes Variacional				Hamiltoniano MC			
	Valor real	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	
ρ_1	0.500	0.176	0.023	0.561	0.383	0.017	0.819	
ρ_2	0.750	0.160	0.022	0.482	0.431	0.025	0.844	
ρ_3	0.850	0.472	0.233	0.700	0.888	0.439	0.983	
ρ_4	0.950	0.210	0.026	0.602	0.856	0.323	0.978	
δ_1	1.200	1.478	1.328	1.636	1.294	0.956	1.650	
μ	0.600	1.682	1.568	1.789	1.290	0.805	1.789	
β_1	1.900	3.168	3.050	3.279	2.246	1.782	2.781	
β_2	-0.500	-0.663	-0.786	-0.547	-0.437	-0.647	-0.261	
β_3	-1.400	-2.294	-2.406	-2.187	-1.704	-2.112	-1.357	
β_4	0.000	-0.003	-0.067	0.055	0.006	-0.060	0.071	
β_5	0.000	0.022	-0.035	0.082	0.035	-0.033	0.169	

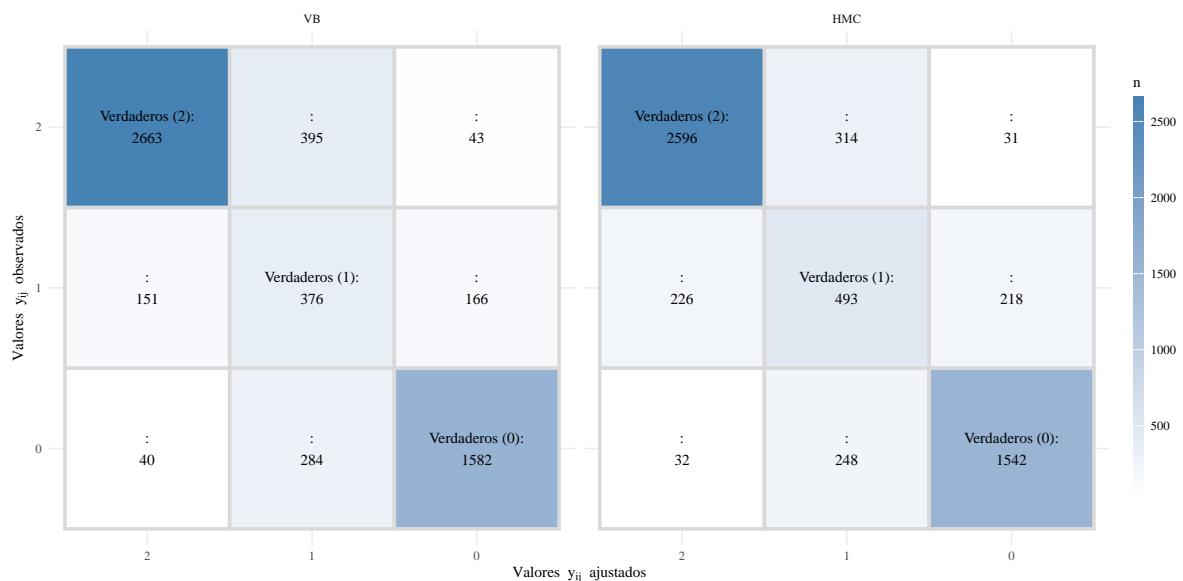


Figura 4.30: Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC

Cuadro 4.29: Métricas del ajuste del modelo probit ordenado sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 5%.

Métrica	Bayes Variacional	Hamiltoniano MC
Exactitud	0.811	0.813
τ de Kendall	0.805	0.809
Tiempo (s)	13.640	28.450

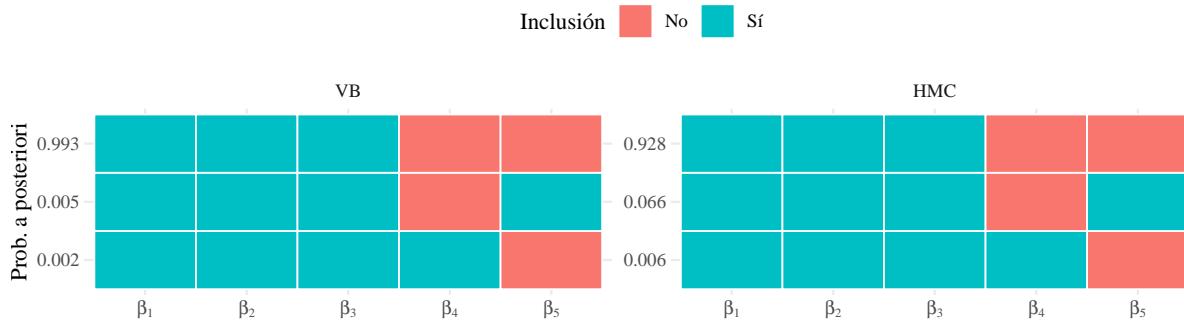


Figura 4.31: Modelo probit ordenado sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 5%.

Cuadro 4.30: Modelo probit ordenado sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 25%.

Par.	Bayes Variacional				Hamiltoniano MC			
	Valor real	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	
ρ_1	0.500	0.140	0.045	0.322	0.351	0.015	0.792	
ρ_2	0.750	0.098	0.021	0.274	0.509	0.041	0.870	
ρ_3	0.850	0.111	0.012	0.383	0.743	0.173	0.932	
ρ_4	0.950	0.093	0.011	0.329	0.857	0.685	0.946	
δ_1	1.200	1.638	1.577	1.701	1.369	1.042	1.602	
μ	0.600	1.839	1.785	1.893	1.436	1.099	1.783	
β_1	1.900	3.097	3.048	3.150	2.146	1.627	2.457	
β_2	-0.500	-0.767	-0.820	-0.717	-0.554	-0.666	-0.419	
β_3	-1.400	-2.270	-2.323	-2.214	-1.570	-1.811	-1.191	
β_4	0.000	0.003	-0.038	0.042	-0.003	-0.053	0.044	
β_5	0.000	0.011	-0.034	0.052	0.012	-0.035	0.061	

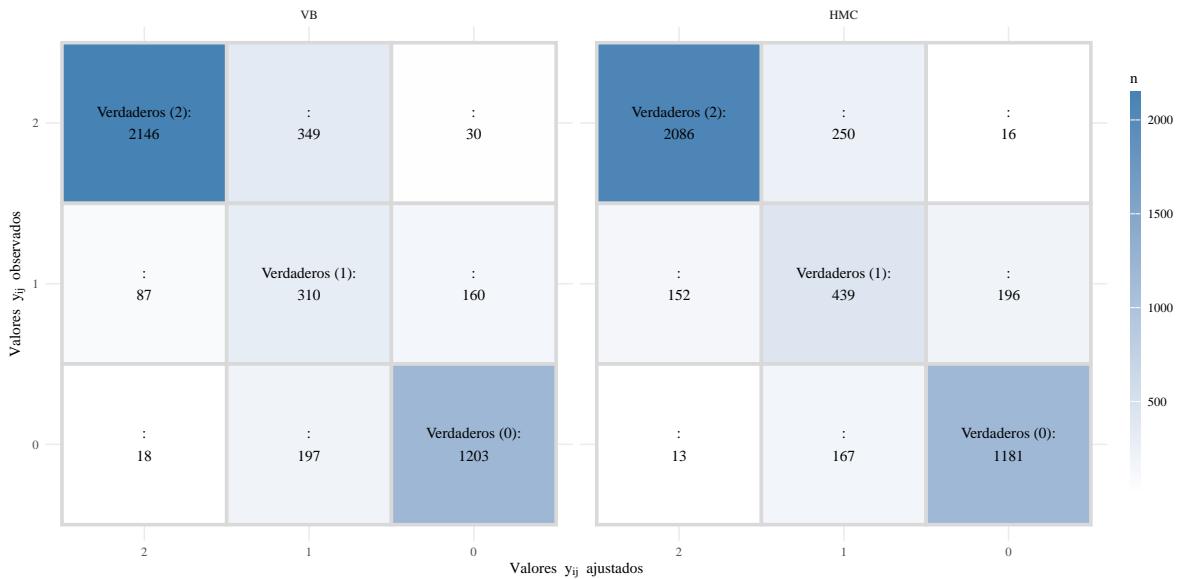


Figura 4.32: Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC

Cuadro 4.31: Métricas del ajuste del modelo probit ordenado sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 25%.

Métrica	Bayes Variacional	Hamiltoniano MC
Exactitud	0.813	0.824
τ de Kendall	0.813	0.826
Tiempo (s)	52.350	150.440

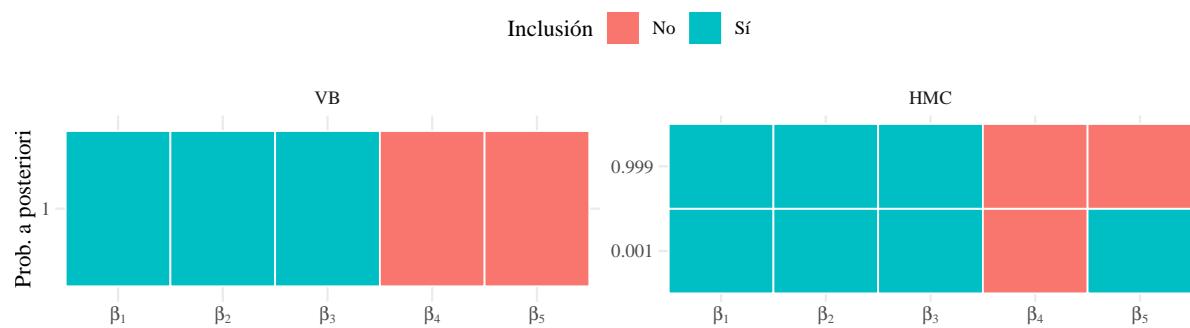


Figura 4.33: Modelo probit ordenado sesgado latente con único intercepto en todas las áreas pequeñas. Porcentaje de muestreo: 25%.

4.6 Análisis descriptivo del conjunto de datos del ICTPC

En la [Figura 4.34](#) se muestra la densidad estimada del log-ICTPC, en la izquierda se agrupan los datos de toda la Ciudad y en la derecha se agrupan de acuerdo al ámbito (rural y urbano). Los datos de toda la ciudad indican que la variable respuesta está sesgada hacia la derecha. Por otro lado, cuando clasificamos de acuerdo al tipo de ámbito, el conjunto de hogares urbanos exhiben mayor sesgo, es decir, se observan valores más grandes, mientras que los ingresos en los hogares rurales están más concentrados.

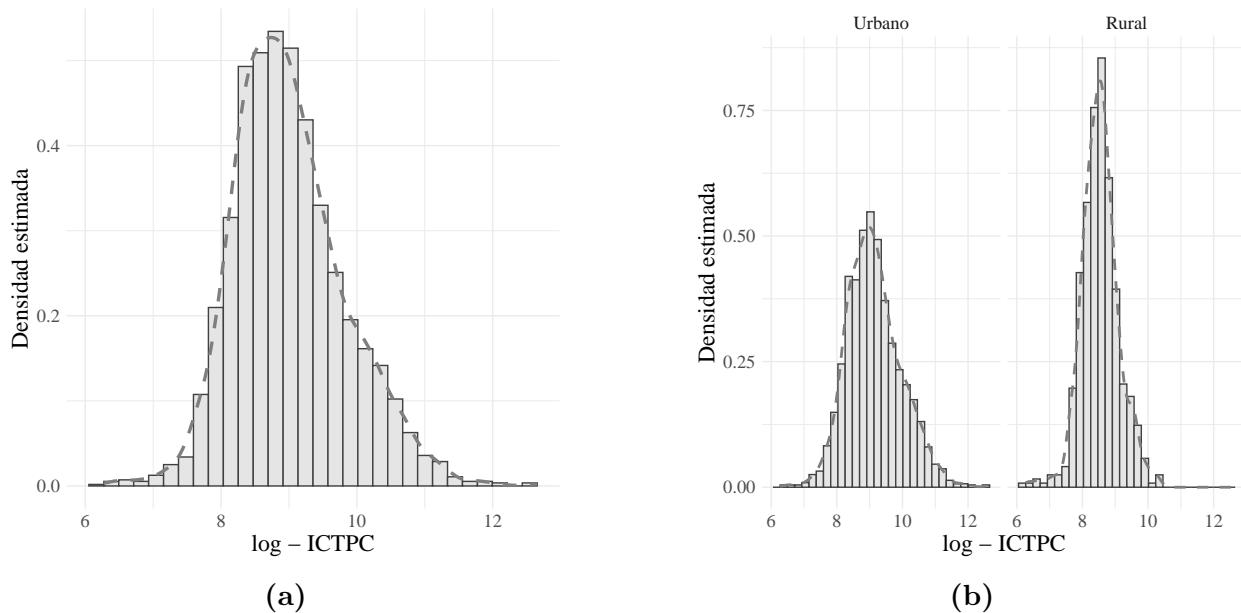


Figura 4.34: Estimación por kernel (suavizado) de la densidad empírica del logaritmo natural de la respuesta (log-ICTPC) e histograma de probabilidad. Se omitieron tres observaciones asociadas a ingresos pequeños (52.66, 187.71 y 367.30 pesos mexicanos). Fuente: elaboración propia.

En el [Cuadro 4.32](#) se muestran estadísticos resumen básicos sobre el ICTPC por alcaldía y por ámbito, notamos que sólo en las demarcaciones Milpa Alta, Tláhuac y Xochimilco se tiene registro de hogares que pertenecen al ámbito rural, sumando a un total de 553, el 21.8% con respecto al total de hogares encuestados (2,329). En el [Cuadro 4.33](#), se muestran los estadísticos resumen agrupados de acuerdo al tipo de ámbito; por ejemplo, la desviación estándar en el ámbito rural es de aproximadamente 17,800 pesos mexicanos, mientras que en el ámbito rural es de 4,000. Así, pese a que el ingreso corriente total per cápita en este

ámbito sea en promedio menor, tienen menos dispersión. Sin embargo, mayor dispersión no implica que exista un mayor grado de desigualdad, por ejemplo, medido con el índice de Gini.

Cuadro 4.32: Estadísticos resumen del ingreso corriente total per cápita por **Alcaldía y ámbito**. Fuente: Elaboración propia a partir de información de la ENIGH 2024.

Alcaldia	Ámbito	Minimo	Mediana	Media	Maximo	Desv. est.	n_i
Azc.	Urbano	2102.456	12 090.792	16 910.051	114 668.00	16 878.831	111
Azc.	Rural						0
Cyc.	Urbano	1551.186	11 696.878	14 677.289	83 211.33	12 202.847	105
Cyc.	Rural						0
CdM.	Urbano	3177.718	8461.970	14 948.206	118 785.32	19 527.767	53
CdM.	Rural						0
GAM.	Urbano	781.046	7787.525	10 774.117	68 570.14	9824.513	252
GAM.	Rural						0
Iztc.	Urbano	631.123	8913.196	13 545.988	61 666.93	12 877.532	105
Iztc.	Rural						0
Iztp.	Urbano	187.717	6073.853	8171.228	48 248.66	7074.950	349
Iztp.	Rural						0
LMC.	Urbano	1504.451	5995.553	10 735.292	72 000.85	12 073.719	68
LMC.	Rural						0
MIA.	Urbano	1277.369	6171.656	7005.241	21 981.60	4342.122	31
MIA.	Rural	529.397	4949.803	5483.564	17 037.63	2809.900	227
ÁlO.	Urbano	1564.368	9434.721	13 817.397	118 931.95	14 118.201	207
ÁlO.	Rural						0
Tlh.	Urbano	1245.704	7960.860	8400.618	21 017.26	4244.966	52
Tlh.	Rural	2531.590	5178.865	6949.424	29 950.02	5877.352	23
Tll.	Urbano	2163.907	9171.878	13 563.610	65 103.11	12 886.281	117
Tll.	Rural	720.764	5180.475	6142.782	21 935.70	3720.346	162
Xch.	Urbano	1430.432	6970.185	10 052.865	56 342.32	9781.255	87

Cuadro 4.32: Estadísticos resumen del ingreso corriente total per cápita por **Alcaldía y ámbito**. Fuente: Elaboración propia a partir de información de la ENIGH 2024.

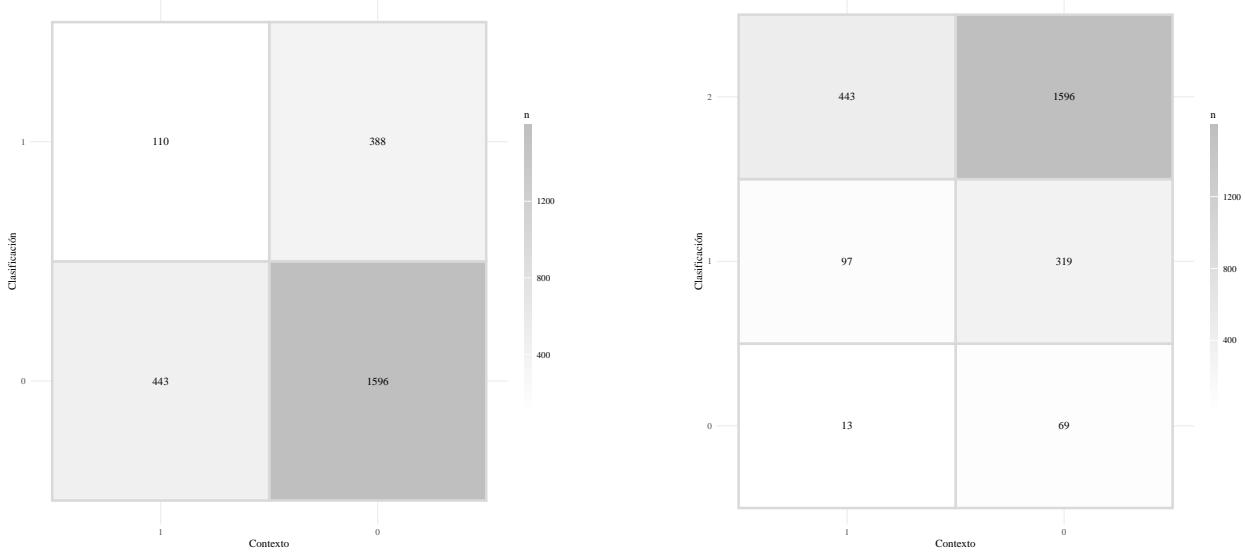
Alcaldia	Ámbito	Minimo	Mediana	Media	Maximo	Desv. est.	n_i
Xch.	Rural	1069.667	4786.025	6545.860	29 559.91	5240.275	141
BnJ.	Urbano	3741.365	20 065.175	25 167.655	176 719.22	22 441.338	135
BnJ.	Rural						0
Cht.	Urbano	921.870	8436.753	17 636.413	313 206.74	35 445.246	115
Cht.	Rural						0
MgH.	Urbano	2290.712	17 156.079	25 668.338	132 651.56	23 268.993	96
MgH.	Rural						0
VnC.	Urbano	52.668	7880.444	15 085.567	297 123.05	32 880.344	101
VnC.	Rural						0

Cuadro 4.33: Estadísticos resumen del ingreso corriente total per cápita por **Alcaldía y ámbito**. Fuente: elaboración propia con información de la ENIGH 2024.

Ámbito	Minimo	Mediana	Media	Maximo	Desv. est.	Conteo
Urbano	52.668	8520.740	13769.613	313206.74	17793.618	1984
Rural	529.397	5032.302	6008.504	29950.02	3979.968	553

En la [Figura 4.35](#) se muestra la distribución de las variables binarias y ordinales obtenidas al discretizar el conjunto de datos del ICTPC, agrupadas de acuerdo al tipo de corte que pertenecen. Para el caso binario, se observa una mayor proporción de observaciones $y_{ij} = 0$, es decir, sin carencias. Lo mismo sucede en el caso ordinal, una mayor parte de observaciones son sin carencias, es decir $y_{ij} = 2$.

Figura 4.35



4.7 Ajuste del Modelo log-normal asimétrico, datos del ICTPC

Cuadro 4.34: Ajuste del Modelo log-normal sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible.

