UNIVERSIDAD NACIONAL DEL LITORAL

Facultad de Ingeniería Química

Reducción de dimensiones para datos composicionales en alta dimensión

Eric Lionel Koplin

Tesis presentada como parte de los requisitos de la

Universidad Nacional del Litoral para la obtención del grado de

Doctor en Ingeniería Matemática

En el campo de: **Estadística**

Institución donde se realizó:

Facultad de Ingeniería Química

(Conicet-Unl)

Directores de Tesis:

Dr. Diego Tomassi y Dra. Liliana Forzani

Defendida ante el jurado compuesto por:

Dr. Diego Cafaro

Dra. Florencia Leonardi

Dra. Daniela Rodriguez

Año de presentación: 2023

Eric Koplin: *Reducción de dimensiones para datos composicionales en alta dimensión,* Tesis, © Octubre 2023

ABSTRACT

Dimension reduction is often included in the analysis of complex high-dimensional data to aid understanding of the underlying phenomena and drive further exploration. In doing so, reduced data should preserve the relevant information for the addressed problem so that any exploratory conclusion is well grounded. Despite methodology for dimension reduction has been widely developed in the statistics and machine learning communities, application to complex data with very special features often does not suit underlying assumptions and non-trivial adaptations are needed to devise proper solutions.

This thesis contributes novel methods to aid understanding of microbiome data. The human microbiome is currently supposed to have a main role in the etiology of some chronic diseases, the efficacy of drug treatments, women fertility or healthy development of infants. From a statistical point of view, microbiome data is high-dimensional count data registering the abundance of micro-organisms living together. Due to technical challenges in such quantification, raw count data is not always meaningful and only relative abundances are informative. Thus, microbiome data is indeed high-dimensional compositional data. Since current technology allows to identify thousands of different species that might be present in a few samples only, the data is also extremely sparse. Moreover, in addition to assess any association between overall microbiome composition and treated group or observed outcome, there is often the need to explain which components of the microbiota drive such association. Currently there is no modeling approach that can address both questions simultaneously in a principled way.

The methodolgy presented in this thesis is inspired by sufficient dimension reductions and uses graphical models as a main ingredient. In particular, we present dimension reduction methods for high-dimensional count/compositional data based on suitable conditional graphical models. Sufficient reductions provide the theoretical background to preserve information when studying associations between microbiome compositions and a given outcome. On the other hand, graphical models allow modeling of complex dependencies among components of the microbiota, better describing them as an ecosystem. However, the large proportion of zeros encountered in microbiome data asks for adapted graphical models able to deal with the excess of zeros.

With this motivation, this thesis introduces new graphical models that accommodate a large proportion of zeros in high-dimensional count/compositional data and presents algorithms for efficient estimation. First, we characterise these multivariate distributions in terms of univariate conditional distributions. We then model predictors that arise from such a pairwise graphical model with excess of zeros as a function of an outcome, and derive the corresponding first order suffi-

cient dimension reduction (SDR). That is, we find linear combinations of the predictors that contain all the information for the regression of the outcome as a function of the predictors. We estimate the SDR by minimizing a divergence with a graph-aware hierarchical penalty that induces structured sparsity to identify key predictors truly associated with the outcome. This method yields consistent estimators of the reduction and can be applied to continuous or categorical outcomes. In addition, by adding variable selection, the obtained reductions help to visualize overall associations while identifying the key components driving them. We illustrate our methods in simulations and by analyzing real microbiome data.

RESUMEN

A menudo se incluye a la reducción de dimensiones en el análisis de datos complejos de alta dimensión para ayudar a comprender los fenómenos subyacentes. Para que resulte efectiva, los datos reducidos deben conservar la información relevante para el problema abordado, para que cualquier conclusión exploratoria esté bien fundamentada. A pesar de que la metodología de reducción de dimensiones ha sido ampliamente desarrollada en las comunidades de estadística y aprendizaje automático, la aplicación a datos complejos con características muy especiales a menudo no se ajusta a las suposiciones subyacentes y se necesitan adaptaciones no triviales para idear soluciones adecuadas.

Esta tesis contribuye con métodos novedosos que ayudan a comprender los datos del microbioma. Actualmente se tienen indicios de que el microbioma humano juega un papel fundamental en la etiología de algunas enfermedades crónicas, sobre la eficacia de los tratamientos con medicamentos, la fertilidad de las mujeres o el desarrollo saludable de niños recién nacidos. Desde un punto de vista estadístico, el microbioma puede describirse como datos de conteo de alta dimensión que registran la abundancia de microorganismos que viven juntos. Debido a que es técnicamente complejo cuantificar dicha abundancia, los datos de conteo brutos no siempre son significativos y solo las abundancias relativas son informativas. Así, los datos de microbioma son efectivamente datos composicionales de alta dimensión. Dado que la tecnología actual permite identificar miles de diferentes especies y subvariantes que podrían estar presentes solo en unas pocas muestras, los datos también presentan una elevadísima cantidad de ceros. Estos perfiles composicionales también muestran una gran variabilidad entre individuos e incluso pueden sufrir variaciones significativas a lo largo del tiempo para un mismo individuo, lo que contribuye a la complejidad del modelado.

El análisis de datos de microbioma tiene por objeto inicial evaluar la existencia de diferencias significativas en la composición general del microbioma asociadas a una variable o grupo de interés, como el grupo tratado o variables fisiológicas en un ensayo clínico. No obstante, a menudo es necesario explicar también qué componentes de la microbiota impulsan dicha asociación. Actualmente no hay un enfoque de modelado que resulte satisfactorio en ambos aspectos.

La metodología presentada en esta tesis está inspirada en reducción suficiente de dimensiones y utiliza modelos gráficos como ingrediente principal. En particular, presentamos métodos de reducción de dimensiones para datos composicionales/de conteo de alta dimensión basados en modelos gráficos condicionales adecuados. La reducción suficiente proporciona el sustento teórico para conservar información al estudiar asociaciones entre composiciones de microbiomas y una respuesta dada. Por otro lado, los modelos gráficos permiten modelar dependencias complejas entre componentes de la microbiota, descri-

biéndolos mejor como un ecosistema. Sin embargo, la gran proporción de ceros presente en los datos requiere modelos gráficos especialmente adaptados.

Por ello, en esta tesis presentamos nuevos modelos gráficos capaces de modelar una gran proporción de ceros en datos de conteo y composicionales de alta dimensión y desarrollamos algoritmos para su estimación. Primero, caracterizamos estas distribuciones multivariadas en términos de distribuciones condicionales univariadas. Luego modelamos los predictores que surgen de tal modelo gráfico de segundo orden con exceso de ceros como función de la respuesta y derivamos la reducción suficiente de dimensiones (RSD) de primer orden correspondiente. Es decir, encontramos combinaciones lineales de los predictores que contienen toda la información de la respuesta como función de los predictores. Estimamos la RSD minimizando una divergencia con una penalización jerárquica que es consistente con el grafo que induce la estructura que permite identificar los predictores asociados con la respuesta. Este método produce estimadores consistentes de la reducción y se puede aplicar a respuestas continuas o categóricas. Al agregar selección de variables, las reducciones obtenidas permiten evaluar exploratoriamente asociaciones globales e identificar simultáneamente los componentes de la microbiota capaces de explicar la variabilidad de la respuesta. Ilustramos nuestros métodos en simulaciones y analizando datos reales de microbioma.

PUBLICACIONES

Algunos resultados y figuras que presentaremos fueron publicados previamente en:

Koplin, Eric, Liliana Forzani, Diego Tomassi y Ruth M. Pfeiffer (2024). «Sufficient dimension reduction for a novel class of zero-inflated graphical models». En: *Computational Statistics Data Analysis* 196, pág. 107959. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2024.107959. URL: https://www.sciencedirect.com/science/article/pii/S0167947324000434.

ÍNDICE GENERAL

I Introducción

1 Reducción de dimensiones y datos composicionales en alta dimensión 3
- 1.1 Motivación 3
  - 1.1.1 Microbioma y sus desafíos 3
  - 1.1.2 El contexto analítico 6
- 1.2 Objetivos y contribución 8
  - 1.2.1 Trabajos relacionados 9
- 1.3 Organización 10

II Modelos y reducción de dimensiones

2 Familia exponencial y modelos gráficos 15
- 2.1 Introducción 15
- 2.2 Contenidos del capítulo 15
- 2.3 Antecedentes 16
  - 2.3.1 Ejemplos 21
- 2.4 Contribución: Modelos gráficos de segundo orden para exceso de ceros 23
- 2.5 Modelos condicionales 26

3 Reducción suficiente de dimensiones 29
- 3.1 Introducción 29
- 3.2 Reducción suficiente de dimensiones en la familia exponencial 30
  - 3.2.1 Selección de una base de funciones para modelar la regresión inversa 31
- 3.3 Estudio bibliográfico 32

III Estimación

4 Divergencias y scores 35
- 4.1 Divergencia de Kullback-Leibler y Máxima verosimilitud 36
- 4.2 Likelihood de composición y Pseudo-likelihood 38
- 4.3 Reglas de score 39
  - 4.3.1 Variables aleatorias continuas 40
  - 4.3.2 Variables aleatorias discretas 41
- 4.4 Divergencia entre modelos condicionales 43
- 4.5 Estimación diferencial 44
- 4.6 M-estimador asociado 45
- 4.7 Métrica local 46
  - 4.7.1 Medida *pullback* local 46
  - 4.7.2 Cálculo de la matriz de información 47
- 4.8 Ejemplos 47
- 4.9 Contribuciones 51

5 Estimación en alta dimensión 53
- 5.1 LASSO 54

5.1.1 Ejemplo pGM 55
5.2 Penalización Jerárquica 56
5.3 Penalización con pesos óptimos 56
5.4 Ejemplo zipGM 57
5.4.1 Penalización jerárquica en el contexto de RSD 57
5.5 Selección del modelo 59
5.5.1 Selección de modelos en el contexto de RSD 60
6 Aprendizaje y Optimización 63
6.1 Gradiente Natural 63
6.1.1 Problemas de la información empírica 64
6.2 Algoritmos proximales 64
6.2.1 Algoritmo proximal SDMM 65
6.2.2 Algoritmos proximales de primer orden y segundo orden 66
6.3 Algoritmo propuesto para el aprendizaje en alta dimensión de modelos zipGM 69
6.4 Contribución 70

IV Resultados numéricos
7 Simulaciones 73
7.1 Proporción de ceros y modelos mal especificados 75
7.1.1 Parámetros poblacionales 75
7.1.2 Resultados 76
7.1.3 Influencia de las direcciones del span en modelos pGM 81
7.2 Efecto de las interacciones en modelos zipGM con penalización jerárquica 82
7.2.1 Parámetros poblacionales 82
7.2.2 Resultados 84
7.3 Datos simulados tipo microbioma 86
7.3.1 Resultados 88
7.4 Comentarios de cierre del capítulo 90
8 Aplicación a datos reales 93
8.1 Human Microbiome Project 94
8.1.1 Resultados 95
8.2 American Gut Project 96
8.2.1 Resultados 98
8.3 Comentarios de cierre del capítulo 99

V Conclusiones
9 Conclusiones 111

VI Anexos
A Anexo al Capítulo 2 117
A.1 Modelos gráficos 117
A.2 Normal-pGM 117
A.3 Fixed-length Poisson-pGM 118
A.4 Zero-inflated pGM 119
B Anexo al Capítulo 3 125
C Anexo al Capítulo 4 127

C.1 Equivalencia divergencia-score 127
C.2 Score propio para distribuciones en el dominio de los enteros, Ejemplo 4.1 128
C.3 Normal-pGM, ejemplo 2.1 129
C.3.1 MLE 129
C.3.2 SME 130
C.4 Poisson-pGM, ejemplo 2.3 131
D Anexo al Capítulo 5 133
D.1 Condiciones de optimalidad para problemas convexos 133
D.1.1 Problemas diferenciables 133
D.1.2 Problemas no diferenciables 135
D.2 Optimalidad en problemas compuestos en estadística 136
E Anexo al Capítulo 6 139
E.1 Optimización del problema penalizado (5.3) para modelos pGM 139
E.2 Optimización del problema penalizado (5.6) para modelos zipGM 140
E.2.1 Operadores proximales asociados al problema de optimización de modelos zipGM 140
E.2.2 Jacobiano y Hessiano de la pseudolikelihood para modelos zipGM 141
F Anexo al Capítulo 7 147
F.1 Estimación de modelos pGM mediante scores propios 147
F.1.1 Parámetros poblacionales 147
F.1.2 Resultados 147
F.2 Proporción de ceros en modelos mal especificados 152
F.2.1 Resultados adicionales 152
F.2.2 Influencia de las direcciones del span en modelos pGM 152
F.3 Efecto de las interacciones en modelos zipGM con penalización jerárquica 159
F.3.1 Resultados adicionales 159
F.4 Datos simulados tipo microbioma 166
G Anexo al Capítulo 8 169
G.1 Human Microbiome Project 169
G.2 American Gut Project 169

Bibliografía 181

ÍNDICE DE FIGURAS

Figura 1.1 Ilustración de la composición de la flora intestinal a nivel de Género para sujetos saludables. Aunque existen más de 250 géneros distintos en el conjunto de datos, incluso los veinte mas abundantes en promedio resultan difíciles de visualizar debido a la disparidad de abundancias y variabilidad entre individuos. 5

Figura 1.2 Representación no supervisada de los datos usando la disimilaridad de Bray-Curtis. 7

Figura 1.3 Resultados de un análisis de abundancia diferencial a nivel de Familias bacterianas. La linea punteada horizontal indica el umbral de significancia individual de los tests. No obstante, ningún resultado es estadísticamente significativo luego de aplicar una corrección de tipo Benjamini-Hochberg sobre los p-valores individuales. 8

Figura 1.4 Diagrama de los contenidos de cada capítulo y su interrelación en función del análisis de datos. Las líneas sólidas indican la relación entre los distintos capítulos que conforman la tesis, mientras que las líneas punteadas implican el uso de datos. 10

Figura 7.1 Interacción de las primeras dos variables generadas por el modelo Poisson-zipGM en relación con la respuesta Y (color) a medida que disminuye la proporción de ceros. 77

Figura 7.3 Medidas de performance en predicción y selección (filas) para los modelos Ising, Normal-pGM y Normal-zipGM con datos generados por el modelo Normal-zipGM detallado en la Sección 7.1.1 para $n \in \{200, 500, 1000\}$ para distintas proporciones de ceros en los datos (columnas). 79

Figura 7.4 Medidas de performance en predicción y selección (filas) para los modelos Ising, Poisson-pGM y Poisson-zipGM con datos generados por el modelo Poisson-zipGM detallado en la Sección 7.1.1 para $n \in \{200, 500, 1000\}$ para distintas proporciiones de ceros en los datos (columnas). 80

Figura 7.6 Medidas de performance en predicción y selección (filas) para los modelos zipGM propuestos con datos generados por el modelo Normal-zipGM detallado en la Sección 7.2.1 para $n \in \{200, 500, 1000\}$ a medida que la fuerza de las interacciones crece (columnas). 85

Figura 7.7 Medidas de performance en predicción y selección (filas) para los modelos zipGM propuestos con datos generados por el modelo Poisson-zipGM detallado en la Sección 7.2.1 para $n \in \{200, 500, 1000\}$ a medida que la fuerza de las interacciones crece (columnas). 87

Figura 7.8 Medidas de performance en predicción y selección de variables (filas) para los modelos zipGM propuestos y el modelo SPLS del paquete `mixOmics`, cuando los datos son generados por un modelo directo $Y|\boldsymbol{\nu}(\boldsymbol{X})$ (columna 1) or $Y|\boldsymbol{X}$ (columna 2) para respuesta continua Y basados en muestras generadas por el paquete `MB-GAN` para $n \in \{200, 500, 1000\}$. 89

Figura 7.9 Medidas de performance en predicción y selección de variables (filas) para los modelos zipGM propuestos y el modelo SPLSDA del paquete `mixOmics`, cuando los datos son generados por un modelo directo $Y|\boldsymbol{\nu}(\boldsymbol{X})$ (columna 1) or $Y|\boldsymbol{X}$ (columna 2) para respuesta binaria Y basados en muestras generadas por el paquete `MB-GAN` para $n \in \{200, 500, 1000\}$. 91

Figura 8.1 Proporción de ceros en los datos HMP para los niveles taxonómicos L2 y L6. 95

Figura 8.2 Reducción en predicción de los datos de microbioma HMP aprendida por el modelo Ising-pGM en los niveles taxonómicos L2 y L6 en relación con el lugar de donde se extrajeron las muestras. 101

Figura 8.3 Reducción en predicción de los datos de microbioma HMP aprendida por el modelo sqPoisson-pGM en los niveles taxonómicos L2 y L6 en relación con el lugar de donde se extrajeron las muestras. 102

Figura 8.4 Reducción en predicción de los datos de microbioma HMP aprendida por los modelos Normal-pGM y Normal-zipGM. En ambos casos consideramos el nivel taxonómicos L6. 103

Figura 8.5 Análisis de asociación entre datos de microbioma de intestino y sexo. Se muestra el número de veces que cada modelo selecciona cada variable en base a 5 particiones independientes del conjunto de datos. Sólo se muestran las variables seleccionadas por algún modelo. El id. de dichas variables se encuentra en la Tabla 8.3. 104

Figura 8.6 Análisis de asociación entre datos de microbioma de intestino y sexo. Se muestra el AUC en predicción en base a 5 particiones independientes del conjunto de datos. 106

Figura 8.7 En gris se muestran los grafos de interacciones entre las varibales descriptivas obtenidos por los modelos zipGM propuestos, los pGM y los algoritmos *SpiecEasi* y *Spring*. En azul se muestran las variables e interacciones seleccionadas por la penalización jerárquica propuesta para los modelos zipGM y pGM al considerar el modelo condicional $\boldsymbol{X} \mid Y$. Los números en cada nodo corresponden a las especies detalladas en la Tabla 8.3, mientras que el tamaño de cada nodo indica el número de veces que la variable fue seleccionada dadas las 5 particiones de entrenamiento. 107

Figura F.1 Medidas de performance en estimación del modelo Ising-pGM mediante la función de costo dada por el score de ratio matching penalizada. Cada columna corresponde a $p \in \{10, 20\}$ variables, mientras que cada fila corresponde al error relativo de estimación del parámetro $\boldsymbol{\eta}$, $\boldsymbol{\Gamma}$, el ángulo entre los subespacios generados por las columnas de $\boldsymbol{\Gamma}$ y $\hat{\boldsymbol{\Gamma}}$ y el error relativo de estimación de $\boldsymbol{\Theta}$ resp. 149

Figura F.2 Medidas de performance en estimación del modelo Poisson-pGM mediante la función de costo dada por el score de ratio matching penalizado. 150

Figura F.3 Medidas de performance en estimación del modelo Normal-pGM mediante la función de costo dada por el score de ratio matching penalizado. 151

Figura F.4 Resultados obtenidos para el ajuste de los modelos Ising, TPoisson-zipGM, y TPoisson-pGM a datos generados a partir del TPoisson-zipGM definido en Sección 7.1.1 cuando el criterio de selección es la AUC. 153

Figura F.5 Resultados obtenidos para el ajuste de los modelos Ising, TPoisson-zipGM, y TPoisson-pGM a datos generados a partir del TPoisson-zipGM definido en Sección 7.1.1 cuando el criterio de selección es el "oráculo". 154
Figura F.6 Resultados obtenidos para el ajuste de los modelos Ising, Normal-zipGM, y Normal-pGM a datos generados a partir del Normal-zipGM definido en Sección 7.1.1 cuando el criterio de selección es el "oráculo". 155
Figura F.7 Resultados obtenidos para el ajuste de los modelos Ising, Poisson-zipGM, y Poisson-pGM a datos generados a partir del Poisson-zipGM definido en Sección 7.1.1 cuando el criterio de selección es el "oráculo". 156
Figura F.8 Resultados obtenidos para el ajuste de los modelos zipGM considerados cuando los datos son generados a partir del modelo Normal-zipGM definido en Sección 7.2.1 cuando el criterio de selección es el "oráculo". Las columnas presentan interacciones con intensidad creciente. 160
Figura F.9 Resultados obtenidos para el ajuste de los modelos zipGM considerados cuando los datos son generados a partir del modelo Poisson-zipGM definido en Sección 7.2.1 cuando el criterio de selección es el "oráculo". Las columnas presentan interacciones con intensidad creciente. 161
Figura F.10 Resultados obtenidos para el ajuste de los modelos zipGM considerados cuando los datos son generados a partir del modelo Normal-zipGM para interacciones entre variables asociadas a la respuesta cuando el criterio de selección es la AUC. Las columnas presentan interacciones con intensidad creciente. 162
Figura F.11 Resultados obtenidos para el ajuste de los modelos zipGM considerados cuando los datos son generados a partir del modelo Normal-zipGM para interacciones entre variables asociadas a la respuesta cuando el criterio de selección es el "oráculo". Las columnas presentan interacciones con intensidad creciente. 163
Figura F.12 Resultados obtenidos para el ajuste de los modelos zipGM considerados cuando los datos son generados a partir del modelo Poisson-zipGM para interacciones entre variables asociadas a la respuesta cuando el criterio de selección es la AUC. Las columnas presentan interacciones con intensidad creciente. 164

Figura F.13 Resultados obtenidos para el ajuste de los modelos zipGM considerados cuando los datos son generados a partir del modelo Normal-zipGM para interacciones entre variables asociadas a la respuesta cuando el criterio de selección es el "oráculo". Las columnas presentan interacciones con intensidad creciente. 165

Figura F.14 Medidas de performance en predicción y selección de variables (filas) utilizando el criterio de selección "oráculo" para los modelos zipGM propuestos y el modelo SPLS del paquete `mixOmics`, cuando los datos son generados por un modelo directo $Y|\boldsymbol{\nu}(\boldsymbol{X})$ (columna 1) or $Y|\boldsymbol{X}$ (columna 2) para respuesta continua Y basados en muestras generadas por el paquete `MB-GAN` para $n \in \{200, 500, 1000\}$. 167

Figura F.15 Medidas de performance en predicción y selección de variables (filas) utilizando el criterio de selección "oráculo" para los modelos zipGM propuestos y el modelo SPLSDA del paquete `mixOmics`, cuando los datos son generados por un modelo directo $Y|\boldsymbol{\nu}(\boldsymbol{X})$ (columna 1) or $Y|\boldsymbol{X}$ (columna 2) para respuesta binaria Y basados en muestras generadas por el paquete `MB-GAN` para $n \in \{200, 500, 1000\}$. 168

Figura G.1 Reducción en predicción de los datos de microbioma HMP aprendida por el modelo Normal-pGM en los niveles taxonómicos L2 y L6 en relación con el lugar de donde se extrajeron las muestras. 171

Figura G.2 Reducción en predicción de los datos de microbioma HMP aprendida por el modelo Poisson-pGM en los niveles taxonómicos L2 y L6 en relación con el lugar de donde se extrajeron las muestras. 172

Figura G.3 Reducción en predicción de los datos de microbioma HMP aprendida por el modelo FPoisson-pGM en los niveles taxonómicos L2 y L6 en relación con el lugar de donde se extrajeron las muestras. 173

Figura G.4 Reducción en predicción de los datos de microbioma HMP aprendida por el modelo Poisson-pGM y Poisson-zipGM. En ambos casos consideramos el nivel taxonómicos L6. 174

Figura G.5 Medidas de predicción (10 folds) de los datos de microbioma HMP (L2) aprendida por el los distintos pGMs considerados en relación con el lugar de donde se extrajeron las muestras. 175

Figura G.6 Medidas de predicción (10 folds) de los datos de microbioma HMP (L6) aprendida por el los distintos pGMs considerados en relación con el lugar de donde se extrajeron las muestras. 176
Figura G.7 Interacciones seleccionadas por los distintos pGMs considerados en relación con el lugar de donde se extrajeron las muestras al ser entrenados con datos de microbioma HMP (L2). 177
Figura G.8 Interacciones seleccionadas por los distintos pGMs considerados en relación con el lugar de donde se extrajeron las muestras al ser entrenados con datos de microbioma HMP (L6). 178
Figura G.9 Cantidad de observaciones por clase en los datos de microbioma HMP. 179
Figura G.10 Grafo de interacciones (aristas) y variables (nodos) seleccionados por la penalización jerárquica (5.6) para el modelo Normal-pGM y pseudo-likelihood cuando es evaluada sobre una grilla regular de parámetros de regularización en el cuadrante positivo $\lambda_{\mathcal{C}}, \lambda_{\mathcal{R}} \geq 0$. La primer columna es computada usando $\lambda_{\mathcal{C}} = 0$, mientras que la última fila es computada con $\lambda_{\mathcal{R}} = 0$. Los círculos azules alrededor de los nodos indican las variables seleccionadas. 179

ÍNDICE DE TABLAS

Tabla 8.1 Scores considerados para cada modelo. 95
Tabla 8.2 Error de predicción (accuracy) de la respuesta Y basándonos en distintas RSDs de los datos de microbioma HMP en niveles taxonómicos L2 y L6 sobre 10 particiones independientes de test mediante validación cruzada. 96
Tabla 8.3 Especies seleccionadas por los modelos considerados junto con su identificador. En todos los casos las especies pertenecen a la familia Bacteria. 105

Tabla 8.4 Error de predicción (accuracy) obtenido al emplear las variables estables estimadas con cada modelo y distintos modelos predictivos sobre 10 particiones independientes de test. Los modelos predictivos consideramos fueron: LR (regresión logística), qda (análisis discriminante cuadrático), svm-poly (clasificador de vector soporte con kernel polinómico), RF (random forest). 106

Tabla F.1 Parámetros poblacionales y score considerado para cada familia de distribuciones considerada en la simulación. 147

GLOSARIO

FE	Familia Exponencial.
MG	Modelo Gráfico.
PGM	MG de segundo orden.
ZIPGM	MG de segundo orden para exceso de ceros.
ES	Estadístico Suficiente.
RSD	Reducción Suficiente de Dimensiones.
MLE	Estimador de Máxima Verosimilitud.
PLE	Estimador de Máxima Pseudo-verosimilitud.
SME	Estimador de Score Matching.
RME	Estimador de Ratio Matching.
LASSO	Least Absolute Shrinkage and Selection Operator, por sus siglas en inglés.
KL	Divergencia de Kullback-Leibler.
PKL	Divergencia inducida por PLE.
DS	Divergencia inducida por score S.
SDP	Matriz Semidefinida Positiva.
DVS	Descomposición en Valores Singulares.
KKT	Condiciones de Karush-Kuhn-Tucker.
AUC	Area Under the Receiver Operating Characteristic Curve, por sus siglas en inglés.

PARÁMETROS

ω	Parámetros naturales de la FE.
η_j	Parámetro del ES lineal.
ξ_j	Parámetro del ES lineal binario.
Θ_{jl}	Interacciones entre ESs j y l.
Λ_{jl}	Interacciones entre ESs binarios j y l.
Φ_{jl}	Cross-interacciones entre ES binario j y ES l.

Parte I

INTRODUCCIÓN

1

REDUCCIÓN DE DIMENSIONES Y DATOS COMPOSICIONALES EN ALTA DIMENSIÓN

1.1 MOTIVACIÓN

La metagenómica y, más concretamente, la investigación sobre el microbioma humano es una disciplina que enfrenta una necesidad concreta de mejores herramientas analíticas. Esta tesis busca extender modelos y proponer estrategias para su exploración e interpretación.

1.1.1 *Microbioma y sus desafíos*

La metagenómica es el estudio de la estructura y función de la información genética de todos los microorganismos que cohabitan en un mismo ambiente. Una aplicación que ha cobrado especial relevancia recientemente es el estudio del microbioma humano; esto es, del conjunto de microorganismos que viven en el interior del cuerpo humano y sobre su superficie (Turnbaugh et al., 2012a). Estos microorganismos aportan una enorme capacidad metabólica e inmunitaria, tanto que alteraciones en la composición de esta flora microbiana tienen efectos importantes sobre la salud y el bienestar de la persona en la que habitan (Guarner, Malagelada et al., 2012). En los últimos 15 años se han descrito asociaciones entre el microbioma humano y la aparición de ciertas enfermedades crónicas, como enfermedades inflamatorias del intestino (Khoruts y Sadowsky, 2017), diabetes (Cani et al., 2014) y síndrome metabólico (Díaz-Rizzolo et al., 2018; Turnbaugh et al., 2012b), o incluso algunos tipos de cáncer (Gopalakrishnan et al., 2017). Estudios más recientes se han concentrado también en la transmisión de perfiles microbianos de la madre al hijo durante el parto y la lactancia, y su posible efecto sobre la salud y el desarrollo infantil (Blaser et al., 2016). Otros estudios han abordado las alteraciones en la flora vaginal y su impacto sobre la fertilidad (Heinemann et al., 2017) o el desencadenamiento de partos prematuros (Fettweis, Serrano, Brooks et al., 2019).