Par.	Bayes Variacional			Hamiltoniano MC				\hat{R}
	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	\hat{R}	
ρ_1	0.766	0.650	0.855	0.836	0.689	0.916	1.005	
ρ_2	0.813	0.725	0.882	0.852	0.714	0.926	1.033	
ρ_3	0.815	0.673	0.913	0.878	0.734	0.939	1.012	
ρ_4	0.807	0.755	0.850	0.869	0.783	0.923	1.006	
ρ_5	0.717	0.566	0.850	0.797	0.482	0.907	1.024	
ρ_6	0.791	0.743	0.831	0.858	0.754	0.916	1.032	
ρ_7	0.862	0.786	0.919	0.909	0.816	0.951	1.015	
ρ_8	0.893	0.868	0.914	0.921	0.873	0.950	1.022	
ρ_9	0.828	0.776	0.873	0.888	0.810	0.934	1.004	
ρ_{10}	0.924	0.890	0.949	0.948	0.907	0.970	1.021	
ρ_{11}	0.833	0.792	0.870	0.878	0.799	0.926	1.066	

Cuadro 4.34: Ajuste del Modelo log-normal sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible.

	Bayes Variacional			Hamiltoniano MC				
Par.	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	\hat{R}	
ρ_{12}	0.790	0.727	0.843	0.868	0.762	0.923	1.010	
ρ_{13}	0.767	0.667	0.848	0.850	0.693	0.919	1.026	
ρ_{14}	0.209	0.021	0.600	0.530	0.036	0.837	1.083	
ρ_{15}	0.727	0.568	0.849	0.812	0.578	0.915	1.039	
ρ_{16}	0.133	0.015	0.424	0.297	0.012	0.711	1.001	
σ^2	0.665	0.649	0.683	0.758	0.655	0.876	1.024	
μ_1	7.281	7.175	7.387	6.425	2.687	9.468	1.014	
μ_2	7.235	7.131	7.347	6.312	2.311	9.387	1.007	
μ_3	7.219	7.081	7.358	6.443	2.531	9.474	1.017	
μ_4	7.257	7.188	7.325	6.164	2.456	8.846	0.999	
μ_5	7.067	6.950	7.182	6.293	2.412	8.947	1.007	
μ_6	7.183	7.126	7.241	6.218	2.086	9.202	1.003	
μ_7	7.267	7.150	7.383	6.116	2.205	8.531	1.005	
μ_8	7.298	7.244	7.355	6.053	2.133	8.500	1.007	
μ_9	7.304	7.235	7.376	6.371	2.558	9.380	1.002	
μ_{10}	7.246	7.161	7.335	6.226	2.194	9.180	1.000	
μ_{11}	7.349	7.290	7.411	6.254	2.280	9.072	1.007	
μ_{12}	7.206	7.130	7.281	6.150	2.453	8.591	1.005	
μ_{13}	7.416	7.312	7.522	6.560	2.823	9.330	1.017	
μ_{14}	6.888	6.755	7.013	6.257	2.372	8.794	1.000	
μ_{15}	7.356	7.244	7.481	6.667	2.761	9.469	1.004	
μ_{16}	6.842	6.706	6.980	6.232	2.431	8.751	1.000	

Cuadro 4.35: Los porcentajes de población bajo la LPI y LPEI son de 18.81% y 2.58% para el método VB, y de 31.14% y 5.60% para el método HMC. La medición estatal oficial es de 25.70% y 4.45%.

Alcaldía	Bayes Variacional		Hamiltoniano MC	
	Población bajo LPI (%)	Población bajo LPEI (%)	Población bajo LPI (%)	Población bajo LPEI (%)
Azcapotzalco	12.351	0.853	12.330	1.185
Coyoacán	14.386	2.823	14.521	2.958
Cuajimalpa de Morelos	17.139	2.203	17.278	2.685
Gustavo A. Madero	25.399	3.681	25.940	3.697
Iztacalco	17.971	2.741	20.075	3.119
Iztapalapa	38.800	7.496	39.523	8.575
La Magdalena Contreras	31.147	5.036	32.311	5.954
Milpa Alta	54.554	11.006	51.490	9.047
Álvaro Obregón	23.016	2.865	23.807	3.815
Tláhuac	45.255	8.788	46.011	9.341
Tlalpan	25.701	3.775	24.807	3.621
Xochimilco	36.142	6.212	34.584	6.746
Benito Juárez	1.472	0.275	1.519	0.320
Cuauhtémoc	11.443	1.244	10.377	1.173
Miguel Hidalgo	4.690	0.280	4.847	0.359
Venustiano Carranza	22.563	3.875	18.675	3.067

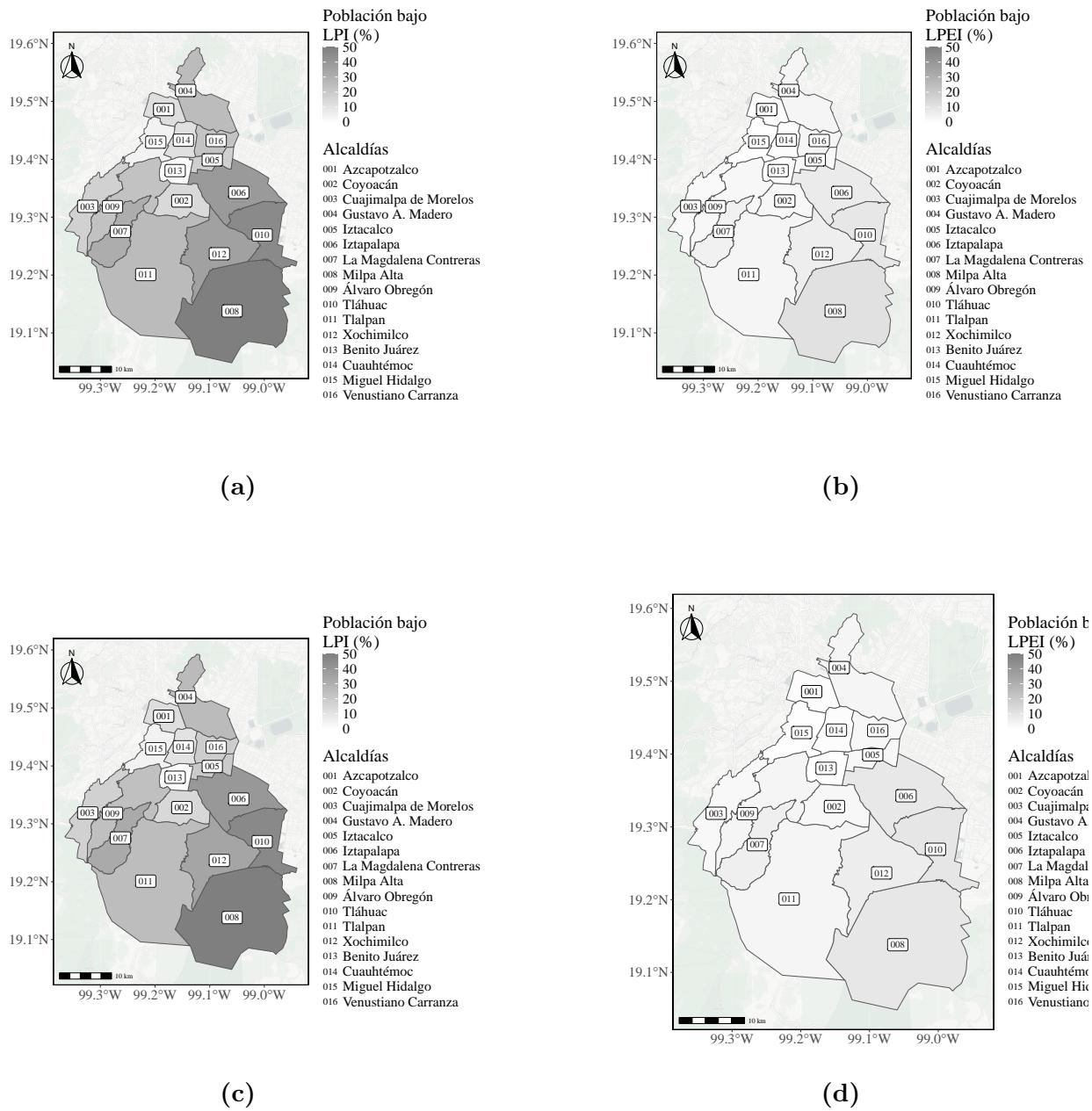


Figura 4.36: Mapas con la estimación del porcentaje de la población bajo la LPI y LPEI, en el primer renglón se muestran las estimaciones a partir del método BV y en el segundo renglón a partir del método HMC.

Cuadro 4.36: Sesgo promedio calculado a partir de la muestra *a posteriori*. El sesgo promedio para el método VB es de 815 pesos mexicanos y 798 para el método HMC. Fuente: elaboración propia.

Alcaldía \ Media	Bayes Variacional			Hamiltoniano MC	
	ICTPC obs.	Ajustados	Dif. abs.	Ajustados	Dif. abs
Álvaro Obregón	13 817.397	14 291.669	13 639.443	474.273	177.954
Azcapotzalco	16 910.051	16 822.117	16 409.086	87.933	500.964
Benito Juárez	25 167.655	25 741.130	24 823.521	573.475	344.135
Coyoacán	14 677.289	15 459.464	14 965.697	782.175	288.408
Cuajimalpa	de 14 948.206	14 228.982	13 706.666	719.225	1241.540
Morelos					
Cuauhtémoc	17 636.413	13 486.522	13 854.120	4149.890	3782.293
Gustavo A. Madero	10 774.117	10 845.581	10 610.828	71.464	163.289
Iztacalco	13 545.988	13 803.319	13 220.033	257.331	325.955
Iztapalapa	8171.228	8311.141	8175.019	139.913	3.791
La Magdalena	10 735.292	10 521.981	10 187.149	213.312	548.144
Contreras					
Miguel Hidalgo	25 668.338	24 660.611	23 727.426	1007.726	1940.911
Milpa Alta	5666.401	6113.938	6044.178	447.537	377.777
Tláhuac	7955.585	8375.282	8236.845	419.697	281.260
Tlalpan	9254.742	8995.394	8950.127	259.348	304.615
Venustiano Carranza	15 085.567	11 719.490	12 629.506	3366.077	2456.061
Xochimilco	7884.059	7968.530	7915.140	84.471	31.081

4.7.1 Validación (entrenamiento-prueba)

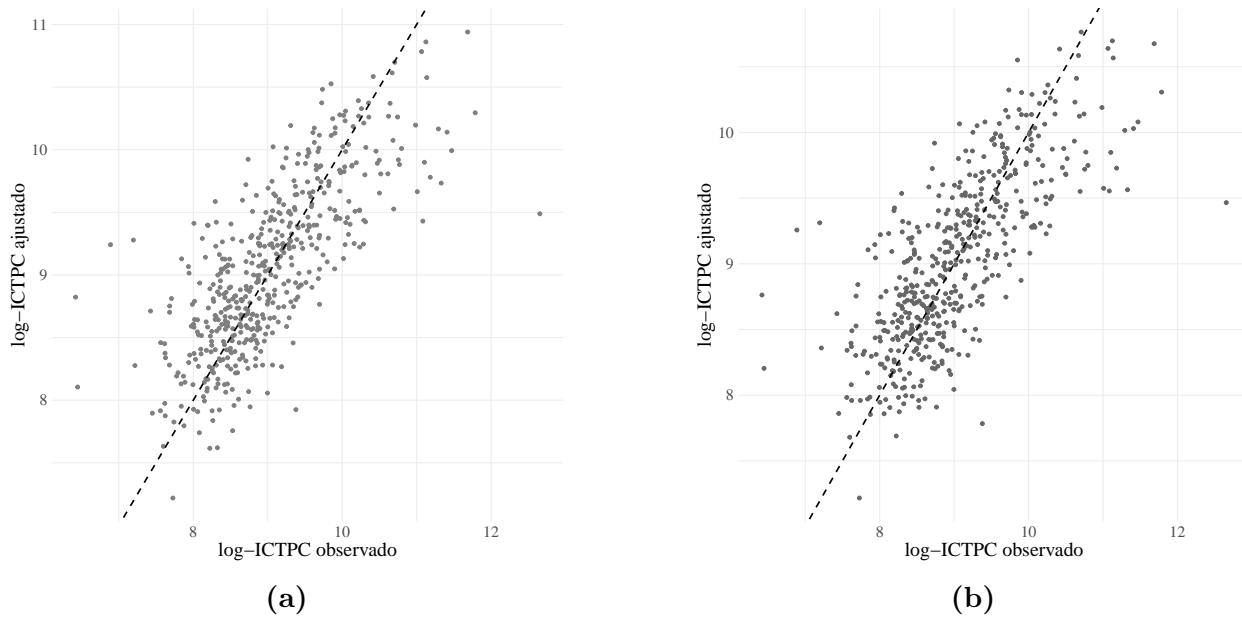


Figura 4.37: Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC.

Cuadro 4.37: Métricas del ajuste entre los valores observados y pronósticos en escala original.
Fuente: elaboración propia.

Métrica	Bayes Variacional	Hamiltoniano MC
MAE	6057.762	6045.277
RMSE	16 962.542	17 143.977
MAPE	0.586	0.573
Tiempo (s)	46.840	1222.350

4.7.2 Pronóstico de nuevas observaciones ($n_i = 0$)

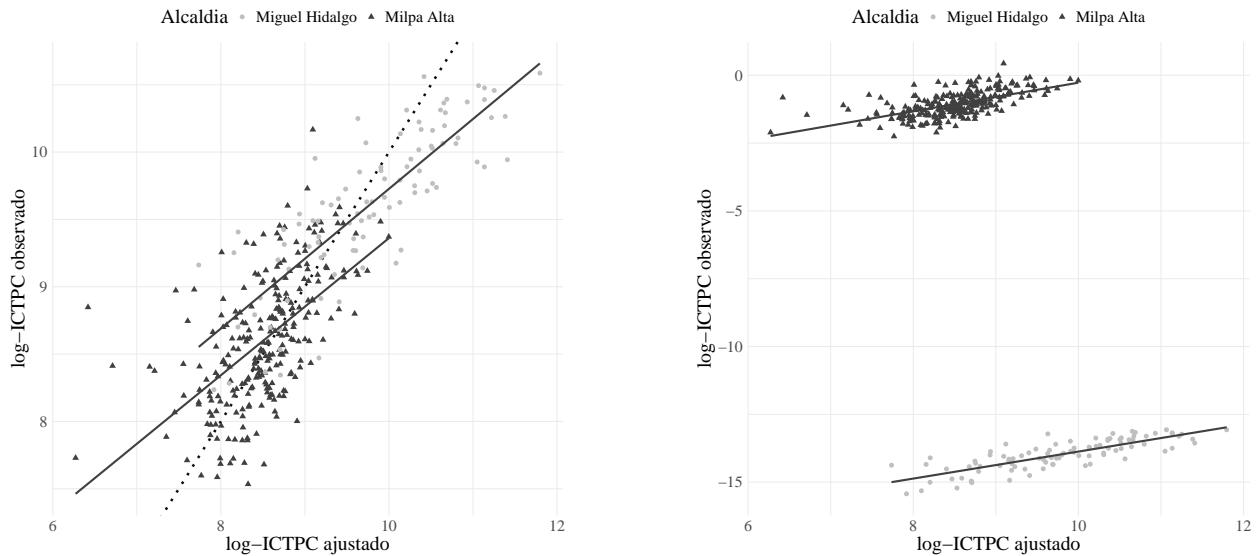


Figura 4.38: Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC. Fuente: elaboración propia.

Cuadro 4.38: Ajuste del Modelo log-normal sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible.

Par.	Bayes Variacional			Hamiltoniano MC				\hat{R}
	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	\hat{R}	
ρ_8	0.119	0.013	0.389	0.792	0.087	0.997	1.035	
ρ_{15}	0.175	0.019	0.542	0.770	0.059	0.995	1.034	
σ^2	0.817	0.796	0.840	0.777	0.661	0.950	0.999	
μ_8	8.175	8.072	8.280	-0.896	-3.411	1.188	1.734	
μ_{15}	8.358	8.195	8.526	-14.589	-42.353	7.186	1.117	

Cuadro 4.39: Métricas del ajuste entre los valores observados y pronósticos en escala original.
Fuente: elaboración propia.

Métrica	Bayes Variacional	Hamiltoniano MC
MAE	5197.489	10 760.931
RMSE	9134.199	18 218.011
MAPE	0.773	0.981
Tiempo (s)	89.650	4992.650

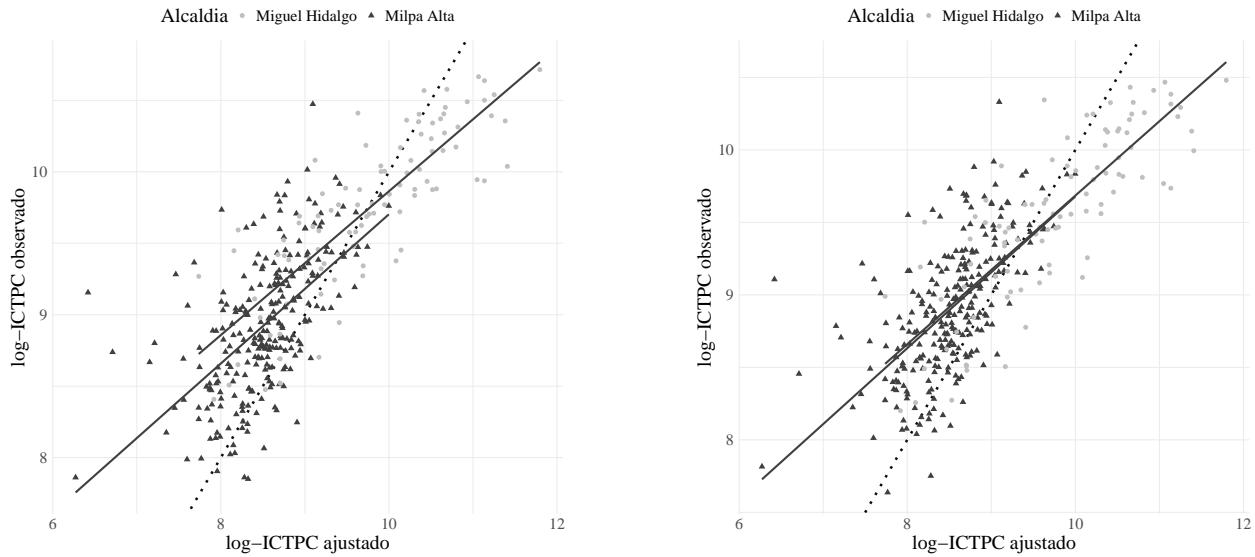


Figura 4.39: Gráficos de dispersión entre la respuesta observada (eje vertical) y la respuesta ajustada (eje horizontal). En el lado izquierdo se muestra el método Bayes variacional y en el derecho Hamiltoniano MC. Fuente: elaboración propia.

Cuadro 4.40: Ajuste del Modelo log-normal sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible.

Par.	Bayes Variacional				Hamiltoniano MC				\hat{R}
	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	\hat{R}		
ρ_8	0.125	0.013	0.425	0.998	0.987	1.000	1.888		
ρ_{15}	0.174	0.022	0.504	0.790	0.063	1.000	1.225		
σ^2	0.878	0.854	0.902	0.657	0.586	0.739	1.011		
μ	8.694	8.672	8.716	9.315	9.031	9.580	1.011		

Cuadro 4.41: Métricas del ajuste entre los valores observados y pronósticos en escala original.
 Fuente: elaboración propia.

Métrica	Bayes Variacional	Hamiltoniano MC
Corr.	0.795	0.754
MAE	8218.140	6297.290
RMSE	11 120.497	10 732.655
MAPE	1.394	0.945
Tiempo (s)	67.850	2889.673

4.8 Ajuste del Modelo probit sesgado con variable latente, datos del ICTPC

Cuadro 4.42: Ajuste del modelo probit sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible.

Par.	Bayes Variacional			Hamiltoniano MC				\hat{R}
	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%		
ρ_1	0.128	0.015	0.411	0.530	0.038	0.891	1.003	
ρ_2	0.150	0.014	0.489	0.589	0.054	0.928	1.007	
ρ_3	0.183	0.018	0.558	0.553	0.051	0.922	0.999	
ρ_4	0.091	0.010	0.314	0.712	0.130	0.921	1.039	
ρ_5	0.136	0.014	0.439	0.815	0.228	0.968	1.026	
ρ_6	0.078	0.008	0.270	0.452	0.025	0.816	1.011	
ρ_7	0.171	0.021	0.549	0.643	0.042	0.943	1.007	
ρ_8	0.083	0.009	0.291	0.358	0.024	0.753	1.017	
ρ_9	0.103	0.011	0.346	0.470	0.025	0.853	1.005	
ρ_{10}	0.142	0.017	0.472	0.462	0.024	0.860	1.001	
ρ_{11}	0.082	0.009	0.283	0.447	0.021	0.821	1.001	
ρ_{12}	0.084	0.010	0.293	0.362	0.015	0.755	1.004	
ρ_{13}	0.130	0.014	0.424	0.553	0.033	0.927	1.026	
ρ_{14}	0.130	0.014	0.472	0.729	0.104	0.952	1.012	

Cuadro 4.42: Ajuste del modelo probit sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible.

	Bayes Variacional				Hamiltoniano MC			
Par.	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	\hat{R}	
ρ_{15}	0.141	0.015	0.443	0.639	0.076	0.940	1.012	
ρ_{16}	0.136	0.013	0.468	0.489	0.028	0.881	1.007	
μ_1	0.333	0.147	0.515	-1.430	-2.672	-0.199	1.008	
μ_2	0.023	-0.189	0.220	-1.519	-3.011	-0.175	1.006	
μ_3	0.519	0.242	0.803	-1.247	-2.667	0.003	1.020	
μ_4	0.433	0.299	0.559	-1.127	-2.636	0.061	1.002	
μ_5	0.323	0.121	0.518	-1.175	-2.921	0.363	1.009	
μ_6	0.552	0.438	0.661	-1.162	-2.199	-0.105	1.016	
μ_7	0.020	-0.234	0.275	-1.428	-3.017	-0.159	1.009	
μ_8	-0.476	-0.609	-0.350	-1.805	-2.741	-0.728	1.026	
μ_9	0.154	0.007	0.304	-1.443	-2.484	-0.327	1.011	
μ_{10}	0.087	-0.161	0.329	-1.451	-2.690	-0.213	1.020	
μ_{11}	-0.524	-0.642	-0.407	-1.995	-3.193	-0.930	1.018	
μ_{12}	-0.151	-0.293	-0.010	-1.668	-2.629	-0.652	1.018	
μ_{13}	-0.256	-0.419	-0.082	-1.634	-3.030	-0.310	1.020	
μ_{14}	0.421	0.230	0.617	-1.229	-2.878	0.217	1.003	
μ_{15}	0.071	-0.128	0.283	-1.585	-3.076	-0.299	1.003	
μ_{16}	0.355	0.150	0.552	-1.369	-2.602	-0.218	1.006	

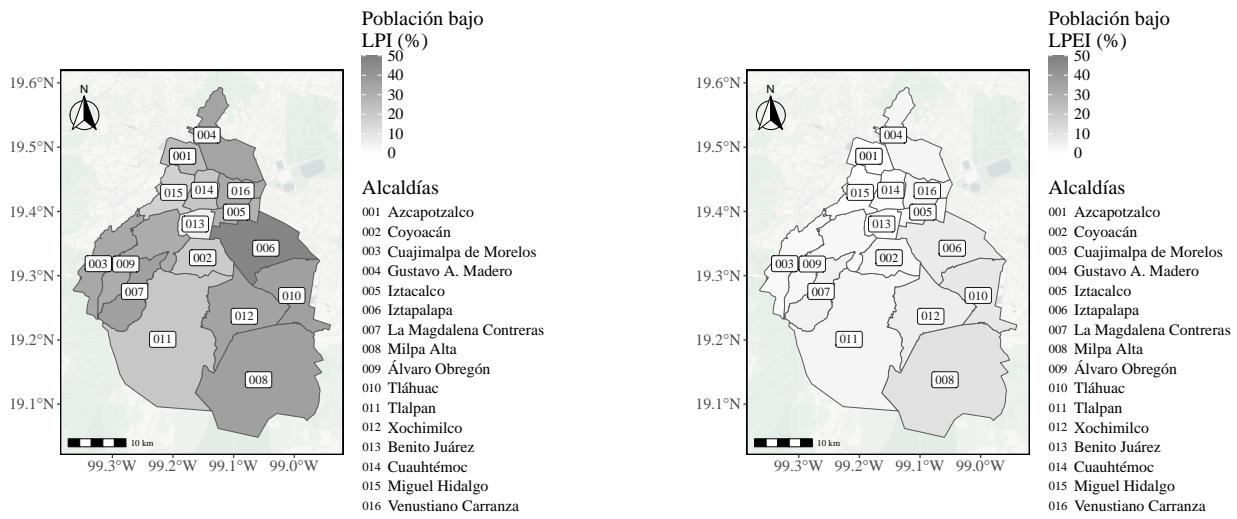
Cuadro 4.43: Los porcentajes de población bajo la LPI son de 31.70% para el método VB y 27.25% para el método HMC.

Alcaldía	Bayes Variacional		Hamiltoniano MC	
		Población bajo	Población bajo LPI (%)	Población bajo
		LPI (%)		LPI (%)
Azcapotzalco		25.328		19.302

Cuadro 4.43: Los porcentajes de población bajo la LPI son de 31.70% para el método VB y 27.25% para el método HMC.

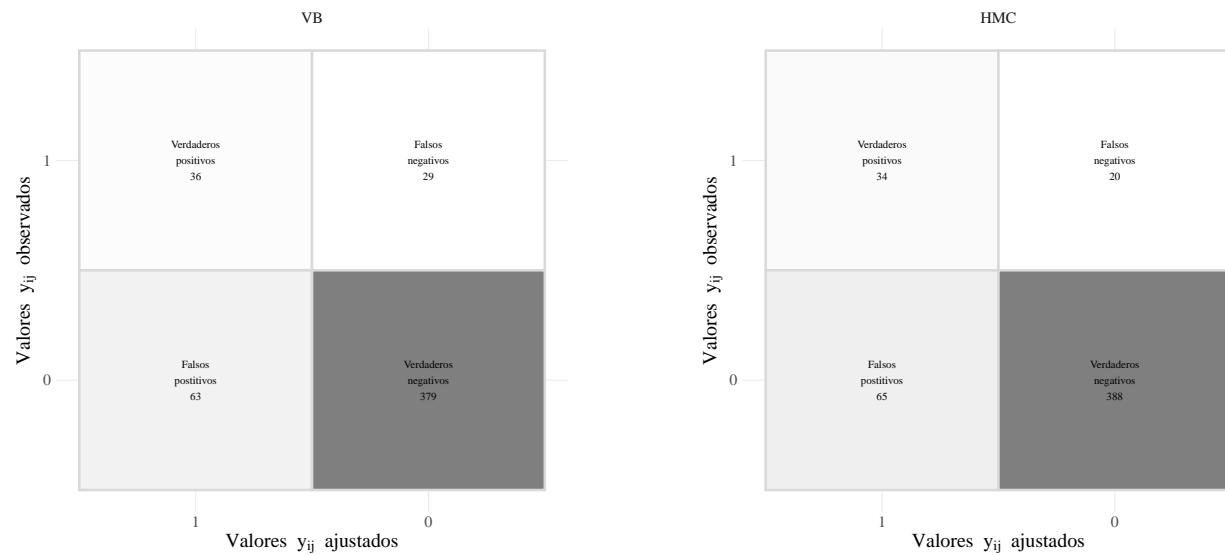
Alcaldía	Bayes Variacional	Hamiltoniano MC
	Población bajo LPI (%)	Población bajo LPI (%)
Coyoacán	19.589	16.330
Cuajimalpa de Morelos	33.091	26.449
Gustavo A. Madero	35.011	32.162
Iztacalco	29.417	29.520
Iztapalapa	47.658	44.382
La Magdalena Contreras	36.487	33.765
Milpa Alta	36.341	25.081
Álvaro Obregón	31.528	24.488
Tláhuac	36.635	29.546
Tlalpan	21.024	12.818
Xochimilco	35.431	27.577
Benito Juárez	5.404	4.910
Cuauhtémoc	21.593	19.171
Miguel Hidalgo	17.448	12.961
Venustiano Carranza	33.021	27.662

Figura 4.40



4.8.1 Validación (entrenamiento-prueba)

Figura 4.41: Matriz de confusión entre la respuesta binaria discretizada observada (columnas) y los valores ajustados (renglones). Fuente: elaboración propia.



Cuadro 4.44: Métricas del ajuste entre los valores binarios observados y pronósticos. Fuente: elaboración propia.

Métrica	Bayes Variacional	Hamiltoniano MC
Acc.	0.819	0.832
TPR	0.554	0.630
TNR	0.554	0.630
F1-Score	0.439	0.444
Tiempo (s)	98.060	5558.220

4.9 Ajuste del Modelo probit ordenado sesgado con variable latente, datos del ICTPC

Cuadro 4.45: Ajuste del modelo probit ordenado sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible.

Par.	Bayes Variacional			Hamiltoniano MC				\hat{R}
	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	\hat{R}	
ρ_1	0.137	0.015	0.449	0.692	0.179	0.920	1.121	
ρ_2	0.138	0.015	0.473	0.471	0.051	0.881	1.020	
ρ_3	0.192	0.024	0.590	0.577	0.091	0.924	1.079	
ρ_4	0.095	0.010	0.340	0.640	0.194	0.849	1.087	
ρ_5	0.148	0.014	0.471	0.683	0.076	0.929	1.004	
ρ_6	0.084	0.008	0.310	0.448	0.049	0.755	1.014	
ρ_7	0.187	0.017	0.592	0.798	0.310	0.939	1.090	
ρ_8	0.086	0.010	0.305	0.395	0.043	0.725	1.002	
ρ_9	0.101	0.011	0.346	0.685	0.218	0.869	1.011	
ρ_{10}	0.154	0.017	0.506	0.353	0.027	0.773	1.003	
ρ_{11}	0.089	0.011	0.313	0.452	0.071	0.832	1.016	
ρ_{12}	0.093	0.011	0.329	0.544	0.075	0.830	1.023	
ρ_{13}	0.132	0.015	0.447	0.397	0.016	0.863	1.160	
ρ_{14}	0.135	0.015	0.426	0.598	0.124	0.876	1.038	
ρ_{15}	0.146	0.016	0.452	0.481	0.045	0.858	1.083	

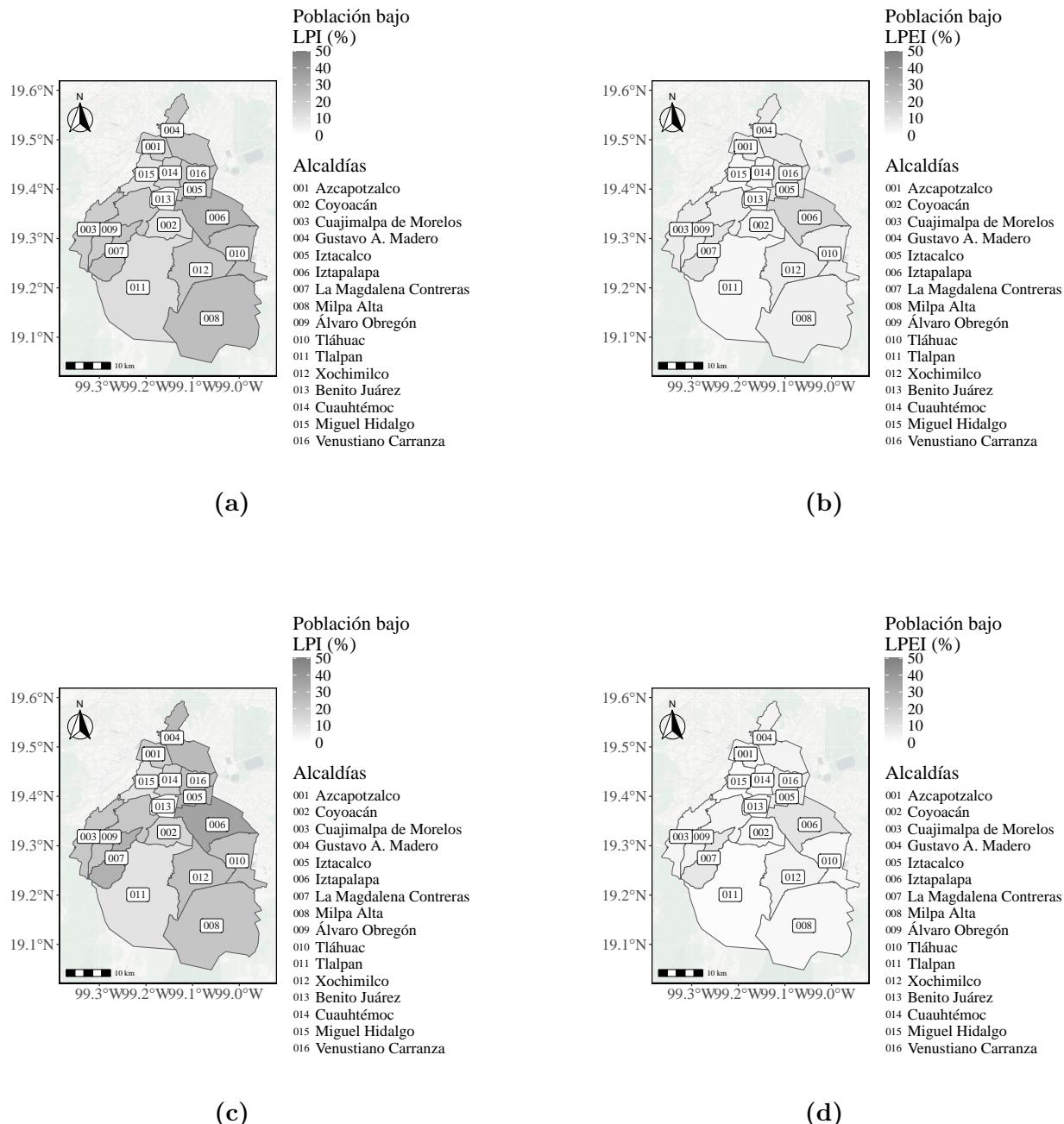
Cuadro 4.45: Ajuste del modelo probit ordenado sesgado con interceptos en cada área pequeña.
Se empleó toda la información disponible.

Par.	Bayes Variacional			Hamiltoniano MC				\hat{R}
	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%		
ρ_{16}	0.132	0.014	0.420	0.510	0.023	0.873	1.085	
δ_1	1.433	1.391	1.476	1.242	1.123	1.359	1.073	
μ_1	0.423	0.217	0.642	2.493	1.002	3.538	1.061	
μ_2	0.687	0.482	0.871	2.713	1.695	3.528	1.001	
μ_3	0.338	0.061	0.601	2.775	1.875	3.478	1.010	
μ_4	0.304	0.174	0.433	2.540	1.522	3.420	1.027	
μ_5	0.205	0.009	0.411	2.361	0.935	3.266	1.002	
μ_6	0.126	0.012	0.239	2.671	2.143	3.371	1.084	
μ_7	0.656	0.409	0.911	2.628	0.906	3.610	1.133	
μ_8	0.976	0.853	1.104	3.224	2.557	3.802	1.012	
μ_9	0.660	0.522	0.804	2.883	1.362	3.761	1.039	
μ_{10}	0.652	0.415	0.879	3.085	2.319	3.714	1.019	
μ_{11}	1.117	0.997	1.245	3.479	2.708	4.124	1.027	
μ_{12}	0.812	0.673	0.950	2.855	1.745	3.721	1.091	
μ_{13}	1.186	1.020	1.354	3.153	2.478	3.705	1.067	
μ_{14}	0.295	0.110	0.508	2.637	1.323	3.603	1.010	
μ_{15}	0.743	0.531	0.954	3.054	2.144	4.027	1.045	
μ_{16}	0.269	0.077	0.475	2.749	2.022	3.512	1.143	

Cuadro 4.46: Ajuste del modelo probit ordenado sesgado con interceptos en cada área pequeña. Se empleó toda la información disponible. Los porcentajes obtenidos con BV son 19.26% y 7.59%, con HMC son 22.61% y 4.88%. Los totales estatales oficiales son 25.7% y 4.5%. Fuente: elaboración propia.

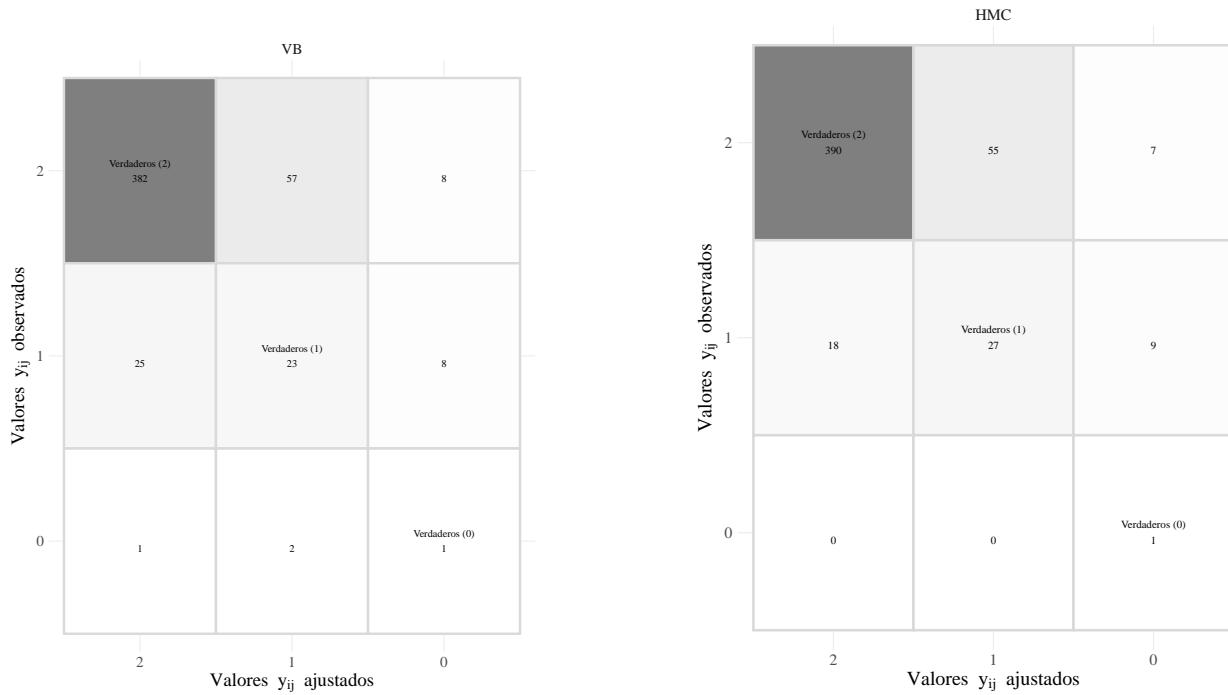
Alcaldía	Bayes Variacional		Hamiltoniano MC	
	Población bajo	Población bajo	Población bajo	Población bajo
	LPI (%)	LPEI (%)	LPI (%)	LPEI (%)
Azcapotzalco	16.1	5.4	18.4	3.2
Coyoacán	11.5	4.3	15.3	3.4
Cuajimalpa de Morelos	18.9	7.1	20.9	3.7
Gustavo A. Madero	21.3	8.0	26.1	4.4
Iztacalco	20.7	8.2	27.5	6.4
Iztapalapa	28.2	15.0	35.4	10.2
La Magdalena Contreras	22.0	9.5	29.6	9.0
Milpa Alta	24.8	5.7	22.1	2.7
Álvaro Obregón	18.7	5.9	21.0	3.5
Tláhuac	22.1	5.9	22.6	2.7
Tlalpan	12.7	3.0	10.6	1.3
Xochimilco	20.8	7.1	23.5	4.5
Benito Juárez	2.3	0.5	2.0	0.4
Cuauhtémoc	14.8	3.4	16.2	2.3
Miguel Hidalgo	10.7	3.0	9.9	1.3
Venustiano Carranza	20.7	9.0	24.4	5.8

Figura 4.42: Estimación del porcentaje de la población bajo la LPI y LPEI, en el primer renglón se muestran las estimaciones a partir del método BV y en el segundo renglón a partir del método HMC.



4.9.1 Validación (entrenamiento-prueba)

Figura 4.43: Matriz de confusión entre la respuesta ordinal observada (columnas) y las categorías ajustadas (renglones). Fuente: elaboración propia.



Cuadro 4.47: Métricas del ajuste entre los valores ordinales observados y pronósticos. Fuente: elaboración propia.

Métrica	Bayes Variacional	Hamiltoniano MC
Acc.	0.801	0.824
τ de Kendall	0.347	0.425
Tiempo (s)	100.280	3147.070

4.10 Estimaciones de β en los tres modelos, datos del ICTPC

Con el propósito de no hacer abrumadora la cantidad de estimaciones de β , se reportan sólo aquellos coeficientes β_k con frecuencia de aparición marginal mayor al 75% en ambos métodos de estimación.