Esta comprensión creciente de la interacciones entre la microbiota y su huésped también tiene el potencial de engendrar aplicaciones de valor comercial, o de tener una relevancia creciente en tecnologías ya existentes (Yadav y Chauhan, 2021). Así, distintas empresas y grupos académicos exploran soluciones naturales no medicamentosas para problemas de obesidad y diabetes, complementos dietarios, propiedades de cosméticos, etc. El potencial de la microbiota para modular la respuesta del organismo a tratamientos oncológicos es también un área de investigación activa.

El análisis de datos referidos al microbioma es un ingrediente fundamental para acelerar estos desarrollos, a fin de poder interpretar

cómo los distintos componentes de la microbiota interactúan entre sí y con el huésped, cómo responden a perturbaciones externas y cómo evolucionan en el tiempo. De lograrlo, la actuación controlada sobre el microbioma podría jugar un papel clave en la medicina personalizada. No obstante, los datos de microbioma presentan un conjunto de desafíos importantes para el modelado estadístico que vuelven inadecuadas muchas herramientas de análisis comunes. Los presentamos brevemente a continuación.

1.1.1.1 *Datos de conteo y composicionales*

La característica de conteo y composicional de los datos de microbioma proviene de la manera en la que se recolectan los datos. Al estudiar una comunidad microbiana como la flora intestinal, interesa conocer qué especies están presentes y en qué cantidad. La secuenciación genómica de alto rendimiento (high-throughput Next Generation Sequencing) permite una cuantificación directa de la abundancia de las distintas especies. El proceso implica una sucesión de procedimientos complejos que típicamente comienzan con la toma de la muestra biológica y terminan con la clasificación de fragmentos de ADN en especies catalogadas en bases de datos de referencia. El análisis bioinformático final simplemente reporta cuántas veces se logró identificar cada especie del catálogo en una muestra. Así, los datos son esencialmente *datos de conteo*. Sin embargo, las distintas etapas de este proceso de cuantificación no son perfectas e introducen variabilidad, tanto que la abundancia absoluta registrada a menudo es menos informativa que la abundancia relativa de las distintas especies. Es decir, interesa más la composición de la comunidad. De esta forma, los datos de microbioma son intrínsecamente *composicionales*.

Desde el punto de vista formal, si $X_1, X_2, \ldots, X_p$ representan la abundancia de las distintas especies encontradas en la comunidad, el perfil completo de abundancias $\mathbf{X} = (X_1, \ldots, X_p)^T$ satisface $X_j \in \mathbb{R}_+ \bigcup \{0\}$ y $\sum_{j=1}^{p} X_j = \kappa$, con κ fijo. Cuando consideramos abundancias relativas, $\kappa = 1$. Alternativamente pueden considerarse valores enteros y κ tomar un valor distinto para cada muestra. Este valor, no obstante, a menudo refleja más la variabilidad técnica que la realidad biológica. Más aún, algunas estrategias de modelado discreto recurren a fijar el valor de κ submuestreando al azar el perfil original. Esto se conoce a menudo como *rarefacción* (Ovaskainen y Knegt, 2017) y, si bien facilita el modelado, puede despreciar la influencia de especies muy poco presentes (McMurdie y Holmes, 2014).

La naturaleza composicional de los datos implica que una comunidad pueda representarse como una distribución de probabilidad puntual. Es así como la entropía de Shannon y la divergencia de Jensen-Shannon son usados frecuentemente para describir la composición de cada comunidad y la similaridad entre ellas, respectivamente (Li, 2015). Más aún, como se puede definir una distancia filogenética entre especies distintas (una ultramétrica), también es posible comparar

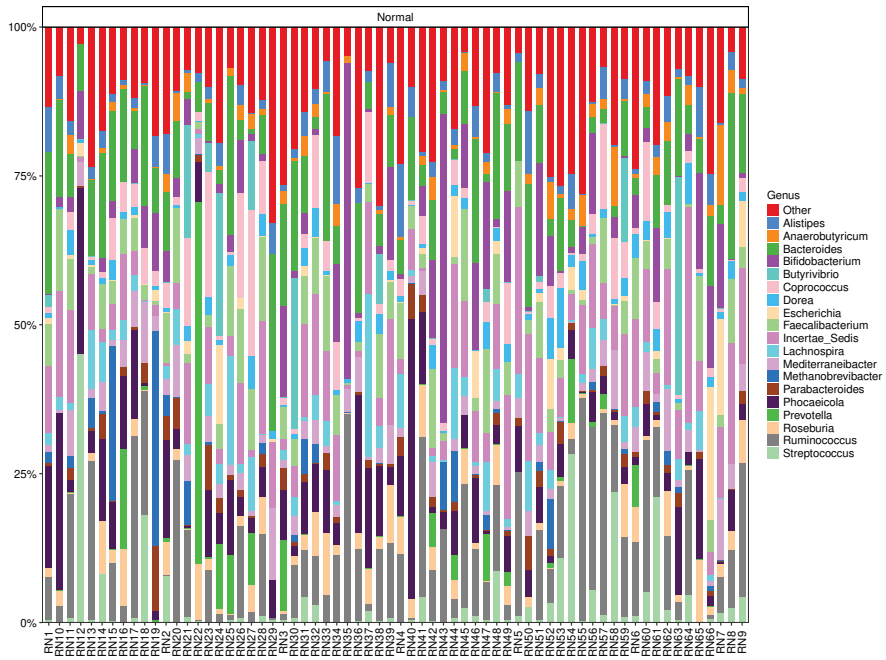


Figura 1.1: Ilustración de la composición de la flora intestinal a nivel de Género para sujetos saludables. Aunque existen más de 250 géneros distintos en el conjunto de datos, incluso los veinte mas abundantes en promedio resultan difíciles de visualizar debido a la disparidad de abundancias y variabilidad entre individuos.

comunidades distintas en términos de distancias de Wassertein [1]. Más aún, muchos métodos estadísticos suponen implícitamente la distancia euclídea como métrica natural subyacente. La aplicación directa de tales métodos sobre datos composicionales resulta inadecuada. En (Aitchison y Bacon-Shon, 1984) se describen transformaciones logarítmicas que pueden aplicarse sobre los perfiles composicionales para que las herramientas estadísticas clásicas resulten más apropiadas. No obstante, las transformaciones con mejores propiedades matemáticas usualmente son útiles para ayudar a identificar diferencias globales en la composición, pero no para explorar la relevancia de componentes individuales.

1.1.1.2 *Alta dimensionalidad y gran proporción de ceros*

La metagenómica permite cuantificar la abundancia de miles de especies en forma simultánea. Más aún, en ocasiones es de interés analizar qué cepas o variantes de una misma especie intervienen en el problema, incrementando así el número de variables. Por el contrario, la cantidad de muestras o casos disponibles típicamente es reducida, debido a los costos de la tecnología y a la dificultad de conseguir voluntarios adecuados para pruebas específicas. Por ejemplo, un escenario común de estudio suele incluir entre 30 y 50 voluntarios por grupo de interés, posiblemente tomando mediciones

1 En el ámbito biológico, la distancia de Wassertein entre composiciones microbianas se conoce como *weighted UniFrac*. Esta relacion se establece en Evans y Matsen (2012).

repetidas a lo largo del tiempo. Esto implica que el escenario de análisis típico tiene muchas más variables que casos y configura un problema de alta dimensión.

Por otra parte, la composición de la microbiota suele presentar diferencias importantes entre una persona y otra (ver Figura 1.1). Esto hace que una especie pueda ser detectada en algunas pocas muestras pero no en el resto. Como resultado, la matriz de datos completa presenta una gran cantidad de ceros, pudiendo alcanzar el 90 % de los valores. Aunque en el análisis a menudo se descartan de antemano componentes que se encuentran sólo en unas pocas muestras, un filtrado excesivo puede perder información clave para interpretar los resultados de un estudio. Más aún, se acepta que algunas especies pueden ejercer una función reguladora sobre la composición del ecosistema microbiano, aún presentando abundancias muy bajas [2].

Vale mencionar que a menudo el análisis se realiza a diferentes niveles taxonómicos. Esto es, la especies se agrupan en géneros, estos en familias, etc. Es claro que la cantidad de variables, al igual que la proporción de elementos nulos en la matriz de datos, disminuye al subir en la jerarquía taxonómica.

1.1.2 *El contexto analítico*

La presentación anterior sugiere que los datos de microbioma recogidos en un estudio clínico cuentan con dos componentes informativas: los *perfiles de abundancia* y los *patrones de co-ocurrencia*. El análisis macroscópico de los datos a fin de encontrar asociaciones globales entre la microbiota y otras variables de interés comúnmente tiene en cuenta ambos aspectos por separado, definiendo para ello nociones diferentes de similitud que resultan relevantes desde un punto de vista ecológico. En este sentido, la disimilaridad de Bray-Curtis es una forma común de tener en cuenta la abundancia relativa de las distintas especies (ver Figura 1.2), mientras que la disimilaridad de Jaccard se enfoca en la presencia/ausencia de especies en las muestras comparadas. Versiones generalizadas de otra noción de disimilaridad conocida como UniFrac permiten una comparación similar, incluyendo las relaciones filogenéticas entre las especies.

Así, es común computar estas medidas de disimilaridad entre pares de observaciones y usar esas matrices para producir representaciones de baja dimensión por escalado multidimensional (MDS). Tales representaciones son no-supervisadas y típicamente capturan una fracción pequeña de la variabilidad total, por lo que no garantizan ser indicativas de la posible relación entre la composición de la microbiota y los grupos o variables en estudio. Las relaciones de disimilaridad entre muestras distintas también se usan para contrastar formalmente si existen diferencias en la composición de la microbiota entre grupos de interés. Tales pruebas estadísticas recurren a permutaciones sobre

2 Estas especies a menudo se denominan *keystone* y su identificación es un objetivo de análisis en sí mismo.

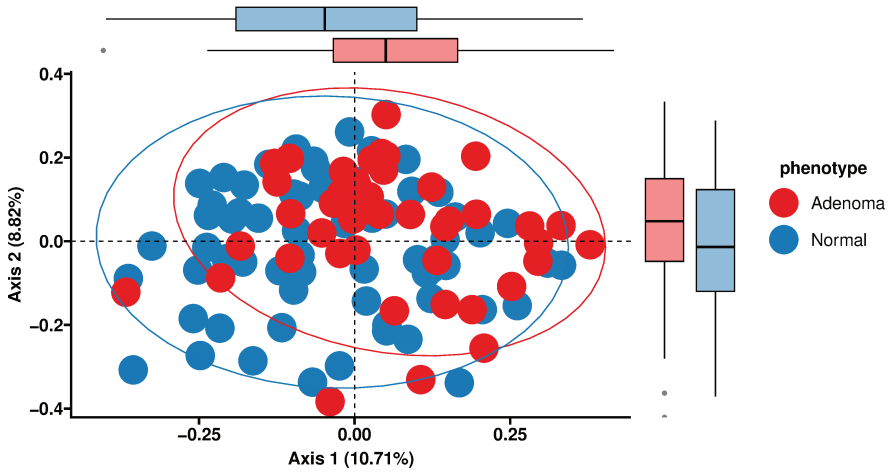


Figura 1.2: Representación no supervisada de los datos usando la disimilaridad de Bray-Curtis.

las distancias o modelos parcialmente lineales que modelan el efecto del microbioma como una función de tales matrices (ver, por ejemplo Zhao et al., 2015). A diferencia de lo que ocurre en pruebas simples donde la inferencia intenta confirmar si la evidencia observada en un análisis exploratorio es siginificativa, los resultados del análisis exploratorio y los tests mencionados pueden ser contradictorios. Más aún, los tests estadísticos no ofrecen una estimación eficaz de tamaño de efecto, por lo que eventualmente resulta difícil apreciar la relevancia clínica de comparaciones estadísticamente significativas.

Otros métodos comunes de estadística multivariada y aprendizaje automático pueden usarse para obtener representaciones de baja dimensión que ayuden a la visualización y comprensión de los datos estudiados, aunque típicamente requieren alguna transformación previa. Por ejemplo, es posible encontrar visualizaciones basadas en PCA, UMAP o t-SNE de datos transformados de forma isométrica (por ejemplo, mediante la transformación ILR, por Isometric Log-Ratio). Estas transformaciones encuentran un nuevo sistema de coordenadas de representación como combinación de las originales. El mayor inconveniente es que dificultan la posibilidad de ponderar qué componentes de la microbiota contribuyen más al efecto observado. Más aún, estas transformaciones logarítmicas a veces se implementan corrigiendo los ceros en los datos mediante la adición de un valor mínimo o *pseudocuentas*. El efecto global de esa corrección, cuya motivación es puramente numérica, no siempre es obvio.

La identificación de cuáles componentes de la microbiota son los verdaderos responsables de un efecto observado a escala macroscópica es fundamental para comprender la interacción entre microbioma, huésped y ambiente y para diseñar actuaciones que puedan inducir un efecto clínico favorable en la persona vía una perturbación controlada de su microbioma. El enfoque usual para esta tarea consiste en contrastar diferencias significativas de abundancia o prevalencia para cada componente utilizando modelos lineales mixtos generalizados, ajustando luego los p-valores obtenidos para controlar la tasa de

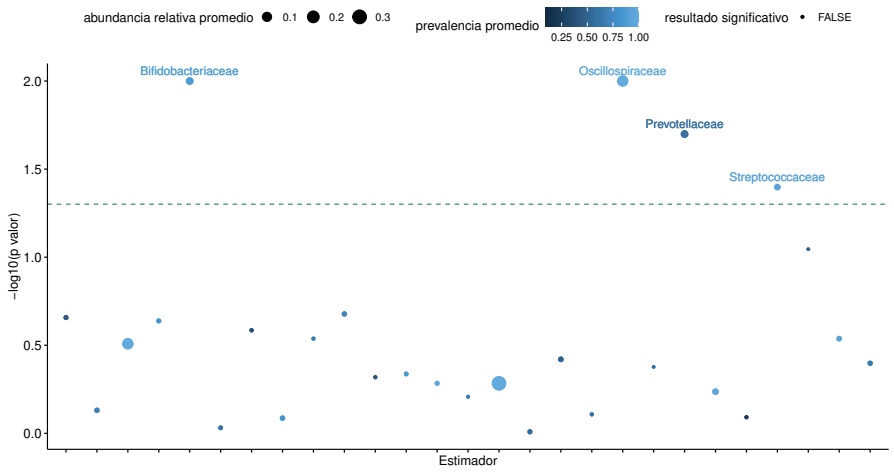


Figura 1.3: Resultados de un análisis de abundancia diferencial a nivel de Familias bacterianas. La linea punteada horizontal indica el umbral de significancia individual de los tests. No obstante, ningún resultado es estadísticamente significativo luego de aplicar una corrección de tipo Benjamini-Hochberg sobre los p-valores individuales.

descubrimientos falsos (FDR). En la práctica, los tamaños muestrales disponibles habitualmente dificultan encontrar resultados estadísticamente significativos a menos que se limite a priori el conjunto de componentes contrastados (ver Figura 1.3). Más aún, la utlización simultánea de diferentes modelos estadísticos muchas veces muestra una gran disparidad de resultados. Esto es, los resultados obtenidos dependen fuertemente de la metodología usada. Por otra parte, el volumen de pruebas estadísticas abordado en simultáneo vuelve impracticable un análisis de residuos. De esta forma, todas las suposiciones de modelado se aceptan a priori. Por lo tanto, aún si se obtienen p-valores con pruebas estadísticas formales, las conclusiones continúan siendo esencialmente exploratorias. Reconocer esta limitación de los tests formales permite escoger alternativas multivariadas que son intrínsecamente exploratorias, pero que pueden revelar tendencias útiles para la comprensión global del fenómeno en estudio.

1.2 OBJETIVOS Y CONTRIBUCIÓN

El objetivo de esta tesis es explorar herramientas multivariadas para el análisis de datos de microbioma que ofrezcan una alternativa de análisis más unificada a las preguntas típicas, evitando la fragmentación de modelado actual en la que enfoques potencialmente muy distintos se usan para responder preguntas relacionadas de un mismo problema. El hilo conductor será la reducción suficiente de dimensiones, como forma de obtener una representación de baja dimensión más efectiva para el problema de interés que motiva el análisis. Basaremos esas reducciones en modelos gráficos condicionales, que nos permitirán modelar simultáneamente relaciones entre las componentes del microbioma y entre ellas y otras variables de interés.

Trabajaremos especialmente en modelos gráficos que pertenezcan a la familia exponencial de distribuciones de probabilidad, que nos ofrecen una plantilla general para hallar reducciones suficientes. La mayor parte del trabajo recae entonces en la definición y adopción de modelos adecuados y en su estimación. La introducción de estrategias de estimación con regularización jerárquica nos permitirá identificar los componentes más importantes del microbioma en relación con la respuesta, dando así una vía más unificada para el análisis de datos de microbioma.

En este trabajo definimos familias paramétricas multivariadas para el modelado de datos composicionales con exceso de ceros. Desarrollamos e implementamos luego distintos métodos de estimación para esos modelos, que sean apropiados para escenarios de alta dimensión. Recurrimos en particular a enfoques basados en scores generalizados y funciones de pseudoverosimilitud penalizados. Además de facilitar la estimación en alta dimensión, la regularización introducida está especialmente estructurada para inducir patrones de selección de variables en el estimador que permitan identificar las variables más relevantes asociadas a la respuesta. Finalmente, estudiamos la aplicación de los métodos propuestos en casos reales.

1.2.1 *Trabajos relacionados*

Existen algunos pocos antecedentes del uso de modelos gráficos para la descripción de datos de microbioma. La mayoría de estos trabajos incorpora en realidad modelos pensados originalmente para el análisis automático de texto, aprovechando la similitud formal entre ambos problemas. En este marco podemos encontrar algunas extensiones multivariadas de modelos Poisson y modelos de variables latentes que culminan con un modelo multinomial en su capa observable. En particular, los modelos Dirichlet-Multinomial se han usado como simuladores paramétricos elementales de datos de microbioma y para estudiar dependencia de la composición del microbioma con factores externos, como la dieta (Chen y Li, 2013). Los modelos jerárquicos Log-Normal-Multinomial también se han usado en este contexto, con la ventaja de permitir relaciones de covarianza más flexibles a nivel de las variables latentes (Xia et al., 2013). La mayor limitación de estos modelos, no obstante, reside en que no ofrecen una traducción simple de los parámetros del modelo a una representación interpretable de los datos. Típicamente tampoco modelan el exceso de ceros, que es una característica distintiva. Modelos gráficos similares a modelos Ising también se han usado para describir patrones de co-ocurrencia, sin ofrecer una vía de visualización o de relación con una variable respuesta.

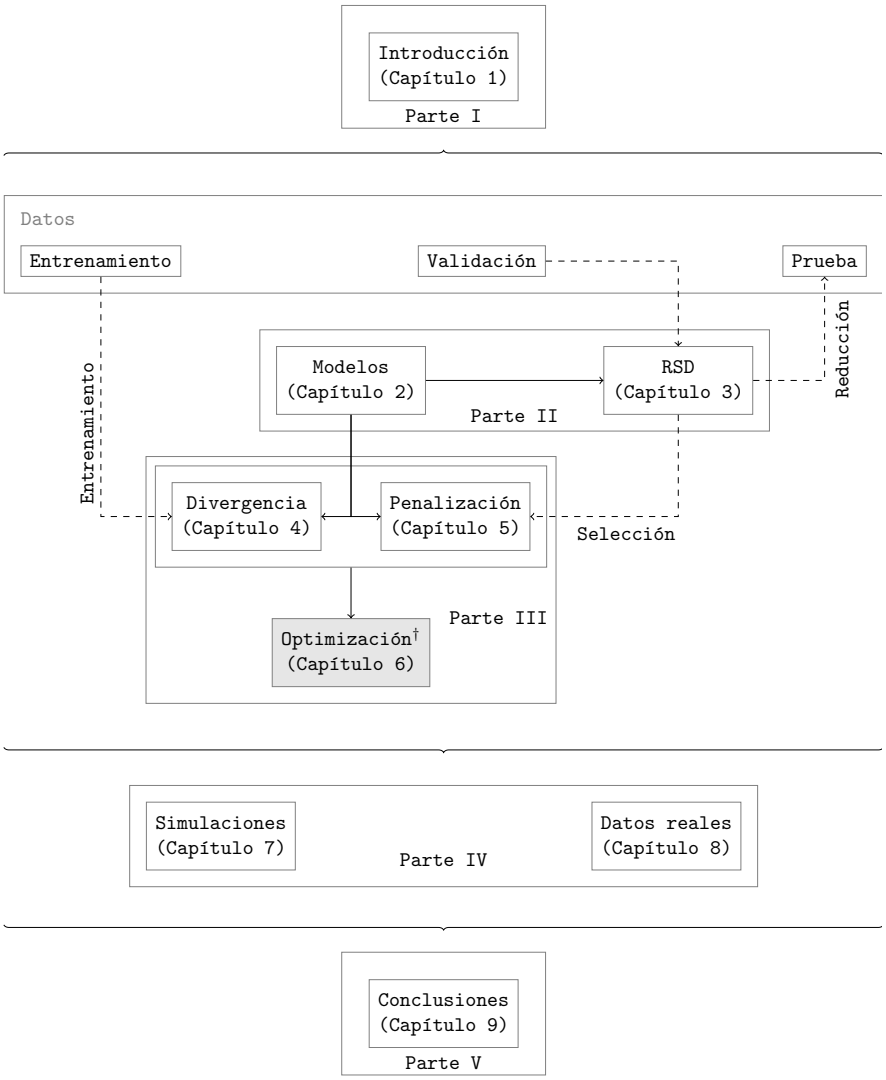


Figura 1.4: Diagrama de los contenidos de cada capítulo y su interrelación en función del análisis de datos. Las líneas sólidas indican la relación entre los distintos capítulos que conforman la tesis, mientras que las líneas punteadas implican el uso de datos.

1.3 ORGANIZACIÓN

La primera parte está dedicada al estudio de la metodología general que fundamenta esta tesis. En particular, la reducción suficiente de dimensiones en familias exponenciales. Además veremos que los modelos gráficos nos permiten modelar distribuciones conjuntas en la familia exponencial, lo que permitirá diseñar distribuciones que representen a los datos con mayor fidelidad. Concluimos con ejemplos

† El Capítulo 6 referido a los métodos de optimización no es necesario para comprender los resultados de esta tesis y se presenta en el texto principal por completitud.

de distintas distribuciones paramétricas que servirán de base para mostrar resultados en los siguientes capítulos.

En una segunda parte estudiaremos distintos métodos de estimación para los modelos presentados. Veremos que la noción fundamental de la *divergencia*, como distancia entre distribuciones, permite interpretar la estimación como un problema de minimización de un *score*. Mostraremos con ejemplos su aplicación en la estimación de modelos en baja dimensión. Seguidamente abordaremos el problema de estimación en el contexto de alta dimensión y mostraremos su aplicación con ejemplos comparativos mediante simulaciones. Concluimos con ejemplos de aplicación a datos composicionales de microbioma.

Los desarrollos presentados toman como punto de partida resultados previos de reducción suficiente de dimensiones, propiedades de la familia exponencial y conceptos de estimación regularizada y optimización. A fin de lograr un documento autocontenido pero a la vez fluido, la división entre antecedentes y contribuciones específicas se intercala a lo largo de los capítulos, remarcando las contribuciones.

En la Figura 1.4 se presenta un diagrama que resume los contenidos correspondientes a cada capítulo, así como su interrelación en función al análisis de datos.

Parte II

MODELOS Y REDUCCIÓN DE DIMENSIONES

La tesis se estructura en torno a la reducción suficiente de dimensiones basada en modelos gráficos condicionales pertenecientes a la familia exponencial y especialmente adaptados a datos composicionales en alta dimensión. La reducción de dimensiones abordada bajo el enfoque de suficiencia estadística persigue formalmente preservar la información sobre la respuesta disponible en los predictores. Abordar el problema a partir de modelos de regresión inversa permite obtener reducciones exhaustivas si el modelo es apropiado. Por otra parte, el esfuerzo dedicado a la estimación de esos modelos de regresión inversa frecuentemente tiene relevancia en sí mismo, ya que sirve también para entender cómo un conjunto de predictores varía en función de covariables. En el caso de datos de microbioma, por ejemplo, permitirían modelar la variación de la composición en función de la dieta, ingesta de medicamentos, estado gestacional, etc. Todo ello se presenta aquí dentro de la familia exponencial. La ventaja de permanecer en la familia exponencial reside en conocer de antemano una estrategia general para deducir la reducción suficiente correspondiente a modelos particulares.

A continuación, en los capítulos 2 y 3 se presentan los conceptos básicos sobre reducción suficiente de dimensiones para regresión, familia exponencial y modelos gráficos, poniendo énfasis en los resultados que forman el punto de partida para los desarrollos originales de esta tesis. Posteriormente introducimos una nueva familia general de modelos gráficos para datos que poseen una proporción de ceros arbitraria, lo que nos ayudará a definir nuevas reducciones suficientes para datos composicionales en alta dimensión.

2

FAMILIA EXPONENCIAL Y MODELOS GRÁFICOS

2.1 INTRODUCCIÓN

La Familia Exponencial (FE) de distribuciones (Definición 2.1) es un modelo estadístico que contiene algunas de las distribuciones más usadas en el modelado estadístico. Entre ellas se encuentran muchas de las distribuciones discretas conocidas, como la Bernoulli, su extensión multivariada conocida como modelo Ising, la multinomial y la Poisson. También contiene las distribuciones continuas usuales, como la Normal y su extensión multivariada, la Beta y la Gamma.

Una subfamilia que ha cobrado importancia en los últimos años es la de los *Modelos Gráficos* MG (Definición 2.3). Esta subfamilia parametriza la estructura de independencia condicional entre las variables (Teorema 2.1). Los modelos más simples que codifican dicha estructura son los MGs de segundo orden (Definición 2.5), siendo la distribución Normal multivariada y el modelo Ising algunos miembros notables. Teorema 2.2 caracteriza los MGs de segundo orden cuando la distribución de una variable dadas las demás está en la FE y su estadístico suficiente es escalar.

2.2 CONTENIDOS DEL CAPÍTULO

El capítulo comienza con un resumen de propiedades básicas de la familia exponencial, hasta llegar a presentar los modelos gráficos. En la Sección 2.3.1 mostramos ejemplos de MG. Finalizado el resumen de estos antecedetes, en la Sección 2.4 introducimos una nueva familia de MG formulada para modelar exceso de ceros en los datos de microbioma, incluyendo una masa de probabilidad en los ejes coordenados. Seguidamente definimos la FE condicional lineal que usaremos en capítulos posteriores para definir y estimar una *reducción suficiente de dimensiones*, concepto que desarrollaremos en el Capítulo 3.

En el material que presentamos a continuación consideramos que las variables aleatorias tienen soporte en un espacio medible $\mathcal{X} \subseteq \mathbb{R}^p$. Notar que cuando el espacio sobre el que se definen las variables aleatorias es numerable, las variables se dicen discretas y llamamos $P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\omega})$ a la función de probabilidad evaluada en $\boldsymbol{x}$; sin embargo, cuando el espacio es continuo, es necesario considerar que la notación $P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\omega})$ se refiere a la densidad de probabilidad evaluada en $\boldsymbol{x}$. A su vez, el cómputo de algunas cantidades que están definidas en términos de la suma $\sum_{\boldsymbol{x}\in\mathcal{X}}(\cdot)$ deben entenderse como la integral $\int_{\boldsymbol{x}\in\mathcal{X}}(\cdot)d\boldsymbol{x}$.