- Para el modelo log-normal sesgado, el método VB determina que 78 de 106 covariables cumplen esta condición, mientras que el método HMC selecciona 24 de 106 covariables.

-
- Por su parte, en el modelo probit sesgado latente se seleccionan 75 y 10 con los métodos BV y HMC.
 - Finalmente, en el modelo probit ordenado sesgado latente se seleccionaron 76 y 12 con los métodos BV y HMC.

Cuadro 4.48: Estimaciones de los coeficientes de regresión en el modelo log-normal sesgado.

β	Bayes Variacional			Hamiltoniano MC			\hat{R}
	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	
β_3	0.253	0.244	0.261	0.053	0.035	0.066	1.026
β_4	0.117	0.107	0.127	0.049	0.039	0.061	1.026
β_7	-0.837	-0.848	-0.826	-0.034	-0.052	-0.003	1.044
β_8	0.603	0.591	0.615	-0.043	-0.058	-0.030	1.002
β_{11}	-0.691	-0.710	-0.672	0.057	0.032	0.080	1.007
β_{18}	0.260	0.234	0.287	0.070	0.043	0.096	1.034
β_{26}	-0.167	-0.197	-0.137	0.079	0.003	0.115	1.033
β_{33}	1.853	1.671	2.034	0.352	0.167	0.498	1.012
β_{34}	-0.271	-0.301	-0.239	0.176	0.128	0.234	1.016
β_{35}	0.102	0.058	0.145	0.109	0.057	0.156	1.044
β_{43}	0.240	0.122	0.343	0.236	0.133	0.359	1.028
β_{60}	-0.928	-0.978	-0.880	0.354	0.280	0.428	1.011
β_{68}	-0.414	-0.437	-0.390	0.128	0.071	0.184	1.004
β_{74}	0.140	0.098	0.185	0.128	0.033	0.206	1.012
β_{81}	0.101	0.045	0.161	0.192	0.088	0.276	1.030
β_{84}	0.331	0.284	0.378	0.323	0.266	0.387	0.999
β_{85}	0.153	0.097	0.211	0.167	0.102	0.230	1.017
β_{98}	-1.237	-1.289	-1.184	2.423	-0.003	6.180	1.001
β_{99}	-1.137	-1.162	-1.113	2.524	0.080	6.270	1.000
β_{100}	-1.013	-1.060	-0.965	2.648	0.194	6.397	1.002
β_{103}	3.636	3.613	3.659	-0.147	-0.212	-0.079	1.077

Cuadro 4.49: Estimaciones de los coeficientes de regresión en el modelo probit sesgado.

β	Bayes Variacional			Hamiltoniano MC			\hat{R}
	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	
β_4	0.278	0.259	0.300	0.147	0.107	0.188	1.045
β_9	-0.285	-0.318	-0.250	-0.232	-0.300	-0.168	1.000
β_{35}	0.544	0.448	0.635	0.327	0.168	0.495	1.032
β_{69}	0.964	0.818	1.110	0.339	-0.003	0.674	1.065
β_{74}	0.375	0.292	0.455	0.364	-0.002	0.641	1.325
β_{78}	0.546	0.326	0.765	0.741	0.327	1.160	1.007
β_{84}	1.356	1.271	1.439	0.756	0.533	1.005	1.073
β_{95}	-1.069	-1.112	-1.026	-0.775	-1.567	0.003	1.074
β_{103}	4.016	3.968	4.063	-0.246	-0.455	0.002	1.041

Cuadro 4.50: Estimaciones de los coeficientes de regresión en el modelo probit ordenado sesgado.

β	Bayes Variacional			Hamiltoniano MC			\hat{R}
	Media p.	I.C. 2.5%	I.C. 97.5%	Media p.	I.C. 2.5%	I.C. 97.5%	
β_4	0.280	0.259	0.300	0.139	0.104	0.185	1.122
β_9	-0.256	-0.292	-0.223	-0.235	-0.283	-0.180	1.018
β_{18}	0.359	0.306	0.412	0.157	0.086	0.232	1.031
β_{27}	-0.313	-0.415	-0.204	-0.296	-0.435	-0.154	0.999
β_{35}	0.434	0.350	0.519	0.263	0.095	0.422	1.048
β_{57}	1.034	0.360	1.695	0.974	-0.060	2.608	1.007
β_{78}	0.813	0.586	1.021	0.653	0.192	1.075	1.335
β_{79}	-0.323	-0.422	-0.223	-0.269	-0.440	-0.093	1.088
β_{84}	1.322	1.238	1.407	0.851	0.662	1.068	1.016
β_{85}	0.476	0.371	0.587	0.267	-0.003	0.569	1.110
β_{105}	-0.846	-1.252	-0.448	-0.499	-1.061	0.004	1.148

CAPÍTULO 5. DISCUSIÓN DE LOS RESULTADOS

Este es el apartado final del documento donde se discutirán las dos clases de resultados obtenidos en el capítulo previo: con simulación y a partir del conjunto de datos real. Este análisis se realiza de acuerdo al orden de aparición de los hallazgos.

5.1 Estudio de simulación

5.1.1 Interacción entre ρ_i y μ_i

Este sencillo experimento busca ilustrar la relación entre el efecto de los parámetros ρ_i y μ_i en la generación de variables y_{ij} binarias, el razonamiento es similar si y_{ij} es ordinal. La intuición dice que si $\rho_i \rightarrow \pm 1$, entonces esperamos una mayor proporción de $y_{ij} = 1$ o $y_{ij} = 0$. No obstante, el proceso para generar y_{ij} depende también del parámetro de localidad μ_i , cuya intuición dicta lo mismo: $\mu_i \rightarrow \pm\infty$ resulta en una mayor proporción de $y_{ij} = 1$ o $y_{ij} = 0$. Es posible observar que

- Cuando $|\mu_i|$ crece, este parámetro domina la proporción de $y_i = 1$ en la simulación.
- Tal como se espera, cuando $\rho \rightarrow 1$, se observa una mayor proporción de $y_i = 1$. $\rho \rightarrow 0$ tiene el efecto contrario.

En el siguientes apartados, se empleó el método del signo para simular variables aleatorias binarias y_{ij} . Esta discusión sugiere que para evitar que todos los y_{ij} colapsen a cero o uno, deben de fijarse con cuidado los valores de μ_i .

5.1.2 Identificación de ρ_i

El segundo experimento realizado compara el ajuste de los tres modelos de regresión propuestos, permitiendo que ρ_i varíe de negativos a positivos. Se considera que, para un modelo en áreas pequeñas, este ajuste es generoso en el sentido de que se entrena con 25%

de la información disponible en cada dominio, lo que mitiga el posible efecto de sesgo.

En general, estos resultados indican que el método Bayesiano variacional de campo medio genera aproximaciones *a posteriori* para ρ_i cuya naturaleza es unimodal, mientras que el método Hamiltoniano MC induce densidades *a posteriori* bimodales: esto sugiere que los valores $-\rho_i$ y ρ_i producen valores de la verosimilitud cercanos entre sí. No obstante, la estructura propuesta para cada modelo de regresión permite que los parámetros sean identificables, es decir, es posible aprenderlos a partir de la muestra.

De manera concreta, para cada modelo de regresión podemos señalar los siguientes aspectos:

- Log-normal sesgado: el método BV puede encontrar el signo de ρ_i cuando su señal o efecto es grande, en caso contrario, sitúa la aproximación en el promedio de los dos extremos, es decir, en cero. Con valores negativos de ρ_i , el método HMC no puede explorar de forma satisfactoria la densidad y colapsa hacia los extremos del soporte.
- Probit sesgado latente: el método BV genera aproximaciones cercanas en magnitud, pero con el sentido contrario.¹ Por su lado, las medias *a posteriori* generadas con HMC recuperan el signo, sin embargo, similar al caso anterior, tampoco genera de forma adecuada la densidad y colapsa hacia los extremos del soporte.
- Probit ordenado sesgado latente: el método VB tiene un comportamiento similar al modelo probit: magnitud cercana pero dirección opuesta. Ahora, el método HMC decide muestrear sólo una región del soporte, pero genera estimaciones con signo contrario.

En general, conforme $|\rho_i|$ se acerca a cero, su efecto sobre y_{ij} disminuye, entonces, si el objetivo es predicción, una estimación pequeña de ρ_i podría reemplazarse de forma segura por $-\rho_i$. Sin embargo, en un modelo cuyo objetivo primordial es la interpretación, el sentido de $|\rho_i|$ adquiere mayor relevancia.

Como se comentó en el [Capítulo 3](#), los coeficientes de correlación $\rho_i \in (-1, 1)$, lo que significa que es posible modelar sesgo tanto a la izquierda como a la derecha. Sin embargo, la aplicación real que nos ocupa sugiere la presencia de sesgo a la derecha. El resultado de

¹Lo cuál recuerda al fenómeno de *label switching*.

estos dos experimentos encausan el resto de las simulaciones: restringimos la atención a los valores de $\rho_i \in (0, 1)$, y para los modelos binarios y ordinales se fijan valores pequeños para μ_i .

5.1.3 Modelo log-normal sesgado

En términos generales, una vez que se han considerado las dos modificaciones previas, este ajuste es el más sencillo y robusto de implementar en comparación con el resto de los modelos. Por tal motivo, podemos considerar el caso general de incluir un intercepto μ_i en cada área pequeña.

- Con ambos ajustes, el método BV encuentra las señales más grandes, que corresponden a ρ_3 y ρ_4 , mientras que el método HMC encuentra todas las señales de ρ_1 a ρ_4 . Por su parte, VB estima con mayor precisión a la varianza σ^2 . En general, el resto de las estimaciones generadas con HMC producen valores más cercanos a los reales.
- Dado que $\rho_1 = 0.5$ puede considerarse un valor pequeño de sesgo, es plausible permitir que la estimación BV encoja esta señal.
- Los gráficos de dispersión de los valores ajustados por ambos métodos contra los valores observados, junto a una recta identidad. A simple vista, no se aprecia diferencia entre ambos ajustes.
- Las métricas de ajuste de ambos métodos son similares. Para el muestreo del 5%, la técnica BV es aproximadamente 5 veces más rápida, mientras que para 25%, la técnica BV es 14 veces más rápida.
- Con ambos porcentajes de muestreo, los dos métodos no exploran muchos modelos, así mismo, son capaces de encontrar el modelo real.

5.1.4 Modelo probit sesgado latente

En términos generales, la aproximación BV de este modelo tiende a encoger las estimaciones de ρ_i . Con el propósito de abordar esta situación y basados en el primer experimento, se ajustan tres tipos de modelos: uno sin interceptos, un único intercepto y varios interceptos.

Cuando se omite este parámetro, se observó que los parámetros de forma ρ_i estimados con el método BV toman valores más grandes, sin embargo, no lograron ser capaces de identificar de forma precisa la señal original. Por otro lado, el método HMC tuvo mejor desempeño al momento de estimar los parámetros ρ_i , μ_i (cuando se incluye) y β en todos los escenarios.

- A pesar de que no se aprecian diferencias evidentes en las matrices de confusión (tablas de valores observados contra ajustados), en algunos casos, las métricas de ajuste favorecen al método BV sobre el método HMC.
- El punto anterior sugiere que la técnica BV aplicada a este modelo prefiere estimar a los parámetros de localidad, es decir, $\mu_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}$, para producir un buen ajuste. Así mismo, tiende a expandir las estimaciones de μ_i (si se incluyen) y de $\boldsymbol{\beta}$.
- La estimación BV siempre fue más rápida, la diferencia más pequeña es en el modelo sin interceptos y muestra del 5%, donde fue aproximadamente 1.6 veces más rápida, mientras que la diferencia más grande es en el modelo con interceptos en cada área pequeña y muestreo 25%: el método BV fue aproximadamente 6 veces más rápida.
- Es posible mejorar la estimación BV de esta colección de parámetros al asignar valores más grandes para `grad_samples` y `elbo_samples` en el método `variational` de la implementación en `cmdstanr`, sin embargo, el tiempo adicional de este ajuste tendría como consecuencia que el muestreo HMC sea más rápido, y entonces se optaría por este método.
- Sólo en el modelo sin interceptos y muestreo 5%, la aproximación VB no fue capaz de encontrar el modelo real. En todos los demás escenarios, esta técnica y HMC le asignaron la probabilidad de aparición más alta al modelo correcto.
- El modelo de regresión probit sesgado latente en áreas pequeñas, ajustado con el método BV, sería preferible sólo ante una gran número de observaciones o de parámetros, es decir, cuando el método HMC sea intratable.

5.1.5 Modelo probit ordenado sesgado latente

Conforme el número de categorías en un modelo ordinal crece sin límite, este se asemeja a un modelo con respuesta continua, de modo que esperamos que la calidad de las estimaciones

se encuentre entre el modelo log-normal sesgado y el probit sesgado latente. Sin embargo, aquí se simula únicamente de dos categorías ordenadas, y los hallazgos son similares a los obtenidos con el modelo binario.

- Las estimaciones de ρ_i se encogen con el método BV, mientras que las estimaciones de μ_i (si se incluyen) y β tienden a expandirse. En contraste, la técnica HMC tiene mejor desempeño para encontrar las señales de ρ_1 a ρ_4 .
- Adicionalmente, la estimación del punto de corte δ_1 es en cierta medida consistente en cada uno de los ajustes presentados. Nuevamente, el método HMC estima de forma más precisa a esta parámetro.
- Todas las métricas de ajuste favorecen al método basado en MCMC, no obstante, el tiempo de ejecución es mayor, en el mejor de los casos, el ajuste BV es aproximadamente 1.8 veces más rápido y 7.6 veces más lento en el escenario contrario.
- De igual modo, todos los ajustes son capaces de encontrar el modelo real. De igual modo, se prefiere al ajuste BV únicamente cuando la alternativa basada en MCMC sea intratable debido al número de observaciones o parámetros.

5.2 Datos del ICTPC

Iniciamos la discusión con un estudio descriptivo acerca del conjunto de datos en la Ciudad de México. En esta entidad, se tienen observaciones en cada uno de los $M = 16$ dominios. A su vez, cada hogar, observado o no, es clasificado de acuerdo al ámbito donde se localiza: urbano o rural. Se observa que únicamente en cuatro alcaldías se tienen registros urbanos y rurales: Milpa Alta, Tlalpan, Tláhuac y Xochimilco.

- En la [Figura 4.34b](#) se muestra la estimación kernel de las densidades del log-ICTPC de acuerdo al tipo de rubro. La densidad estimada en los hogares rurales muestra menor dispersión, lo que se traduce en una mayor concentración de esta variable, esto se confirma en el [Cuadro 4.33](#) con los estadísticos descriptivos por tipo de corte. Así mismo, dentro de este ámbito, la distancia entre la media y la mediana podrían sugerir

que la presencia de sesgo en este ámbito no está muy marcada.

- Por su lado, la [Figura 4.34b](#) también muestra que los hogares urbanos exhiben mayor dispersión, por lo que no se observa un pico en su densidad, además, se nota un ligero sesgo a la derecha. Nuevamente, los estadísticos descriptivos del [Cuadro 4.33](#) confirman que en promedio, los ingresos en este ámbito están más dispersos.
- De forma complementaria, en el [Cuadro 4.33](#) se muestra que la dispersión en el ICTPC de los hogares urbanos es aproximadamente cuatro veces menor a la de los hogares rurales. Adicionalmente, el ICTPC medio en los hogares urbanos es aproximadamente el doble con respecto a los hogares rurales. Esto indica que los ingresos en el contexto rural son más homogéneos pero pequeños.
- Similar a lo anterior, el [Cuadro 4.32](#) muestra que el ICTPC promedio más alto corresponde a Miguel Hidalgo con \$26,668, mientras que el más pequeño se encuentra en el contexto rural de Milpa Alta con \$5,483.
- La [Figura 4.34a](#) muestra la estimación kernel de la densidad para todos los hogares en la Ciudad, este suavizado reesambla una densidad normal asimétrica con sesgo positivo o a la derecha.

En los siguientes apartados, se analizan los resultados obtenidos a partir del ajuste de los tres modelos propuestos. Para el modelo con respuesta continua, se consideró un análisis adicional que es recurrente en la estimación de áreas pequeñas, el cuál consiste en realizar predicciones en algún dominio del cuál no se dispone de información ($n_i = 0$).

5.2.1 Modelo log-normal sesgado

En este modelo, se encontró un desempeño similar entre ambos métodos de inferencia, especialmente en la estimación de σ^2 . Se observó que el método HMC tiende a expandir ρ_i y μ_i con respecto al método BV. Este hecho se refleja al momento de calcular los porcentajes de población por debajo de las líneas de pobreza por ingresos: HMC reporta porcentajes más grandes -especialmente en la LPI- con respecto al método BV.

Con respecto al ajuste del 80% entrenamiento - 20% prueba, basados en las métricas

calculadas y en el gráfico de dispersión, ambos procedimientos tienen desempeños similares, no obstante, se observa una reducción del tiempo con el método VB de aproximadamente 50 veces: en general, en aplicaciones con datos reales, el muestreo HMC se realiza con más de una cadena paralela y con un mayor número de iteraciones, lo que significa mayor tiempo de cómputo. Sin embargo, con el número de iteraciones que se definió, la mayoría de los valores de $\hat{R} < 1.05$ no sugieren falta de convergencia severa.

Por otro lado, cuando $n_i = 0$, la inferencia *a posteriori* representa un reto mayor. En ambos escenarios, el método BV encoge $\tilde{\rho}_i$ hacia cero, lo que implica pronósticos más simétricos de y_{ij}^* . Cuando se usa un intercepto para cada región, las estimaciones $\tilde{\mu}_8 = -0.896$ y $\tilde{\mu}_8 = -14.589$ obtenidas con el método HMC indican problemas en el muestreo, aunado a los valores \hat{R} de 1.7 y 1.1. En el primer escenario, las métricas favorecen al método VB, especialmente en los tiempos de ejecución. Cuando se ajuste el modelo con un único intercepto, este parámetro es informado directamente por el resto de áreas pequeñas, por lo que se corrigen las estimaciones. En este segundo escenario, las métricas de ajuste que corresponden al error, MAE, RMSE y MAPE, favorecen al método HMC.

5.2.2 Modelo probit sesgado latente

En este escenario, la estimación sobre los parámetros ρ_i generada con ambos métodos, obtuvo resultados diferentes. Al igual que en el caso de simulación, los ρ_i estimados con el método BV tienden a encogerse. Las estimaciones negativas de los μ_i , obtenidas con ambos algoritmos, corresponden a que, una vez discretizada la respuesta, la mayoría de las observaciones de la variable de estudio son $y_{ij} = 0$, es decir, la población con ICTPC por encima de la línea de pobreza por ingresos (LPI).

Sin embargo, las métricas de ajuste son bastante similares para ambos métodos, siendo que el tiempo de ejecución del método BV sea aproximadamente 45 veces más rápido. Incluso, las estimaciones del porcentaje de la población bajo la LPI son semejantes entre sí. Por tal motivo, el modelo ajustado con BV es relevante si la intención principal es realizar pronósticos acerca de nuevas observaciones y_{ij}^* binarias.

Por otro lado, los experimentos de simulación indican que es posible mejorar las estimaciones de ρ_i al emplear un único intercepto, o bien, prescindir de este. Sin embargo, esta alternativa sería más adecuada si se busca ganar interpretación sobre los parámetros de correlación/forma.

De forma general, al considerar un mayor número de muestras del gradiente en la implementación variacional de ADVI, es decir, incrementar el argumento `grad_samples`, se pueden obtener estimaciones más precisas en las estimaciones de μ_i y ρ_i , a costa de un aumento considerable en el tiempo de ejecución, no obstante, aún menor al muestreo con el método HMC.

5.2.3 Modelo probit ordenado sesgado latente

Este ajuste tiene un comportamiento similar al modelo binario: el método BV encoge las estimaciones de ρ_i . Aún así, ambos métodos generan valores similares para el umbral δ_1 . En este caso, las estimaciones de μ_i obtenidas con HMC son notablemente mayores a las obtenidas con el método BV, y están centradas en torno a 2.0. Nuevamente, esta estimación refleja que la mayor proporción de valores y_{ij} observados están en la categoría dos, es decir, sin carencias, por lo que estas observaciones exceden el umbral $\tilde{\delta}_1 \simeq 1.24$. A pesar de esto, los valores de $\hat{R}^2 > 1.3$ para las estimaciones de μ_i sugieren falta de convergencia severa.

En contraste con el método anterior, aquí las proporciones de la población bajo las líneas de pobreza son discrepantes. Los porcentaje estimados con el método BV son aproximadamente la mitad de los porcentajes estimados con el método HMC. Además, al comparar estas proporciones con respecto a las obtenidas con el modelo log-normal sesgado, los porcentajes obtenidos aquí son más conservadores, en el sentido de que encogen ciertos porcentajes de cada alcaldía.

CAPÍTULO 6. CONCLUSIONES Y RECOMENDACIONES

El presente estudio tuvo como objetivo principal presentar la estimación de dos clases de modelos de regresión Bayesiana en áreas pequeñas, empleando un enfoque variacional. La primera clase de modelos es de respuesta continua, donde se asume que los errores siguen una distribución normal asimétrica. Así mismo, el segundo tipo de modelos son de clasificación binaria y ordinal, a los cuáles denominamos probit sesgado y probit ordenado sesgado, el atributo sesgado se refiere a que la función liga es la distribución normal sesgada con escala unitaria. Para esta clase de modelos, se consideró el enfoque de variable latente, lo que facilita la implementación y ofrece una interpretación sencilla.

La principal aportación de este trabajo, consistió en implementar un método de inferencia Bayesiana variacional (BV) para la estimación de diversos modelos de regresión en áreas pequeñas. Este paradigma de estimación convierte el problema de muestreo de la densidad *a posteriori* a un problema de optimización, donde se busca la familia de densidades más próximas a la verdadera distribución en términos de la divergencia Kullback-Leibler. La motivación central de este método es aliviar el costo computacional de realizar inferencia Bayesiana, y particularmente, reducir los tiempos de cómputo.

Para esta aplicación en concreto, los experimentos de simulación mostraron que

- Cuando se emplean respuestas continuas, el método BV es bastante competitivo con respecto al método HMC, tanto para encontrar la señal de los parámetros verdaderos como en las métricas de ajuste para el caso de validación-prueba.
- En los modelos de clasificación, el método BV tiende a encoger las estimaciones de los parámetros de forma/correlación, y en cambio expande los parámetros de localidad μ_i y β . Sin embargo, esto no compromete las métricas de ajuste obtenidas en el caso de validación-prueba.

Por su parte, en la aplicación con datos reales se observó que

-
- Con respuestas continuas, ambos métodos de estimación exhiben un comportamiento similar, tanto en estimaciones, métricas y porcentajes de la población bajo alguna línea de pobreza por ingresos.
 - En el modelo binario, al igual que en las simulaciones, el método BV encoge las estimaciones de los parámetros de forma/correlación. No obstante, los dos algoritmos obtienen porcentajes similares de la población bajo la línea de pobreza por ingresos (LPI).
 - Los estadísticos de ajuste indican que, en general, el modelo continuo y binario tienen buen desempeño en el escenario de validación-prueba. Las métricas para el pronóstico generado con el modelo log-normal favorecen al método BV, aunque encoge las estimaciones ρ_8 y ρ_{15} ; por su lado, HMC no puede muestrear de forma satisfactoria el parámetro de localidad μ_{15} , lo que perjudica de forma importante la calidad de las predicciones.
 - En el modelo ordinal es posible estimar porcentajes de la población bajo los umbrales LPI y LPEI. A diferencia del caso previo, los porcentajes obtenidos en la LPI discrepan casi en proporción 1:2. Así mismo, este modelo presentó el rendimiento más bajo en cuanto a los estadísticos de ajuste.

Los métodos variacionales ofrecen un alternativa al muestreo de la distribución *a posteriori*, sin embargo, esta aproximación no siempre puede remplazar la calidad de ajuste obtenida con métodos MCMC, particularmente la estimación de parámetros sensibles o de interés.

En el escenario de respuesta continua, la aproximación BV mostró un desempeño sobresaliente, es decir, es útil para predicción e interpretación, ya que, como se observó los experimentos de simulación, recupera la magnitud de todos los parámetros. Para los casos de clasificación binaria y ordinal, la aproximación BV redujo los tiempos de ejecución y produce pronósticos razonables - basado en las métricas de ajuste-; sin embargo, la desventaja principal es que el proceso de optimización opta por encoger la estimación de los parámetros de correlación/forma.

La implementación Bayesiana variacional que ofrece Stan es automática, lo cuál minimiza el trabajo analítico y de programación. Sin embargo, su propósito general significa que

es posible encontrar alternativas especializadas que proporcionen mejores resultados, por ejemplo, en las estimaciones de los modelos de clasificación, o bien, permitiendo que el parámetro de forma tome valores tanto positivos como negativos para alguna aplicación más general.

6.1 Recomendaciones

- Desarrollar una alternativa Bayesiana Variacional híbrida usando el supuesto de campo medio para todos los parámetros y variables latentes, a excepción de los parámetros de forma o correlación ρ_i y de varianza σ^2 , este último resulta tampoco ser conjugado cuando se centran los parámetros de localidad y escala. Es decir, en lugar de implementar los modelos con el algoritmo de forma fija con densidades gausianas implícitas mediante el algoritmo ADVI de Stan, implementar este esquema híbrido, cuya base es el Ejemplo 3 de la [Sección 2.3](#). Si no se desea calcular derivadas de forma manual, puede emplearse diferenciación automática junto a integración Monte Carlo para estimar la esperanza de los gradientes, en cuyo caso aún se aprovecha el algoritmo CAVI para actualizar los parámetros conjugados μ_i y β . De este modo, se obtienen aproximaciones analíticas que relajan el costo computacional y sobretodo, mejoran la precisión y calidad de la aproximación por medio de pasos analíticos adicionales.
- Emplear alguna estructura *a priori* jerárquica que permita compartir información para mejorar las estimaciones de ρ_i , μ_i y las predicciones y_{ij}^* para dominios donde $n_i = 0$. O bien, fuera del enfoque Bayesiano objetivo, puede emplearse una *a priori* informativa y estudiar su efecto en las estimaciones y pronósticos.
- Por otro lado, como una actividad complementaria, puede estudiarse el planteamiento de los modelos con independencia marginal, es decir, considerar una variable latente por cada observación en el proceso de truncamiento oculto, lo cuál es útil para escenarios fuera de la estimación en áreas pequeñas, por ejemplo, estudiar realizaciones independientes e idénticamente distribuidas.

Bibliografía

- (2025). en. URL: <https://mc-stan.org/docs/reference-manual/mcmc.html#hamiltonian-monte-carlo>.
- (s.f.). en. URL: https://mc-stan.org/docs/cmdstan-guide/variational_config.html.
- Albert, J. H. y S. Chib (1993). "Bayesian Analysis of Binary and Polychotomous Response Data". En: *Journal of the American Statistical Association* 88.422, págs. 669-679. ISSN: 01621459, 1537274X. URL: <http://www.jstor.org/stable/2290350>.
- Arellano-Valle, R. B. y A. Azzalini (2008). "The centred parametrization for the multivariate skew-normal distribution". En: *Journal of Multivariate Analysis* 99.7. Special Issue: Multivariate Distributions, Inference and Applications in Memory of Norman L. Johnson, págs. 1362-1382. ISSN: 0047-259X. DOI: <https://doi.org/10.1016/j.jmva.2008.01.020>. URL: <https://www.sciencedirect.com/science/article/pii/S0047259X08000341>.
- Arellano-Valle, R. B. y A. Azzalini (abr. de 2022). "Some properties of the unified skew-normal distribution". En: *Statistical Papers* 63.2, págs. 461-487. ISSN: 1613-9798. DOI: 10.1007/s00362-021-01235-2. URL: <https://doi.org/10.1007/s00362-021-01235-2>.
- Arnold, B. C., R. J. Beaver et al. (jun. de 2002). "Skewed multivariate models related to hidden truncation and/or selective reporting". En: *Test* 11.1, págs. 7-54.
- Arnold, B. C. y R. J. Beaver (2000). "Hidden Truncation Models". En: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 62.1, págs. 23-35. ISSN: 0581572X. URL: <http://www.jstor.org/stable/25051286> (visitado 13-11-2025).
- Azevedo, C. L., H. Bolfarine y D. F. Andrade (2011). "Bayesian inference for a skew-normal IRT model under the centred parameterization". En: *[Computational Statistics & Data Analysis]* 55.1, págs. 353-365. ISSN: 0167-9473. DOI: <https://doi.org/10.1016/j.csda.2010.05.003>. URL: <https://www.sciencedirect.com/science/article/pii/S0167947310001908>.
- Azzalini, A. (1985). "A Class of Distributions Which Includes the Normal Ones". En: *Scandinavian Journal of Statistics* 12.2, págs. 171-178. ISSN: 03036898, 14679469. URL: <http://www.jstor.org/stable/4615982> (visitado 09-06-2025).
- Azzalini, A. y A. Capitanio (sep. de 1999). "Statistical Applications of the Multivariate Skew Normal Distribution". En: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61.3, págs. 579-602. ISSN: 1467-9868. DOI: [10.1111/1467-9868.00194](https://dx.doi.org/10.1111/1467-9868.00194). URL: <https://dx.doi.org/10.1111/1467-9868.00194>.
- Azzalini, A. y A. D. Valle (dic. de 1996). "The multivariate skew-normal distribution". En: *Biometrika* 83.4, págs. 715-726. ISSN: 0006-3444. DOI: [10.1093/biomet/83.4.715](https://academic.oup.com/biomet/article-pdf/83/4/715/702865/83-4-715.pdf). eprint: <https://academic.oup.com/biomet/article-pdf/83/4/715/702865/83-4-715.pdf>. URL: <https://doi.org/10.1093/biomet/83.4.715>.
- Azzalini, A. (2005). "The Skew-Normal Distribution and Related Multivariate Families". En: *Scandinavian Journal of Statistics* 32.2, págs. 159-188. ISSN: 03036898, 14679469. URL: <http://www.jstor.org/stable/4616868> (visitado 09-06-2025).

-
- Azzalini, A. (2013). *The Skew-Normal and Related Families*. Institute of Mathematical Statistics Monographs. Cambridge University Press.
- Battese, G. E., R. M. Harter y W. A. Fuller (1988). “An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data”. En: *Journal of the American Statistical Association* 83.401, págs. 28-36. ISSN: 01621459, 1537274X. URL: <http://www.jstor.org/stable/2288915> (visitado 21-06-2025).
- Baydin, A. G. et al. (2018). “Automatic differentiation in machine learning: a survey”. En: arXiv: 1502.05767 [cs.SC]. URL: <https://arxiv.org/abs/1502.05767>.
- Bayes, C. L. y M. D. Branco (2007). “Bayesian inference for the skewness parameter of the scalar skew-normal distribution”. En: *Brazilian Journal of Probability and Statistics* 21.2, págs. 141-163. ISSN: 01030752, 23176199. URL: <http://www.jstor.org/stable/43601095> (visitado 16-07-2025).
- Ben Hcine, M. y R. Bouallegue (dic. de 2014a). “Fitting the Log Skew Normal to the Sum of Independent Lognormals Distribution”. En: *[Computer Science & ; Information Technology (CS & IT)]*. NeTCoM 2014. [Academy & Industry Research Collaboration Center (AIRCC)], págs. 54-68. DOI: 10 . 5121 / csit . 2014 . 41305. URL: <http://dx.doi.org/10.5121/csit.2014.41305>.
- Ben Hcine, M. y R. Bouallegue (dic. de 2014b). “Highly Accurate Log Skew Normal Approximation to the Sum of Correlated Lognormals”. En: *[Computer Science & Information Technology (CS & IT)]*. NeTCoM 2014. [Academy & Industry Research Collaboration Center (AIRCC)], págs. 41-52. DOI: 10 . 5121 / csit . 2014 . 41304. URL: <http://dx.doi.org/10.5121/csit.2014.41304>.
- Berger, J. O. y J. M. Bernardo (ago. de 1992). “On the Development of Reference Priors*”. En: *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting, Dedicated to the memory of Morris H. DeGroot, 1931–1989*. Oxford University Press. ISBN: 9780198522669. DOI: 10 . 1093 / oso / 9780198522669 . 003 . 0003. eprint: <https://academic.oup.com/book/0/chapter/422209447/chapter-pdf/52447147/isbn-9780198522669-book-part-3.pdf>. URL: <https://doi.org/10.1093/oso/9780198522669.003.0003>.
- Bertsekas, D. P. y J. N. Tsitsiklis (2018). *Introduction to Probability*. 2^a ed. Athena Scientific.
- Betancourt, M. J. y M. Girolami (2013). *Hamiltonian Monte Carlo for Hierarchical Models*. arXiv: 1312.0906 [stat.ME]. URL: <https://arxiv.org/abs/1312.0906>.
- Betancourt, M. (2018). “A Conceptual Introduction to Hamiltonian Monte Carlo”. En: arXiv: 1701.02434 [stat.ME]. URL: <https://arxiv.org/abs/1701.02434>.
- Bishop, C. (2006). eng. 1th ed. Springer.
- Blei, D. M., A. Kucukelbir y J. D. McAuliffe (abr. de 2017). “Variational Inference: A Review for Statisticians”. En: *Journal of the American Statistical Association* 112.518, págs. 859-877. ISSN: 1537-274X. DOI: 10 . 1080 / 01621459 . 2017 . 1285773. URL: <http://dx.doi.org/10.1080/01621459.2017.1285773>.
- Brooks, S. et al. (mayo de 2011). *Handbook of Markov Chain Monte Carlo*. DOI: 10.1201/b10905. URL: <https://doi.org/10.1201/b10905>.
- Bunch, J. R., C. P. Nielsen y D. C. Sorensen (mar. de 1978). “Rank-one modification of the symmetric eigenproblem”. En: *Numerische Mathematik* 31.1, págs. 31-48. ISSN: 0945-3245. DOI: 10.1007/BF01396012. URL: <https://doi.org/10.1007/BF01396012>.

-
- Carpenter, B. et al. (2017). "Stan: A Probabilistic Programming Language". En: *Journal of Statistical Software* 76.1, págs. 1-32. DOI: [10.18637/jss.v076.i01](https://doi.org/10.18637/jss.v076.i01). URL: <https://www.jstatsoft.org/index.php/jss/article/view/v076i01>.
- Casella, G. y E. I. George (1992). "Explaining the Gibbs Sampler". En: *The American Statistician* 46.3, págs. 167-174. ISSN: 00031305. URL: <http://www.jstor.org/stable/2685208> (visitado 15-06-2025).
- Chib, S. y E. Greenberg (1995). "Understanding the Metropolis-Hastings Algorithm". En: *The American Statistician* 49.4, págs. 327-335. ISSN: 00031305. URL: <http://www.jstor.org/stable/2684568> (visitado 15-06-2025).
- Christen, J. y C. Fox (2010). "A general purpose sampling algorithm for continuous distributions (the t-walk)". En: DOI: [10.1214/10-BA603](https://doi.org/10.1214/10-BA603).
- Coneval (2021). *Metodología para la medición de la pobreza en los municipios de México, 2020*. 3^a ed. México: Coneval.
- Coneval (2023). *Metodología para la medición multidimensional de la pobreza en México*. 3^a ed. México: Coneval.
- DeGroot, M. H. y M. J. Schervish (2014). *Probability and Statistics*. 4^a ed. Pearson.
- Diallo, M. S. y J. N. K. Rao (2018). "Small area estimation of complex parameters under unit-level models with skew-normal errors". En: *Scandinavian Journal of Statistics* 45.4, págs. 1092-1116. DOI: <https://doi.org/10.1111/sjos.12336>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/sjos.12336>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12336>.
- Domínguez-Molina, J. A. et al. (2007). "A Matrix Variate Closed Skew-Normal Distribution with Applications to Stochastic Frontier Analysis". En: *Communications in Statistics - Theory and Methods* 36.9, págs. 1691-1703. DOI: [10.1080/03610920601126126](https://doi.org/10.1080/03610920601126126). eprint: <https://doi.org/10.1080/03610920601126126>. URL: <https://doi.org/10.1080/03610920601126126>.
- Durante, D. (ago. de 2019). "Conjugate Bayes for probit regression via unified skew-normal distributions". En: *Biometrika* 106.4, págs. 765-779. ISSN: 1464-3510. DOI: [10.1093/biomet/asz034](https://doi.org/10.1093/biomet/asz034). URL: <http://dx.doi.org/10.1093/biomet/asz034>.
- Gelman, A., J. B. Carlin y D. B. Rubin (2025). 3rd ed. Springer. URL: <https://sites.stat.columbia.edu/gelman/book/BDA3.pdf>.
- Geman, S. y D. Geman (nov. de 1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6, págs. 721-741. DOI: [10.1109/tpami.1984.4767596](https://doi.org/10.1109/tpami.1984.4767596). URL: <https://doi.org/10.1109/tpami.1984.4767596>.
- George, E. I. y R. E. McCulloch (1993). "Variable Selection Via Gibbs Sampling". En: *Journal of the American Statistical Association* 88.423, págs. 881-889. ISSN: 01621459, 1537274X. URL: <http://www.jstor.org/stable/2290777> (visitado 17-06-2025).
- Ghosh, J. K., M. Delampady y T. Samanta (2006). *An Introduction to Bayesian Analysis: Theory and Methods*. Springer.
- Gilks, W., S. Richardson y D. Spiegelhalter (dic. de 1995). *Markov Chain Monte Carlo in practice*. DOI: [10.1201/b14835](https://doi.org/10.1201/b14835). URL: <https://doi.org/10.1201/b14835>.
- Gómez-Déniz, E. y E. Calderín-Ojeda (2020). "On the Usefulness of the Logarithmic Skew Normal Distribution for Describing Claims Size Data". En: *Mathematical Problems in*

-
- Engineering* 2020.1. 1420618, pág. 1420618. ISSN: 1024-123X. DOI: [10.1155/2020/1420618](https://doi.org/10.1155/2020/1420618). URL: <https://doi.org/10.1155/2020/1420618>.
- Greenberg, E. (2012). *Introduction to Bayesian Econometrics*. 2^a ed. Cambridge University Press.
- Griewank, A. (2003). “A mathematical view of automatic differentiation”. En: *Acta Numerica* 12, págs. 321-398. DOI: [10.1017/S0962492902000132](https://doi.org/10.1017/S0962492902000132).
- Härdle, W. K. y L. Simar (2015). *Applied Multivariate Statistical Analysis*. 3^a ed. Springer.
- Hedlin, D. (2008). “SMALL AREA ESTIMATION: A PRACTITIONER’S APPRAISAL”. En: *Rivista Internazionale di Scienze Sociali* 116.4, págs. 407-417. ISSN: 0035676X, 18277918. URL: <http://www.jstor.org/stable/41625217> (visitado 23-06-2025).
- Hoffman, M. D. y A. Gelman (2014). “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo”. En: *Journal of Machine Learning Research* 15.47, págs. 1593-1623. URL: <http://jmlr.org/papers/v15/hoffman14a.html>.
- INEGI (ago. de 2025). *Análisis de los resultados de la medición de la pobreza multidimensional, 2024*. Inf. téc. Reporte de Resultados 27/25. Instituto Nacional de Estadística y Geografía. URL: https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2025/pm/pm2025_RR_08.pdf.
- Jens, S. (2023). “A Tutorial on Parametric Variational Inference”. En: arXiv: [2301.01236 \[stat.ML\]](https://arxiv.org/abs/2301.01236). URL: <https://arxiv.org/abs/2301.01236>.
- Johnson, R. A. y D. W. Wichern (2007). eng. 6th ed. Upper Saddle River: Prentice Hall.
- Kass, R. E. y L. Wasserman (1996). “The Selection of Prior Distributions by Formal Rules”. En: *Journal of the American Statistical Association* 91.435, págs. 1343-1370. DOI: [10.1080/01621459.1996.10477003](https://doi.org/10.1080/01621459.1996.10477003). eprint: <https://www.tandfonline.com/doi/pdf/10.1080/01621459.1996.10477003>. URL: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10477003>.
- Kucukelbir, A. et al. (2017). “Automatic Differentiation Variational Inference”. En: *Journal of Machine Learning Research* 18.14, págs. 1-45. URL: <http://jmlr.org/papers/v18/16-107.html>.
- Lehmann, E. L. y G. Casella (1998). *Theory of Point Estimation*. 2nd ed. Springer. DOI: <https://doi.org/10.1007/b98854>.
- Liu, J. S. (1994). “The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem”. En: *Journal of the American Statistical Association* 89.427, págs. 958-966. ISSN: 01621459, 1537274X. URL: <http://www.jstor.org/stable/2290921> (visitado 16-06-2025).
- López, A. R. (2024). “Estimación Bayesiana de un modelo de áreas pequeñas con distribución asimétrica en el error”. Tesis doct. Colegio de Postgraduados.
- Margossian, C. C. (mar. de 2019). “A review of automatic differentiation and its efficient implementation”. En: *WIREs Data Mining and Knowledge Discovery* 9.4. ISSN: 1942-4795. DOI: [10.1002/widm.1305](https://doi.org/10.1002/widm.1305). URL: <http://dx.doi.org/10.1002/WIDM.1305>.
- Martínez-Flórez, G., S. Vergara-Cardozo y L. M. González (2013). “The Family of Log-Skew-Normal Alpha-Power Distributions using Precipitation Data”. En: *Revista Colombiana de Estadística* 36.1, págs. 43-57.
- Meent, J.-W. van de et al. (2021). *An Introduction to Probabilistic Programming*. arXiv: [1809.10756 \[stat.ML\]](https://arxiv.org/abs/1809.10756). URL: <https://arxiv.org/abs/1809.10756>.