También haremos uso de dos operadores de concatenación: $(\cdot,\cdot) : \mathbb{R}^{a_1\times b_1} \times \mathbb{R}^{a_2\times b_2} \to \mathbb{R}^{a_1 b_1 + a_2 b_2}$ devuelve el vector columna resultante

de concatenar las columnas del primer argumento seguido de las columnas del segundo; mientras que $(\cdot;\cdot) : \mathbb{R}^{a_1 \times b} \times \mathbb{R}^{a_2 \times b} \to \mathbb{R}^{(a_1+a_2)\times b}$ devuelve una matriz resultado de concatenar los argumentos verticalmente.

2.3 ANTECEDENTES

Familia exponencial

Definición 2.1 (Familia Exponencial)**.** *Una variable aleatoria pertenece a la FE si su densidad o función de probabilidad P puede expresarse como*

$$P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\omega}) = \exp\{\langle \boldsymbol{\omega}, \boldsymbol{T}(\boldsymbol{x})\rangle + h(\boldsymbol{x}) - Z(\boldsymbol{\omega})\}, \qquad \forall \boldsymbol{x} \in \mathcal{X}, \tag{2.1}$$

donde $\boldsymbol{T} : \mathbb{R}^p \to \mathbb{R}^m$ es el Estadístico Suficiente (ES), $\boldsymbol{\omega} \in \mathbb{R}^m$ el parámetro natural, $h : \mathbb{R}^p \to \mathbb{R}$ la medida base y $\exp\{Z(\boldsymbol{\omega})\} = \sum_{x\in\mathcal{X}} \exp\{\langle \boldsymbol{\omega}, \boldsymbol{T}(\boldsymbol{X})\rangle + h(\boldsymbol{x})\}$ la función de partición.

Fijando el ES, $\boldsymbol{T}(\cdot)$, cada $\boldsymbol{\omega} \in \Omega \subseteq \mathbb{R}^m$ indexa un miembro particular de $P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\omega})$, i. e. una densidad o función de probabilidad en la familia exponencial.

Definición 2.2 (Espacio paramétrico)**.** *Los parámetros naturales $\boldsymbol{\omega}$ en la familia exponencial (Definición 2.1) pertenecen al espacio de parámetros*

$$\Omega := \{\boldsymbol{\omega} \in \mathbb{R}^m \mid Z(\boldsymbol{\omega}) < +\infty\}.$$

Además, una familia exponencial se dice *regular* si su espacio de parámetros $\boldsymbol{\Omega}$ es abierto en $\mathbb{R}^m$ y su representación es *minimal* si su estadístico suficiente $\boldsymbol{T}(\boldsymbol{x})$ es de rango completo para casi todo $\boldsymbol{x}$. Esta condición asegura que cada miembro de la familia está indexado por un único parámetro natural $\boldsymbol{\omega}$.

Una propiedad fundamental de la familia exponencial es la convexidad de la función de partición, enunciada en la siguiente Proposición.

Proposición 2.1 (Función de partición)**.** *La función de partición $Z(\boldsymbol{\omega})$ de una familia exponencial regular tiene las siguientes propiedades (Wainwright, Jordan et al., 2008):*

- *Tiene derivadas de todos los órdenes en Ω. Las primeras dos derivadas pueden escribirse como la esperanza y la varianza del ES $\boldsymbol{T}(\boldsymbol{X})$, esto es:*

$$\begin{aligned} \frac{\partial Z}{\partial \boldsymbol{\omega}} &= \mathrm{E}_{\boldsymbol{\omega}}[\boldsymbol{T}(\boldsymbol{X})], \\ \frac{\partial^2 Z}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^\top} &= \mathrm{var}_{\boldsymbol{\omega}}[\boldsymbol{T}(\boldsymbol{X})]. \end{aligned} \tag{2.2}$$

- *Z es una función convexa de $\boldsymbol{\omega}$ en Ω; además es convexa en sentido estricto si la representación es minimal.*

Además, la esperanza del estadístico suficiente (2.2) define una parametrización equivalente de la familia exponencial (Amari, 2016).

Modelos gráficos

En lo que sigue, se utiliza la notación $\boldsymbol{X}_S$ y $\boldsymbol{X}_{\setminus S}$ para referirse al subconjunto de variables indexado por S y a su complemento respectivamente.

Definición 2.3 (Modelo Gráfico alias *Markov Random Field*)**.** *Una distribución con función de probabilidad o densidad de probabilidad P es un Modelo Gráfico (MG) respecto del grafo $G = (V, E)$, donde V es el conjunto de nodos y E de aristas, si P factoriza por cliques, i. e.*

$$P(\boldsymbol{X} = \boldsymbol{x}) = \prod_{C \in cl(G)} \phi_C(\boldsymbol{x}_C), \qquad \forall \boldsymbol{x} \in \mathcal{X},$$

donde $cl(G)$ es el conjunto de cliques (subgrafo completo) de G y $\phi_C(\cdot)$ son potenciales no negativos.

La siguiente definición es más técnica, pero necesaria para la validez de la mayoría de los resultados. Su postulado es central para la construcción de extensiones multivariadas y las propiedades de independencia condicional.

Definición 2.4 (v.a. positiva)**.** *$\boldsymbol{X}$ es una v.a. positiva si y solo si su función de probabilidad o densidad de probabilidad satisface*

$$P(\boldsymbol{X} = \boldsymbol{x}) > 0 \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

Proposición 2.2 (Modelos gráficos $\subset$ familia exponencial)**.** *Todo modelo gráfico positivo está en la familia exponencial.*

La demostración puede encontrarse en el Apéndice A.1.

INDEPENDENCIA CONDICIONAL Los modelos gráficos representan la estructura de independencia condicional entre las variables a partir de un grafo de conectividad. Dado un grafo $G = (V, E)$ y variables aleatorias $\{X_j,\ j \in V\}$, una medida de probabilidad P obedece las propiedades de Markov

DE A PARES: si todas las variables no adyacentes en G son condicionalmente independientes

$$X_j \perp\!\!\!\perp X_l \mid \boldsymbol{X}_{\setminus\{j,l\}}$$

LOCAL: si una variable es condicionalmente independiente a las demás dados sus vecinos en G

$$X_j \perp\!\!\!\perp X_{\setminus N(j)} \mid \boldsymbol{X}_{N(j)}$$

GLOBAL: si los conjuntos de variables son condicionalmente independientes dado un conjunto que las separe

$$X_A \perp\!\!\!\perp X_B \mid \boldsymbol{X}_S,$$

si cada camino entre A y B pasa por S en G.

En general, si P satisface la propiedad Markov global, también satisface Markov local y a su vez si satisface Markov local, también satisface Markov de a pares. Pero si las variables son positivas, las propiedades de Markov son equivalentes (Lauritzen, 1996).

Teorema 2.1 (Hammersley–Clifford (Lauritzen, 1996))**.** *Dado un grafo $G = (V, E)$, un conjunto de variables aleatorias positivas forman un MG respecto a G si y solo si satisfacen la propiedad de Markov de a pares.*

Modelos gráficos de segundo orden

La siguiente subfamilia de los modelos gráficos resulta de considerar momentos de hasta segundo orden, lo que permite explorar relaciones de independencia condicional entre las variables, dando lugar a la siguiente forma paramétrica:

pGM

Definición 2.5 (Modelo gráfico de segundo orden alias pGM)**.** *Una distribución con función o densidad de probabilidad P es un MG de segundo orden (pGM) si*

$$P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{\Theta}) = \exp\left\{\boldsymbol{\eta}^\top \boldsymbol{T}(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{T}(\boldsymbol{x})^\top \boldsymbol{\Theta}\boldsymbol{T}(\boldsymbol{x}) + h(\boldsymbol{x}) - Z(\boldsymbol{\eta}, \boldsymbol{\Theta})\right\}, \tag{2.3}$$

donde $(\boldsymbol{\eta}, \boldsymbol{\Theta})$ son los parámetros naturales, $(\boldsymbol{T}(\boldsymbol{x}), \boldsymbol{T}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{x})^\top)$ el estadístico suficiente, $\exp Z(\boldsymbol{\eta}, \boldsymbol{\Theta}) = \sum_{\boldsymbol{x}\in\mathcal{X}} \exp\left\{\boldsymbol{\eta}^\top \boldsymbol{T}(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{T}(\boldsymbol{x})^\top \boldsymbol{\Theta}\boldsymbol{T}(\boldsymbol{x}) + h(\boldsymbol{x})\right\}$ la función de partición y $h(\boldsymbol{x}) = \sum_{j=1}^{p} h_j(x_j)$ la medida base del espacio producto. El espacio de parámetros $\boldsymbol{\Omega}$ impone las restricciones de simetría sobre la matriz de interacciones $\boldsymbol{\Theta}$.

Observación 2.1. *Notar que la integral que define la función de partición en (2.3) es p-dimensional, dificultando su cómputo en problemas de aplicación en alta dimensión.*

La Proposición 2.3 enunciada a continuación muestra que cuando $\boldsymbol{X}$ se distribuye de acuerdo a (2.3), las condicionales $X_j \mid X_{\setminus j}$ pertenecen a la FE y su parámetro natural resulta de una combinación lineal de los estadísticos suficientes de las variables condicionales.

Proposición 2.3 (distribución condicional (Yang et al., 2015))**.** *Si $\boldsymbol{X}$ se distribuye de acuerdo con (2.3), la distribución condicional de la variable X_j dadas las demás está dada por:*

$$P(X_j = x_j \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{x}_{\setminus j}) = \exp\left\{\eta_{j|\setminus j} T(x_j) + \frac{\Theta_{jj}}{2} T(x_j)^2 + h_j(x) - Z(\eta_{j|\setminus j})\right\}, \tag{2.4}$$

$$\eta_{j|\setminus j} := \eta_j + \sum_{j\neq i} \Theta_{i,j} T(x_j), \tag{2.5}$$

donde $\exp\{Z(\eta_{j|\setminus j})\} = \sum_{x_j\in\mathcal{X}_j} \exp\left\{\eta_{j|\setminus j} T(x_j) + \frac{\Theta_{jj}}{2} T(x_j)^2 + h_j(x_j)\right\}$.

Observación 2.2. *Notar que la integral que define la función de partición en (2.4) es unidimensional, por lo que el cómputo de dichas distribuciones escala linealmente con la dimensión p del problema.*

Recíprocamente, el Teorema 2.2 permite caracterizar los MGs de segundo orden a partir de las condicionales. Esta equivalencia, sumada a la facilidad de cómputo de la función de partición de algunos modelos condicionales, motiva el uso del estimador de *pseudolikelihood* que estudiaremos en Capítulo 4.

Algoritmo 1: Gibbs sampling (Koller y Friedman, 2009)

Datos: $P(\boldsymbol{X} = \boldsymbol{x})$, $P^{(0)}(\boldsymbol{X} = \boldsymbol{x})$, $\tilde{T}$
Resultado: $\boldsymbol{x}^{(0)}, \cdots, \boldsymbol{x}^{(T)}$
Generar $\boldsymbol{x}^{(0)}$ a partir de $P^{(0)}(\boldsymbol{X} = \boldsymbol{x})$
para $t = 1, \cdots, \tilde{T}$ **hacer**
  $\boldsymbol{x}^{(t)} := \boldsymbol{x}^{(t-1)}$
  para $j = 1, \cdots, p$ **hacer**
    Generar $x_j^{(t)}$ a partir de $P(X_j = x_j \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{x}_{\setminus j}^{(t)})$
  fin
fin

Yang et al. (2015) establece un teorema de construcción de modelos gráficos de orden K, asumiendo que las condicionales $X_j \mid \boldsymbol{X}_{\setminus i}$ están en la familia exponencial y poseen un estadístico suficiente escalar. Enunciamos el teorema a continuación para MG de orden 2:

Teorema 2.2 (Caracterización (Yang et al., 2015))**.** *Supongamos que la distribución conjunta $P(\boldsymbol{X})$ factoriza sobre cliques de grado 2 de un grafo $G = (V, E)$, donde cada nodo representa una variable X_j. $\boldsymbol{X}$ es un vector aleatorio cuyas distribuciones condicionales $P(X_j \mid \boldsymbol{X}_{\setminus j})$ están especificadas por una distribución en la familia exponencial con estadístico suficiente escalar $T(X_j)$ y parámetro natural $\eta_{j|\setminus j}$, el cual depende de todas las variables excepto X_j. Entonces, $\eta_{j|\setminus j}$ satisface (2.5) y la distribución conjunta de $\boldsymbol{X}$ está dada por (2.3).*

Observar que la condición $P(\boldsymbol{X})$ factoriza sobre cliques de grado 2 de una grafo G es equivalente a que la distribución P resulta del producto de potenciales, donde cada factor depende de a lo sumo dos componentes de $\boldsymbol{X}$.

MUESTREO Tanto para simular datos como para aproximar algunas esperanzas, usamos en esta tesis que los MG son *generativos*, ya que permiten generar muestras a partir de una distribución multivariada $P(\boldsymbol{X} = \boldsymbol{x})$. En particular, el método llamado *Gibbs sampling* presentado en el Algoritmo 1 permite generar muestras de $P(\boldsymbol{X} = \boldsymbol{x})$ partiendo de una distribución inicial $P^{(0)}(\boldsymbol{X} = \boldsymbol{x})$ para luego entrar en una fase de *burn-in* donde se reemplaza el valor de cada una de las variables por una muestra generada a partir de la distribución condicional dado el valor de las demás. Este proceso se repite un número $\tilde{T}$ veces que asegure la convergencia.

Este proceso define una *cadena de Markov* que consiste en un espacio de estados $\mathcal{X}$ y las probabilidades de transición $\mathcal{T}(\boldsymbol{x} \to \boldsymbol{x}')$ de un estado $\boldsymbol{x} \in \mathcal{X}$ a un estado $\boldsymbol{x}' \in \mathcal{X}$. Para asegurar que este algoritmo genere muestras de la distribución objetivo $P(\boldsymbol{X} = \boldsymbol{x})$ es necesario asegurar dos condiciones:

1. que la distribución objetivo $P(\boldsymbol{X})$ sea estacionaria (ver Definición 2.6 debajo),

2. que la cadena de Markov sea *ergódica*, i. e. tiene una única distribución estacionaria que puede ser alcanzada partiendo de cualquier distribución inicial $P^{(0)}$.

A continuación definimos estas condiciones en el contexto de modelos gráficos.

En particular estamos interesados en el comportamiento asintótico de la cadena de Markov, lo que define una distribución estacionaria $\pi(\boldsymbol{X})$ en la cual la probabilidad de estar en un estado $\boldsymbol{x}$ es la misma que la probabilidad de caer en el estado $\boldsymbol{x}$ a partir de un predecesor aleatorio $\boldsymbol{x}'$, i. e.

Definición 2.6 (Distribución estacionaria (Koller y Friedman, 2009))**.** *Una distribución $\pi(\boldsymbol{X})$ se dice estacionaria para una cadena de Markov $\mathcal{T}$ si satisface*

$$\pi(\boldsymbol{X} = \boldsymbol{x}') = \sum_{\boldsymbol{x} \in \mathcal{X}} \pi(\boldsymbol{X} = \boldsymbol{x}) \mathcal{T}(\boldsymbol{x} \to \boldsymbol{x}').$$

Para el caso de modelos gráficos, si consideramos el estado $(x_j, \boldsymbol{x}_{\setminus j}) \in \mathcal{X}$ y definimos las probabilidades de transición

$$\mathcal{T}_j \Big((x_j, \boldsymbol{x}_{\setminus j})) \to (x'_j, \boldsymbol{x}_{\setminus j}) \Big) = P(X_j = x_j \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{x}_{\setminus j}),$$

es posible demostrar que la distribución $P(\boldsymbol{X})$ es estacionaria para la cadena de markov $\mathcal{T}_j$, $j = 1, \cdots, p$ partiendo de la observación de que para $\mathcal{T}_j$, la distribución marginal de $\boldsymbol{X}_{\setminus j}$ es invariante ya que no cambia el valor de $\boldsymbol{x}_{\setminus j}$ y que X_j es muestreada de la distribución condicional $P(X_j = x_j \mid \boldsymbol{X}_{\setminus j = \boldsymbol{x}_{\setminus j}})$. Como la distribución condicional y la marginal definen la distribución conjunta $P(\boldsymbol{X})$, esta resulta invariante (Bishop y Nasrabadi, 2006).

Definición 2.7 (Cadena de Markov regular (Koller y Friedman, 2009))**.** *Una cadena de Markov se dice regular si existe un número k tal que para cualquier $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, la probabilidad de ir de $\boldsymbol{x}$ a $\boldsymbol{x}'$ en exactamente k pasos es positiva.*

En el contexto de modelos gráficos, esta condición se asegura requiriendo que los potenciales que definen $P(\boldsymbol{X})$ sean positivos o equivalentemente, que las distribuciones condicionales $P(X_j = x_j \mid \boldsymbol{X}_{\setminus j = \boldsymbol{x}_{\setminus j}})$ sean positivas. De esta manera, es posible asegurar que cualquier punto $\boldsymbol{x}'$ del espacio de estados $\mathcal{X}$ puede ser alcanzado a partir de otro punto $\boldsymbol{x} \in \mathcal{X}$ en un número finito de pasos que involucren la actualización de cada una de las variables.

Teorema 2.3 (Regularidad implica ergodicidad (Koller y Friedman, 2009))**.** *Si una cadena de Markov con estados finitos $\mathcal{X}$ es regular, entonces tiene una unica distribución estacionaria.*

El número de iteraciones $\tilde{T}$ en el Algoritmo 1 define la fase de *burn-in*. Para especificarlo, es necesario considerar la velocidad de convergencia hacia la distribución estacionaria (Jones y Hobert, 2001). Otro enfoque basado en el hecho que las distribuciones condicionales son suficientes para caracterizar la distribución conjunta (Casella

y George, 1992), ver Teorema2.2 por ejemplo, consiste en evaluar la convergencia de las distribuciones marginales de $\boldsymbol{X}$ a partir de la evaluación de distintos momentos de las variables a lo largo de la fase de *burn-in*.

2.3.1 *Ejemplos*

A continuación mostramos ejemplos de modelos gráficos de segundo orden que fueron analizados por Inouye et al. (2017).

Ejemplo 2.1 (Normal-pGM alias *Normal Multivariada*)**.** *La distribución Normal Multivariada es la generalización multivariada para $\boldsymbol{X} \in \mathbb{R}^p$ de la distribución Normal, que en su parametrización de momentos está definida como*

$$P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \tag{2.6}$$

donde $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{X}] \in \mathbb{R}^p$ y $\boldsymbol{\Sigma} = \mathrm{cov}[\boldsymbol{X}] \in \mathbb{R}^{p\times p}$ es simétrica definida positiva y es conocida como matriz de covarianza.

Podemos escribir (2.6) en la familia exponencial, definiendo los parámetros naturales

$$\begin{aligned}\boldsymbol{\Theta} &= -\boldsymbol{\Sigma}^{-1} \\ \boldsymbol{\eta} &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu},\end{aligned}$$

lo que resulta en

$$P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{\Theta}) = \exp\{\boldsymbol{\eta}^\top \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Theta}\boldsymbol{x} - Z(\boldsymbol{\eta}, \boldsymbol{\Theta})\}, \tag{2.7}$$

donde

$$Z(\boldsymbol{\eta}, \boldsymbol{\Theta}) = \frac{1}{2}\boldsymbol{\eta}^\top \boldsymbol{\Theta}\boldsymbol{\eta} - \frac{p}{2}\log(2\pi) + \frac{1}{2}\log|-\boldsymbol{\Theta}|. \tag{2.8}$$

es el logaritmo de la función de partición. Ver detalles en el Apéndice A.2.

Cuando se trabaja con datos composicionales, una transformación usual es la logarítmica (Aitchison y Bacon-Shon, 1984), dada por

$$Z_j(\boldsymbol{X}) = \log(X_j/X_k), \qquad j = 1, 2, \cdots, k-1, k+1, \cdots, p, \tag{2.9}$$

En este caso, el vector aleatorio pierde una variable y para hacerlo evidente consideramos los parámetros reducidos $\tilde{\boldsymbol{\eta}} \in \mathbb{R}^{p-1}, \tilde{\boldsymbol{\Theta}} \in \mathbb{R}^{p-1\times p-1}$ en lugar de $\boldsymbol{\eta}$ y $\boldsymbol{\Theta}$ respectivamente.

Ejemplo 2.2 (Binary-pGM alias *Ising*)**.** *En el caso $\boldsymbol{X} \in \{0,1\}^p$, el modelo Ising generaliza la distribución Bernoulli univariada. Su forma funcional corresponde a (2.3), con $h(\boldsymbol{x}) = 0$, i. e.*

$$P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{\Theta}) = \exp\left\{\boldsymbol{\eta}^\top \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Theta}\boldsymbol{x} - Z(\boldsymbol{\eta}, \boldsymbol{\Theta})\right\}, \tag{2.10}$$

donde $\Theta_{ii} = 0$, $i = 1, \cdots, p$ y $\exp Z(\boldsymbol{\eta}, \boldsymbol{\Theta}) = \sum_{\boldsymbol{x}\in\{0,1\}^p} \exp\left\{\boldsymbol{\eta}^\top \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Theta}\boldsymbol{x}\right\}$ es intratable para problemas de algunas decenas de variables.

Ejemplo 2.3 (Poisson-pGM). *Es la generalización multivariada cuyas condicionales son Poisson. En particular, consideramos* $\boldsymbol{X} \in (\mathbb{N} \cup \{0\})^p$ *y su distribución corresponde a* (2.3) *con estadístico suficiente* $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{x}$, *medida base* $h(\boldsymbol{x}) = -\sum_{i=1}^{p} \log(x_j!)$, *aunque el espacio paramétrico agrega la restricción* $\Theta_{i,j} \leq 0 \, \forall i \neq j$ *para asegurar la sumabilidad, además* $\Theta_{ii} = 0$, $i = 1, \cdots, p$. *Esta restricción únicamente permite modelar competencia, pero no colaboración. Nuevamente, la constante de normalización es intratable para problemas de algunas decenas de variables.*

Ejemplo 2.4 (TPoisson-pGM). *Motivados por modelos biológicos de conteos, se considera la generalización multivariada de una distribución Poisson truncada superiormente en* T^*, *con lo cual consideramos* $\boldsymbol{X} \in \{0, 1, \cdots, T^*\}^p$. *En este caso es posible relajar la restricción de negatividad de interacciones.*

Inouye, Ravikumar y Dhillon (2015) proponen una generalización de la distribución multinomial que logra incorporar interacciones entre las variables (Definición 2.8). Dicha distribución se obtiene de agregar la restricción $\sum_{i=1}^{p} X = m_{\boldsymbol{X}}$ al modelo Poisson-pGM (Ejemplo 2.3). Dicha restricción resulta natural para los datos composicionales pero vuelve al modelo no identificable.

Definición 2.8 (FPoisson-pGM (Inouye, Ravikumar y Dhillon, 2015))**.** *Decimos que* $\boldsymbol{X} \in \{0, 1, \cdots\}^p$ *distribuye conjuntamente FPoisson-pGM*$(\boldsymbol{\eta}, \boldsymbol{\Theta})$ *si para cada* $\boldsymbol{x} \in \{0, 1, \cdots\}^{p-1}$,

$$P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{\Theta}) \propto \exp\left(\boldsymbol{\eta}^T \boldsymbol{x} + \frac{1}{2} \boldsymbol{x}^\top \boldsymbol{\Theta} \boldsymbol{x} - \sum_{j=1}^{p} \log(x_j!)\right) I(\mathbf{1}^T \boldsymbol{x} = m_{\boldsymbol{x}}), \tag{2.11}$$

donde $\boldsymbol{\eta} \in \mathbb{R}^p$ *y* $\boldsymbol{\Theta}$ *es una matriz simétrica con diagonal nula de dimensión* $p \times p$.

Para resolver dicho problema, se propone un modelo reducido (Definición 2.5) obtenido luego de considerar la reparametrización $X_k = m_{\boldsymbol{X}} - \mathbf{1}^\top \boldsymbol{X}_{\setminus k}$. En la Proposición A.1 dada en el Apéndice A.3, identificamos los parámetros del modelo reducido y proponemos un modelo de regresión inversa identificable basado en (2.11).

Contribución original

Ejemplo 2.5 (FPoisson-pGM$_{\setminus k}$)**.** *Decimos que* $\boldsymbol{X}_{\setminus k} \mid m_{\boldsymbol{X}}$ *distribuye conjuntamente FPoisson-pGM*$_{\setminus k}(m_{\boldsymbol{x}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\Theta}})$ *si para cada* $\boldsymbol{x} \in \{0, 1, \cdots\}^{p-1}$,

$$P(\boldsymbol{X}_{\setminus k} = \boldsymbol{x}_{\setminus k} \mid m_{\boldsymbol{x}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\Theta}}) \propto \exp\Bigg(\tilde{\boldsymbol{\eta}}^T \boldsymbol{x}_{\setminus k} + \frac{1}{2} \boldsymbol{x}_{\setminus k}^T \tilde{\boldsymbol{\Theta}} \boldsymbol{x}_{\setminus k} - \sum_{j \neq k} \log(x_j!) - \log(m_{\boldsymbol{x}} - \mathbf{1}^T \boldsymbol{x}_{\setminus k})!\Bigg) I(\mathbf{1}^T \boldsymbol{x}_{\setminus k} \leq m_{\boldsymbol{x}}), \tag{2.12}$$

donde $\tilde{\boldsymbol{\eta}} \in \mathbb{R}^{p-1}$ *y* $\tilde{\boldsymbol{\Theta}}$ *es una matriz simétrica de dimensión* $p - 1 \times p - 1$.

Además, las condicionales $X_j \mid \boldsymbol{X}_{\setminus j}, m_{\boldsymbol{x}}$ *también pertenecen a la familia FPoisson-pGM*$_{\setminus k}$ *(Proposición A.2).*

Además, el Corolario A.1 permite definir un modelo de regresión (ver Definición 2.12 presentada más adelante) para $\boldsymbol{X} \mid Y \sim$*FPoisson-pGM (Definición 2.8) y encontrar el modelo reducido FPoisson-pGM*$_{\setminus k}$ *equivalente.*

Ejemplo 2.6 (sqPoisson-pGM). *Inouye, Ravikumar y Dhillon (2016) propusieron un modelo basado en el Poisson-pGM (Ejemplo 2.3) que permite modelar interacciones generales. Para ello consideraron la distribución correspondiente a (2.3) sobre $\boldsymbol{X} \in (\mathbb{N} \cup \{0\})^p$ con estadístico suficiente $\boldsymbol{T}(\boldsymbol{x}) = \sqrt{\boldsymbol{x}}$ y la misma medida base que el modelo Poisson-pGM $h(\boldsymbol{x}) = -\sum_{i=1}^{p} \log(x_j!)$. De este modo asegura la sumabilidad de la función de partición $Z(\boldsymbol{\eta}, \boldsymbol{\Theta})$ aún cuando las interacciones Θ_{jl} toman valores positivos y donde la diagonal de $\boldsymbol{\Theta}$ puede tomar valores positivos.*

Las distribuciónes condicionales de este modelo están dadas por (Proposición 2.3)

$$P(X_j = x_j \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{x}_{\setminus j}, \boldsymbol{\eta}, \Theta) \propto \exp\left\{ \Theta_{jj} x_j + \left(\eta_j + 2\boldsymbol{\Theta}_{j\setminus j}\sqrt{\boldsymbol{x}_{\setminus j}}\right)\sqrt{x_j} - \log(x_j!) \right\}, \quad (2.13)$$

donde $\boldsymbol{\Theta}_{j\setminus j}$ corresponde a la j-ésima fila de $\boldsymbol{\Theta}$ luego de haberle quitado su j-ésima columna. En el caso independiente, y cuando $\eta_j = 0$, las condicionales (2.13) se reducen a la distribución Poisson.

2.4 CONTRIBUCIÓN: MODELOS GRÁFICOS DE SEGUNDO ORDEN PARA EXCESO DE CEROS

Cuando los datos contienen muchos ceros pueden ser muy mal especificados por los modelos de conteo usuales, como el Poisson. En el caso de distribuciones continuas como la Normal, el efecto es incluso más notable, ya que en estos la probabilidad puntual es nula. Para remediar este problema se definen distribuciones que permitan mayor grado de acumulación en el origen. En particular consideramos el modelo Hurdle, que agrega una masa de probabilidad en el origen y puede definirse como la mezcla entre la variable idénticamente nula y una variable condicional a no ser cero.

Para definirlo, consideramos un nuevo estadístico suficiente que resulta de concatenar el estadístico suficiente $T(x)$ con la indicadora $\nu(x)$ que toma el valor cero cuando $x = 0$ y el valor unitario en otro caso. De igual modo, en el caso multivariado consideramos el estadístico suficiente $\boldsymbol{T}(\boldsymbol{x}) \in \mathbb{R}^p$ ampliado por $\boldsymbol{\nu}(\cdot)$ que aplica $\nu(\cdot) \in \{0,1\}^p$, cuyas coordenadas están dadas por $\nu(x_j)$.

Contribución original

Definición 2.9 (Hurdle univariado). *Decimos que $X \in \mathcal{X}_0$ se distribuye Hurdle con estadístico suficiente $\boldsymbol{t}(x)$ si*

$$P(X = x|\boldsymbol{\theta}, \tau) = \exp\{\boldsymbol{\theta}^\top \boldsymbol{t}(x) + \nu(x)[\tau + \tau_0 - Z^+(\boldsymbol{\theta})] + h(x) - Z(\tau, \tau_0)\}. \quad (2.14)$$

donde $Z^+(\boldsymbol{\theta}) = \log \sum_{x \neq 0} \exp\{\boldsymbol{\theta}^\top \boldsymbol{t}(x) + h(x)\}$ es el logaritmo de la función de partición condicionado a que $X \neq 0$, $\tau_0 = \boldsymbol{\theta}^\top \boldsymbol{t}(0) + h(0)$, $Z(\tau, \tau_0) = \tau_0 + Z(\tau)$, y $Z(\tau) = \log\{1 + \exp(\tau)\}$.

Observación 2.3. *Notar que si $h(0) = 0$ y $\boldsymbol{t}(0) = \boldsymbol{0}$ en (2.14), entonces $\tau_0 = 0$ y $Z(\tau, \tau_0) = Z(\tau)$.*

La función de probabilidad (2.14) admite una factorización como el producto de una v.a. Bernoulli ν y una condicional $\boldsymbol{X} \mid \nu$ en la familia exponencial (Proposición 2.4 a continuación), lo que facilita el muestreo o generación de datos.

Muestreo

Proposición 2.4 (Factorización). *Si X se distribuye de acuerdo con (2.14), entonces $\nu(X)$ distribuye como una v.a. Bernoulli de parámetro $p = 1/\{1 + \exp(-\tau)\}$. Además, la distribución condicional de X dado $\nu(X)$ es*

$$P(X = x \mid \nu(x), \boldsymbol{\theta}) = \delta_{\nu,0}\delta_{x,0} + \delta_{\nu,1} \exp\{\boldsymbol{\theta}^T \boldsymbol{t}(x) + h(x) - Z^+(\boldsymbol{\theta})\}, \quad (2.15)$$

donde $\delta_{x,y} = I(x = y)$ es la delta de Kronecker.

La demostración se encuentra en el Apéndice A.4. La generalización multivariada de (2.14) se presenta a continuación.

zipGM
Contribución original

Definición 2.10 (Hurdle multivariado alias zipGM). *Una distribución con función de probabilidad P es un MG de segundo orden inflado en cero (zipGM) si*

$$P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{\omega}) \propto \exp\{\boldsymbol{\eta}^\top \boldsymbol{T}(\boldsymbol{x}) + \boldsymbol{\xi}^\top \boldsymbol{\nu}(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{T}(\boldsymbol{x})^\top \boldsymbol{\Theta} \boldsymbol{T}(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{\nu}(\boldsymbol{x})^\top \boldsymbol{\Lambda} \boldsymbol{\nu}(\boldsymbol{x}) + \boldsymbol{\nu}(\boldsymbol{x})^\top \boldsymbol{\Phi} \boldsymbol{T}(\boldsymbol{x}) + \mathbf{1}^\top \boldsymbol{h}(\boldsymbol{x})\}, \quad (2.16)$$

donde $(\boldsymbol{\nu}(\boldsymbol{x}), \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\nu}(\boldsymbol{x})\boldsymbol{\nu}(\boldsymbol{x})^\top, \boldsymbol{\nu}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{x})^\top, \boldsymbol{T}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{x})^\top)$ es el estadístico suficiente, con $\boldsymbol{T}(\boldsymbol{x}) = (T(x_1), \ldots, T(x_p))^\top$ y $\boldsymbol{h}(\boldsymbol{X}) = (h(X_1), \cdots, h(X_p))^\top$ el vector de medidas base. Los parámetros naturales están dados por

$$\boldsymbol{\omega} = (\boldsymbol{\eta}, \boldsymbol{\xi}, \mathrm{vec}(\boldsymbol{\Lambda}), \mathrm{vec}(\boldsymbol{\Phi}), \mathrm{vec}(\boldsymbol{\Theta})) \in \mathbb{R}^{2p+3p^2}, \quad (2.17)$$

y el espacio de parámetros $\boldsymbol{\Omega}$ incorpora restricciones de simetría sobre las matrices $\boldsymbol{\Theta}$ y $\boldsymbol{\Lambda}$, mientras que $\boldsymbol{\Phi}$ es una matriz general.

Teorema 2.4 (Caracterización). *Suponiendo que la distribución conjunta de $\boldsymbol{X} = (X_1, \cdots, X_k)^\top$ es positiva y factoriza sobre los cliques de orden 2 de un grafo $G = (V, E)$ i. e. $P(\boldsymbol{X}|\boldsymbol{\omega}) = \prod_{i>j} G_{ij}(X_j, X_l)$, para funciones G_{jl} para todo $\boldsymbol{\omega} \in \boldsymbol{\Omega}$. Luego, $P(\boldsymbol{X}|\boldsymbol{\omega})$ tiene la forma (2.16) sii para todo $j = 1, \ldots, p$,*

$$P(X_j = x_j \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{x}_{\setminus j}, \boldsymbol{\omega}) = \exp\{\xi_{j|\setminus j}\nu(x_j) + \eta_{j|\setminus j}T(x_j) + \frac{1}{2}\Theta_{jj}T(x_j)^2 + h(x_j) - Z(\xi_{j|\setminus j}, \eta_{j|\setminus j}, \Theta_{jj})\}, \quad (2.18)$$

con estadístico suficiente $(\nu(x_j), T(x_j), T(x_j)^2/2)$, parámetros naturales $(\xi_{j|\setminus j}, \eta_{j|\setminus j}, \Theta_{jj})$, medida base $h(x_j)$ y función de partición

$$\exp Z(\xi_{j|\setminus j}, \eta_{j|\setminus j}, \Theta_{jj}) = 1 + \exp\left[\xi_{j|\setminus j} - \eta_{j|\setminus j}T(0) - \frac{1}{2}\Theta_{jj}T(0)^2 + Z^+(\eta_{j|\setminus j}, \Theta_{jj})\right], \quad (2.19)$$

donde $Z^+(\eta_{j|\setminus j}, \Theta_{jj}) = \log \sum_{x \neq 0} \eta_{j\setminus j}T(x) + \frac{1}{2}\Theta_{jj}T(x)^2 + h(x)$ es la función de partición del modelo dado que $X \neq 0$. Los parámetros condicionales $\xi_{j|\setminus j}$ y $\eta_{j|\setminus j}$ dependen de $\boldsymbol{X}_{\setminus j}$ a través de

$$\xi_{j|\setminus j} = \xi_j + \boldsymbol{\Lambda}_{j,\setminus j}\boldsymbol{\nu}_{\setminus j} + \boldsymbol{\Phi}_{j,\setminus j}\boldsymbol{T}(\boldsymbol{x}_{\setminus j}) \quad (2.20)$$

$$\eta_{j|\setminus j} = \eta_j + \boldsymbol{\nu}_{\setminus j}^\top \boldsymbol{\Phi}_{\setminus j,j} + \boldsymbol{\Theta}_{j,\setminus j}\boldsymbol{T}(\boldsymbol{x}_{\setminus j}), \quad (2.21)$$

donde $\mathbf{\Lambda}_{j,\setminus j}$ corresponde a la fila j de la matrix $\mathbf{\Lambda}$ luego de remover la j-ésima columna, la misma definición aplica a las otras matrices.

La prueba se ofrece en el Apéndice A.4.

Observación 2.4. *Esta caracterización corresponde a una generalización del Teorema 2.4, válido para distribuciones condicionales con estadístico suficiente escalar. En este caso consideramos distribuciones condicionales Hurdle (Definición 2.9) con $\boldsymbol{t}(x) = (T(x), T(x)^2)$, cuyo estadístico suficiente $(T(x), \nu(x))$ es bidimensional. Notar que en este caso, la distribución condicional $\nu(X_j) \mid \boldsymbol{X}_{\setminus j}$ es Bernoulli con*

$$P(\nu(X_j) = 1 \mid \boldsymbol{X}_{\setminus j}, \boldsymbol{\omega}) = \left(1 + \exp\{-\xi_{j|\setminus j} - Z^+(\eta_{j|\setminus j}, \Theta_{jj})\}\right)^{-1}, \quad (2.22)$$

donde $Z^+(\eta_{j|\setminus j}, \Theta_{jj})$ es la función de partición del modelo condicional $X_j \mid \nu(X_j) = 1, \boldsymbol{X}_{\setminus j}$ y $\xi_{j|\setminus j}, \eta_{j|\setminus j}$ son los parámetros condicionales dadas en (2.20) y (2.21) resp.

El Teorema 2.4 permite definir modelos zipGM (2.16) a partir de las distribuciones condicionales $X_j \mid \boldsymbol{X}_{\setminus j}$ dadas en (2.18). Equivalentemente dichas distribuciones condicionales quedan determinadas por $\nu(X_j) \mid \boldsymbol{X}_{\setminus j}$ dada en 2.22 y por $X_j \mid \nu(X_j) = 1, \boldsymbol{X}_{\setminus j}$ (Proposición 2.4). En lo que sigue, especificando esta distribución condicional mostramos ejemplos en la familia zipGM (2.16) que luego estudiaremos en mayor profundidad.

Ejemplo 2.7 (Normal-zipGM). *Decimos que una v.a. X se distribuye de acuerdo con el modelo Normal-zipGM si tiene una densidad dada por (2.16) y cuyas distribuciones condicionales corresponden a (2.18) en el Teorema 2.4 con $T(x) = x$, $\Theta_{jj} < 0$ y $h(x) = 0$ sobre un espacio muestral continuo $\boldsymbol{X} \in \mathbb{R}^p$. Equivalentemente, el modelo queda caracterizado por las condicionales*

$$(X_j \mid \nu(X_j) = 1, \boldsymbol{X}_{\setminus j}) \sim \mathcal{N}(-\Theta_{jj}^{-1}\eta_{j|\setminus j}, -\Theta_{jj}^{-1}), \quad (2.23)$$

y $\nu(X_j) \mid \boldsymbol{X}_{\setminus j}$ en 2.22 donde la función Z^+ está dada por

$$Z^+(\eta_{j|\setminus j}, \Theta_{jj}) = -\eta_{j|\setminus j}^2/2\Theta_{jj} - 1/2\log(-\Theta_{jj}) + 1/2\log(2\pi). \quad (2.24)$$

De este modo recuperamos el modelo Hurdle-normal (McDavid et al., 2019).

Contribución original

Ejemplo 2.8 (Poisson-zipGM). *Decimos que una v.a. X se distribuye de acuerdo con el modelo Poisson-zipGM si su función de probabilidad satisface (2.16) y cuyas distribuciones condicionales corresponden a (2.18) en el Teorema 2.4 con $T(x) = x$, $\theta = 0$ y $h(x) = -\log(x!)$ sobre un espacio muestral contable $\boldsymbol{X} \in (\mathbb{N} \cup \{0\})^p$. Equivalententemente, la distribución condicional de $X_j \mid \nu(\boldsymbol{X}_j) = 1, \boldsymbol{X}_{\setminus j}$ es Poisson truncada en cero con parámetros naturales $\eta_{j|\setminus j}$ y soporte en los enteros:*

$$(X_j \mid \nu(X_j) = 1, \boldsymbol{X}_{\setminus j}) \sim \textit{Poisson}^+\{\exp(\eta_{j|\setminus j})\}. \quad (2.25)$$

Mientras que $\nu(X_j) \mid \boldsymbol{X}_{\setminus j}$ está dada en 2.22 donde el logaritmo de la función de partición Z^+ resulta

$$Z^+(\eta_{j|\setminus j}) = \log\left(\sum_{t=1}^{\infty} \exp(\eta_{j|\setminus j} t)/t!\right) = \log[\exp\{\exp(\eta_{j|\setminus j})\} - 1].$$

Para evitar problemas numéricos cuando $\eta_{j|\setminus j}$ es grande, usamos la formulación equivalente

$$\begin{aligned} Z^+(\eta_{j|\setminus j}) &= \exp(\eta_{j|\setminus j}) + \log\{P(\tilde{X} \geq 1)\} \\ &= \exp(\eta_{j|\setminus j}) + \log[1 - \exp\{-\exp(\eta_{j|\setminus j})\}], \end{aligned} \tag{2.26}$$

donde definimos una variable auxiliar $\tilde{X}$ que distribuye como una Poisson con parámetro $\exp(\eta_{j|\setminus j})$.

Notar que al igual que en el Ejemplo 2.3, es necesario incorporar la restricción $\boldsymbol{\Theta} \leq 0$ al espacio de parámetros para asegurar que la función de partición del modelo (2.16) sea finita.

Ejemplo 2.9 (TPoisson-zipGM)**.** *Restringiendo el espacio muestral $\boldsymbol{X} \in \{x \in \mathbb{N} \cup \{0\} : x \leq T^*\}^p$, decimos que $\boldsymbol{X}$ distribuye de acuerdo con el modelo TPoisson-zipGM si satisface (2.16) y tiene distribuciones condicionales que corresponden a (2.18) en el Teorema 2.4 con $T(x) = x$, $\theta = 0$ y $h(x) = -\log(x!)$.*

Equivalententemente, la distribución condicional de $X_j \mid \nu(\boldsymbol{X}_j) = 1, \boldsymbol{X}_{\setminus j}$ es Poisson truncada al intervalo $\{x : 1 \leq x \leq T^*\}$, y su función de partición condicional resulta

$$\begin{aligned} Z^+(\eta_{j|\setminus j}) &= \log\left\{\sum_{t=1}^{T^*} \frac{\exp(\eta_{j|\setminus j} t)}{t!}\right\} \\ &= \log\left\{\exp\{\exp(\eta_{j|\setminus j})\} - \sum_{t=T^*+1}^{\infty} \frac{\exp(\eta_{j|\setminus j} t)}{t!} - 1\right\} \\ &= \exp(\eta_{j|\setminus j}) + \log\{P(1 \leq \tilde{X} \leq T^*)\}, \end{aligned} \tag{2.27}$$

donde definimos una variable auxiliar $\tilde{X}$ que distribuye como una Poisson con parámetro $\exp(\eta_{j|\setminus j})$.

2.5 MODELOS CONDICIONALES

En el contexto de aprendizaje supervisado, el conjunto de aprendizaje está compuesto de pares $(\boldsymbol{x}, y)$ y en predicción se intenta predecir y luego de haber medido $\boldsymbol{x}$. Reducción suficiente de dimensiones, como veremos en Capítulo 3, permite estimar una transformación de los datos hacia una representación de menor dimensión $R(\boldsymbol{X})$ que conserva toda la información de la respuesta Y que se encuentra en los predictores $\boldsymbol{X}$. Dicha reducción puede ser utilizada para predecir la respuesta Y habiendo medido $\boldsymbol{X}$. Es posible obtener dicha reducción a partir del modelo condicional $\boldsymbol{X} \mid Y$ de los predictores dada la respuesta (Cook, 2007), ver también (Li, 1991). En particular, dado y, consideramos $\boldsymbol{X}$ en la familia exponencial (Definición 2.1) donde los parámetros naturales $\boldsymbol{\omega}$ dependen linealmente de una función conocida de Y que llamaremos $\boldsymbol{f}(Y) \in \mathbb{R}^r$, lo que resulta en la que llamamos FE condicional lineal (Bura, Duarte y Forzani, 2016):

Definición 2.11 (Familia exponencial condicional lineal)**.** *Dado $Y = y$, decimos que $\boldsymbol{X}$ sigue una familia exponencial condicional lineal si $P(\boldsymbol{X} = \boldsymbol{x} \mid Y = y)$ corresponde a (2.1) con $\omega(y) = \omega_0 + \boldsymbol{\Gamma}_{\boldsymbol{\omega}} \boldsymbol{f}(y)$ con $\boldsymbol{\Gamma}_{\boldsymbol{\omega}} \in \mathbb{R}^{m \times r}$.*

Un caso particular que estudiaremos en profundidad consiste en definir un modelo gráfico de segundo orden, (Definición 2.5), para $\boldsymbol{X} \mid Y$, donde únicamente los parámetros naturales vinculados al estadístico suficiente $\boldsymbol{T}(\boldsymbol{X})$ en (2.3) dependan de la covariable, es decir:

Definición 2.12 (Modelo gráfico de segundo orden condicional lineal en $\boldsymbol{T}(\boldsymbol{X})$)**.** *Dado $Y = y$, decimos que $\boldsymbol{X}$ sigue un modelo gráfico de segundo orden condicional lineal en $\boldsymbol{T}(\boldsymbol{X})$ si $P(\boldsymbol{X} = \boldsymbol{x} \mid Y = y)$ corresponde a* (2.3) *con*

$$\boldsymbol{\eta}(y) = \boldsymbol{\eta}_0 + \boldsymbol{\Gamma}_{\boldsymbol{\eta}} \boldsymbol{f}(y) \tag{2.28}$$

y donde $\boldsymbol{\Gamma}_{\boldsymbol{\eta}} \in \mathbb{R}^{p \times r}$.

3

REDUCCIÓN SUFICIENTE DE DIMENSIONES

3.1 INTRODUCCIÓN

En el contexto de regresión y clasificación, estamos interesados en predecir una respuesta Y que puede ser continua, categórica o incluso multivariada a partir de un conjunto de mediciones $\boldsymbol{X} \in \mathbb{R}^p$. La distribución condicional evaluada en una muestra $\boldsymbol{x}$

$$Y \mid (\boldsymbol{X} = \boldsymbol{x}), \tag{3.1}$$

contiene la información predictiva sobre la variable de interés Y. Cuando el número de mediciones p crece, se vuelve más difícil encontrar modelos para Y como función de $\boldsymbol{x}$ en (3.1), perdiendo interpretabilidad. El objetivo de Reducción Suficiente de Dimensiones es encontrar una representación $\boldsymbol{R}(\boldsymbol{X}) \in \mathbb{R}^d$ de menor dimensión ($d \ll p$) de los predictores $\boldsymbol{X}$ que conserve *toda* la información contenida en $\boldsymbol{X}$ necesaria para predecir la respuesta Y.

Definición 3.1 (RSD). *$\boldsymbol{R} : \mathbb{R}^p \to \mathbb{R}^q$ es una Reducción Suficiente de Dimensiones (RSD) para Y basada en $\boldsymbol{X}$ si $F(Y \mid \boldsymbol{X}) = F(Y \mid \boldsymbol{R}(\boldsymbol{X}))$, donde $F(\cdot \mid \cdot)$ es la función de distribución condicional.*

El trabajo fundacional (Cook, 2007) establece las siguientes equivalencias:

Teorema 3.1 (RSDs equivalentes). *Las siguientes condiciones son equivalentes para una RSD $\boldsymbol{R}(\boldsymbol{X})$:*

$$Y \mid \boldsymbol{X} \overset{d}{=} Y \mid \boldsymbol{R}(\boldsymbol{X}), \qquad \textit{(Reducción directa)} \tag{3.2}$$

$$\boldsymbol{X} \mid (Y, \boldsymbol{R}(\boldsymbol{X})) \overset{d}{=} X \mid \boldsymbol{R}(\boldsymbol{X}), \qquad \textit{(Reducción inversa)} \tag{3.3}$$

$$(\boldsymbol{X} \perp\!\!\!\perp Y) \mid \boldsymbol{R}(\boldsymbol{X}). \qquad \textit{(Reducción conjunta)} \tag{3.4}$$

La condición (3.4) implica una relación Markoviana de independencia condicional y será representada más adelante por el diagrama de la Figura 5.1a. La condición (3.3) permite simplificar el problema de regresión multivariada (3.1) permitiendo modelar p regresiones univariadas $\boldsymbol{X}_j \mid Y,\ j = 1, \cdots, p$. En particular, el Teorema de Factorización derivado de (Casella y Berger, 2002) para modelos en la FE enunciado a continuación permite encontrar RSDs a partir del modelo condicional de $\boldsymbol{X} \mid Y$:

Teorema 3.2 (Tomassi et al. (2019)). *Considerando que el modelo condicional $\boldsymbol{X} \mid Y$ se distribuye de acuerdo con $P(\boldsymbol{X} \mid Y = y)$. Una reducción $\boldsymbol{R} : \mathbb{R}^p \to \mathbb{R}^d$ es una RSD para la regresión de Y en $\boldsymbol{X}$ si y sólo si existen funciones $g(\boldsymbol{t}, y)$ y $h(\boldsymbol{x})$ tales que*

$$P(\boldsymbol{X} = \boldsymbol{x} \mid Y = y) = g(\boldsymbol{R}(\boldsymbol{x}), y)h(\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \mathcal{X},\ y \in \mathcal{Y}, \tag{3.5}$$

MINIMALIDAD Vimos que una RSD $R(\boldsymbol{X}) \in \mathbb{R}^q$ es transformación de los predictores que conserva la capacidad predictiva respecto de una respuesta Y. El concepto de *minimalidad* está relacionado con la dimensión de la reducción q, en particular nos interesa obtener la máxima reducción de los datos, dando lugar a la siguiente definición:

Definición 3.2 (RSD minimal)**.** *Una RSD $\boldsymbol{R}^* : \mathbb{R}^p \to \mathbb{R}^d$ se dice minimal si para cualquier otra RSD $\boldsymbol{R} : \mathbb{R}^p \to \mathbb{R}^q$ existe una función $\boldsymbol{S} : \mathbb{R}^q \to \mathbb{R}^d$ tal que*

$$\boldsymbol{R}^*(\boldsymbol{X}) = \boldsymbol{S}(\boldsymbol{R}(\boldsymbol{X})). \tag{3.6}$$

El Teorema siguiente basado en el Teorema Lehmann-Scheffé (Casella y Berger, 2002), permite caracterizar las RSD minimales:

Teorema 3.3 (Bura, Duarte y Forzani (2016))**.** *Suponiendo que $\boldsymbol{X} \mid Y$ se distribuye de acuerdo a $P(\boldsymbol{X} \mid Y = y)$, una RSD $\boldsymbol{R}(\boldsymbol{X})$ para la regresión de Y en $\boldsymbol{X}$ es minimal si satisface*

$$\frac{P(\boldsymbol{X} = \boldsymbol{x}_1 \mid Y = y)}{P(\boldsymbol{X} = \boldsymbol{x}_2 \mid Y = y)} = c(\boldsymbol{x}_1, \boldsymbol{x}_2) \iff \boldsymbol{R}(\boldsymbol{x}_1) = \boldsymbol{R}(\boldsymbol{x}_2), \tag{3.7}$$

para todo $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ y donde $c(\boldsymbol{x}_1, \boldsymbol{x}_2)$ no depende de y.

En la sección 3.2 estudiamos cómo obtener RSDs minimales cuando $\boldsymbol{X} \mid Y$ se encuentra en la FE y en particular encontramos RSDs para las familias de MG de segundo orden (pGM) y MG de segundo orden para exceso de ceros (zipGM) (Definiciones 2.5 y 2.10 respectivamente).

3.2 REDUCCIÓN SUFICIENTE DE DIMENSIONES EN LA FAMILIA EXPONENCIAL

En el caso de FE, la siguiente proposición es consecuencia del Teorema 3.3 y se observa que en este caso, la RSD minimal es lineal en el estadístico suficiente $\boldsymbol{T}(\boldsymbol{X})$.

Proposición 3.1 (RSD para familias exponenciales (Bura, Duarte y Forzani, 2016))**.** *Suponiendo que $\boldsymbol{X} \mid Y$ se distribuye de acuerdo con una FE condicional lineal (Definición 2.11), una RSD minimal $R(\boldsymbol{X})$ para la regresión de Y en $\boldsymbol{X}$ está dada por*

$$R(\boldsymbol{X}) = \boldsymbol{\alpha}^\top \boldsymbol{T}(\boldsymbol{X}), \tag{3.8}$$

donde $\boldsymbol{\alpha} \in \mathbb{R}^{m \times r}$ es una base de $\operatorname{span}\{\boldsymbol{\omega}(y) - \boldsymbol{\omega}_0, y \in \mathcal{Y}\} = \operatorname{span}\{\boldsymbol{\Gamma}_\omega\}$, con $\mathcal{Y}$ el espacio muestral de Y.

La demostración se encuentra en Bura, Duarte y Forzani (2016) y se reproduce en el Apéndice B por completitud. La siguiente proposición especifica la RSD minimal para modelos gráficos de segundo orden:

Corolario 3.1. *Una RSD para modelos gráficos condicionales de segundo orden condicional lineal en $\boldsymbol{T}(\boldsymbol{X})$ (Definición 2.12) está dada por*

$$\boldsymbol{R}(\boldsymbol{X}) = \boldsymbol{\alpha}^T T(\boldsymbol{X}), \tag{3.9}$$

donde $\boldsymbol{\alpha} \in \mathbb{R}^{p \times d}$ es una base de $\operatorname{span}\{\boldsymbol{\eta}_y - \boldsymbol{\eta}_0, y \in \mathcal{Y}\} = \operatorname{span}\{\boldsymbol{\Gamma}\}$.

En el caso de los modelos zipGM (Definición 2.10) consideramos la Definición 2.12 con $T(\boldsymbol{X}) = (\boldsymbol{X}, \boldsymbol{\nu}(\boldsymbol{X}))$ y potenciales lineales dados por

$$\boldsymbol{\eta}(y) = \boldsymbol{\eta}_0 + \boldsymbol{\Gamma} \boldsymbol{f}(y), \tag{3.10a}$$
$$\boldsymbol{\xi}(y) = \boldsymbol{\xi}_0 + \Psi \boldsymbol{f}(y). \tag{3.10b}$$

Con esto, es posible definir dos reducciones (ver demostraciones en el Apéndice B):

Contribución original

Corolario 3.2 (RSD conjunta)**.** *Asumiendo que $\boldsymbol{X}|Y$ se distribuye de acuerdo con (2.16), con potenciales (3.10a) y (3.10b). Si la matriz de dimensión $2p \times r$ resultante de la concatenación $(\boldsymbol{\Gamma}; \boldsymbol{\Psi})$, obtenida apilando las columnas de $\boldsymbol{\Gamma}$ en (3.10b) y $\boldsymbol{\Psi}$ en (3.10a), tiene rango $d_J \leq \min\{r, 2p\}$, luego una RSD para $Y|\boldsymbol{X}$ de dimensión d_J está dada por*

$$\boldsymbol{R}(\boldsymbol{X}) = \boldsymbol{\kappa}^T \begin{bmatrix} T(\boldsymbol{X}) \\ \nu(\boldsymbol{X}) \end{bmatrix} \in \mathbb{R}^{d_J}, \tag{3.11}$$

donde $\boldsymbol{\kappa}$ es una base de $\operatorname{span}\{(\boldsymbol{\eta}_y - \boldsymbol{\eta}_0; \boldsymbol{\xi}_y - \boldsymbol{\xi}_0),\ y \in \mathcal{Y}\} = \operatorname{span}\{(\boldsymbol{\Gamma}; \boldsymbol{\Psi})\}$.