-
- Mitchell, T. J. y J. J. Beauchamp (1988). "Bayesian Variable Selection in Linear Regression". En: *Journal of the American Statistical Association* 83.404, págs. 1023-1032. DOI: [10.1080/01621459.1988.10478694](https://doi.org/10.1080/01621459.1988.10478694). eprint: <https://www.tandfonline.com/doi/pdf/10.1080/01621459.1988.10478694>. URL: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478694>.
- Mood, A. M., F. A. Graybill y D. C. Boes (1974). *Introduction to the Theory of Statistics*. 3^a ed. McGraw-Hill.
- Morán-Vásquez, R. A., A. D. Giraldo-Melo y M. A. Mazo-Lopera (2023). "Quantile Estimation Using the Log-Skew-Normal Linear Regression Model with Application to Children's Weight Data". En: *Mathematics* 11.17. ISSN: 2227-7390. DOI: [10.3390/math11173736](https://doi.org/10.3390/math11173736). URL: <https://www.mdpi.com/2227-7390/11/17/3736>.
- Murphy, K. P. (2007). "Conjugate Bayesian analysis of the Gaussian distribution". En: URL: <https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf>.
- Nguyen, D. (2013). "An in Depth Introduction to Variational Bayes Note". En: *SSRN*. URL: <https://dx.doi.org/10.2139/ssrn.4541076>.
- Onorati, P. y B. Liseo (2025). "An Extension of the Unified Skew-Normal Family of Distributions and its Application to Bayesian Binary Regression". En: *Journal of Computational and Graphical Statistics* 34.4, págs. 1291-1304. DOI: [10.1080/10618600.2024.2444313](https://doi.org/10.1080/10618600.2024.2444313). eprint: <https://doi.org/10.1080/10618600.2024.2444313>. URL: <https://doi.org/10.1080/10618600.2024.2444313>.
- Ormerod, J. T. y M. P. Wand (2010). "Explaining Variational Approximations". En: *The American Statistician* 64.2, págs. 140-153. ISSN: 00031305. URL: <http://www.jstor.org/stable/20799885> (visitado 23-05-2025).
- Pérez-Rodríguez, P. (2008). "Algunos Aspectos de Inferencia Estadística en la Distribución Normal Asimétrica". Tesis doct. Colegio de Postgraduados.
- Pérez-Rodríguez, P., R. Acosta-Pech et al. (mayo de 2018). "A Bayesian Genomic Regression Model with Skew Normal Random Errors". En: *G3 Genes/Genomes/Genetics* 8.5, págs. 1771-1785. ISSN: 2160-1836. DOI: [10.1534/g3.117.300406](https://doi.org/10.1534/g3.117.300406). eprint: <https://academic.oup.com/g3journal/article-pdf/8/5/1771/40573868/g3journal1771.pdf>. URL: <https://doi.org/10.1534/g3.117.300406>.
- Pérez-Rodríguez, P., J. A. Villaseñor et al. (ene. de 2017). "Bayesian Estimation for the Centered Parameterization of the Skew-Normal Distribution". En: *Revista Colombiana de Estadística* 40.1, págs. 123-140. DOI: [10.15446/rce.v40n1.55244](https://doi.org/10.15446/rce.v40n1.55244). URL: <https://revistas.unal.edu.co/index.php/estad/article/view/55244>.
- Petersen, K. B. y M. S. Pedersen (2012). *The Matrix Cookbook*. Version: November 15, 2012. URL: <https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf>.
- Pichler, C., J. Jewson y A. Avalos-Pacheco (2025). *Probabilistic Programming with Sufficient Statistics for faster Bayesian Computation*. arXiv: [2502.04990 \[stat.CO\]](https://arxiv.org/abs/2502.04990). URL: <https://arxiv.org/abs/2502.04990>.
- Rao, J. e I. Molina (ago. de 2015). *Small area estimation*. DOI: [10.1002/9781118735855](https://doi.org/10.1002/9781118735855). URL: <https://doi.org/10.1002/9781118735855>.
- Robert, C. P. (ene. de 2007). *The Bayesian choice*. DOI: [10.1007/0-387-71599-1](https://doi.org/10.1007/0-387-71599-1). URL: <https://doi.org/10.1007/0-387-71599-1>.

-
- Robert P., C. y G. Casella (2004). 2nd ed. Springer. DOI: <https://doi.org/10.1007/978-1-4757-4145-2>.
- Roberts, G. O. y J. S. Rosenthal (ene. de 2004). “General state space Markov chains and MCMC algorithms”. En: *Probability Surveys* 1.none. ISSN: 1549-5787. DOI: <10.1214/154957804100000024>. URL: <http://dx.doi.org/10.1214/154957804100000024>.
- Rohde, D. y M. P. Wand (2016). “Semiparametric Mean Field Variational Bayes: General Principles and Numerical Issues”. En: *Journal of Machine Learning Research* 17.172, págs. 1-47. URL: <http://jmlr.org/papers/v17/15-276.html>.
- Sáenz Vela, H. M. (2020). “Metodología AF: Pobreza multidimensional en México, 2008 y 2018”. En: *Economía Informa* 420.enero-febrero, págs. 48-62. URL: <https://www.economia.unam.mx/assets/pdfs/econinfo/420/05MetodologiaAF.pdf>.
- Spade, D. A. (2020). “Chapter 1 - Markov chain Monte Carlo methods: Theory and practice”. En: *Principles and Methods for Data Science*. Ed. por A. S. Srinivasa Rao y C. Rao. Vol. 43. Handbook of Statistics. Elsevier, págs. 1-66. DOI: <https://doi.org/10.1016/bs.host.2019.06.001>. URL: <https://www.sciencedirect.com/science/article/pii/S0169716119300379>.
- Stan Development Team (2025a). *Stan Reference Manual, Version 2.37*. en. URL: <https://mc-stan.org/docs/reference-manual/>.
- Stan Development Team (2025b). *Stan User’s Guide, Version 2.37*. en. URL: <https://mc-stan.org/docs/stan-users-guide/>.
- Tran, M.-N., T.-N. Nguyen y V.-H. Dao (2021). “A practical tutorial on Variational Bayes”. En: arXiv: [2103.01327 \[stat.CO\]](2103.01327). URL: <https://arxiv.org/abs/2103.01327>.
- Wand, M. P. (2014). “Fully Simplified Multivariate Normal Updates in Non-Conjugate Variational Message Passing”. En: *Journal of Machine Learning Research* 15.39, págs. 1351-1369. URL: <http://jmlr.org/papers/v15/wand14a.html>.
- Wand, M. P. et al. (2011). “Mean Field Variational Bayes for Elaborate Distributions”. En: *Bayesian Analysis* 6.4, págs. 847-900. DOI: <10.1214/11-BA631>. URL: <https://doi.org/10.1214/11-BA631>.
- Zhang, L. et al. (2022). “Pathfinder: Parallel quasi-Newton variational inference”. En: *Journal of Machine Learning Research* 23.306, págs. 1-49. URL: <http://jmlr.org/papers/v23/21-0889.html>.

ANEXOS

Anexo A

Este apartado presenta brevemente el concepto de identificabilidad en un modelo estadístico. Además, proporciona varias pruebas con referencia al proceso de truncamiento oculto y su relación con la distribución normal sesgada.

Identificabilidad

La identificabilidad es una propiedad importante de un modelo estadístico, determina cuando los parámetros del modelo pueden recuperarse a partir de los datos observados¹. Lehmann y Casella (1998) plantean que el modelo $p(\mathbf{x} | \boldsymbol{\theta})$ es identifiable si el mapeo $\boldsymbol{\theta} \rightarrow p(\mathbf{x} | \boldsymbol{\theta})$ es uno-a-uno, es decir que

$$\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta, [p(\mathbf{x} | \boldsymbol{\theta}_1) = p(\mathbf{x} | \boldsymbol{\theta}_2) \Rightarrow \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2, \quad \forall \mathbf{x}], \quad (\text{A.1})$$

esto quiere decir que diferentes valores de $\boldsymbol{\theta}$ generan diferentes distribuciones, de este modo podemos estimar de forma única a $\boldsymbol{\theta}$ a partir de los datos.

Normal sesgada, caso univariado. Método *i*

Primero, usando un teorema conocido sobre las distribuciones condicionales de la normal multivariada, sean \mathbf{X}_1 y \mathbf{X}_2 dos vectores aleatorios de dimensiones $p_1 \times 1$ y $p_2 \times 2$ tales que se distribuyen conjuntamente como $[\mathbf{X}_1, \mathbf{X}_2]^T \sim N_{p_1+p_2}([\boldsymbol{\mu}_1, \boldsymbol{\mu}_2]^T, \Sigma)$, además (Johnson

¹<https://www.sciencedirect.com/topics/mathematics/identifiability>

y [Wichern 2007](#); [Härdle y Simar 2015](#))

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad (\text{A.2})$$

entonces $\mathbf{X}_1 | (\mathbf{X}_2 = \mathbf{x}_2) \sim N_{p_1}(\boldsymbol{\mu}^*, \Sigma^*)$, donde

$$\begin{aligned} \boldsymbol{\mu}^* &= \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}, \\ \Sigma^* &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned} \quad (\text{A.3})$$

Recordemos que se plantea

$$\begin{bmatrix} V \\ W \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \quad (\text{A.4})$$

y se define $U \equiv V | (W > 0)$. Ahora, veamos que

$$f_{U,W}(u, w) \equiv f_{U|W}(u|w) f_W(w) \quad (\text{A.5})$$

luego, al usar la definición de U podemos escribir

$$= \left[\frac{f_{V|W}(v|w)}{\mathbb{P}(W > 0)} I_{(0, \infty)}(w) \right] f_W(w), \quad (\text{A.6})$$

en la última expresión cambiamos el soporte de W de \mathbb{R} a \mathbb{R}^+ , por lo que es necesario normalizar esta densidad a fin de que integre uno; como $\mathbb{P}(W > 0) = 1/2$ se tiene que

$$f_{U,W}(u | w) = 2f_{V|W}(v|w) f_W(w) I_{(0, \infty)}(w). \quad (\text{A.7})$$

Note que tanto las densidades $f_{V|W}(v | w)$ como $f_W(w)$ se obtienen inmediatamente al usar los resultados previos. $V | (W = w) \sim N(\mu, \sigma^2[1 - \rho^2])$ y $W \sim N(0, \sigma^2)$; note que $2f_W(w)I_{(0, \infty)}(w)$ es una densidad normal truncada a la derecha en cero. Ahora probaremos

que la densidad marginal de U en

$$f_{U,W}(u, w) = 2\phi_{U|W}(u|\mu - \rho w, \sigma^2[1 - \rho^2]) \phi_W(w|0, \sigma^2) I_{(0, \infty)}(w) \quad (\text{A.8})$$

es $SN(u | \mu, \sigma^2, \lambda)$, donde $\lambda = \rho/\sqrt{1 - \rho^2}$. Ahora, escribimos

$$\begin{aligned} f_U(u) &\equiv \int_{\mathbb{R}} 2\phi_{U|W}(u|\mu - \rho w, \sigma^2(1 - \rho^2)) \phi_W(w|0, \sigma^2) I_{(0, \infty)}(w), dw \\ &= 2 \int_0^\infty \phi_{U|W}(u|\mu - \rho w, \sigma^2(1 - \rho^2)) \phi_W(w|0, \sigma^2)(w) dw \\ &= 2 \frac{1}{\sqrt{2\sigma^2(1 - \rho^2)}} \frac{1}{\sqrt{2\sigma^2}} \int_0^\infty \exp\left\{-\frac{(u - \mu - w\rho)^2}{2\sigma^2(1 - \rho^2)}\right\} \exp\left\{-\frac{w^2}{2\sigma^2}\right\} dw \end{aligned}$$

desarollando el cuadrado en w

$$\begin{aligned} &= 2 \frac{1}{\sqrt{2\sigma^2(1 - \rho^2)}} \frac{1}{\sqrt{2\sigma^2}} \int_0^\infty \exp\left\{-\frac{(u - \mu)^2 + w^2\rho^2 - 2\rho(u - \mu)w + (1 - \rho^2)w^2}{2\sigma^2(1 - \rho^2)}\right\} dw \\ &= 2 \underbrace{\frac{1}{\sqrt{2\sigma^2(1 - \rho^2)}} \frac{1}{\sqrt{2\sigma^2}}}_{C} \int_0^\infty \exp\left\{-\frac{(u - \mu)^2 + w^2 - 2\rho(u - \mu)w}{2\sigma^2(1 - \rho^2)}\right\} dw \end{aligned}$$

completamos el término al cuadrado en w , sumamos y restamos $\rho^2(u - \mu)^2$

$$\begin{aligned} &= 2C \int_0^\infty \exp\left\{-\frac{(u - \mu)^2 + w^2 - 2\rho(u - \mu)w + \rho^2(u - \mu)^2 - \rho^2(u - \mu)^2}{2\sigma^2(1 - \rho^2)}\right\} dw \\ &= 2C \int_0^\infty \exp\left\{-\frac{(u - \mu)^2 + [w - \rho(u - \mu)]^2 - \rho^2(u - \mu)^2}{2\sigma^2(1 - \rho^2)}\right\} dw \end{aligned}$$

factorizando $(u - \mu)^2$

$$= 2C \int_0^\infty \exp\left\{-\frac{[w - \rho(u - \mu)]^2 + (1 - \rho^2)(u - \mu)^2}{2\sigma^2(1 - \rho^2)}\right\} dw,$$

podemos extraer el término constante,

$$= 2 \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(u - \mu)^2\right\} \int_0^\infty \exp\left\{-\frac{[w - \rho(u - \mu)]^2}{2\sigma^2(1 - \rho^2)}\right\} dw,$$

fuera de la integral podemos identificar una densidad normal en u , con media μ y varianza σ^2 . Así mismo, podemos identificar el argumento en la integral como el kernel de una densidad normal en w , con media $\rho(u - \mu)$ y varianza $\sigma^2(1 - \rho^2)$, de hecho, note que las constantes en C normalizan ambas densidades Gaussianas, para mejor visualización, re-acomodamos términos y escribimos

$$\begin{aligned} &= 2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(u-\mu)^2\right\} \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2(1-\rho^2)}} \exp\left\{-\frac{[w-\rho(u-\mu)]^2}{2\sigma^2(1-\rho^2)}\right\} dw, \\ &= 2\phi_U(u|\mu, \sigma^2) \int_0^\infty \phi_W(w|\rho(u-\mu), \sigma^2[1-\rho^2]) dw, \end{aligned}$$

podemos identificar la integral como la probabilidad $\mathbb{P}(W > 0)$, donde $W \sim N(\rho(u - \mu), \sigma^2[1 - \rho^2])$, para calcular esta probabilidad, estandarizamos la densidad por su localidad y escala, así, podemos escribir

$$\begin{aligned} &= 2\phi_U(u|\mu, \sigma^2) \mathbb{P}\left(\frac{W - \rho(u - \mu)}{\sigma\sqrt{1 - \rho^2}} > -\frac{\rho(u - \mu)}{\sigma\sqrt{1 - \rho^2}}\right) \\ &= 2\phi_U(u|\mu, \sigma^2) \left[1 - \Phi\left(-\frac{\rho(u - \mu)}{\sigma\sqrt{1 - \rho^2}}\right)\right], \end{aligned}$$

por simetría de la función de distribución Φ , $1 - \Phi(-x) = \Phi(x)$, por tanto escribimos

$$\begin{aligned} &= 2\phi_U(u|\mu, \sigma^2) \Phi\left(-\frac{\rho(u - \mu)}{\sigma\sqrt{1 - \rho^2}}\right) \\ &= 2\phi\left(\frac{u - \mu}{\sigma}\right) \Phi\left(\lambda \frac{u - \mu}{\sigma}\right) \quad \square \end{aligned}$$

Normal sesgada, caso univariado. Método *ii*

Alternativamente, se muestra un camino más corto para mostrar que la densidad de X obtenida mediante el proceso de truncamiento oculto es normal asimétrica. En general, usamos la definición de condicionar una o más variables aleatorias a un evento de estas (Bertsekas y Tsitsiklis 2018; DeGroot y Schervish 2014; Mood, Graybill y Boes 1974). Así,

definimos $(X, W) \equiv (U, W) \mid (W > 0)$. de este modo, podemos escribir que

$$f_{(U, W) \mid (W > 0)}(u, w) \equiv \frac{f_{U, W}(u, w)}{\mathbb{P}(W > 0)} 1(w > 0),$$

así, es posible obtener la densidad de $U \mid (W > 0)$ marginalizando sobre W , es decir

$$\begin{aligned} f_{U \mid (W > 0)}(u) &= \int \frac{f_{U, W}(u, w)}{\mathbb{P}(W > 0)} 1(w > 0) dw \\ &= f_U(u) \int \frac{f_{W \mid U}(w \mid u)}{\mathbb{P}(W > 0)} 1(w > 0) dw \end{aligned}$$

note que $\mathbb{P}(W > 0) = 1/2$ es constante, además, $W \mid U$ tiene ley Gaussiana dada por

$$f_{W \mid U=u}(w \mid u) = N(w \mid \rho(u - \mu), \sigma^2(1 - \rho^2)),$$

juntando estas dos partes podemos escribir

$$\begin{aligned} &= 2f_U(u) \int_0^\infty N(w \mid \rho(u - \mu), \sigma^2(1 - \rho^2)) dw \\ &= 2N(u \mid \mu, \sigma^2) \Phi\left(\lambda \frac{u - \mu}{\sigma}\right), \end{aligned}$$

con $\lambda = \rho/\sqrt{1 - \rho^2}$.

Normal sesgada, caso univariado. Método *iii*

Otro método directo para encontrar la densidad marginal de X consiste en emplear el teorema de Bayes a la densidad de $U \mid (W > 0)$:

$$f_{U \mid (W > 0)}(u) = \frac{\mathbb{P}(W > 0 \mid U = u) f_U(u)}{\mathbb{P}(W > 0)},$$

ya mostramos que $W \mid (U = u)$ tiene densidad Gaussiana, por tanto

$$\mathbb{P}(W > 0 \mid U = u) \equiv 1 - \Phi\left(-\frac{\rho(u - \mu)}{\sqrt{\sigma^2(1 - \rho^2)}}\right) = \Phi\left(\lambda \frac{u - \mu}{\sigma}\right),$$

juntando estas partes, la densidad de $X \equiv U \mid (W > 0)$ puede ser escrita como

$$f_X(x) \equiv f_{U \mid (W > 0)}(u) = 2N(u \mid \mu, \sigma^2)\Phi\left(\lambda \frac{u - \mu}{\sigma}\right).$$

En general, se evita el uso de la representación directa de la densidad normal asimétrica, es decir, la densidad marginal de X , y en cambio se considera la representación conjunta de (X, W) generada a partir del proceso de truncamiento oculto. Así, a partir de las expresiones previas, esta se obtiene como

$$\begin{aligned} f_{(X, W)}(x, w) &\equiv \frac{f_{(U, W)}(u, w)}{\mathbb{P}(W > 0)} 1(w > 0) \\ &= 2f_{U \mid (W=w)}(u)f_W(w) 1(w > 0), \end{aligned}$$

note que $U \mid (W = w)$ tiene ley Gaussiana dada por

$$f_{U \mid (W=w)}(u, w) = N(u \mid \mu + \rho w, \sigma^2(1 - \rho^2)),$$

así, la densidad conjunta (X, W) está dada por

$$f_{X, W}(x, w) = 2N(u \mid \mu + \rho w, \sigma^2(1 - \rho^2)) N(w \mid 0, \sigma^2) I_{(0, \infty)}(w).$$

En general, en la última expresión reemplazamos la variable u por x .