Corolario 3.3 (RSD separada)**.** *Asumiendo las mismas condiciones que en la Proposición 3.2, y considerando $d_X = \operatorname{rango}(\boldsymbol{\Gamma}) \leq \min\{r, k\}$ y $d_\nu = \operatorname{rango}(\boldsymbol{\Psi}) \leq \min\{r, k\}$. Una RSD para $Y|\boldsymbol{X}$ de dimensión $d_X + d_\nu$ está dada por*

$$\boldsymbol{R}_s(\boldsymbol{X}) = \begin{bmatrix} \boldsymbol{\alpha} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\zeta} \end{bmatrix}^T \begin{bmatrix} T(\boldsymbol{X}) \\ \nu(\boldsymbol{X}) \end{bmatrix} \in \mathbb{R}^{d_X + d_\nu}, \tag{3.12}$$

donde $\boldsymbol{\alpha}$ es una base de $\operatorname{span}\{\boldsymbol{\eta}_y - \boldsymbol{\eta}_0,\ y \in \mathcal{Y}\} = \operatorname{span}\{\boldsymbol{\Gamma}\}$ *y $\boldsymbol{\zeta}$ es una base de* $\operatorname{span}\{\boldsymbol{\xi}_y - \boldsymbol{\xi}_0,\ y \in \mathcal{Y}\} = \operatorname{span}\{\boldsymbol{\Psi}\}$.

La estimación de la RSD *separada* utilizando la descomposición en valores singulares para el cálculo del span $\operatorname{diag}(\boldsymbol{\alpha}, \boldsymbol{\zeta})$ en (3.12) es más estable numéricamente, en especial cuando la varianza de $\boldsymbol{\Gamma}^\top \boldsymbol{X}$ y la de $\boldsymbol{\Psi}^\top \boldsymbol{\nu}(\boldsymbol{X})$ tienen magnitudes muy distintas, mientras que la *conjunta* en (3.11) obtiene la mayor reducción, ya que $\operatorname{rango}(\boldsymbol{\Gamma}; \boldsymbol{\Psi}) \leq \operatorname{rango}(\boldsymbol{\Gamma}) + \operatorname{rango}(\boldsymbol{\Psi})$.

3.2.1 *Selección de una base de funciones para modelar la regresión inversa*

En (Cook, 1998) se establece la metodología para obtener reducciones suficientes a partir de la reducción inversa (ver Teorema 3.1). En particular, en la selección de $\boldsymbol{f} : \mathcal{Y} \to \mathbb{R}^r$, se busca que cada parámetro natural sea lineal en $\boldsymbol{f}(y)$, de manera de poder caracterizar un subespacio de reducción lineal.

En general, para respuestas Y continuas, se suele considerar polinomios centrados de Y, donde la estimación de la matrices de regresión Γ (y Ψ en los modelos zipGM) corresponden a los coeficientes de dichos polinomios que modulan a los potenciales lineales correspondientes a cada variable. En cambio, en casos de respuesta discreta $Y \in \{0, \cdots, r\}$, es posible definir $\boldsymbol{f} : \mathcal{Y} \to \mathbb{R}^r$ de manera que

$f_j(y) = \delta_{y,j} - \hat{p}_j$ para todo $j = 1, \cdots, r$, donde $\delta_{y,j} = I(y = j)$ es la delta de Kronecker y $\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} \delta_{y^{(i)},j}$, es la probabilidad muestral de observar la respuesta j-ésima. De esta manera se evita incorporar restricciones adicionales sobre las matrices de regresión, manteniendo al problema identificable. El caso binario $Y \in \{0, 1\}$ resulta un caso particular del caso discreto donde $r = 1$ y $f(y)$ se simplifica en $y - \hat{E}Y$.

3.3 ESTUDIO BIBLIOGRÁFICO

En el campo de la aplicación al análisis de datos de microbioma que motiva esta tesis, los métodos de reducción dimensional más usados son no supervisados. En particular, es frecuente encontrar versiones de escalado multidimensional que utilizan algún criterio relevante desde el punto de vista ecológico para ponderar la similitud entre dos composiciones diferentes (conocidos comúnmente como índices de β-diversidad). El uso de métodos supervisados de reducción dimensional es más limitado, aunque suelen encontrarse aplicaciones de variantes sparse de cuadrados míñimos parciales (sparse PLS), tanto con respuesta continua como categórica (Rohart et al., 2017). El uso de modelos gráficos como estrategia de reducción dimensional y visualización para datos de microbioma ha sido poco explorado, principalmente por dificultades en su estimación para escenarios de valor práctico. Tomassi et al. (2019) propone un enfoque de RSD basado en modelos gráficos de tipo Poisson, aprovechando la semejanza formal entre datos de microbioma y datos de tipo texto. Esta relación permite aprovechar y extender modelos propuestos para la descripción de tópicos en documentos de texto (Inouye et al., 2017). No obstante, la estimación de la reducción propuesta en Tomassi et al. (2019) se basa en una versión penalizada de la función de verosimilitud y no escala adecuadamente con el número de variables presentes en los datos experimentales más recientes. Tampoco se incluye ninguna adaptación particular para el exceso de ceros. Más recientemente se ha propuesto un enfoque basado en regresión inversa que considera posibles dependencias de la composición con otros factores o respuestas, como así también factores adicionales que ayudan a modelar el exceso de ceros en forma indirecta (Pang, Zhao y Wang, 2023). Aunque esta estrategia permite conservar cierta información estructural que puede perderse en una RSD común, no permite una interpretación directa de las reducciones obtenidas en términos de patrones de abundancia o de co-ocurrencia de las distintas especies. Otros modelos estadísticos multivariados para describir perfiles de composición completos frecuentemente recurren a un enfoque jerárquico, en el que una capa de variables latentes modela la probabilidad de observar cada especie y las correlaciones entre ellas, mientras que una capa externa de tipo multinomial o Poisson modela los datos de conteo observables a partir de tales variables latentes (por ejemplo, Chen y Li, 2013; Xia et al., 2013). Ejemplos recientes de esta estrategia se encuentran, por ejemplo, en Chiquet, Mariadassou y Robin (2021). No obstante, tales modelos jerárquicos no pertenecen a la FE y no conducen a una RSD.

Parte III

ESTIMACIÓN

En los siguientes capítulos estudiaremos distintos métodos que permiten estimar los parámetros de los modelos presentados en la Parte II, tanto en baja como en alta dimensión. Los estimadores que presentaremos están basados en divergencias. Una *divergencia* entre distribuciones representa el grado de separación entre ellas. Es posible *aprender* la distribución de los datos al minimizar una divergencia entre la distribución empírica y el modelo. Además, cada divergencia define una métrica de Riemman en el espacio de distribuciones, induciendo una métrica euclídea local en el espacio de parámetros. En el Capítulo 5 veremos cómo esta métrica nos ayuda a definir penalizaciones óptimas y en el Capítulo 6 veremos cómo es usada por algunos algoritmos de aprendizaje.

4

DIVERGENCIAS Y SCORES

En el Capítulo 2 mostramos ejemplos de familias de distribuciones $f(\boldsymbol{\omega})$ en la FE, donde cada $\boldsymbol{\omega}$ en el espacio de parámetros Ω define una distribución en la familia exponencial. En la definición de divergencia entre distribuciones que introducimos a continuación consideramos distribuciones f y g en la FE, con coordenadas $\boldsymbol{\omega}_f$ y $\boldsymbol{\omega}_g$ en el espacio de parámetros Ω.

Definición 4.1 (Divergencia)**.** *Una divergencia (Amari, 2016) entre dos distribuciones pertenecientes a FE, $D[f : g]$, es una función diferenciable de $\boldsymbol{\omega}_f$ y $\boldsymbol{\omega}_g$ que satisface las siguientes propiedades:*

1. $D[f : g] \geq 0$
2. $D[f : g] = 0$ *si y solo si* $f = g$
3. *En un entorno de* $\boldsymbol{\omega}$,

$$D[f(\boldsymbol{\omega}) : f(\boldsymbol{\omega} + d\boldsymbol{\omega})] = \frac{1}{2} d\boldsymbol{\omega}^\top \mathcal{I}(\boldsymbol{\omega}) d\boldsymbol{\omega} + O(\|d\boldsymbol{\omega}\|^3),$$

donde $\mathcal{I} : \Omega \to \mathbb{R}^{m \times m}$ es definida positiva.

En un espacio de Riemman, la matriz $\mathcal{I}$ define una métrica local dada por

$$ds^2 := 2D[f(\boldsymbol{\omega}) : f(\boldsymbol{\omega} + d\boldsymbol{\omega})] = d\boldsymbol{\omega}^\top \mathcal{I}(\boldsymbol{\omega}) d\boldsymbol{\omega} =: \|d\boldsymbol{\omega}\|^2_{\mathcal{I}}. \tag{4.1}$$

En lo que sigue llamaremos *métrica local* a la matriz $\mathcal{I}$.

Dada una muestra aleatoria $\{\boldsymbol{X}^{(i)}\}_{i=1}^n$, $\boldsymbol{X}^{(i)} \in \mathcal{X}$, con distribución empírica $\hat{f} \in \mathcal{P}$, y una divergencia D, definimos al estimador $\hat{g}$ de g como el minimizante de la divergencia D entre la distribución empírica y el modelo g:

$$\hat{g} = \arg\min_{g \in \mathcal{P}} D[\hat{f} : g]. \tag{4.2}$$

Las divergencias D que satisfacen

$$D[\hat{f} : g] \propto E_{\boldsymbol{X} \sim \hat{f}} S(\boldsymbol{X}, g(\boldsymbol{\omega})), \tag{4.3}$$

para algún $S : \mathcal{X} \times \mathcal{P} \to \mathbb{R}$ definen un *score* $S(\cdot, \cdot)$, una función de costo que cuantifica la exactitud con la que una distribución $g \in \mathcal{P}$ de la variable aleatoria $\boldsymbol{X}$ explica la observación $\boldsymbol{x}$. Con esto, podemos definir al estimador como el minimizante del promedio del score dados los datos de entrenamiento:

$$\hat{g} = \arg\min_{\boldsymbol{\omega} \in \Omega} \frac{1}{n} \sum_{i=1}^n S(\boldsymbol{x}^{(i)}, g(\boldsymbol{\omega})) =: \hat{E} S(\boldsymbol{X}, g(\boldsymbol{\omega})). \tag{4.4}$$

Observación 4.1. *Notar que la expresión (4.4), a diferencia de (4.2), está bien definida incluso cuando la distribución empírica $\hat{f} \notin \mathcal{P}$ y dicha minimización se interpreta como la proyección de la distribución empírica al espacio de distribuciones $\mathcal{P}$ (Dawid, 1979).*

Cuando el modelo está bien especificado y contamos con mediciones independientes, es posible inducir estimadores insesgados (Sección 4.3).

En la Sección 4.1 definiremos la Divergencia de Kullback-Leibler (KL) y veremos que induce la métrica local dada por la matriz de información de Fisher. Además, su minimizante con respecto a la distribución empírica de los datos define al Estimador de Máxima Verosimilitud (MLE). También discutiremos algunas de sus principales características y dificultades en su aplicación. Seguidamente, en la Sección 4.2, estudiaremos alternativas a la KL. En particular, veremos que definiendo la divergencia de KL sobre eventos marginales o condicionales, se define una divergencia de composición, cuya minimización con respecto a una distribución empírica tiene como caso particular a los estimadores pseudolikelihood, los cuales permiten simplificar su aplicación a modelos multivariados.

En la Sección 4.3 estudiaremos distintas formas de definir *reglas de score*, brindando mayor flexibilidad a la hora de definir una divergencia, las cuales podrían permitir mayor control entre robustez, eficiencia y costo computacional. En la Sección 4.4 consideramos la divergencia entre modelos condicionales (Definición 2.12 por ejemplo) y en 4.5 mostramos una divergencia para modelos diferenciales, lo que permite estimar la RSD (Proposición 3.1) sin necesidad de estimar las interacciones cuando la respuesta es binaria.

En la Sección 4.6 definimos las ecuaciones de estimación inducidas por el problema (4.4), caracterizando un *m-estimador*, de esto se obtiene su distribución asintótica y pruebas para determinar la dimensión de la Reducción Suficiente de Dimensiones (RSD).

Haciendo uso de la propiedad 3 de la Definición 4.1, veremos en la Sección 4.7 que las divergencias traducen localmente distancias entre distribuciones $f(\boldsymbol{\omega})$ a distancias entre los parámetros $\boldsymbol{\omega}$. En particular considerando el mapeo

$$\boldsymbol{\omega} \to f(\boldsymbol{\omega}), \tag{4.5}$$

donde $f(\boldsymbol{\omega})$ es una distribución en la FE parametrizada por $\boldsymbol{\omega}$, la divergencia nos permite definir una métrica *pullback* que traduce localmente al espacio de parámetros Ω algunas estructuras universales de las distribuciones $f(\boldsymbol{\omega})$. En el Capítulo 5 utilizaremos esta propiedad para definir penalizaciones óptimas.

Finalmente en la Sección 4.8 mostramos ejemplos de scores locales para las familias paramétricas estudiadas en el Capítulo 2.

4.1 DIVERGENCIA DE KULLBACK-LEIBLER Y MÁXIMA VEROSIMILITUD

Veamos como KL induce el MLE y la métrica de Fisher.

Definición 4.2 (Kullback-Leibler). *La divergencia de KL de la distribución f a la distribución positiva g está dada por*

$$KL[f : g] := \sum_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) \log \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}.$$

Cuando hacemos inferencia estamos interesados en estimar la distribución g indexada por los parámetros $\boldsymbol{\omega} \in \Omega$ que esté más cerca de la distribución objetivo f. Dado un conjunto de datos de entrenamiento $\{\boldsymbol{X}^{(i)}\}_{i=1}^n$ y considerando que $\hat{f}$ es la distribución empírica que coloca una masa de probabilidad $1/n$ en cada observación, la KL de $\hat{f}$ al modelo $g(\boldsymbol{\omega})$ resulta equivalente a la esperanza muestral del *score* $\log g(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\omega})$ (Martens, 2020), ya que

$$KL[\hat{f} : g] \propto \frac{1}{n} \sum_{\boldsymbol{x} \in \{\boldsymbol{x}^{(i)}\}_{i=1}^n} [-\log g(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\omega}) - \log n]. \tag{4.6}$$

El Estimador de Máxima Verosimilitud (MLE) de $\boldsymbol{\omega}$ se obtiene de minimizar (4.6), la KL entre la distribución empírica de los datos y el modelo que se está aprendiendo, o equivalentemente maximizando la esperanza muestral del score $\log g(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\omega})$. Sin embargo, históricamente se definió el MLE a partir del concepto de verosimilitud:

Definición 4.3 (Estimador de Máxima Verosimilitud). *La verosimilitud $\mathcal{L}$ es la probabilidad conjunta de las observaciones $\{\boldsymbol{X}^{(i)}\}_{i=1}^n$ vista como función de los parámetros $\boldsymbol{\omega}$:*

$$\mathcal{L}(\boldsymbol{\omega} \mid \{\boldsymbol{X}^{(i)}\}_{i=1}^n) = \prod_{i=1}^n g(\boldsymbol{x} \mid \boldsymbol{\omega}), \tag{4.7}$$

El MLE se obtiene maximizando (4.7) con $\boldsymbol{\omega} \in \Omega$:

$$\hat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega} \in \Omega} \mathcal{L}(\boldsymbol{\omega} \mid \{\boldsymbol{X}^{(i)}\}_{i=1}^n). \tag{4.8}$$

Observación 4.2. *La evaluación de (4.6) o (4.7) requiere la evaluación de la función de partición del modelo que involucra una integral (o una sumatoria) p-dimensional sobre el espacio muestral.*

La métrica inducida por la divergencia de KL es la única métrica invariante por estadísticos suficientes salvo un factor de escala (Amari, 2016) y está dada por la Matriz de información de Fisher:

Observación 4.3 (Matriz de información de Fisher). *Bajo condiciones de regularidad (Casella y Berger, 2002, eq. (2.5.16), Lemma 5.3, p.116), la métrica local $\mathcal{I}$ inducida por la KL (Definición 4.1) cuando $\boldsymbol{X} \sim g(\boldsymbol{\omega})$ resulta*

$$\mathcal{I}(\boldsymbol{\omega}) := E_{\boldsymbol{\omega}^*} \left[\frac{\partial}{\partial \boldsymbol{\omega}} \log g(\boldsymbol{X} \mid \boldsymbol{\omega}) \frac{\partial}{\partial \boldsymbol{\omega}^\top} \log g(\boldsymbol{X} \mid \boldsymbol{\omega}) \right] \tag{4.9}$$

$$= -E_{\boldsymbol{\omega}} \frac{\partial^2}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^\top} \log g(\boldsymbol{X} \mid \boldsymbol{\omega}). \tag{4.10}$$

De la primera expresión resulta que $\mathcal{I}$ es una Matriz Semidefinida Positiva (SDP) por ser la esperanza de un producto externo. Además,

por (4.10), resulta equivalente a la esperanza del Hessiano del score logarítmico.

La sensibilidad con la cual $g(\boldsymbol{x} \mid \boldsymbol{\omega})$ depende de una muestra $\boldsymbol{x}$ es cuantificada por la matriz de información de Fisher. Cuando se cuenta con n muestras de entrenamiento, la varianza de cualquier estimador está acotada por la cota de Cramér-Rao.

Proposición 4.1 (Cota inferior de Cramér-Rao (Vaart, 2000))**.** *Supongamos que contamos con n muestras i.i.d;* $\boldsymbol{X}^{(i)} \sim g(\boldsymbol{X} \mid \boldsymbol{\omega}^*)$ *para estimar el parámetro* $\boldsymbol{\omega}$ *de la familia* $g(\boldsymbol{X} \mid \boldsymbol{\omega})$ *y se satisfacen las condiciones de regularidad detalladas en Vaart (2000, Theorem 5.21) y además la matriz de información Fisher (Definición 4.3) evaluada en el parámetro poblacional* $\boldsymbol{\omega}^*$ *es inversible, entonces la varianza* $V_{i,j} = E\left[(\hat{\omega}_i - \omega_i^*)(\hat{\omega}_j - \omega_j^*)\right]$ *de cualquier estimador insesgado* $\hat{\boldsymbol{\omega}}$ *de* $\boldsymbol{\omega}^*$ *está acotada por la inversa de la matriz de Fisher evaluada en* $\boldsymbol{\omega}^*$*:*

$$\boldsymbol{V} \geq \frac{1}{n}\mathcal{I}(\boldsymbol{\omega}^*)^{-1}, \tag{4.11}$$

donde $\mathcal{I}$ *es la matriz de información de Fisher (4.9) y la desigualdad implica que* $\boldsymbol{V} - \frac{1}{n}\mathcal{I}(\boldsymbol{\omega}^*)^{-1}$ *es Matriz Semidefinida Positiva (SDP).*

Más aún, el Estimador de Máxima Verosimilitud (MLE) resulta asintóticamente eficiente, i. e. alcanza la cota (4.11).

Notar que el concepto fundamental de la cota de Cramér-Rao radica en la suavidad del mapeo $\boldsymbol{\omega} \to f(\boldsymbol{\omega})$, o equivalentemente en $\boldsymbol{\omega} \to KL[g(\boldsymbol{\omega}^*) : g(\boldsymbol{\omega})]$, en el sentido que una estimación es plausible de tener poca varianza si un pequeño cambio en los parámetros no puede provocar un cambio muy grande en la distribución resultante. Veremos en el Capítulo 5 que, en general, la inversa de la métrica local permite cuantificar la curvatura del espacio paramétrico, información que puede ayudar a mejorar la convergencia de los algoritmos de aprendizaje.

4.2 LIKELIHOOD DE COMPOSICIÓN Y PSEUDO-LIKELIHOOD

Con el fin de evitar el cálculo de la función de partición del modelo ($Z(\boldsymbol{\omega})$ en la Definición 2.1), cuando resulta intratable, o requiere aproximaciones, Varin, Reid y Firth (2011) construyen una *verosimilitud de composición* como el producto de verosimilitudes de eventos condicionales o marginales, obteniendo un estimador insesgado. Formalmente,

Definición 4.4 (Verosimilitud de composición)**.** *Dado un conjunto de eventos marginales o condicionales* $\{\mathcal{A}_1, \cdots, \mathcal{A}_K\}$*, con verosimilitud asociada* $\mathcal{L}_k(\boldsymbol{\omega} \mid \boldsymbol{x}) \propto g(\boldsymbol{x} \in \mathcal{A}_k \mid \boldsymbol{\omega})\}$*, la verosimilitud de composición es el producto pesado*

$$\mathcal{L}_C(\boldsymbol{\omega} \mid \boldsymbol{x}) = \prod_{k=1}^{K} \mathcal{L}_k(\boldsymbol{\omega} \mid \boldsymbol{x})^{w_k}, \tag{4.12}$$

donde $w_k > 0$*.*

Además, podemos definir una divergencia de composición como la combinación lineal de KL respecto al conjunto de eventos:

Definición 4.5 (Divergencia de composición). *Dado un conjunto de eventos marginales o condicionales* $\{\mathcal{A}_1, \cdots, \mathcal{A}_K\}$*, la divergencia de composición de la distribución* f *a la* g *está dada por*

$$CKL[f : g] = \sum_{k=1}^{K} w_k E_f[\log f(\boldsymbol{x} \in \mathcal{A}_k) - \log g(\boldsymbol{x} \in \mathcal{A}_k))]. \tag{4.13}$$

Un caso particular resulta de considerar los eventos condicionales $\mathcal{A}_k = f(X_j = x_j \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{x}_{\setminus j})$, $k = 1, \cdots, p$, referido a continuación:

Definición 4.6. *El Estimador de Máxima Pseudo-verosimilitud (PLE) es el argumento que maximiza la verosimilitud de composición respecto de los eventos condicionales (Besag, 1975):*

$$\mathcal{L}_P(\boldsymbol{\omega} \mid \boldsymbol{x}) = \prod_{k=1}^{p} g(X_k = x_k \mid \boldsymbol{X}_{\setminus k} = \boldsymbol{x}_{\setminus k}, \boldsymbol{\omega}). \tag{4.14}$$

Equivalentemente, definimos la divergencia asociada como

$$PKL[f : g] = \sum_{k=1}^{p} w_k E_f[\log f(X_j = x_j \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{x}_{\setminus j}) - \log g(X_j = x_j \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{x}_{\setminus j})]. \tag{4.15}$$

En el Apéndice C.1 demostramos la equivalencia (4.3) para la divergencia (4.15).

Observación 4.4. *En* (4.14) *y* (4.15) *se evalúan únicamente las funciones de partición condicionales, evitando el costo de evaluar la función de partición del modelo multivariado.*

Dada una divergencia de composición CKL, Definición 4.5, Lindsay, Yi y Sun (2011) definen un sustituto para las densidades tales que la CKL resulta efectivamente la divergencia de Kullback-Leibler del modelo, Definición 4.2. Esto presenta una ventaja conceptual, ya que considerando el modelo sustituto, el estimador de verosimilitud de composición tiene las mismas propiedades que el MLE. En particular, el estimador de pseudo-verosimilitud induce una densidad sustituta donde las suposiciones se limitan a los primeros momentos, y por lo tanto será menos afectada por la mala especificación del modelo. Partiendo de la factorización $g(\boldsymbol{X}) = g(x_1)g(X_2 \mid X_1)g(X_3 \mid X_2, X_1) \cdots g(X_p \mid \boldsymbol{X}_{\setminus p})$, la densidad sustituto puede escribirse como el producto de las condicionales:

$$\tilde{g}(\boldsymbol{X} = \boldsymbol{x}) = \prod_{j=1}^{p} g(X_j = x_j \mid \boldsymbol{X}_{<j} = \boldsymbol{x}_{<j}), \tag{4.16}$$

donde $\boldsymbol{X}_{<j}$ representa el conjunto $\{X_1, \cdots, X_{j-1}\}$.

4.3 REGLAS DE SCORE

A partir de (4.3), observamos que un *score* $S(\boldsymbol{x}, g) : \mathcal{X} \times \mathcal{P} \to \mathbb{R}$ es una función de costo que cuantifica la exactitud con la que una

distribución $g \in \mathcal{P}$ de la variable aleatoria $\boldsymbol{X}$ explica la observación $\boldsymbol{x}$. Un score se dice *propio* si es honesto en el sentido que la esperanza del score $E_{\boldsymbol{X}\sim f} S(\boldsymbol{X}, g)$ se minimiza con $g = f$ y es *propio estricto* si el mínimo es único, induciendo estimadores consistentes (Dawid, Lauritzen, Parry et al., 2012; Parry, Dawid, Lauritzen et al., 2012). Además, dos scores S_1, S_2 son equivalentes si $S_1(\boldsymbol{x}, g) = aS_2(\boldsymbol{x}, g) + b(\boldsymbol{x})$ con $a > 0$ y b una función medible.

Cada regla de score S define una entropía (medida de incerteza) y una divergencia:

Definición 4.7 (Entropía inducida por un score propio)**.** *La entropía H inducida por un score propio S está definida como*

$$H[f] = E_{X\sim f} S(X, f) \quad \textit{con } f \in \mathcal{P}, \tag{4.17}$$

que es una función cóncava de f.

Definición 4.8 (Divergencia inducida por un score propio)**.** *La divergencia D inducida por el score propio S satisface*

$$D[f : g] = E_{X\sim f} S(\boldsymbol{X}, g) - H[f] \quad \textit{con } f, g \in \mathcal{P}, \tag{4.18}$$

con $D[f : g] \geq 0$ con igualdad cuando $f = g$ y sólo en este caso si S es propio estricto.

SCORES LOCALES Un score es *local* si depende de la densidad $g(\cdot)$ sólo a través del valor $g(\boldsymbol{x})$ en $\boldsymbol{x}$. Dawid, Lauritzen, Parry et al. (2012) demostraron que cualquier score local es equivalente al score logarítmico $S(\boldsymbol{x}, g) = -\log g(\boldsymbol{x})$ inducido por la divergencia de Kulback-Leibler, Definición 4.2.

En lo que sigue, mostraremos que relajando la noción de localidad, es posible obtener scores *homogéneos* en la densidad g, induciendo estimadores para modelos no normalizados, o equivalentemente, que no requieran evaluar la función de partición del modelo. En la Sección 4.3.1 veremos que permitiendo dependencia en un número m de derivadas de g, es posible definir scores locales eficientes para variables aleatorias continuas. Por otro lado, en la Sección 4.3.2 tratamos el caso discreto, donde se generaliza el concepto de localidad a un entorno discreto de $\boldsymbol{x}$.

4.3.1 *Variables aleatorias continuas*

En el caso de variables aleatorias absolutamente continuas, un score se dice *m-local* si depende de la distribución objetivo $g(x)$ solamente a través de un número $m < \infty$ de derivadas de $g(\cdot)$ en x. En general, Dawid, Lauritzen, Parry et al. (2012) muestran que existe una forma de construir scores propios a partir de una función homogénea m-local para m par, pero no existe para m impares. Prueban además a que para $m \geq 2$ los scores son homogéneos. En particular, Hyvärinen (2005) encontró un score homogéneo de orden 2 al considerar la Divergencia de Fisher:

Definición 4.9 (Divergencia de Fisher (Lyu, 2012)). *La divergencia de Fisher de la densidad f a la g se define como*

$$DF[f:g] = \frac{1}{2}\int_{x\in\mathcal{X}} f(x)\left\|\frac{\partial}{\partial \boldsymbol{x}}\log f(x) - \frac{\partial}{\partial \boldsymbol{x}}\log g(x)\right\|^2. \tag{4.19}$$

Hyvärinen (2005) mostró, usando integración por partes, la siguiente proposición que relaciona la divergencia de Fisher con un score propio:

Proposición 4.2 (Estimador de Score Matching (Hyvärinen, 2005)). *Cuando $\mathcal{X} = \mathbb{R}^p$ y $g \in \mathcal{P}$ son distribuciones con densidad diferenciable de segundo orden respecto de $\boldsymbol{x}$ tales que $\frac{\partial}{\partial \boldsymbol{x}}\log g(x) \to 0$ cuando $\|\boldsymbol{x}\| \to \infty$, minimizar la divergencia (4.19) equivale a minimimizar la esperanza del score*

$$S(x,g) = \frac{\partial^2}{\partial \boldsymbol{x}\partial \boldsymbol{x}^\top}\log g(x) + \frac{1}{2}\left\|\frac{\partial}{\partial \boldsymbol{x}}\log g(x)\right\|^2. \tag{4.20}$$

Además, S es un score propio estricto (Dawid, Lauritzen, Parry et al., 2012).

El Estimador de Score Matching (SME) se obtiene de minimizar la esperanza empírica del score (4.20) o equivalentemente de la divergencia de Fisher (4.19) partiendo de la distribución empírica $\hat{f}$.

Observación 4.5. *Como $\frac{\partial}{\partial \boldsymbol{x}}\log g(x)$ en (4.19) no depende de la función de partición del modelo, el estimador de Score Matching no requiere la evaluación de la función de partición, permitiendo su aplicación sobre modelos no normalizados.*

4.3.2 *Variables aleatorias discretas*

La extensión al caso discreto (Dawid, Lauritzen, Parry et al., 2012; Gneiting y Raftery, 2007) requiere la definición de una familia de scores, uno por cada evento en el espacio muestral. Esto se conoce como Represención de Savage, la cual enunciamos a continuación:

Definición 4.10 (Represención de Savage). *Consideremos un espacio muestral discreto $\mathcal{X} = \{1, \cdots, m\}$ que consiste de m eventos mutuamente exclusivos asociados al vector de probabilidad $(p_1, \cdots, p_m) \in \boldsymbol{P}_m$, el simplex de dimensión m. Un score propio puede identificarse con una colección de m funciones*

$$S_i(\boldsymbol{p}) : \mathcal{P}_m \to \mathbb{R} \quad i = 1, \cdots, m, \tag{4.21}$$

que asigna el costo $S_i(\boldsymbol{p})$ cuando ocurre el evento i. En lo que sigue utilizaremos la notación $S(\boldsymbol{x}, \boldsymbol{p})$ para hacer referencia al score $S_x(\boldsymbol{p})$.

Notemos que cuando el espacio muestral es discreto, es posible definir un score o-homogéneo, i. e. $S(x, \boldsymbol{p}) = S(\boldsymbol{x}, \lambda\boldsymbol{p})$ para todo $\lambda > 0$, a partir de un score S considerando $S(x, \tilde{\boldsymbol{p}}) := S(x, \tilde{\boldsymbol{p}}/\sum_{i=1}^m \tilde{p}_i)$.

El siguiente teorema enuncia que todo score diferenciable en un espacio muestral finito es el gradiente de una función cóncava homogénea:

Teorema 4.1 (McCarthy, Savage). *$H : \mathcal{P} \to \mathbb{R}$ es una función cóncava 1-homogénea y $\boldsymbol{S}(\boldsymbol{p}) := \nabla H$ un supergradiente de H si y sólo si $S(x, \boldsymbol{p}) : \mathcal{X} \times \mathcal{P} \to \mathbb{R}$, correspondiente a la componente x de $\boldsymbol{S}(\boldsymbol{p})$, es un score propio y $H(\boldsymbol{p})$ la entropía asociada en $\boldsymbol{p}$.*

Una consecuencia del Teorema 4.1 es que si S es un score propio o-homogéneo,

$$\frac{\partial S(x, \tilde{\boldsymbol{p}})}{\partial \tilde{p}_y} = \frac{\partial S(y, \tilde{\boldsymbol{p}})}{\partial \tilde{p}_x} \quad \forall x, y.$$

Dawid, Lauritzen, Parry et al. (2012) consideraron scores que dependen de $\boldsymbol{p}$ solamente en un entorno N_x de x: $S(\boldsymbol{x}, \boldsymbol{p}) = S(\boldsymbol{x}, \boldsymbol{p}_{N_x})$ y por lo tanto, si $c \notin N_y$,

$$\frac{\partial S(x, \tilde{\boldsymbol{p}})}{\partial \tilde{p}_y} = \frac{\partial S(y, \tilde{\boldsymbol{p}})}{\partial \tilde{p}_x} = 0,$$

equivalentemente podemos pedir que $y \notin N_x$, lo que induce una relación simétrica y por lo tanto determinada por un grafo no dirigido $\mathcal{G}$. Decimos que $y \in N_x$ sii $x - y$ i. e. x e y son vecinos en $\mathcal{G}$. En este caso llamamos score $\mathcal{G}$-local. Además, un score $\mathcal{G}$-local o-homogéneo sólo depende de $\boldsymbol{p}$ a través de la distribución condicional $\boldsymbol{p}_{|N_x}$ de X dado $X \in N_x$: $S(\boldsymbol{x}, \boldsymbol{p}) = S(\boldsymbol{x}, \boldsymbol{p}_{|N_x})$.

Ejemplo 4.1 (Score propio para distribuciones en el dominio de los enteros). *Dawid, Lauritzen, Parry et al. (2012) consideran un grafo de conexión entre enteros sucesivos, cuyo conjunto de cliques es $C_x = \{(x, x+1) : x = 0, 1, \cdots\}$. Una entropía C_x-local está dada por*

$$H_x(p_x, p_{x+1}) = p_x G_x(p_{x+1}/p_x),$$

con G_x concava y $\{p_x : p_x > 0, x \in \mathcal{X}\}$ es un vector de probabilidades no normalizadas. El score propio inducido resulta

$$S(x) = G'_{x-1}(u_{x-1}) I(0 < x) + G_x(u_x) - u_x G'_x(u_x) \quad x = 0, 1, \cdots \tag{4.22}$$

donde $u_x = p_{x+1}/p_x$ y la indicadora $I(0 < x)$ hace nulo al primer término para $x = 0$.

En particular, podemos considerar $G_x(v) = -(x+1)^a \frac{v^m}{m(m-1)}$ para $m \neq 0, 1$. Con lo cual, el score obtenido para p_x correspondiente a la distribución Poisson de media μ resulta:

$$S(x) = -x^{a-m+1} \frac{\mu^{m-1}}{m-1} + (x+1)^{a-m} \frac{\mu^m}{m}, \quad x = 0, 1, \cdots \tag{4.23}$$

En el Apéndice C.2 se encuentran las demostraciones de (4.22) y (4.23).

Score para distribuciones multivariadas discretas

En el caso multivariado, es posible definir un score de composición, similar a la pseudo-verosimilitud, Definición 4.6, donde cada término corresponde a un score propio en lugar del log-score. En particular, Dawid, Lauritzen, Parry et al. (2012) consideraron que el espacio discreto es un espacio producto $\mathcal{X} = \mathcal{X}_1 \times, \cdots, \times, \mathcal{X}_p$ y definieron la relación simétrica de vecindad $\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}$, en $\mathcal{X}$, si para algún j, $\boldsymbol{x}^{(1)}_{\setminus j} =$

$\boldsymbol{x}_{\setminus j}^{(2)}$. Luego, para cada clique en el conjunto de cliques maximales $C = C_{j,\boldsymbol{\pi}_{\setminus j}} := \{\boldsymbol{x} : \boldsymbol{x}_{\setminus j} = \boldsymbol{\pi}_{\setminus j}\}$ del grafo asociado $\boldsymbol{G}$, donde sólo $X_j \in \mathcal{X}_j$ puede variar ,se define un score, que depende de g a través de los eventos condicionales $g(\cdot \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{\pi}_{\setminus j})$:

Definición 4.11 (Score de composición para espacios producto (Dawid, Lauritzen, Parry et al., 2012)). *Partiendo de un conjunto de scores propios univariados $\{S_c\}_{c\in C}$, se define el score de composición para espacios producto como*

$$S(\boldsymbol{X}, g) = \sum_{c\in C} S_c(x_j, g(\cdot \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{\pi}_{\setminus j}))1_{\{\boldsymbol{x}_{\setminus j}=\boldsymbol{\pi}_{\setminus j}\}}, \tag{4.24}$$

En particular, si consideramos el mismo score S_j para todos los eventos condicionales $\boldsymbol{\pi}_{\setminus j}$, el score de composición para espacios producto se simplifica en

$$S(\boldsymbol{X}, g) = \sum_{j=1}^{p} S_j(x_j, g(\cdot \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{x}_{\setminus j})), \tag{4.25}$$

donde S_j es un score propio para variables y distribuciones definidas sobre $\mathcal{X}_j$. Además, la divergencia asociada resulta

$$D[f : g] = \sum_{j=1}^{p} E_{\boldsymbol{X}\sim f} D_j[f(\cdot \mid \boldsymbol{X}_{\setminus j}) : g(\cdot \mid \boldsymbol{X}_{\setminus j})], \tag{4.26}$$

donde D_j es la divergencia asociada a S_j.

Teorema 4.2 (Score de composición para espacios producto (Dawid, 1979)). *Si f, g son distribuciones positivas y S_j es un score propio estricto para $X_j \mid \boldsymbol{X}_{\setminus j}$ para todo $j = 1, \cdots, p$, entonces el score multivariado S dado en la Definición 4.11 es un score propio estricto, además $D[f : g] = 0$ sii $f = g$.*

4.4 DIVERGENCIA ENTRE MODELOS CONDICIONALES

Al considerar un problema de aprendizaje supervisado, podemos suponer que contamos con muestras $\{(\boldsymbol{x}^{(y=i)}, y^{(i)})\}_{i=1}^{n}$, donde $(\boldsymbol{X}, Y) \sim f_{\boldsymbol{X},Y} = f_{\boldsymbol{X}|Y} f_Y$. Luego la divergencia de $f_{\boldsymbol{X},Y}$ a $g_{\boldsymbol{X},Y}$ se define como

$$\mathrm{KL}[f_{\boldsymbol{X},Y} : g_{\boldsymbol{X},Y}] = \int f_{\boldsymbol{X},Y}(\boldsymbol{x}, y) \log \frac{f_{\boldsymbol{X},Y}(\boldsymbol{x}, y)}{g_{\boldsymbol{X},Y}(\boldsymbol{x}, y)} d\boldsymbol{x}dy. \tag{4.27}$$

Como vimos en el Capítulo 3, cuando el objetivo es estimar una RSD, solamente es necesario estimar el modelo condicional $g_{\boldsymbol{X}|Y}$ en lugar del modelo conjunto $g_{\boldsymbol{X},Y}$. En un contexto similar, Martens (2020) observa que al considerar $g(y) = f(y)$, es posible utilizar (4.27) para estimar modelos condicionales, ya que

$$\mathrm{KL}[f_{\boldsymbol{X},Y} : g_{\boldsymbol{X},Y}] = \int f_{\boldsymbol{X},Y}(\boldsymbol{x}, y) \log \frac{f_{\boldsymbol{X}|Y}(\boldsymbol{x}, y) f_Y(y)}{g_{\boldsymbol{X}|Y}(\boldsymbol{x}, y) f_Y(y)} d\boldsymbol{x}dy \tag{4.28}$$

$$= \int f_Y(y) \int f_{\boldsymbol{X}|Y}(\boldsymbol{x} \mid y) \log \frac{f_{\boldsymbol{X}|Y}(\boldsymbol{x}, y)}{g_{\boldsymbol{X}|Y}(\boldsymbol{x}, y)} d\boldsymbol{x}dy \tag{4.29}$$

$$= E_{f_Y} \mathrm{KL}[f_{\boldsymbol{X}|Y} : g_{\boldsymbol{X}|Y}]. \tag{4.30}$$

Con lo cual es posible aproximar la esperanza (4.30) por la media muestral, al considerar la distribución empírica $\hat{f}_Y$ basada en n muestras $\{y^{(i)}\}_{i=1}^n$, i. e.

$$E_{\hat{f}_Y}\mathrm{KL}[f_{\boldsymbol{X}|Y} : g_{\boldsymbol{X}|Y}] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}[f_{\boldsymbol{X}|Y=y^{(i)}} : g_{\boldsymbol{X}|Y=y^{(i)}}], \tag{4.31}$$

Operando del mismo modo, es posible definir la misma cantidad cuando reemplazamos la KL por otra divergencia basada en un score S, y por lo tanto la divergencia inducida, Definición 4.8, resulta equivalente a

Contribución original: extensión a scores propios

Proposición 4.3 (Divergencia entre modelos condicionales)**.** *La divergencia entre modelos condicionales $f_{\mathcal{X}|Y}$ y $g_{\mathcal{X}|Y}$ está dada por*

$$D[f : g] = E_{f_Y}D[f_{\boldsymbol{X}|Y} : g_{\boldsymbol{X}|Y}]. \tag{4.32}$$

Asumiendo además que f_Y y g_Y están caracterizadas por la distribución empírica $\hat{f}_Y$, se obtiene que

$$D[f : g] = E_{\hat{f}_Y}D[f_{\boldsymbol{X}|Y} : g_{\boldsymbol{X}|Y}] = \frac{1}{n}\sum_{i=1}^{n}D[f_{\boldsymbol{X}|Y=y^{(i)}} : g_{\boldsymbol{X}|Y=y^{(i)}}]. \tag{4.33}$$

4.5 ESTIMACIÓN DIFERENCIAL

En el caso $Y \in \{0, 1\}$, un modelo gráfico condicional, Definición 2.12, es equivalente a un modelo gráfico diferencial (Kim, Liu y Kolar, 2021; Zhao, Cai y Li, 2014) basado en la partición de la muestra en $\{\boldsymbol{x}^{(t)} \mid y^{(t)} = 0\}_{t=1}^n$ y $\{\boldsymbol{x}^{(t)} \mid y^{(t)} = 1\}_{t=1}^n$. En particular, si consideramos que $\boldsymbol{X} \mid Y = 0 \sim f_{\boldsymbol{X}|Y=0}$ y $\boldsymbol{X} \mid Y = 1 \sim f_{\boldsymbol{X}|Y=1}$ siguen modelos gráficos de segundo orden (Definición 2.5) podemos estimar el modelo gráfico diferencial caracterizado por el parámetro natural $\Delta\boldsymbol{\omega} = \boldsymbol{\eta}_{\boldsymbol{X}|Y=1} - \boldsymbol{\eta}_{\boldsymbol{X}|Y=0} = \boldsymbol{\Gamma}_{\boldsymbol{\eta}} \in \mathbb{R}^p$, ya que $\boldsymbol{\eta}_0$ y $\boldsymbol{\Theta}$ son considerados constantes en el modelo condicional, i. e.

Definición 4.12 (Modelo gráfico diferencial de segundo orden condicional lineal en T(X))**.** *Definimos al modelo gráfico diferencia basado en la partición binaria de las muestras dada por la respuesta Y al cociente*

$$\begin{aligned} r(\boldsymbol{X} = \boldsymbol{x} \mid \Delta\boldsymbol{\omega} = \boldsymbol{\Gamma}) &= \frac{P(\boldsymbol{X} = \boldsymbol{x} \mid Y = 1)}{P(\boldsymbol{X} = \boldsymbol{x} \mid Y = 0)} \\ &= \frac{\exp\left\{\boldsymbol{\eta}_{\boldsymbol{X}|Y=1}^\top \boldsymbol{T}(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{T}(\boldsymbol{x})^\top\boldsymbol{\Theta}\boldsymbol{T}(\boldsymbol{x}) + h(\boldsymbol{x}) - Z(\boldsymbol{\eta}_{\boldsymbol{X}|Y=1}, \boldsymbol{\Theta})\right\}}{\exp\left\{\boldsymbol{\eta}_{\boldsymbol{X}|Y=0}^\top \boldsymbol{T}(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{T}(\boldsymbol{x})^\top\boldsymbol{\Theta}\boldsymbol{T}(\boldsymbol{x}) + h(\boldsymbol{x}) - Z(\boldsymbol{\eta}_{\boldsymbol{X}|Y=0}, \boldsymbol{\Theta})\right\}} \\ &= \exp\left\{\boldsymbol{\Gamma}^\top\boldsymbol{T}(\boldsymbol{x}) - Z(\boldsymbol{\Gamma})\right\}, \end{aligned}$$

donde $\exp Z(\boldsymbol{\Gamma}) = \frac{\exp Z(\boldsymbol{\eta}_{\boldsymbol{X}|Y=1}, \boldsymbol{\Theta})}{\exp Z(\boldsymbol{\eta}_{\boldsymbol{X}|Y=0}, \boldsymbol{\Theta})} = \frac{\exp Z(\boldsymbol{\eta}_{\boldsymbol{X}|Y=0} + \boldsymbol{\Gamma}, \boldsymbol{\Theta})}{\exp Z(\boldsymbol{\eta}_{\boldsymbol{X}|Y=0}, \boldsymbol{\Theta})}$ *considera a* $f_{\boldsymbol{X}|Y=0}$ *como medida base.*

La Definición 4.12 permite la factorización $f_{\boldsymbol{X}|Y=1}(\boldsymbol{X}) = r(\boldsymbol{X})f_{\boldsymbol{X}|Y=0}(\boldsymbol{X})$ y por lo tanto el parámetro diferencial $\Delta\boldsymbol{\omega} = \boldsymbol{\Gamma}$ puede ser caracterizado como el minimizante de la KL entre $f_{\boldsymbol{X}|Y=1}(\boldsymbol{X})$ y $r(\boldsymbol{X})f_{\boldsymbol{X}|Y=0}(\boldsymbol{X})$:

$$\mathrm{KL}[f_{\boldsymbol{X}|Y=1}(\boldsymbol{X}) : r(\boldsymbol{X}) f_{\boldsymbol{X}|Y=0}(\boldsymbol{X})] = \int_{\boldsymbol{x}\in\mathcal{X}} f_{\boldsymbol{X}|Y=1}(\boldsymbol{X}) \log \frac{f_{\boldsymbol{X}|Y=1}(\boldsymbol{X})}{r(\boldsymbol{X}) f_{\boldsymbol{X}|Y=0}(\boldsymbol{X})}, \tag{4.34}$$

Si consideramos únicamente los términos que dependen de $\Delta\boldsymbol{\omega}$, el minimizante de (4.34) coincide con el minimizante de

$$-E_{\boldsymbol{X}|Y=1}[\boldsymbol{\Gamma}^\top T(\boldsymbol{X})] + \log E_{\boldsymbol{X}|Y=0}[\exp\{\boldsymbol{\Gamma}^\top T(\boldsymbol{X})\}], \tag{4.35}$$

donde consideramos que $\int_{\boldsymbol{x}\in\mathcal{X}} r(\boldsymbol{x}) f_{\boldsymbol{X}|Y=0}(\boldsymbol{X}) = 1$ y $Z(\boldsymbol{\Gamma}) = \log E_{\boldsymbol{X}|Y=0}[\exp\{\boldsymbol{\Gamma}^\top T(\boldsymbol{X})\}]$. Finalmente, considerando estimaciones de las esperanzas $E_{\boldsymbol{X}|Y=0}$ y $E_{\boldsymbol{X}|Y=1}$ basadas en las muestras de entrenamiento, es posible aproximar la función de costo (4.35), definiendo el método

Definición 4.13 (Kullback-Leibler Importance Estimation Procedure)**.** *El estimador Kullback-Leibler Importance Estimation Procedure (KLIEP) (Kim, Liu y Kolar, 2021) del modelo diferencial, Definición 4.12, se obtiene como el minimizante de la función convexa*

$$\begin{aligned}\ell_{KLIEP}(\boldsymbol{\Gamma}) = &-\frac{1}{|\{\boldsymbol{x}^{(t)} \mid y^{(t)}=1\}_{t=1}^n|} \sum_{\boldsymbol{x}\in\{\boldsymbol{x}^{(t)}|y^{(t)}=1\}_{t=1}^n} \boldsymbol{\Gamma}^\top T(\boldsymbol{X}) \\ &+ \log\left\{ \frac{1}{|\{\boldsymbol{x}^{(t)} \mid y^{(t)}=0\}_{t=1}^n|} \sum_{\boldsymbol{x}\in\{\boldsymbol{x}^{(t)}|y^{(t)}=0\}_{t=1}^n} \exp\{\boldsymbol{\Gamma}^\top T(\boldsymbol{X})\} \right\}.\end{aligned} \tag{4.36}$$

4.6 M-ESTIMADOR ASOCIADO

Al estimar un parámetro $\boldsymbol{\omega}$ a partir de una muestra $\{\boldsymbol{X}^{(i)}\}_{i=1}^n$, se busca minimizar la esperanza de un score propio $S(\boldsymbol{x}, \boldsymbol{\omega})$, o equivalentemente, la divergencia asociada entre la distribución empírica y el modelo. En particular,

Definición 4.14 (M-estimador)**.** *Un m-estimador se define como el argumento que maximiza una función convexa*

$$\hat{ES}(\boldsymbol{X}, \boldsymbol{\omega}) : \Omega \to \mathbb{R}, \tag{4.37}$$

o equivalentemente, la solución al sistema de ecuaciones de estimación:

$$\frac{\partial}{\partial \boldsymbol{\omega}} \hat{ES}(\boldsymbol{X}, \boldsymbol{\omega}) = \mathbf{0}. \tag{4.38}$$

La consistencia de un *m*-estimador basado en score propio, es consecuencia de la ley débil de los grandes números. Más aún, si asumimos que la esperanza del score $ES(X, \boldsymbol{\omega})$ es una función Lipschitz continua de $\boldsymbol{\omega}$ en un entorno del valor poblacional $\boldsymbol{\omega}^*$, podemos asegurar la normalidad asintótica del estimador (Vaart, 2000):

Proposición 4.4 (Normalidad asintótica)**.** *Bajo las condiciones de regularidad enunciadas en (Vaart, 2000, Theorem 5.21), el estimador $\hat{\boldsymbol{\omega}}$ obtenido de minimizar la esperanza muestral de un score propio $\hat{ES}(\boldsymbol{X}, \boldsymbol{\omega})$ para un modelo gráfico de segundo orden, Definición 2.5, es consistente y su distribución asintótica es normal. En particular, tenemos que*

$$\sqrt{n}(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}) \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{V}), \tag{4.39}$$

donde

$$\boldsymbol{V} = \left(E\frac{\partial^2 S}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^T}\right)^{-1} E\frac{\partial S}{\partial \boldsymbol{\omega}}\frac{\partial S}{\partial \boldsymbol{\omega}^T}\left(E\frac{\partial^2 S}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^T}\right)^{-1}. \tag{4.40}$$

En el caso del MLE, la varianza asintótica se simplifica y es igual a la inversa de la información de Fisher, Definición 4.3.

Observación 4.6. *Notar que como $\boldsymbol{V} \in \mathbb{R}^{\dim(\omega)^2}$, i.e. es de orden p^4, su evaluación no escala con la dimensión del problema, por lo que es necesario asumir alguna estructura para el cálculo de dicha matriz.*

Corolario 4.1. *El estimador $\hat{\boldsymbol{\Gamma}}$ obtenido de minimizar la esperanza muestral de un score propio $\hat{ES}_{\boldsymbol{X}|Y}((\boldsymbol{x}; y), (\boldsymbol{\eta}; \boldsymbol{\Gamma}; \boldsymbol{\Theta}))$ para un modelo gráfico de segundo orden condicional lineal en $T(\boldsymbol{X})$, Definición 2.12, es consistente y su distribución asintótica es normal.*

4.7 MÉTRICA LOCAL

La métrica local $\mathcal{I}$ en la Definición 4.1 que resulta la matriz de información en el caso de considerar la divergencia de Kullback-Leibler, expresa la curvatura local del espacio de distribuciones $\mathcal{F} = \{f(\boldsymbol{\omega}) : \boldsymbol{\omega} \in \Omega\}$ inducida por una divergencia D. Una métrica *pullback* definida sobre el espacio de parámetros Ω resulta localmente independiente a la parametrización si está definida por la expansión de Taylor de dicha divergencia (Amari, 2016; Martens, 2020).

4.7.1 *Medida pullback local*

En la Definición 4.1, vimos que toda divergencia induce una métrica local dada por la matriz de información $\mathcal{I}(\boldsymbol{\omega})$, la cual expresa la curvatura local del espacio paramétrico o variedad Ω.

En otras palabras, la métrica local $\mathcal{I}(\boldsymbol{\omega})$ define un mecanismo para trasladar localmente la geometría del espacio de distribuciones inducida por la divergencia D, independiente de la parametrización, a la del espacio de parámetros $\boldsymbol{\Omega}$ con su geometría Euclídea usual. Este mecanismo define una métrica *pullback* local independiente a la parametrización $\|\boldsymbol{\omega}\|_{\mathcal{I}}^2 = \boldsymbol{\omega}^T \mathcal{I} \boldsymbol{\omega}$ para $\boldsymbol{\omega}$ en un entorno del vector nulo y donde $\mathcal{I}$ es la métrica local.

En las secciones 5.3 y 5.4 explotaremos esta propiedad para definir una penalización con pesos óptimos quasi-independientes a la parametrización y mostraremos su aplicación en los ejemplos estudiados. Además, en la Sección 6.1 veremos que la curvatura dada por la matriz de información de Fisher permite direccionar al gradiente de la

función objetivo en la dirección de descenso más rápido del espacio de distribuciones, idea que fue implementada al desarrollar un algoritmo de aprendizaje en alta dimensión dada en la Sección 6.3.

4.7.2 *Cálculo de la matriz de información*

En la Definición 4.3, vimos que cuando consideramos la divergencia KL, la matriz de información puede ser evaluada tanto como la esperanza del producto externo del gradiente del score como la esperanza del Hessiano del score. Ésta es la métrica universal en el espacio de distribuciones (Amari, 2016).

En general, considerando (4.3), la métrica local $\mathcal{I}$ puede evaluarse como la esperanza del Hessiano del score:

$$\mathcal{I}(\boldsymbol{\omega}_0) = E_{\boldsymbol{X}}\left[\frac{\partial^2}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T} S(\boldsymbol{X}, f(\boldsymbol{\omega}))\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}\right], \tag{4.41}$$

donde la esperanza es respecto a la distribución $f_{\boldsymbol{X}}(\boldsymbol{\omega}_0)$.

4.7.2.1 *Modelos condicionales*

Al considerar modelos condicionales $f_{\boldsymbol{X}|Y}$, Definición 2.12, y asumiendo una distribución f_Y para la respuesta, la divergencia está dada por (4.32). Considerando la igualdad (4.4) y como f_Y no depende de los parámetros $\boldsymbol{\omega}$, se tiene que

$$\mathcal{I}(\boldsymbol{\omega}_0) = E_Y E_{\boldsymbol{X}|Y}\left[\frac{\partial^2}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T} S(\boldsymbol{X}, f(\boldsymbol{X} \mid \boldsymbol{\omega}, Y))\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}\right], \tag{4.42}$$

donde la esperanza externa es respecto a la distribución f_Y y la interna respecto a la condicional $f_{\boldsymbol{X}|Y}(\boldsymbol{\omega}_0)$.

Al considerar la distribución empírica $\hat{f}_Y$, Martens (2020) encuentra la expresión

$$\mathcal{I}(\boldsymbol{\omega}_0) = \frac{1}{n}\sum_{i=1}^{n} E_{\boldsymbol{X}|Y=y^{(i)}}\left[\frac{\partial^2}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T} S(\boldsymbol{X}, f(\boldsymbol{X} \mid \boldsymbol{\omega}, Y = y^{(i)}))\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}\right]. \tag{4.43}$$

4.8 EJEMPLOS

En los próximos ejemplos suponemos que contamos con una muestra i.i.d. de entrenamiento $\{\boldsymbol{x}^{(i)}\}_{i=1}^n$ y buscamos estimar los parámetros naturales de las distintas familias presentadas en los ejemplos 2.1-2.9. En el caso supervisado, consideramos la muestra aleatoria $\{\boldsymbol{x}^{(i)}, y^{(i)}\}_{i=1}^n$ y el modelo gráfico de segundo orden condicional lineal en $T(\boldsymbol{X})$ (Definición 2.12) correspondiente a cada ejemplo.

Para ello, primero construimos scores (presentados en la Sección 4.3), cuya esperanza define una función de costo convexa que debemos minimizar. Equivalentemente, dicha minimización está caracterizada por un sistema de ecuaciones, lo que define un m-estimador (Definición 4.14). Cuando este sistema de ecuaciones es lineal, es posible encontrar estimadores en forma cerrada.