Normal sesgada, caso multivariado. Método *i*

Ahora estudiemos el caso multivariado para obtener la densidad de \mathbf{X}_n , sea

$$\begin{bmatrix} \mathbf{U} \\ W \end{bmatrix} \sim N_{n+1} \left(\begin{bmatrix} \boldsymbol{\mu} \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 \bar{\Sigma}_U & \sigma^2 \rho \mathbf{1}_n \\ \sigma^2 \rho \mathbf{1}_n^T & \sigma^2 \end{bmatrix} \right),$$

donde $\bar{\Sigma}_U = \sigma^2 ((1 - \rho^2)I_n + \rho^2 J_n)$, vale la pena notar que $(\bar{\Sigma}_U)_{ij} = \sigma^2$ si $i = j$ y $\sigma^2 \rho^2$ si $i \neq j$. Ahora, aunque la elección de esta matriz de covarianzas para \mathbf{U} pueda resultar extraña, garantiza que $\mathbf{X} \equiv \mathbf{U} \mid W > 0$ tenga ley normal asimétrica multivariada y particularmente cada $X_j \equiv U_j \mid (W > 0)$ tiene ley normal asimétrica univariada. A continuación mostraremos estas dos afirmaciones y para ello, empleamos una estrategia similar para el caso bivariado: veamos que

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &\equiv f_{(\mathbf{U}) \mid (W > 0)}(\mathbf{u}, w) \\ &= \frac{\mathbb{P}(W > 0 \mid (\mathbf{U} = \mathbf{u})) f_{\mathbf{U}}(\mathbf{u})}{\mathbb{P}(W > 0)}, \end{aligned}$$

nuevamente, $\mathbb{P}(W > 0) = 1/2$ y necesitamos calcular la densidad de $W \mid (\mathbf{U} = \mathbf{u})$, la cuál sabemos es Gausiana:

$$\begin{aligned} f_{W \mid (\mathbf{u} = \mathbf{u})}(w \mid \mathbf{u}) &= N(w \mid 0 + \sigma^2 \rho \mathbf{1}_n^T (\sigma^2 \bar{\Sigma}_U)^{-1} (\mathbf{u} - \boldsymbol{\mu}), \sigma^2 - \sigma^2 \rho \mathbf{1}_n (\sigma^2 \bar{\Sigma}_U)^{-1} \sigma^2 \rho \mathbf{1}_n^T) \\ &= N(w \mid \rho \mathbf{1}_n^T \bar{\Sigma}_U^{-1} (\mathbf{u} - \boldsymbol{\mu}), \sigma^2 (1 - \rho^2 \mathbf{1}_n \bar{\Sigma}_U^{-1} \mathbf{1}_n^T)), \end{aligned}$$

dada la estructura propuesta de $\bar{\Sigma}_U$, no es difícil calcular (1) la inversa $\bar{\Sigma}_U^{-1}$ y (2) la forma cuadrática $\mathbf{1}_n^T \bar{\Sigma}_U^{-1} \mathbf{1}_n$ asociada. Para la primera tarea, la fórmula de Sherman–Morrison, un caso particular de la fórmula Sherman–Morrison–Woodbury: sea A una matriz con entradas reales e invertible, sean \mathbf{a}, \mathbf{b} dos vectores columna, entonces $A + \mathbf{a}\mathbf{b}^T$ es invertible si y sólo si $1 + \mathbf{b}^T A^{-1} \mathbf{a} \neq 0$. Así, se tiene que

$$(A + \mathbf{a}\mathbf{b}^T)^{-1} = A^{-1} - \frac{A^{-1} \mathbf{a} \mathbf{b}^T A^{-1}}{1 + \mathbf{b}^T A^{-1} \mathbf{a}}. \quad (\text{A.9})$$

En nuestro caso podemos identificar $A = (1 - \rho^2)I_n$ y $\mathbf{a} = \mathbf{b} = \rho\mathbf{1}_n$. De este modo, escribimos

$$\begin{aligned} [(1 - \rho^2)I_n + \rho^2\mathbf{1}_n\mathbf{1}_n^T]^{-1} &= \frac{1}{1 - \rho^2}I_n - \frac{\frac{1}{1-\rho^2}I_n\rho\mathbf{1}_n\rho\mathbf{1}_n^T\frac{1}{1-\rho^2}I_n}{1 + \rho\mathbf{1}_n^T\frac{1}{1-\rho^2}I_n\rho\mathbf{1}_n} \\ &= \frac{1}{1 - \rho^2}I_n - \frac{\rho^2/(1 - \rho^2)^2}{1 + n\rho^2/(1 - \rho^2)}J_n \\ &= \frac{1}{1 - \rho^2}I_n - \frac{1}{1 - \rho^2}\frac{\rho^2}{(n - 1)\rho^2 + 1}J_n \\ &= \frac{1}{1 - \rho^2}\left(I_n - \frac{\rho^2}{(n - 1)\rho^2 + 1}J_n\right) = \bar{\Sigma}_U^{-1}, \end{aligned}$$

luego, para la tarea (2) calculamos la forma cuadrática como sigue

$$\begin{aligned} \mathbf{1}_n^T\bar{\Sigma}_U^{-1}\mathbf{1}_n &= \mathbf{1}_n^T\frac{1}{1 - \rho^2}\left(I_n - \frac{\rho^2}{(n - 1)\rho^2 + 1}(\mathbf{1}_n\mathbf{1}_n^T)\right)\mathbf{1}_n \\ &= \frac{1}{1 - \rho^2}\left(\cancel{\mathbf{1}_n^T}I_n\cancel{\mathbf{1}_n} - \frac{\rho^2}{(n - 1)\rho^2 + 1}(\cancel{\mathbf{1}_n^T}\cancel{\mathbf{1}_n})(\cancel{\mathbf{1}_n^T}\cancel{\mathbf{1}_n})\right)^{n^2} \\ &= \frac{n}{1 - \rho^2}\left(1 - \frac{n\rho^2}{(n - 1)\rho^2 + 1}\right) \\ &= \frac{n}{(n - 1)\rho^2 + 1}, \\ \Rightarrow 1 - \rho^2\mathbf{1}_n^T\Sigma_U^{-1}\mathbf{1}_n &= \frac{1 - \rho^2}{(n - 1)\rho^2 + 1} \end{aligned}$$

por lo tanto, podemos escribir la densidad de $W \mid (\mathbf{U} = \mathbf{u})$ como

$$f_{W|\mathbf{U}_n}(w \mid \mathbf{u}_n) = N\left(w \mid \rho\mathbf{1}_n^T\bar{\Sigma}_U^{-1}(\mathbf{u}_n - \boldsymbol{\mu}_n), \sigma^2\frac{1 - \rho^2}{(n - 1)\rho^2 + 1}\right),$$

desarrollando análogamente a la media de $W \mid (\mathbf{U}_n = \mathbf{u}_n)$,

$$\begin{aligned} \mathbf{1}_n^T\frac{1}{1 - \rho^2}\left(I_n - \frac{\rho^2}{(n - 1)\rho^2 + 1}(\mathbf{1}_n\mathbf{1}_n^T)\right)^{-1} &= \frac{1}{1 - \rho^2}\left(\cancel{\mathbf{1}_n^T}I_n\cancel{\mathbf{1}_n} - \frac{\rho^2}{(n - 1)\rho^2 + 1}(\cancel{\mathbf{1}_n^T}\cancel{\mathbf{1}_n})^n\mathbf{1}_n^T\right) \\ &= \frac{1}{1 - \rho^2}\mathbf{1}_n^T\left(1 - \frac{n\rho^2}{(n - 1)\rho^2 + 1}\right) \\ &= \frac{1}{(n - 1)\rho^2 + 1}\mathbf{1}_n^T \\ \Rightarrow f_{W|(\mathbf{U}=\mathbf{u})} &= N\left(w \mid \rho\frac{1}{(n - 1)\rho^2 + 1}\mathbf{1}_n^T(\mathbf{u} - \boldsymbol{\mu}), \sigma^2\frac{1 - \rho^2}{(n - 1)\rho^2 + 1}\right), \end{aligned}$$

note que es posible emplear la densidad donde hemos desarrollado la media de $\mathbf{W} \mid (\mathbf{U} = \mathbf{u})$, pero para mantener más simple la notación, continuamos así. Ahora, escribimos

$$\begin{aligned}\mathbb{P}(W > 0 \mid (\mathbf{U} = \mathbf{u})) &= 1 - \left(\frac{0 - \rho \mathbf{1}_n^T \bar{\Sigma}_U^{-1}(\mathbf{u} - \boldsymbol{\mu})}{\sqrt{\sigma^2 \frac{1-\rho^2}{(n-1)\rho^2+1}}} \right) \\ &= \Phi \left(\frac{\rho \sqrt{(n-1)\rho^2 + 1}}{\sqrt{1-\rho^2}} \mathbf{1}_n^T \bar{\Sigma}_U^{-1}(\mathbf{u} - \boldsymbol{\mu}) \right) \\ &= \Phi \left(\lambda \sqrt{(n-1)\rho^2 + 1} \mathbf{1}_n^T \bar{\Sigma}_U^{-1}(\mathbf{u} - \boldsymbol{\mu}) \right), \quad (\lambda = \rho / \sqrt{1-\rho^2})\end{aligned}$$

y si preferimos desarrollar $\mathbf{1}_n^T \bar{\Sigma}_U^{-1}$:

$$\begin{aligned}\mathbf{1}_n^T \bar{\Sigma}_U^{-1} &= \mathbf{1}_n^T \frac{1}{1-\rho^2} \left(I_n - \frac{\rho^2}{(n-1)\rho^2+1} J_n \right) \\ &= \mathbf{1}_n^T \frac{1}{1-\rho^2} \left(1 - \frac{n\rho^2}{(n-1)\rho^2+1} \right) \\ &= \lambda \mathbf{1}_n^T \frac{1}{1-\rho^2} \frac{1-\rho^2}{(n-1)\rho^2+1} \\ &= \lambda \mathbf{1}_n^T \frac{1}{(n-1)\rho^2+1} \\ &\Rightarrow \Phi \left(\frac{\lambda}{\sqrt{(n-1)\rho^2+1}} \mathbf{1}_n^T(\mathbf{u} - \boldsymbol{\mu}) \right),\end{aligned}$$

¿es posible encontrar alguna relación entre el término $1/\sqrt{(n-1)\rho^2+1}$ con $\bar{\Sigma}_U$? es decir, si $1/\sqrt{(n-1)\rho^2+1}$ resulta ser una potencia de $\bar{\Sigma}_U$, entonces sería posible escribir el parámetro de forma como $\boldsymbol{\lambda}_n \equiv \lambda \mathbf{1}_n$. Calculemos $\bar{\Sigma}_U^{1/2}$ recurriendo a la descomposición espectral: sea A una matriz cuadrada con entradas reales, definimos la pareja $(\mathbf{v}_i, \lambda_i)$ como un eigenvector-eigenvalor de A si satisfacen la ecuación

$$A\mathbf{v}_i = \mathbf{v}_i \lambda_i,$$

de este modo, A tiene exactamente n eigenvectores-eigenvalores (no necesariamente diferentes). La descomposición espectral de A está dada por

$$A = \sum_{i=1}^n \lambda_i \mathbf{v}_i.$$

Una ventaja de la descomposición espectral es que podemos calcular potencias de A de forma simple. Ahora, encontraremos las n parejas de eigenvectores-eigenvalores para Σ_U , en (Bunch, Nielsen y Sorensen 1978, Teorema 1. Sección (2.3)-(2.4)) se muestra como encontrar todos los eigenvalores para una perturbación de rango uno de una matriz simétrica, $\bar{\Sigma}_U$ en nuestro caso. Sin embargo, simplificamos este proceso mediante propuestas apropiadas de \mathbf{v}_i y λ_i : sea $\mathbf{v}_1 = \frac{1}{\sqrt{n}} \mathbf{1}_n$, note que

$$\begin{aligned}\bar{\Sigma}_U \mathbf{v}_1 &= ((1 - \rho^2)I_n + \rho^2 J_n) \frac{1}{\sqrt{n}} \mathbf{1}_n \\ &= ((1 - \rho^2)\mathbf{1}_n + n\rho^2 \mathbf{1}_n) \frac{1}{\sqrt{n}} \\ &= ((n - 1)\rho^2 + 1) \frac{1}{\sqrt{n}} \mathbf{1}_n \\ &= ((n - 1)\rho^2 + 1) \mathbf{v}_1,\end{aligned}$$

es decir, $\lambda_1 = (n - 1)\rho^2 + 1$ es el eigenvalor asociado a \mathbf{v}_1 . Ahora, encontremos la segunda pareja de eigenvalores-eigenvectores: partiendo nuevamente de la definición

$$\begin{aligned}\bar{\Sigma}_U \mathbf{v}_2 &= ((1 - \rho^2)I_n + \rho^2 J_n) \mathbf{v}_2 \\ &= (1 - \rho^2)\mathbf{v}_2 + \rho^2(J_n \mathbf{v}_2),\end{aligned}$$

ahora, queremos encontrar la pareja $(\lambda_2, \mathbf{v}_2)$ tal que $\bar{\Sigma}_U \mathbf{v}_2 = \mathbf{v}_2 \lambda_2$, por ejemplo, si tomamos $\lambda_2 = 1 - \rho^2$, entonces escribimos

$$\begin{aligned}\bar{\Sigma}_U \mathbf{v}_2 &= (1 - \rho^2)\mathbf{v}_2 &&\iff \\ (1 - \rho^2)\mathbf{v}_2 + \rho^2(J_n \mathbf{v}_2) &= (1 - \rho^2)\mathbf{v}_2 &&\iff \\ \rho^2(J_n \mathbf{v}_2) &= \mathbf{0} &&\iff\end{aligned}$$

$$J_n \mathbf{v}_2 = \mathbf{0}$$

es decir, \mathbf{v}_2 es tal que $J_n \mathbf{v}_2 = \mathbf{1}_n (\mathbf{1}_n^T \mathbf{v}_2) = \mathbf{0}$, en otras palabras, \mathbf{v}_2 es cualquier vector cuyas entradas suman cero. Ahora, sustituyendo la condición previa, se tiene que

$$\bar{\Sigma}_U \mathbf{v}_2 = (1 - \rho^2) \mathbf{v}_2 + \rho^2 (J_n \mathbf{v}_2) \rightarrow \mathbf{0}$$

así, de acuerdo con la definición, $\lambda_2 = 1 - \rho^2$. Así, λ_1 es de multiplicidad uno y λ_2 es de multiplicidad $n - 1$, como tenemos únicamente dos eigenvalores diferentes, podemos escribir

$$\bar{\Sigma}_U = \lambda_1 P + \lambda_2 Q,$$

donde P, Q son matrices de proyección definidas como $P = \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$ y $Q = I_n - P$. Debido a que son matrices de proyección, $\sum_i^n \lambda_i \mathbf{v}_i \mathbf{v}_i^T = I_n$; luego, esta suma puede escribirse como $\lambda_1 \mathbf{1}_n \mathbf{1}_n^T + \sum_{i=2}^n \lambda_i \mathbf{v}_i \mathbf{v}_i^T$ y de ahí obtenemos que $Q = \sum_{i=2}^n \lambda_i \mathbf{v}_i \mathbf{v}_i^T = I - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T = I - P$. Por tanto, se tiene que

$$\begin{aligned} \bar{\Sigma}_U &= ((n-1)\rho^2 + 1) \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T + (1 - \rho^2)(I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T) \\ \Rightarrow \bar{\Sigma}_U^{1/2} &= ((n-1)\rho^2 + 1)^{1/2} \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T + (1 - \rho^2)^{1/2}(I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T) \\ \Rightarrow (\Sigma_U^{1/2})_{ii} &= \frac{\sqrt{(n-1)\rho^2 + 1}}{n} + (1 - \rho^2) \frac{n-1}{n}, \end{aligned}$$

es decir, la entrada (i, i) de $\bar{\Sigma}_U^{1/2}$ no reensambla el término $\sqrt{(n-1)\rho^2 + 1}$, esto significa que el parámetro de forma debe absorber esta constante. Esto es, $\boldsymbol{\lambda}_n = \rho / \sqrt{(n-1)\rho^2 + 1} \sqrt{1 - \rho^2} \mathbf{1}_n$ y por lo tanto, $\boldsymbol{\sigma} = \text{diag}\{\sigma, \dots, \sigma\}$. De esta forma, podemos escribir la densidad marginal de X como

$$f_{\mathbf{X}}(\mathbf{x}) = 2N(\mathbf{u}_n \mid \boldsymbol{\mu}_n, \sigma^2 \bar{\Sigma}_U) \Phi(\boldsymbol{\lambda}_n^T \boldsymbol{\sigma}^{-1}(\mathbf{u}_n - \boldsymbol{\mu}_n)). \quad (\text{A.10})$$

Normal sesgada, caso multivariado. Método *ii*

Es posible aplicar la misma idea de completar una densidad normal en la variable w , para el caso de n observaciones $\mathbf{U} = (U_1, \dots, U_n)^T$ para mostrar que la densidad marginal de \mathbf{U}_n es normal asimétrica multivariada. Entonces, nuevamente se plantea

$$f_{\mathbf{X}_n, W}(\mathbf{x}_n, w) \equiv f_{X_n|W}(\mathbf{x}_n | w) f_W(w) = \left[\frac{f_{\mathbf{U}_n|(W=w)}(\mathbf{u}_n | w)}{\mathbb{P}(W > 0)} I_{(0, \infty)} \right] f_W(w), \quad (\text{A.11})$$

es inmediato obtener la densidad $p(\mathbf{u}_n | w)$:

$$\begin{aligned} p(\mathbf{u}_n | w) &= N(\mathbf{u}_n | \boldsymbol{\mu}_n + \sigma^2 \rho \mathbf{1}_n (\sigma^2)^{-1} (w - 0), \sigma^2 (1 - \rho^2) I_n + \underbrace{\sigma^2 \rho^2 J_n}_{=\sigma^2 \rho^2 \mathbf{1}_n \mathbf{1}_n^T} \xrightarrow{0}) \\ &= N(\mathbf{u}_n | \boldsymbol{\mu}_n + w \rho \mathbf{1}_n, \sigma^2 (1 - \rho^2) I_n), \end{aligned} \quad (\text{A.12})$$

aquí se hace evidente el por qué de la inclusión del término $\rho^2 J_n$, de otra manera, la matriz de covarianza $\sigma^2((1 - \rho^2)I_n - \rho^2 J_n)$ no es definida positiva para ciertas elecciones de $\rho \in (-1, 1)$. Dado que la $\sigma^2(1 - \rho^2)I_n$ es diagonal, es posible factorizar esta distribución con elementos condicionalmente independientes, es decir

$$f_{\mathbf{U}_n|W}(\mathbf{u}_n | w) = 2 \prod_{i=1}^n N(u_i | \mu_i + \rho w, \sigma^2(1 - \rho^2)), \quad (\text{A.13})$$

por lo tanto,

$$f_{\mathbf{U}_n, W}(\mathbf{u}_n, w) = 2 \prod_{i=1}^n N(u_i | \mu_i + \rho w, \sigma^2(1 - \rho^2)) N(w | 0, \sigma^2) I_{(0, \infty)}(w), \quad (\text{A.14})$$

ahora, obtenemos la densidad marginal de \mathbf{U} al integrar fuera w :

$$f_{\mathbf{U}}(\mathbf{u}) = \frac{2}{ab^n} \int_0^\infty \exp\left\{-\frac{\sum_{i=1}^n (u_i - \mu_i - \rho w)^2}{2\sigma^2(1 - \rho^2)}\right\} \exp\left\{-\frac{w^2}{2\sigma^2}\right\} dw,$$

aquí, $a = 1/\sqrt{2\pi\sigma^2}$ y $b = 1/\sqrt{2\pi\sigma^2(1-\rho^2)}$; no obstante, por simplicidad no haremos especial énfasis en determinar exactamente las constantes. Ahora, escribimos

$$= \frac{2}{ab^n} \int_0^\infty \exp\left\{-\frac{\sum_{i=1}^n ((u_i - \mu_i)^2 + \rho^2 w^2 - 2\rho w(u_i - \mu_i)) + (1 - \rho^2)w^2}{2\sigma^2(1 - \rho^2)}\right\} dw,$$