Continuación del Ejemplo 2.1 (Normal-pGM alias *Normal Multivariada*)**.** *El estimador MLE, Definición 4.3, para el modelo Normal-pGM resulta de minimizar la verosimilitud*

$$\ell_{MLE}(\boldsymbol{\eta}, \boldsymbol{\Theta}) = \log \prod_{i=1}^{n} P(\boldsymbol{X} = \boldsymbol{x}^{(i)}) = \sum_{i=1}^{n} \log P(\boldsymbol{X} = \boldsymbol{x}^{(i)}), \quad (4.44)$$

o equivalentemente, la esperanza muestral del score

$$S_{MLE}(\boldsymbol{x}, (\boldsymbol{\eta}; \boldsymbol{\Theta})) = \boldsymbol{\eta}^\top \boldsymbol{x} + \frac{1}{2} \boldsymbol{x}^\top \boldsymbol{\Theta} \boldsymbol{x} - Z(\boldsymbol{\eta}, \boldsymbol{\Theta}), \quad (4.45)$$

donde $Z(\boldsymbol{\eta}, \boldsymbol{\Theta}) = -\frac{1}{2}\boldsymbol{\eta}^\top \boldsymbol{\Theta}^{-1} \boldsymbol{\eta} - \frac{1}{2} \log(|-\boldsymbol{\Theta}|) + \frac{p}{2} \log(2\pi)$ *es el logaritmo de la función de partición. Es posible obtener una solución en forma cerrada a partir de las ecuaciones de estimación, Apéndice C.3.1, de donde se obtiene que* $\hat{\boldsymbol{\Theta}} = -\hat{\boldsymbol{S}}^{-1}$ *y* $\hat{\boldsymbol{\eta}} = -\hat{\boldsymbol{\Theta}} \hat{E}\boldsymbol{x}$*, y se define la matriz de covarianza muestral como* $\hat{\boldsymbol{S}} = \hat{E}\boldsymbol{x}\boldsymbol{x}^\top - \hat{E}\boldsymbol{x}\hat{E}\boldsymbol{x}^\top$*. Notar que el MLE en la parametrización canónica es* $\hat{\boldsymbol{\mu}} = \hat{E}\boldsymbol{x}$ *y* $\Sigma = \hat{\boldsymbol{S}}$ *y por su propiedad de invarianza a la parametrización, resulta coincidente con el MLE en la parametrización natural de la familia exponencial.*

El estimador PLE, Definición 4.6, para el modelo Normal-pGM se obtiene al minimizar la pseudo-verosimilitud

$$\begin{aligned} \ell_{PLE}(\boldsymbol{\eta}, \boldsymbol{\Theta}) &= \log \prod_{i=1}^{n} \prod_{j=1}^{p} P(X_j = x_j^{(i)} \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{x}_{\setminus j}^{(i)}) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{p} \log P(X_j = x_j^{(i)} \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{x}_{\setminus j}^{(i)}), \end{aligned}$$

o equivalentemente, resultado de la Proposición C.1 detallada en el Apéndice C.1, la esperanza muestral del score

$$S_{PLE}(\boldsymbol{x}, (\boldsymbol{\eta}; \boldsymbol{\Theta})) = \sum_{j=1}^{p} \eta_{j|\setminus j} x_j + \Theta_{jj} x_j^2 - Z(\eta_{j|\setminus j}, \Theta_{jj}), \quad (4.46)$$

donde $\eta_{j|\setminus j}$ *es el parámetro natural de la distribución condicional presentado en (2.5),* y $Z(\eta_{j|\setminus j}, \Theta_{jj}) = -\frac{\eta_{j|\setminus j}^2}{2\theta_{jj}} - \frac{1}{2}\log(-\Theta_{jj}) + \frac{1}{2}\log(2\pi)$ *la función de partición univariada del modelo condicional.*

El estimador SME, Proposición 4.2, para el modelo Normal-pGM se obtiene de minimizar la esperanza muestral del score obtenido al evaluar (4.20) para la densidad $g(x)$ *dada en (2.7), lo que resulta*

$$S_{SME}(\boldsymbol{x}, (\boldsymbol{\eta}; \boldsymbol{\Theta})) = tr(\boldsymbol{\Theta}) + \frac{1}{2}\|\boldsymbol{\eta} + \boldsymbol{\Theta}\boldsymbol{x}\|^2. \quad (4.47)$$

Hyvärinen (2005) muestra que dicha estimación coincide con el MLE para esta familia paramétrica. Además, si se considera el modelo condicional asociado, Definición 2.12, el SME y el MLE están dados por $\hat{\boldsymbol{\Theta}} = -S_{\boldsymbol{x}|\boldsymbol{f}_y}^{-1}$, $\hat{\boldsymbol{\Gamma}} = -\boldsymbol{\Theta} S_{\boldsymbol{x}\boldsymbol{f}_y} S_{\boldsymbol{f}_y}^{-1}$ *y* $\hat{\boldsymbol{\eta}}_0 = -\boldsymbol{\Theta}\hat{E}\boldsymbol{x}$*, donde* $S_{\boldsymbol{x}|\boldsymbol{f}_y} = S_n - S_{\boldsymbol{x}\boldsymbol{f}_y} S_{\boldsymbol{f}_y}^{-1} S_{\boldsymbol{f}_y \boldsymbol{x}}$ *es la matriz de convarianza muestal condicional, con* $S_{\boldsymbol{x}\boldsymbol{f}_y} = \hat{E}(\boldsymbol{x} - \hat{E})(\boldsymbol{f}_y - \hat{E}\boldsymbol{f}_y)^T$ *y* $S_{\boldsymbol{f}_y} = \hat{E}(\boldsymbol{f}_y - \hat{E}\boldsymbol{f}_y)(\boldsymbol{f}_y - \hat{E}\boldsymbol{f}_y)^T$*. La condición de Sylvester asegura unicidad en la solución.*

En el Apéndice C.3 se derivan los resultados presentados en el ejemplo.

Continuación del Ejemplo 2.2 (Binary-pGM alias *Ising*)**.** *En este caso, el estimador MLE es intratable, por lo que consideramos el estimador PLE que se obtiene de minimimzar la esperanza muestral del score de composición inducido por la pseudo-verosimilitud:*

$$S_{PLE}(\boldsymbol{x}, (\boldsymbol{\eta}; \boldsymbol{\Theta})) = \sum_{j=1}^{p} \eta_{j|\setminus j} x_j - Z(\eta_{j|\setminus j}), \tag{4.48}$$

donde $\eta_{j|\setminus j}$ es el parámetro natural del modelo condicional, dado en (2.5), y $Z(\eta_{j|\setminus j}) = \log(1 + \exp\{\eta_{j|\setminus j}\})$ la función de partición univariada del modelo Bernoulli condicional.

Por otro lado, al considerar el score de Brier $S(x, \boldsymbol{p}) = \|\boldsymbol{p}\|^2 - 2\boldsymbol{p}^\top \boldsymbol{e}_x$, donde $\boldsymbol{p} \in \mathcal{P}^2$ está en el simplex de 2 componentes de acuerdo con la representación de Savage (Definición 4.10) sobre los eventos condicionales $\boldsymbol{X}_j \mid \boldsymbol{X}_{\setminus j}$, es posible construir un score multivariado para espacios producto (Teorema 4.2) a partir de (4.25), obteniéndose el score

$$S_{RM}(\boldsymbol{x}, (\boldsymbol{\eta}; \boldsymbol{\Theta})) = \sum_{j=1}^{p} (x_j - s(\eta_j + \boldsymbol{\Phi}_j \boldsymbol{x}))^2, \tag{4.49}$$

donde $\boldsymbol{\Phi}_j \in \mathbb{R}^{1\times p}$ es la j-ésima fila de la matriz $\boldsymbol{\Phi}$ y $s(x) = (1 + \exp\{-x\})^{-1}$ es la función logística.

Minimizando la esperanza muestral del score (4.49), recuperamos el Estimador de Ratio Matching (RME) (Hyvärinen, 2007).

Continuación del Ejemplo 2.3 (Poisson-pGM)**.** *El estimador PLE se obtiene de minimimzar la esperanza muestral del score de composición inducido por la pseudo-verosimilitud (Proposición C.1):*

$$S_{PLE}(\boldsymbol{x}, (\boldsymbol{\eta}; \boldsymbol{\Theta})) = \sum_{j=1}^{p} \eta_{j|\setminus j} x_j - \log x_j! - Z(\eta_{j|\setminus j}), \tag{4.50}$$

donde $\eta_{j|\setminus j} = \eta_j + \boldsymbol{\Theta}_j \boldsymbol{x}$ es el parámetro natural del modelo condicional $X_j \mid \boldsymbol{X}_{\setminus j}$, ecuación (2.5), y $\log Z(\eta_{j|\setminus j}) = \exp\{\eta_{j|\setminus j}\}$ la función de partición univariada del modelo Poisson condicional.

Por otro lado, la extensión multivariada (4.25) dada por el score de composición (Teorema 4.2), considerando cada score S_j definido a partir de la entropía $G_x(v) = -(x+1)^a \frac{v^m}{m(m-1)}$ con $m = a = 2$, correspondiente al Ejemplo 4.1, permite obtener el siguiente score local para el modelo Poisson-pGM (ver detalles en el Apéndice C.4):

$$S(\boldsymbol{x}, (\boldsymbol{\eta}; \boldsymbol{\Theta})) = \sum_{j=1}^{p} \left(\frac{\exp\{\eta_{j|\setminus j}\}}{2} - x_j \right) \exp\{\eta_{j|\setminus j}\}. \tag{4.51}$$

Contribución original

Continuación del Ejemplo 2.4 (TPoisson-pGM). *En este caso, el estimador PLE coincide con* (4.50)*, donde* $\exp Z(\eta_{j|\backslash j}) = \sum_{x=0}^{T^*} \exp\{\eta_{j|\backslash j} x - \log x!\}$*. Si definimos una variable auxiliar* $\tilde{X} \sim$ *Poisson*$(\exp\{\eta_{j|\backslash j}\})$*,*

$$
\begin{aligned}
Z(\eta_{j|\backslash j}) &= \log \sum_{x=0}^{T^*} \frac{\exp\{\eta_{j|\backslash j} x\}}{x!} && (4.52)\\
&= \log \left(\sum_{x=0}^{\infty} \frac{\exp\{\eta_{j|\backslash j} x\}}{x!} - \sum_{T^*+1}^{\infty} \frac{\exp\{\eta_{j|\backslash j} x\}}{x!} \right) && (4.53)\\
&= \log \left(\exp\exp\{\eta_{j|\backslash j}\} - \sum_{T^*+1}^{\infty} \frac{\exp\{\eta_{j|\backslash j} x\}}{x!} \right) && (4.54)\\
&= \exp\{\eta_{j|\backslash j}\} - P(\tilde{X} \leq T^*), && (4.55)
\end{aligned}
$$

y este último término puede evaluarse como $\Gamma(T^*+1, \exp\{\eta_{j|\backslash j}\})/\Gamma(T^*+1)$ *y donde* $\Gamma(\cdot,\cdot)$ *es la función gamma incompleta.*

Por otro lado, restringiendo el dominio a $x \in \{0, \cdots, T^*\}$*, en el Ejemplo 4.1, el score inducido por* G_x *resulta en el siguiente score local:*

$$
S_x = G'_{x-1}(\rho_{x-1}) I(0 < x) + \left(G_x(\rho_x) - \rho_x G'_x(\rho_x) \right) I(x < T^*), \quad (4.56)
$$

donde $I(\cdot)$ *es la función indeicadora. Considerando la entropía* G_x *como en el ejemplo anterior, obtenemos el siguiente score de composición a partir del Teorema 4.2:*

$$
S(\boldsymbol{x}, (\boldsymbol{\eta}; \boldsymbol{\Theta})) = \sum_{j=1}^{p} \left(\eta_{j|\backslash j} x_j + \frac{1}{2} \eta_{j|\backslash j} I(x_j < T^*) \right). \quad (4.57)
$$

Continuación del Ejemplo 2.5 (FPoisson-pGM$_{\backslash k}$). *El estimador PLE que se obtiene de minimimzar la esperanza muestral del score de composición inducido por la pseudo-verosimilitud sobre el modelo reducido:*

$$
S_{PLE}(\boldsymbol{x}, (\tilde{\boldsymbol{\eta}}; \tilde{\boldsymbol{\Theta}})) = \sum_{j \neq k} \tilde{\eta}_{j|\backslash j} x + \frac{\tilde{\Theta}_{jj}}{2} x^2 - \log(m_{\boldsymbol{x}|\boldsymbol{x}_{\backslash j}}!) - Z(m_{\boldsymbol{x}|\boldsymbol{x}_{\backslash j}}, \tilde{\eta}_{j|\backslash j}, \tilde{\Theta}_{jj}), \quad (4.58)
$$

donde

$$
\begin{aligned}
m_{\boldsymbol{x}|\boldsymbol{x}_{\backslash j}} &= m_{\boldsymbol{x}} - \sum_{l \neq j, l} x_l, && (4.59)\\
\tilde{\eta}_{j|\backslash j} &= \tilde{\eta}_j + \sum_{l \neq j} \tilde{\Theta}_{jl} x_l. && (4.60)
\end{aligned}
$$

La función de partición del modelo condicional resulta

$$
\log Z(m_{\boldsymbol{x}|\boldsymbol{x}_{\backslash j}}, \tilde{\eta}_{j|\backslash j}, \tilde{\Theta}_{jj}) = \sum_{x=0}^{m_{\boldsymbol{x}|\boldsymbol{x}_{\backslash j}}} \exp\{\tilde{\eta}_{j|\backslash j} x + \frac{\tilde{\Theta}_{jj}}{2} x^2 - \log(m_{\boldsymbol{x}|\boldsymbol{x}_{\backslash j}}!)\}, \quad (4.61)
$$

la cual puede evaluarse como una suma cuadrática generalizada de Gauss. Ver Proposición A.2 en el Apéndice A.3.

Continuación del Ejemplo 2.6 (sqPoisson-pGM). *En este caso, donde las distribuciones condicionales* (2.13) *no son Poisson en general, no existe una*

forma cerrada para computar la función de partición del modelo condicional $X_j \mid \boldsymbol{X}_{\setminus j}$ *y por lo tanto, el cómputo del estimador PLE no resulta eficiente. En su lugar, consideramos un estimador basado en el score propio definido en el Ejemplo 4.1 para* $m = a = 2$*. Para ello consideramos el Teorema 4.2 y construimos el siguiente score multivariado a partir de (4.25):*

$$S(\boldsymbol{x}, (\boldsymbol{\eta}; \boldsymbol{\Theta})) = \sum_{j=1}^{p} -x_j \exp\left\{\Theta_{jj} + \eta_{j\setminus j}\left(\sqrt{x_j} - \sqrt{x_j - 1}\right)\right\} + \frac{1}{2}\exp\left\{2\Theta_{jj} + 2\eta_{j\setminus j}\left(\sqrt{x_j + 1} - \sqrt{x_j}\right)\right\} \quad (4.62)$$

donde $\eta_{j|\setminus j} = \eta_j + 2\boldsymbol{\Theta}_{j\setminus j}\sqrt{\boldsymbol{x}_{\setminus j}}$ *es un parámetro natural del modelo condicional* $X_j \mid \boldsymbol{X}_{\setminus j}$*.*

Continuación del Ejemplo 2.7 (Normal-zipGM)**.** *Como consecuencia del Teorema 2.4, el estimador PLE se obtiene de minimizar la esperanza empírica del score*

$$S_{PLE}(\boldsymbol{x}, (\boldsymbol{\eta}, \boldsymbol{\psi}, \boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Lambda})) = \sum_{j=1}^{p} \xi_{j|\setminus j}\nu(x_j) + \eta_{j|\setminus j}x_j + \frac{1}{2}\Theta_{jj}x_j^2 - Z(\xi_{j|\setminus j}, \eta_{j|\setminus j}, \Theta_{jj}), \quad (4.63)$$

donde $\eta_{j|\setminus j}$ *y* $\xi_{j|\setminus j}$ *son parámetros condicionales de* $X_j \mid \boldsymbol{X}_{\setminus j}$ *y se obtienen evaluando (2.21) y (2.20) respectivamente. La función de partición* $Z(\xi_{j|\setminus j}, \eta_{j|\setminus j}, \Theta_{jj})$ *se evalúa a través de (2.19) siendo* $Z^+(\eta_{j|\setminus j}, \Theta_{jj})$ *la función de partición del modelo condicional* $X_j \mid \boldsymbol{X}_{\setminus j}, \nu(X_j) = 1$*, definida en la ecuación (2.24).*

Continuación del Ejemplo 2.8 (Poisson-zipGM)**.** *El estimador PLE resulta de minimizar la esperanza empírica del score*

$$S_{PLE}(\boldsymbol{x}, (\boldsymbol{\eta}, \boldsymbol{\psi}, \boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Lambda})) = \sum_{j=1}^{p} \xi_{j|\setminus j}\nu(x_j) + \eta_{j|\setminus j}x_j - \log x_j! - Z(\xi_{j|\setminus j}, \eta_{j|\setminus j}), \quad (4.64)$$

como consecuencia del Teorema 2.4. Nuevamente, la función de partición $Z(\xi_{j|\setminus j}, \eta_{j|\setminus j})$ *se evalúa a través de (2.19), siendo* $Z^+(\eta_{j|\setminus j})$ *la función de partición del modelo condicional* $X_j \mid \boldsymbol{X}_{\setminus j}, \nu(X_j) = 1$*, que corresponde a la ecuación (2.26).*

Continuación del Ejemplo 2.9 (TPoisson-zipGM)**.** *El estimador PLE para el modelo coincide con (4.64) y donde* $Z^+(\eta_{j|\setminus j})$*, la función de partición del modelo condicional* $X_j \mid \boldsymbol{X}_{\setminus j}, \nu(X_j) = 1$*, se obtiene evaluando (2.27).*

4.9 CONTRIBUCIONES

La principal contribución del capítulo, además de definir estimadores eficientes para los distintos modelos considerados, es la definición de la métrica local basada en la noción de scores y su cómputo en problemas condicionales. Este resultado nos permitirá en el próximo capítulo definir penalizaciones pesadas por la métrica inducida.

5

ESTIMACIÓN EN ALTA DIMENSIÓN

Llamamos de *alta dimensión* al problema por el cual el número de variables p es del orden, o mayor al número de muestras n. En este contexto, existe un conjunto de soluciones $\hat{\omega}$ al problema de estimación (4.2), las cuales sobreajustan a los datos, perdiendo generalización e interpretabilidad. Una solución ampliamente adoptada se basa en inducir ceros en los coeficientes que parametrizan al modelo y parte de dos premisas fundamentales (Hastie, Tibshirani y Wainwright, 2015):

1. El compromiso *sesgo-varianza* por el cual un estimador sesgado, pero con menor varianza, puede ser preferible. Un ejemplo se encuentra en predicción, donde se busca una generalización del modelo que evite el sobreajuste a los datos de entrenamiento.

2. Interpretabilidad, un modelo con pocos parámetros no nulos permite interpretarlo más fácilmente. La estructura inducida permite por ejemplo seleccionar variables o estimar un grafo de independencias condicionales.

En lo que sigue, introducimos el estimador Least Absolute Shrinkage and Selection Operator, por sus siglas en inglés (LASSO) para modelos lineales generalizados y veremos su aplicación en la estimación de grafos de independencia condicional en el contexto de MGs Sección 5.1). Seguidamente, en la Sección 5.2, incorporamos mayor estructura al modelo mediante penalización jerárquica, permitiendo por ejemplo el modelado de interacciones de orden superior sólo cuando las interacciones de menor grado están presentes. Basándonos en las condiciones de optimalidad inducidas por el problema de optimización, en la Sección 5.3 analizamos penalizaciones pesadas por grupos. Observamos que la norma ℓ_2 pesada por la matriz de información de Fisher logra calibrar la fuerza de la penalización sobre cada grupo de parámetros de manera tal que cada uno de ellos tenga la misma probabilidad de hacerse cero bajo el modelo independiente. En la Sección 5.4 definimos una penalización jerárquica para el problema de RSD cuando los predictores siguen un modelo Hurdle multivariado (Definición 2.10) y la reducción está dada por los Corolarios 3.2 y 3.3, permitiendo interacciones únicamente entre las variables relacionadas con la respuesta. Más adelante, en el Capítulo 6 estudiaremos algoritmos que permiten optimizar el problema inducido por esta penalización y en la Parte IV estudiaremos su comportamiento mediante simulaciones y datos reales. Finalmente, en la Sección 5.5 estudiamos métodos que permiten seleccionar el parámetro de penalización óptimo en el contexto de RSD.

5.1 LASSO

Al penalizar la función objetivo con la norma ℓ_1, el estimador LASSO induce ceros en algunos parámetros penalizados, permitiendo mayor interpretación y generalización en contextos de alta dimensión. Comenzaremos analizando su aplicación en modelos lineales, modelos lineales generalizados y en modelos gráficos.

Asumiendo un modelo lineal

$$y^{(i)} = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}^{(i)} + \varepsilon^{(i)}, \tag{5.1}$$

donde β_0 y $\boldsymbol{\beta} = (\beta_1; \cdots; \beta_p)$ son parámetros desconocidos y $\varepsilon^{(i)}$ es un término de error, el método de mínimos cuadrados permite estimar los parámetros minimizando la función objetivo

$$\ell_{\text{LS}}(\beta_0, \boldsymbol{\beta}) = \hat{E}\left(Y - \beta_0 - \boldsymbol{\beta}\boldsymbol{X}\right)^2. \tag{5.2}$$

En general ninguno de los coeficientes de $\hat{\boldsymbol{\beta}}$ es nulo, dificultando la interpretación del modelo final cuando p es grande. Más aún, si $p > n$, el estimador de mínimos cuadrados no es único. El estimador Least Absolute Shrinkage and Selection Operator, por sus siglas en inglés (LASSO), al regularizar la función de costo (5.2) con la norma ℓ_1 de los parámetros, minimiza ℓ_{LS} sujeto a que $\|\boldsymbol{\beta}\|_1 \leq t$, donde t es un parámetro de regularización. La función de costo dual o Lagrangiano del problema convexo (Definición D.3 en el Apéndice D.1) resulta $\ell_{\text{LS}}(\beta_0, \boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$. Minimizando dicha función se inducen ceros en $\hat{\boldsymbol{\beta}}$, seleccionando un conjunto de variables $\boldsymbol{X}_{\hat{S}} = \{X_j : \hat{\beta}_j \neq 0\}_{j=1}^p$. Además, respecto al valor poblacional $\boldsymbol{\beta}^*$, la estimación $\hat{\boldsymbol{\beta}}$ satisface la cota de error en predicción $\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2/n \lesssim \|\boldsymbol{\beta}^*\|_1\sqrt{\log(p)/n}$ para $n, p \to \infty$, con lo cual si $\|\boldsymbol{\beta}^*\|_1$ es del orden $\sqrt{n/log(p)}$, entonces $\boldsymbol{\beta}^*$ es sparse relativo al cociente $n/\log(p)$, por lo que LASSO resulta consistente en predicción bajo ciertas condiciones.

Notar que la penalización $\sum_{j=1}^p |\beta_j|^q$ resulta convexa para $q \geq 1$, correspondiente a la norma ℓ_q; mientras que para $q \leq 1$ esta penalización induce modelos sparse, es por eso que la norma ℓ_1 es ampliamente aceptada como penalizante cuando el objetivo es inducir modelos interpretables con pocos parámetros no nulos.

Cuando la respuesta y no es cuantitativa, pudiendo ser por ejemplo binaria o categórica, es necesaria una generalización del modelo (5.1), dando lugar a los modelos lineales generalizados, los cuales modelan una respuesta Y mediante una distribución en la FE. En particular, se considera que la función monótona g de los parámetros naturales es lineal en los predictores $\boldsymbol{X}$, quedando definido el modelo lineal $P(Y = y \mid \boldsymbol{X} = \boldsymbol{x}^{(i)}) = \exp\{g^{-1}(\beta_0 + \boldsymbol{\beta}\boldsymbol{x}^{(y)})y + h(y) - Z(\beta_0, \boldsymbol{\beta}, \boldsymbol{x}^{(i)})\}$, En este caso, LASSO considera la función de costo ℓ inducida por MLE aumentada por el término de penalización $\|\boldsymbol{\beta}\|_1$. Notar que cuando g es la identidad y $Y \mid \boldsymbol{X} \sim \mathcal{N}((\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}))/\sigma^2, \sigma^2)$ recuperamos (5.2), un modelo lineal generalizado Normal. Otros casos particulares incluyen el modelo lineal generalizado Bernoulli y Poisson.

Otra generalización se obtiene al considerar J grupos disjuntos de predictores $\boldsymbol{Z}_j \in \mathbb{R}_j^p$, $j = 1, \cdots J$ con $\sum_{j=1}^J p_j = p$ en un modelo de

regresión lineal o lineal generalizado donde $\boldsymbol{\beta}_j$ corresponde a los coeficientes de regresión del grupo j. En este caso, se considera la regularización de la norma $\ell_{1,2}$ dada por $\sum_{j=1}^{J} \|\boldsymbol{\beta}_j\|_2$ y el estimador es conocido como LASSO por grupos. Además Yuan y Lin (2006) recomiendan pesar los grupos de acuerdo al tamaño como $\sum_{j=1}^{J} \sqrt{p_j}\|\boldsymbol{\beta}_j\|_2$, aunque los pesos sólo son óptimos en el caso $(\boldsymbol{Z}_j^{(i)})_{j=1}^{n}$ ortonormales.

En el contexto de modelos gráficos de segundo orden (Definición 2.5) el Teorema 2.1 nos dice que los ceros de la matriz de interacciones definen un grafo de independencia condicional entre las variables. Dichos conjuntos de variables condicionalmente independientes pueden ser inducidos penalizando por la norma ℓ_1 el conjunto de interacciones entre cada par de variables.

Más específicamente, para el modelo Normal del ejemplo 2.1, si penalizamos ℓ_{MLE}, dada en (4.44), con la norma ℓ_1 de las interacciones, i. e. consideramos el problema regularizado $\ell_{\text{MLE}}(\boldsymbol{\eta}, \boldsymbol{\Theta}) + \lambda \sum_{j \neq l} |\Theta_{jl}|$, la solución del problema convexo resultante de minimizar la función de costo regularizada se conoce como el estimador *graphical lasso* (Hastie, Tibshirani y Wainwright, 2015). Además, usando $\lambda = 2\sqrt{\log(p)/n}$, el estimador satisface con alta probabilidad la cota de error $\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_2 \lesssim \sqrt{d^2 \log(p)/n}$, donde d es el grado máximo de los nodos del grafo de independencia condicional, lo que permite controlar el error en predicción.

5.1.1 *Ejemplo pGM*

Continuando con los ejemplos 2.1-2.5, consideramos los modelos gráficos de segundo orden condicional lineal en $T(\boldsymbol{X})$ (Definición 2.12) que llamaremos $f(\boldsymbol{X} \mid Y, \boldsymbol{\omega})$, con el fin de estimar la reducción de dimensiones dada por el Corolario 3.1. Para su estimación en contextos de alta dimensión, consideramos una penalización basada en Lin, Drton y Shojaie (2016), quienes estudiaron el efecto de penalizar las interacciones $\boldsymbol{\Theta}$ del modelo por la norma ℓ_1 la función de costo correspondiente al Estimador de Score Matching (SME). Además, Tomassi et al. (2019) consideraron la norma compuesta $\ell_{1,2}$ sobre cada fila de la matriz de regresión $\boldsymbol{\Gamma}$. De esta manera, es posible inducir la selección de variables en la reducción, y conjuntos de independencia condicional de a pares sobre las variables descriptivas $\boldsymbol{X}$. La función de costo a minimizar está dada por:

$$\hat{ES}(\boldsymbol{X}, f(\boldsymbol{X} \mid Y, \boldsymbol{\omega})) + \lambda_{\boldsymbol{\Gamma}} \sum_{j=1}^{k} \|\boldsymbol{\Gamma}_j\|_2 + \lambda_{\boldsymbol{\Theta}} \sum_{j=1}^{k} \sum_{l \neq j} |\Theta_{jl}|. \tag{5.3}$$

donde $\boldsymbol{\Gamma}_j$ corresponde a la fila j-ésima de la matriz de regresión $\boldsymbol{\Gamma}$ definida en (2.28), correspondiente a los coeficientes de la contribución lineal de Y en los potenciales lineales de X_j y donde $\hat{ES}(\boldsymbol{X}, f(\boldsymbol{X} \mid Y, \boldsymbol{\omega}))$ es la esperanza empírica de un score S propio.

5.2 PENALIZACIÓN JERÁRQUICA

Otra variante convexa de LASSO por grupos se obtiene de considerar grupos solapados de manera jerárquica, por ejemplo permitiendo interacciones solamente cuando los efectos lineales son seleccionados.