expandimos la suma y factorizamos w^2

$$= \frac{2}{ab^n} \int_0^\infty \exp\left\{-\frac{\sum_{i=1}^n (u_i - \mu_i)^2 + ((n-1)\rho^2 + 1)w^2 - 2\rho w \sum_{i=1}^n (u_i - \mu_i)}{2\sigma^2(1 - \rho^2)}\right\} dw,$$

completamos el cuadrado en w^2 , para ello factorizamos $(n-1)\rho^2 + 1$, luego sumamos y restamos $\rho^2(\sum_{i=1}^n (u_i - \mu_i))^2/((n-1)\rho^2 + 1)^2$,

$$= \frac{2}{ab^n} \int_0^\infty \exp\left\{-\frac{\sum_{i=1}^n (u_i - \mu_i)^2 + ((n-1)\rho^2 + 1) \left[\left(w - \frac{\rho \sum_{i=1}^n (u_i - \mu_i)}{(n-1)\rho^2 + 1}\right)^2 - \frac{\rho^2 (\sum_{i=1}^n (u_i - v_i))^2}{((n-1)\rho^2 + 1)^2} \right]}{2\sigma^2(1 - \rho^2)}\right\} dw,$$

en la función exp identificamos dos términos, uno no depende de w y en el otro podemos identificar un kernel gaussiano en w , es decir $p(w) = N(w | \mu_w, \sigma_w^2)$, donde

$$\begin{aligned} \mu_w &= \frac{\rho \sum_{i=1}^n (u_i - v_i)}{((n-1)\rho + 1)}, \\ \sigma_w^2 &= \frac{\sigma^2(1 - \rho^2)}{(n-1)\rho^2 + 1}, \end{aligned} \tag{A.15}$$

luego, sacamos los términos que no dependen de w , y por simplicidad ignoramos las constantes $1/ab^n$ ya que no las necesitamos. Ahora, escribimos

$$\propto 2 \exp\left\{-\frac{\sum [u_i - \mu_i]^2 - \frac{\rho^2 (\sum [u_i - \mu_i])^2}{((n-1)\rho^2 + 1)}}{2\sigma^2(1 - \rho^2)}\right\} \int_0^\infty \exp\left\{-\frac{1}{2\sigma_w^2}(w - \mu_w)^2\right\} dw,$$

note que $\sum_{i=1}^n (u_i - \mu_i)^2 = (\mathbf{u} - \boldsymbol{\mu})^T(\mathbf{u} - \boldsymbol{\mu})$, además, $\sum_{i=1}^n (u_i - \mu_i) = \mathbf{1}^T(\mathbf{u} - \boldsymbol{\mu})$; así, de esta última igualdad se tiene que $(\sum_{i=1}^n (u_i - v_i))^2 = \mathbf{1}^T(\mathbf{u} - \boldsymbol{\mu}) \mathbf{1}^T(\mathbf{u} - \boldsymbol{\mu}) = (\mathbf{u} - \boldsymbol{\mu})^T \mathbf{1} \mathbf{1}^T(\mathbf{u} - \boldsymbol{\mu}) = (\mathbf{u} - \boldsymbol{\mu})^T J(\mathbf{u} - \boldsymbol{\mu})$. Ahora, nos concentramos en la integral en w : así como en el caso univariado, podemos identificar la integral como $\mathbb{P}[W > 0]$, mediante normalización, podemos escribir esta probabilidad como

$$\begin{aligned}\mathbb{P}[W > 0] &= 1 - \Phi\left(-\frac{\rho \sum_{i=1}^n (u_i - \mu_i)/(n-1)\rho + 1}{\sigma \sqrt{1-\rho^2}/\sqrt{(n-1)\rho^2+1}}\right) \\ &= \Phi\left(\frac{\lambda \mathbf{1}_n^T(\mathbf{u} - \boldsymbol{\mu})}{\sigma \sqrt{(n-1)\rho^2+1}}\right) \\ &= \Phi(\boldsymbol{\lambda}_n^T \boldsymbol{\sigma}^{-1}(\mathbf{u} - \boldsymbol{\mu})), \quad (\boldsymbol{\lambda}_n^T \triangleq \rho / (\sqrt{1-\rho^2} \sqrt{(n-1)\rho^2+1}) \mathbf{1}_n^T)\end{aligned}$$

donde hemos definido $\boldsymbol{\sigma} \triangleq \text{diag}(\sigma, \dots, \sigma)$. Ahora, usando los resultados previos, escribimos

$$f_{\mathbf{U}}(\mathbf{u}) \propto \exp\left(-\frac{1}{2\sigma^2(1-\rho^2)}(\mathbf{u} - \boldsymbol{\mu})^T \left[I - \frac{\rho^2}{(n-1)\rho^2+1}J\right](\mathbf{u} - \boldsymbol{\mu})\right) \Phi(\boldsymbol{\lambda}^T \boldsymbol{\sigma}^{-1}(\mathbf{u} - \boldsymbol{\mu})),$$

ahora, dado que estamos parametrizando con la covarianza, entonces, este parámetro está dado por

$$\Sigma_U \equiv \left(\frac{1}{\sigma^2(1-\rho^2)} \left[I - \frac{\rho^2}{(n-1)\rho^2+1}J\right]\right)^{-1} = \sigma^2(1-\rho^2) \left[I - \frac{\rho^2}{(n-1)\rho^2+1}J\right]^{-1}, \quad (\text{A.16})$$

por tanto, ahora la tarea es encontrar esta inversa y se emplea nuevamente la fórmula Sherman–Morrison de la [Ecuación A.9](#), aquí identificamos $A = I_n$, $\mathbf{a} = \rho / \sqrt{(n-1)\rho^2+1} \mathbf{1}_n$, $\mathbf{b} = -\mathbf{a}$. Por tanto, se tiene que

$$\begin{aligned}\Sigma_U^{-1} &= I + \frac{\rho^2}{(n-1)\rho^2+1} \times \frac{I_n J_n I_n}{1 - \frac{\rho^2}{(n-1)\rho^2+1} (\mathbf{1}_n^T I_n \mathbf{1}_n)} \\ &= I_n + \frac{\rho^2}{(n-1)\rho^2+1} \times \frac{1}{1 - \frac{n\rho^2}{(n-1)\rho^2+1}} J_n \\ &= I_n + \frac{\rho^2}{(n-1)\rho^2+1} \times \frac{(n-1)\rho^2+1}{1-\rho^2} = I_n + \frac{\rho^2}{1-\rho^2} J_n,\end{aligned} \quad (\text{A.17})$$

por lo tanto, la matriz de covarianza está dada por

$$\Sigma_U = \sigma^2(1 - \rho^2) \left[I_n + \frac{\rho^2}{1 - \rho^2} \right] = \sigma^2 \left[(1 - \rho^2)I_n + \rho^2 J_n \right], \quad (\text{A.18})$$

por lo tanto, $f_U(\mathbf{u})$ tiene kernel de una densidad Gaussiana multivariada, así, podemos concluir que

$$f_{\mathbf{U}_n}(\mathbf{u}_n) = 2\phi_{\mathbf{U}_n}(\mathbf{u}_n \mid \boldsymbol{\mu}_n, \Sigma_U) \Phi(\boldsymbol{\lambda}_n^T \boldsymbol{\sigma}^{-1}(\mathbf{u}_n - \boldsymbol{\mu}_n)) \equiv SN(\mathbf{u}_n \mid \boldsymbol{\mu}_n, \Sigma_u, \boldsymbol{\lambda}_n^T). \quad (\text{A.19})$$

Aunque la definición de $\boldsymbol{\sigma}$ pueda parecer arbitraria, garantiza que sea definida positiva, además de que con $n = 1$, es decir, en el caso univariado, se recupere correctamente el proceso de truncamiento oculto. Por ejemplo, si se emplea $\bar{\Sigma}_U = \sigma^2 I_n$, entonces, siguiendo el desarrollo al inicio del anexo, se tiene que

$$p(w \mid \mathbf{u}_n) = N(w \mid \rho \mathbf{1}_n^T(\mathbf{u}_n - \boldsymbol{\mu}_n), \sigma^2(1 - n\rho^2)), \quad (\text{A.20})$$

efectivamente, con $n = 1$ se recupera la varianza $\sigma^2(1 - \rho^2)$, sin embargo, $|\rho| < \frac{1}{\sqrt{n}}$, ya que de otro modo se obtiene una varianza negativa.

Anexo B

En este apartado se muestran algunos resultados en cuanto al método campo medio de inferencia Bayesiana variacional.

Restricción Campo Medio: prueba

El método basado en el supuesto Campo Medio aproxima la densidad *a posteriori* con

$$q(\boldsymbol{\theta}) = q(\boldsymbol{\theta}_1) q(\boldsymbol{\theta}_2), \dots, q(\boldsymbol{\theta}_G),$$

donde G es el número de grupos: si $\boldsymbol{\theta}$ es de dimensión p , entonces $G \leq p$. Ahora, de acuerdo con Bishop 2006; Ormerod y Wand 2010, la justificación usual sobre la forma de las densidades óptimas $q^*(\boldsymbol{\theta}_i)$ emplea gradientes y multiplicadores de Lagrange. Así mismo, Blei, Kucukelbir y McAuliffe 2017 considera un argumento meramente probabilístico. Recordando que el límite inferior de la evidencia para la aproximación q está dado por

$$\begin{aligned}\text{ELBO}(q) &= \mathbb{E}_q[\log p(\mathbf{y}, \boldsymbol{\theta})] - \mathbb{E}_q[\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_q[\log p(\mathbf{y}, \boldsymbol{\theta}_{-i}, \boldsymbol{\theta}_i)] - \mathbb{E}_q[\log q(\boldsymbol{\theta}_{-i}, \boldsymbol{\theta}_i)],\end{aligned}$$

ahora, usamos la esperanza iterada en el primer término y en el segundo término desarrollamos, por tanto escribimos

$$\text{ELBO}(q) = \mathbb{E}_i [\mathbb{E}_{-i}[\log p(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i}, \mathbf{y})]] - \mathbb{E}_i[\log q_i(\boldsymbol{\theta}_i)] - \underbrace{\mathbb{E}_{-i}[\log q_{-i}(\boldsymbol{\theta}_{-i})]}_c,$$

note que el último término es constante con respecto a $\boldsymbol{\theta}_i$. Finalmente, escribimos esta expresión únicamente como función del factor variacional $q(\boldsymbol{\theta}_i)$

$$\text{ELBO}(q_i) = \mathbb{E}_i [\mathbb{E}_{-i}[\log p(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i}, \mathbf{y})]] - \mathbb{E}_i[\log q_i(\boldsymbol{\theta}_i)] + c,$$

ahora bien, salvo por una constante, es posible identificar la última expresión como la divergencia KL negativa entre $q(\boldsymbol{\theta}_i)$ y $q^*(\boldsymbol{\theta}_i)$ -obtenida de forma analítica-, ya que

$$\begin{aligned}-\text{KL}(q(\boldsymbol{\theta}_i) \| q^*(\boldsymbol{\theta}_i)) &= - \int q(\boldsymbol{\theta}_i) \log \frac{q(\boldsymbol{\theta}_i)}{q^*(\boldsymbol{\theta}_i)} d\boldsymbol{\theta}_i \\ &= -\mathbb{E}_{q(\boldsymbol{\theta}_i)}[\log q(\boldsymbol{\theta}_i)] + \mathbb{E}_{q(\boldsymbol{\theta}_i)}[\log q^*(\boldsymbol{\theta}_i)] \\ &= -\mathbb{E}_{q(\boldsymbol{\theta}_i)}[\log q(\boldsymbol{\theta}_i)] + \mathbb{E}_{q(\boldsymbol{\theta}_i)}[\log \mathbb{E}_{q(\boldsymbol{\theta}_{-i})}[\log p(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i}, \mathbf{y})]],\end{aligned}$$

de esta manera, el límite inferior de la evidencia con respecto a q_i se maximiza cuando hacemos que $q_i(\boldsymbol{\theta}_i) = q^*(\boldsymbol{\theta}_i)$, este es un resultado básico acerca de la divergencia KL.

Anexo C

En este apartado se incluye la lista de covariables empleadas para el ajuste de los modelos de regresión en áreas pequeñas usando el conjunto de datos del ICTPC. Las primeras 52 covariables son continuas, y el resto son binarias. En general, la presencia de algún indicador se codifica como uno y cero en caso contrario.

1. Número de activos A en el hogar (microondas, lavadora y computadora)
2. Número de activos B en el hogar (radio, tv y refrigerador)
3. Cantidad de activos de comunicación en el hogar
4. Cantidad de bienes (radio, tv, refrigerador, lavadora y auto) en el hogar
5. Cantidad de bienes seleccionados
6. Número de cuartos del hogar
7. Cantidad de equipamientos (tinaco boiler cisterna regadera) en el hogar
8. Promedio de habitantes por cuarto
9. Número de hijas
10. Número de hijos
11. Número de hombres en el hogar
12. Carencias en la vivienda
13. Indicador de presencia de mujeres con ingresos mayores a los de los hombres en el hogar
14. Número de bienes de equipamiento (regadera, tinaco, cisterna, calentador de agua, bomba de agua, aire acondicionado)
15. Escolaridad del jefe
16. Escolaridad relativa del jefe del hogar
17. Edad del jefe del hogar
18. Escolaridad relativa estandarizada del jefe
19. Nivel educativo del cónyuge del jefe
20. Número de personas entre 16 y 64 años en el hogar
71. Cónyuge hablante de lengua indígena
72. Cónyuge con rezago educativo
73. Cónyuge con carencia por acceso a los servicios de salud
74. Forma en que desechan la basura
75. Indicador de presencia de menores de 6 a 15 años que no asisten a la escuela
76. Hogar ampliado
77. Hogar compuesto
78. Hogar corresidente
79. Hogar nuclear
80. Hogar unipersonal
81. Algún adulto tuvo poca variedad en sus alimentos
82. Algún adulto dejó de desayunar, comer o cenar
83. Algún adulto comió menos de lo que debería comer
84. Algún adulto se quedó sin comida
85. Algún adulto sintió hambre, pero por falta de dinero no comió
86. Algún adulto comió sólo una vez al día, o dejó de comer todo un día
87. Indicador de carencia de salud en el hogar
88. Indicador de carencia por calidad y espacios de la vivienda
89. Indicador de carencia por servicios básicos de la vivienda
90. Indicador de carencia por hacinamiento

-
- hogar
21. Máximo nivel académico en el hogar
22. Número de personas de 65 o más años en el hogar
23. Número de menores de 16 años en el hogar
24. Número de mujeres en el hogar
25. Número de personas con 60 años o más
26. Número de personas que saben leer y escribir en el hogar
27. Número de hijos(as) fallecidos de mujeres entre 12 y 49 años
28. Número de hijos(as) fallecidos
29. Número de hijos(as) que tiene 16 años o más y que está en la pea
30. Número de hijos(as) nacidos (as) vivos en el hogar
31. Número de hijos(as) menor(es) de 16 años
32. Número de personas hablantes de lengua indígena en el hogar
33. Número de menores de 12 años en el hogar
34. Número de personas menores de 50 años
35. Número de personas ocupadas en el hogar
36. Número de perceptores de ingresos ocupados
37. Personas mayores de 16 años con rezago educativo
38. Personas menores de 16 años con rezago educativo
39. Personas con carencia en servicios de salud en el hogar
40. Porcentaje de personas de 65 años o más en el hogar
41. Porcentaje de hombres en el hogar
42. Porcentaje de personas indígenas en el hogar
43. Porcentaje de menores de 12 años en el hogar
44. Porcentaje de mujeres en el hogar
91. Indicador de carencia de muros
92. Indicador de carencia de pisos
93. Indicador de carencia de techos
94. Indicador de presencia de menores de 18 años en el hogar
95. Dispone de conexión a internet el hogar
96. Indicador de carencia de agua
97. Indicador de carencia de combustible para cocinar
98. Indicador de carencia de drenaje
99. Indicador de carencia de luz
100. Prestaciones laborales del jefe Afore
101. Prestaciones laborales del jefe Aguinaldo
102. Jefe del hogar que no sabe leer y escribir
103. Jefe desocupado
104. Jefe hablante de lengua indígena
105. jefe con posición de independiente
106. Jefe ocupado
107. Jefe con carencia por rezago educativo
108. Sexo del jefe
109. Prestaciones laborales del jefe Servicios Médicos
110. Jefe con posición de subordinado
111. Indicador se personas jubiladas o pensionadas en el hogar
112. Prestaciones laborales del jefe Utilidades
113. Prestaciones laborales del jefe Prima vacacional
114. Dispone de al menos una lavadora el hogar
115. Dispone de al menos un microondas el hogar
116. Dispone de al menos un radio el hogar
117. Indicador de presencia de red de apoyo femenino
118. Dispone de al menos un refrigerador el hogar

-
45. Porcentaje de ocupados en el hogar
46. Porcentaje de perceptores ocupados
47. Porcentaje de personas mayores de 16 con rezago educativo
48. Porcentaje de personas menores de 16 con rezago educativo
49. Porcentaje de personas con carencia en salud en el hogar
50. Relación de dependencia
51. Tasa de mortalidad
52. Proxy de mortalidad infantil
53. Dispone de aire acondicionado el hogar
54. Indicador de posesión de automóvil en el hogar
55. Condición de recepción de ingresos por apoyo de otro hogar
56. Condición de recepción de ingresos por programas de gobierno en el hogar
57. Dispone de boiler el hogar
58. Dispone de bomba de agua el hogar
59. Dispone de calentador solar de agua el hogar
60. Dispone de teléfono móvil o celular el hogar
61. Hogar con todos sus integrantes mayores de 60 años (envejecido)
62. Hogar nuclear con algún hijo en la PEA (en proceso de fisión)
63. Hogar nuclear sin hijos con sus integrantes menores a 50 años (en formación)
64. Hogar nuclear con hijos menores, sin hijos en la PEA
65. Dispone de cisterna el hogar
66. Disponibilidad de cuarto para cocinar
67. Tipo de combustible para cocinar
68. Identificador de combustible: Leña o carbón o cuenta con chimenea estufa (fogón) de leña o carbón
119. Dispone de al menos un regadera el hogar
120. Condición de recepción de ingresos por remesas en el hogar
121. Sanitario para uso exclusivo
122. Indicador de disponibilidad de sanitario
123. Dispone de sanitario con descarga directa a la red en el hogar
124. Dispone de línea telefónica fija en el hogar
125. Vivienda con otra situación
126. Tenencia propia de la vivienda
127. Vivienda rentada
128. Hogar biparental
129. Hogar uniparental
130. Dispone de tinaco en el hogar
131. Indicador de presencia de mayor de 64 años trabajador
132. Indicador de presencia de menor de 16 años trabajador
133. Dispone de televisión en el hogar
134. Tipo de combustible para cocinar 2
135. Forma en que dispone de agua 1
136. Condición de actividad del jefe del hogar
137. Condición de actividad del jefe con ocupados subordinados o independientes
138. Sector económico del jefe del hogar
139. Tenencia de la vivienda
140. Tipo de unidad doméstica

69. Dispone de al menos una computadora el hogar

70. Indicador de cónyuge

Anexo D

En este apartado se facilitan los códigos empleados para realizar el ajuste de los modelos de regresión Bayesiana en áreas pequeña propuestos -log-normal sesgado, y probit (ordenado) sesgado con variable latente-, y para reproducir las figuras mostradas en el documento.

Se prefiere no colocar explicitamente los códigos empleados por comodidad del usuario interesado en reproducir los resultados o modificarlos a su conveniencia: parece que es más sencillo acceder a ellos de forma sistemática y libre de errores de copiado. Por tanto, se dispone de ellos en la plataforma *Git-Hub* en el siguiente proyecto: <https://github.com/Demian-33/MSc-Thesis-Code>.

Por otro lado, los análisis se realizaron en el lenguaje Stan por medio de la librería `cmdstanr` del lenguaje R. A continuación describimos también aspectos sobre aquellas librerías y versiones empleadas que son más relevantes. En este caso, se debe tener especial cuidado con realizar previamente la instalación de `Stan` a través de su página web oficial, ya que hasta la fecha actual, no se encuentra en el la lista de paquetes del CRAN.

Librería	Función(es)	Versión
<code>sn</code>	<code>psn</code> , <code>selm</code> , <code>dmsn</code>	1.2
<code>cmdstanr</code>	<code>\$compile</code> , <code>\$variational</code> , <code>\$sample</code>	2.37
<code>ggplot2</code>		