Zhao, Rocha y Yu (2009) definen la familia de penalización Composite Absolute Penalty (CAP) que permite incluir variables en el modelo en un orden dado. Por ejemplo, si X_1 debe ser incluida antes que X_2, debemos definir dos grupos: $\mathcal{G}_1 = \{1, 2\}$ y $\mathcal{G}_2 = \{2\}$, recuperando la penalización $\mathbb{T}(\boldsymbol{\beta}) = \|(\beta_1; \beta_2)\|_\gamma + \|\beta_2\|_1$ con $\gamma \geq 1$. Cuando $\gamma > 1$ se inducen ceros en β_2 pero no en β_1, cuando $\gamma = 1$, es menos probable que se agregue X_2 al modelo, ya que está más penalizada. Esto se sigue de las condiciones KKT (Teorema D.2 en el Apéndice D.1) para el problema de minimización de una función de costo convexa $\mathbb{J}(\beta)$ penalizada por $\mathbb{T}(\boldsymbol{\beta})$ (Proposición D.9 en el Apéndice D.2), por las cuales,

$$\begin{cases} \dfrac{\partial}{\partial \beta_j}\mathbb{J}(\boldsymbol{\beta}) = -\dfrac{\partial}{\partial \beta_j}\mathbb{T}(\boldsymbol{\beta}), & \text{si } \beta_j \neq 0 \qquad (5.4a) \\ \left|\dfrac{\partial}{\partial \beta_j}\mathbb{J}(\boldsymbol{\beta})\right| \leq \left|\dfrac{\partial}{\partial \beta_j}\mathbb{T}(\boldsymbol{\beta})\right|, & \text{si } \beta_j = 0, \qquad (5.4b) \end{cases}$$

en particular para $\gamma > 1$,

$$\frac{\partial}{\partial \beta_1}\mathbb{T}(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_1}\|\boldsymbol{\beta}\|_\gamma = \operatorname{sign}(\beta_1)\left(\frac{|\beta_1|}{\|\boldsymbol{\beta}\|_\gamma}\right)^{(\gamma-1)}.$$

Cuando $\beta_2 > 0$ y $\gamma > 1$, β_1 está localmente no penalizada en el origen y sólo tomará este valor si el gradiente de $\mathbb{J}$ es nulo en dicho punto, por lo que con probabilidad 1 vale la condición (5.4a) para $j = 1$. Por otro lado, si $\beta_2 = 0$, $\frac{\partial}{\partial \beta_1} = \operatorname{sign}(\beta_1)$ y el lado derecho de (5.4b) se vuelve constante para β_1 y por lo tanto, si el coeficiente contribuye menos que un cierto umbral en la reducción del costo, se vuelve cero.

5.3 PENALIZACIÓN CON PESOS ÓPTIMOS

Lee y Hastie (2015) observa que una forma de calibrar los pesos en LASSO por grupos de forma equitativa resulta de que cada grupo tenga la misma posibilidad de ser no nulo bajo el modelo independiente donde $\boldsymbol{\beta}_j = \mathbf{0}\, j = 1, \cdots, J$ que llamaremos $\boldsymbol{\omega}_0$. Partiendo de las condiciones de optimalidad KKT se observa que el grupo de parámetros $\boldsymbol{\beta}_j$ es no nulo si (Proposición D.9 en el Apéndice D.2)

$$\left\|\frac{\partial \ell}{\partial \boldsymbol{\beta}_j}\right\| > \lambda w_j,$$

con ω_j el peso dado al grupo j. Asumiendo el modelo independiente $\boldsymbol{\omega}_0$, pueden existir grupos que tengan más posibilidades de ser seleccionados, por ejemplo aquellos de mayor tamaño p_j con $\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}$. Entonces la calibración propuesta busca los pesos w_j tales que

$$E_{\boldsymbol{\omega}_0}\left\|\frac{\partial \ell}{\partial \boldsymbol{\beta}_j}\right\| = c w_j \qquad \forall j = 1, \cdots, J,$$

para una constante $c > 0$ y donde la esperanza se toma respecto al modelo independiente.

Por otro lado, McDavid et al. (2019) propone otra metodología para balancear la penalización por grupos por la cual se considera una norma anisométrica o norma $\ell_{1,2}$ pesada sobre cada grupo de parámetros, $\sum_{j=1}^{J}(\boldsymbol{\beta}_j^\top H_j \boldsymbol{\beta}_j)^{1/2}$ o equivalentemente, $\sum_{j=1}^{J} \|\boldsymbol{\beta}_j\|_{H_j}$, donde $H = \operatorname{diag}(H_1, \cdots, H_J)$ es una matriz diagonal por bloques definida positiva que permite a su vez acomodar la escala de los predictores como la correlación entre los componentes de $\boldsymbol{\beta}_j$. Tomando H como la matriz de Fisher bajo el modelo independiente $\mathcal{I}(\boldsymbol{\omega}^0)$ (ver Definición 4.3), esta da como resultado una selección de parámetros que aproxima un test estadístico: A partir de las condiciones KKT se observa que $\boldsymbol{\beta}_j$ es nulo si (Proposición D.9 en el Apéndice D.2)

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}_j^T}[\mathcal{I}(\boldsymbol{\omega}^0)_j]^{-1}\frac{\partial \ell}{\partial \boldsymbol{\beta}_j} = \left\|\frac{\partial \ell}{\partial \boldsymbol{\beta}_j}\right\|^2_{[\mathcal{I}(\boldsymbol{\omega}^0)_j]^{-1}} < \lambda^2, \tag{5.5}$$

donde ℓ es la log-likelihood de una variable dadas las demás, $\mathcal{I}(\boldsymbol{\omega}^0)_j$ es el bloque correspondiente al parámetro $\boldsymbol{\beta}_j$ de la matriz de información de Fisher $\mathcal{I}(\boldsymbol{\omega}^0)$ y $\boldsymbol{\beta}_j$ es un grupo de parámetros. Observando que el lado izquierdo de (5.5) se distribuye de acuerdo con una χ^2 con $\dim\left(\mathcal{I}(\boldsymbol{\omega}^0)_j\right)$ grados de libertad, los autores concluyen que esta penalización aproxima el siguiente test de hipótesis para cada $j = 1, \cdots, J$:

$$\begin{cases} H_0: & \boldsymbol{\omega} = \boldsymbol{\omega}_0, \\ H_1: & \boldsymbol{\beta}_j \neq \mathbf{0}. \end{cases}$$

En la siguiente sección derivaremos una generalización jerárquica de esta penalización en el contexto de reducción de dimensiones.

5.4 EJEMPLO ZIPGM

Continuando con los ejemplos 2.7-2.9, consideramos los modelos gráficos de segundo orden condicional lineal en $\boldsymbol{T}(\boldsymbol{X})$ (Definición 2.12) que satisfacen (3.10) para estimar la reducción de dimensiones dadas por las Proposiciones 3.2 y 3.3.

Además, para permitir su estimación en alta dimensión, implementamos una penalización jerárquica con pesos óptimos dados por la matriz de información (4.43). Esto permite simplificar el problema computacional, como también brindar mayor interpretación al modelo resultante.

5.4.1 *Penalización jerárquica en el contexto de RSD*

Contribución original: penalización jerárquica en el contexto de RSD

Se propone una penalización jerárquica que modela únicamente las interacciones entre variables relacionadas con la respuesta, de este modo se evita el problema de estabilidad en la selección de variables por LASSO por el cual, el estimador no puede elegir de manera estable entre dos variables que se encuentren muy correlacionadas (Xu, Caramanis y Mannor, 2011).

Considerando $f(\boldsymbol{X} \mid Y, \boldsymbol{\omega})$ la densidad o función de probabilidad condicional de los miembros de la familia paramétrica zipGM (Definición 2.10) junto con (3.10) y la función de costo dada por la esperanza empírica de un score propio $\hat{ES}(\boldsymbol{X}, f(\boldsymbol{X} \mid Y, \boldsymbol{\omega}))$, sumada a la penalización jerárquica propuesta, resulta en la siguiente expresión:

$$\hat{ES}(\boldsymbol{X}, f(\boldsymbol{X} \mid Y, \boldsymbol{\omega})) + \lambda_{\mathcal{R}} \sum_{j=1}^{k} \|(\boldsymbol{\Gamma}_j, \boldsymbol{\Psi}_j, (\Theta_{jl}, \Phi_{lj}, \Phi_{jl}, \Lambda_{jl})_{l \neq j})\|_{\mathcal{R}_j} + \lambda_{\mathcal{C}} \sum_{j=1}^{k} \sum_{l \neq j} \|(\Theta_{jl}, \Phi_{lj}, \Phi_{jl}, \Lambda_{jl})\|_{\mathcal{C}_{jl}}, \quad (5.6)$$

donde los vectores $\boldsymbol{\Gamma}_j$ y $\boldsymbol{\Psi}_j$ corresponden a la fila j-ésima de las matrices $\boldsymbol{\Gamma}$ y $\boldsymbol{\Psi}$ definidas en (3.10a) and (3.10b) respectivamente, las cuales corresponden a los coeficientes de la contribución lineal de Y en los potenciales lineales de $T(X_j)$.

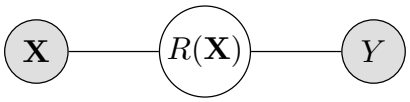
Notar que se introdujo la norma ℓ_2 pesada dada por $\|\boldsymbol{b}\|_{\mathcal{R}_j} = \sqrt{\boldsymbol{b}^T \mathcal{R}_j \boldsymbol{b}}$ y donde $\mathcal{R}_j$ es una matriz simétrica definida positiva de dimsión $2r + 4(p-1) \times 2r + 4(p-1)$; de manera similar, se definió la norma $\|\boldsymbol{b}\|_{\mathcal{C}_{jl}}$ al considerar la matriz de pesos $\mathcal{C}_{jl} \in \mathbb{R}^{4\times 4}$.

En la figura 5.1 se muestra el efecto de la penalización. En particular en la Figura 5.1a se muestra el diagrama de suficiencia, por el cual podemos asegurar que $R(\boldsymbol{X})$ es una reducción suficiente de dimensiones, mientras que en la figura 5.1b se desarrolla el modelo, donde cada arista representa una interacción. Al penalizar con el tercer término de (5.6), se inducen conjuntos de independencia condicional. como se muestra en la Figura 5.1c, mientras que el segundo término de (5.6) induce variables independientes de la respuesta, como se observa en la Figura 5.1d.

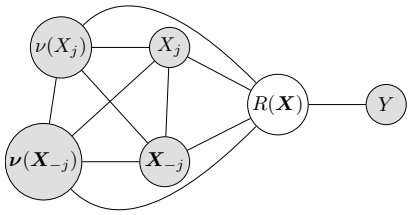
El efecto conjunto de la penalización resulta en la selección de las variables correlacionadas con la respuesta, así como también las que están fuertemente correlacionadas entre sí, pudiendo resultar en falsos positivos. Este compromiso permite obtener un problema más estructurado, con menor grados de libertad y más estable. Mediante un proceso posterior es posible eliminar los falsos positivos en un problema de baja dimensionalidad. Notar que a medida que se consideran sucesiones crecientes en $\lambda_{\mathcal{C}}$ se obtienen modelos con menos interacciones o equivalentemente más cercanos a la independencia entre las variables. Como consecuencia, el conjunto de variables seleccionadas es menos sensible a las interacciones y así menos propenso a agregar falsos positivos, mientras que sucesiones crecientes de $\lambda_{\mathcal{R}}$ seleccionan menos variables vinculadas a la respuesta. La selección de los parámetros óptimos de regularización $\lambda_{\mathcal{R}}^*, \lambda_{\mathcal{C}}^*$ será tratada en la Sección 5.5.

Contribución original: penalización jerárquica con pesos óptimos

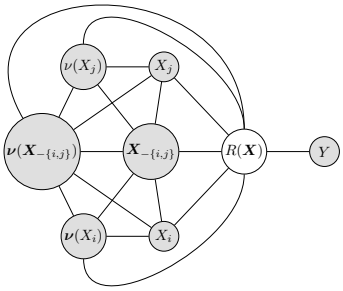
PENALIZACIÓN CON PESOS ÓPTIMOS Las matrices $\mathcal{R}_j$ y $\mathcal{C}_{jl}$ usadas en (5.6) son bloques o submatrices principales de la matriz de información de Fisher $\mathcal{I}(\boldsymbol{\omega}^0)$ (Definición 4.3) calculadas asumiendo el modelo completamente factorizado o independiente, donde se asume independencia entre $\boldsymbol{X}$ e Y así como también entre to-



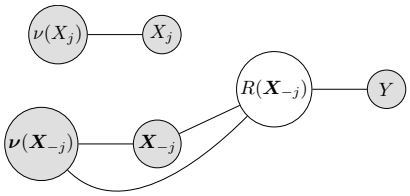
(a) SDR: $Y \perp\!\!\!\perp \boldsymbol{X} \mid R(\boldsymbol{X})$.



(b) zipGM denso (grafo completo)



(c) Penalizando los grupos $(\Theta_{ji}; \Lambda_{ji}; \Phi_{ji}; \Phi_{ij})$ se induce independencia condicional $X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{-\{i,j\}}, R(\boldsymbol{X})$.



(d) Penalizando los grupos $(\gamma_j, \boldsymbol{\psi}_j, (\Theta_{ji}; \Lambda_{ji}; \Phi_{ji}; \Phi_{ij})_{i \neq j})$ se induce independencia $Y \perp\!\!\!\perp X_j$, $R(\boldsymbol{X}) \perp\!\!\!\perp X_j$ and $\boldsymbol{X}_{-j} \perp\!\!\!\perp X_j$.

Figura 5.1: Relaciones de independencia condicional dado por la SDR $R(X)$ obtenida por la familia zipGM y la estructura inducida por la penalización jerárquica.

das las componentes de $\boldsymbol{X}$, i. e. corresponde al modelo (2.16) con $\Gamma_{ij} = \Psi_{ij} = \Theta_{ij} = \Lambda_{ij} = \Phi_{ij} = \Phi_{ji} = 0$ para todo $i \neq j$. La densidad de $\boldsymbol{X}$ asumiendo el modelo independiente zipGM con parámetros $\boldsymbol{\omega}^0$ está dado por

$$
\begin{aligned}
P_{\boldsymbol{X}}(\boldsymbol{\omega}^0) &= \prod_{j=1}^{k} P(X_j = x_j) \qquad (5.7)\\
&= \prod_{j=1}^{k} \exp\left\{ \eta_j^0 X_j + \xi_j^0 \nu(X_j) + \frac{1}{2}\Theta_{jj}^0 X_j^2 - A(\eta_j^0, \xi_j^0, \operatorname{diag}(\Theta_{jj}^0)) \right\}.
\end{aligned}
$$

De este modo, entrenar el modelo independiente (5.7) es equivalente a optimizar (5.6) con $\lambda_{\mathcal{R}}, \lambda_{\mathcal{C}} \to \infty$.

Observación 5.1. *Notar que el problema* (5.6) *admite una parametrización en* $\tilde{\boldsymbol{\omega}} = \mathcal{I}(\boldsymbol{\omega}^0)^{1/2}\boldsymbol{\omega}$.

5.5 SELECCIÓN DEL MODELO

En este capítulo vimos que existe un compromiso sesgo-varianza y que modelos más simples en el sentido de tener pocos parámetros no nulos brindan mayor interpretación. También vimos que regularizando con la norma ℓ_1 es posible obtener modelos *sparse* y evaluando los

Y	$\dim R(\boldsymbol{X})$	MEDIDA PREDICTIVA
binaria	1	Area Under the Receiver Operating Characteristic Curve, por sus siglas en inglés (AUC)
binaria	>1	score de regresión logística
categórica	>1	score de regresión logística multivariada
categórica	>1	accuracy
continua	>0	error cuadrático de un estimador basado en kernel (Adragni y Cook, 2009)

Tabla 5.1: Medidas de predicción para distintos tipos de repuesta Y dependiendo de la dimensión de la reducción $R(\boldsymbol{X})$.

parámetros de regularización sobre una grilla podemos obtener una colección de modelos candidatos. Para seleccionar el modelo óptimo de una colección es posible definir métodos basados en *predicción*, que buscan evaluar la generalización del modelo, mientras que otros métodos basados en una *divergencia*, seleccionan los modelos que mejor aproximan a la verdadera distribución de los datos.

En particular, en el contexto de RSD, las medidas predictivas cobran mayor significancia, ya que permiten cuantificar el poder de predicción de la reducción, y de esta manera elegir el modelo óptimo (Adragni y Cook, 2009). Por otro lado, Konishi y Kitagawa (2008) demostraron que los métodos de selección predictivos son asintóticamente equivalentes a los basados en divergencias.

5.5.1 *Selección de modelos en el contexto de RSD*

El objetivo del modelado estadístico y en particular de RSD es la *generalización*, i. e. obtener información de los datos que puede aparecer en el futuro en contraste con los datos que se usan para aprender el modelo. Por lo tanto, la selección de modelo desde un punto de vista predictivo implica la evaluación del modelo basada en datos futuros obtenidos de manera independiente de los datos observados.

La medida de performance en predicción elegida va a depender de la naturaleza de la respuesta y de la dimensión $R(\boldsymbol{X})$, en la Tabla 5.1 mostramos ejemplos.

PARTICIÓN DE VALIDACIÓN Dada una partición I sobre el conjunto de datos $\{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$, se definen los conjuntos de entrenamiento $\mathbb{E} := \{(\boldsymbol{x}^{(i)}, y^{(i)}) : i \notin I\}$ y de validación $\mathbb{V} := \{(\boldsymbol{x}^{(i)}, y^{(i)}) : i \in I\}$, donde I es un conjunto de indices. Se utiliza el conjunto de entrenamiento para obtener un conjuto de modelos candidato basados en una grilla de penalización y se selecciona el modelo con mejor performance en predicción sobre el conjunto de validación.

VALIDACIÓN CRUZADA En la práctica resulta difícil obtener datos independientes a los que son empleados en el aprendizaje del modelo para definir una partición de validación de tamaño representativo. Aún cuando es posible obtenerlos, es posible obtener una mejor estimación si se utilizan esos datos junto a los de entrenamiento para el aprendizaje.

En este sentido, *validación cruzada* resulta una técnica que permite evitar este inconveniente. Permitie la evaluación de modelos desde un punto de vista predictivo únicamente basado en la muestra de entrenamiento, intentando preservar lo más posible la exactitud de la estimación.

Esta técnica consiste en definir una partición equitativa $\{I_k\}_{k=1}^{K}$, de tamaño K del conjunto de índices $\{1, \cdots, n\}$. Para cada $k = 1, \cdots, K$, considera como conjunto de validación o test $\mathbb{V}_k = \{(\boldsymbol{x}^{(i)}, y^{(i)}) : i \in I_k\}$, mientras que $\mathbb{E}_k = \{(\boldsymbol{x}^{(i)}, y^{(i)}) : i \notin I_k\}$ define el conjunto de entrenamiento. Estas particiones permiten disponer de K mediciones independientes del error de predicción, por lo que su promedio resulta insesgado. La estimación de dicho error para cada modelo indexado en $(\lambda_{\mathcal{R}}, \lambda_{\mathcal{C}})$, permite seleccionar el mejor modelo en predicción promedio $(\hat{\lambda}_{\mathcal{R}}, \hat{\lambda}_{\mathcal{C}})$. Una vez seleccionado el modelo, se reestima usando la totalidad de datos disponibles $\mathbb{V}_k \cup \mathbb{E}_k$.

6

APRENDIZAJE Y OPTIMIZACIÓN

En este capítulo, exploraremos conceptos fundamentales que nos permitirán optimizar una función objetivo penalizada en problemas a escala real. Estos problemas se caracterizan por tener un gran número de variables, p, y una cantidad de muestras disponibles para su estimación n, no muy grande. A esto se lo conoce como problemas de alta dimensión.

En la Sección 6.1, presentamos el gradiente natural, una generalización del método de Newton que permite obtener direcciones descendentes en espacios de parámetros vinculados a una función de probabilidad. Luego, en la Sección 6.2 revisaremos los algoritmos proximales, los cuales permiten entre otras cosas optimizar una composición de funciones convexas no diferenciables, como la inducida por la normal ℓ_1. En la Sección 6.3 detallamos el algoritmo proximal que nos permite optimizar la función de costo (5.6). Dicho algoritmo será utilizado en la Parte IV para analizar, mediante simulaciones, el comportamiento de los modelos propuestos en la Parte II en el contexto de RSD, utilizando funciones de costo y penalización estudiadas en los Capítulos 4 y 5.

6.1 GRADIENTE NATURAL

En relación a la métrica sobre el espacio de distribuciones inducida por una divergencia (ver Definición 4.1) y en el concepto de medida pullback presentado en la Sección 4.7.1, Amari (2016) definió al *gradiente natural* como

$$\tilde{\nabla}\mathbb{J} = \mathcal{I}^{-1}\nabla\mathbb{J}, \tag{6.1}$$

donde $\mathcal{I}$ es la matriz de información de Fisher y $\mathbb{J}$ es la función objetivo. El gradiente negativo $-\nabla\mathbb{J}$ apunta en la dirección de descenso más rápido de la función objetivo $\mathbb{J}$, maximizando la reducción de la función objetivo $\mathbb{J}$ por unidad de cambio en el parámetro $\boldsymbol{\omega}$ medido por la norma usual ℓ_2, i. e.

$$\frac{-\nabla\mathbb{J}}{\|\nabla\mathbb{J}\|} = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\operatorname*{arg\,min}_{d\boldsymbol{\omega}:\|d\boldsymbol{\omega}\|\leq\varepsilon}\mathbb{J}(\boldsymbol{\omega}+d\boldsymbol{\omega}), \tag{6.2}$$

En cambio, el gradiente natural negativo $-\tilde{\nabla}\mathbb{J}$ apunta en la dirección de descenso más rápido respecto a una métrica intrínseca a la distribución que se está modelando. En particular, considera la métrica local (4.1) inducida por la divergencia KL (Martens, 2020):

$$-\sqrt{2}\frac{\tilde{\nabla}\mathbb{J}}{\|\tilde{\nabla}\mathbb{J}\|_{\mathcal{I}^{-1}}} = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\operatorname*{arg\,min}_{d\boldsymbol{\omega}:\mathrm{KL}[f(\boldsymbol{\omega}+d\boldsymbol{\omega}):f(\boldsymbol{\omega})]\leq\varepsilon}\mathbb{J}(\boldsymbol{\omega}+d\boldsymbol{\omega}). \tag{6.3}$$

Ésta es la dirección de descenso más rápido en el espacio de distribuciones.

En general, dado que la divergencia de KL induce una métrica absoluta en el espacio de distribuciones (Amari, 2016), la misma geometría es inducida por una divergencia general D en (6.3). La principal diferencia radica en la eficiencia del estimador logrado y en la pérdida de la propiedad de invarianza a la parametrización que ofrece la divergencia de KL. Dicha invarianza podría ser aproximada al considerar otra divergencia D elegida de manera que su evaluación sea más eficiente.

6.1.1 *Problemas de la información empírica*

Es importante notar la diferencia entre la matriz de información y el Hessiano de la función objetivo o la esperanza empírica del score, la cual tiene la forma de la ecuación (4.41) o (4.42) para modelos condicionales. Cuando las esperanzas se reemplazan por sus versiones empíricas basadas en la muestra de entrenamiento, a esta matriz se la llama *información empírica*.

Kunstner, Hennig y Balles (2019) y Martens (2020) discutieron los problemas que conlleva considerar la matriz de información empírica en lugar de la matriz de información cuando se optimizan modelos locales de segundo orden. Se observa que no es posible asegurar convergencia local en el caso de considerar la información empírica salvo con constantes de aprendizaje que converjan a cero, haciendo el aprendizaje muy lento.

6.2 ALGORITMOS PROXIMALES

Los *algoritmos proximales* permiten optimizar problemas convexos con restricciones, con funciones objetivo no derivables, con muchas variables. Además, permiten agilizar los cálculos, ya que se puede separar el problema original en pequeños subproblemas que se evalúan de manera paralela. Dichos algoritmos se basan en la solución iterada de un *operador proximal* (Definición 6.1), lo que involucra un problema de optimización sencillo que en algunos casos puede obtenerse en forma cerrada y en otros podría requerir de un método iterativo para la solución del subproblema.

Definición 6.1 (Operador proximal (Parikh, Boyd et al., 2014))**.** *Sea $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ una función propia cerrada convexa, i. e. su epígrafo $\boldsymbol{epi}\, f = \{(\boldsymbol{\omega}, t) \in \mathbb{R}^m \times \mathbb{R} \mid f(\boldsymbol{\omega}) \leq t\}$ es un conjunto convexo no vacío. El operador proximal $\boldsymbol{prox}_f(\boldsymbol{v})$ de f se define como la única solución al siguiente problema fuertemente convexo:*

$$\underset{f}{\boldsymbol{prox}}(\boldsymbol{v}) = \underset{\boldsymbol{\omega} \in \mathbb{R}^m}{\arg\min} \left(f(\boldsymbol{\omega}) + \frac{1}{2} \|\boldsymbol{\omega} - \boldsymbol{v}\|_2^2 \right). \tag{6.4}$$

Algoritmo 2: Punto fijo (Parikh, Boyd et al., 2014)

Datos: f, $\boldsymbol{\omega}^{[0]}$, $\lambda > 0$
Resultado: $\boldsymbol{\omega}^{[k]}$
para $k = 0, 1, \cdots$ **hacer**

$$\boldsymbol{\omega}^{[k+1]} := \underset{\lambda f}{\mathbf{prox}}(\boldsymbol{\omega}^{[k]})$$

fin

Los operadores proximales son firmemente no expansivos[1] y su punto fijo es precisamente el conjunto que minimiza f. Por lo que los algoritmos proximales utilizan los operadores proximales como una extensión a los operadores de proyección empleados en los algoritmos de factibilidad convexa.

El algoritmo proximal más sencillo está definido por la iteración repetida del operador proximal y su convergencia está garantizada para la minimización de una función propia cerrada convexa $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$. En el Algoritmo 2 se detalla el procedimiento donde $\boldsymbol{\omega}^{[k]}$ corresponde a la k-ésima iteración. Este simple algoritmo garantiza la convergencia de la sucesión $\boldsymbol{\omega}^1, \cdots, \boldsymbol{\omega}^k$, como se detalla en la siguiente Proposición:

Proposición 6.1 (Convergencia (Bauschke y Combettes, 2011))**.** *Sea $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ una función propia cerrada convexa. Si f tiene un mínimo, la sucesión $\{\boldsymbol{\omega}^{[k]}\}_{k\in\mathbb{N}}$ dada por el Algoritmo 2 converge al conjunto que minimiza f y $f(\boldsymbol{\omega}^k)$ converge a su valor óptimo.*

6.2.1 *Algoritmo proximal SDMM*

El método toma su nombre por las siglas en inglés Simultaneous-Direction Method of Multiplier (SDMM) (Combettes y Pesquet, 2011) y permite resolver problemas compuestos.

Consideremos el siguiente problema compuesto de minimización que involucra transformaciones lineales en cada uno de sus términos:

Problema 6.1 (Problema compuesto de composición lineal)**.**

$$\min_{\boldsymbol{\omega}\in\mathbb{R}^m} g_1(\boldsymbol{L}_1\boldsymbol{\omega}) + \cdots + g_l(\boldsymbol{L}_l\boldsymbol{\omega}), \tag{6.5}$$

donde $g_i : \mathbb{R}^q \to \mathbb{R} \cup \{+\infty\}$ son funciones convexas semicontinuas inferiormente con dominio no vacío y para cada $i = 1, \cdots, l$ el span de $\boldsymbol{L}_i \in \mathbb{R}^{q\times m}$ está en el interior relativo[2] al dominio de g_i. Además se

1 Los operadores firmemente no expansivos son un caso particular de los operadores no expansivos (Lipschitz continuos con constante 1) y su iteración converge a un punto fijo (Parikh, Boyd et al., 2014).

2 El interior relativo **relint**(C) de un conjunto convexo $C \subseteq \mathbb{R}^q$ se define como **relint**$(C) := \{x \in C : \forall y \in C$, existe algún $\lambda > 0$ tal que $\lambda x + (1-\lambda)y \in C\}$ (Bertsekas, 2016).

Algoritmo 3: SDMM (Combettes y Pesquet, 2011)

Datos: g_i, L_i, $\boldsymbol{v}_i^{[0]}$, $\boldsymbol{\varrho}_i^{[0]}$, $\gamma > 0 \quad i = 1, \cdots, l$
Resultado: $\boldsymbol{\omega}^{[k]}$
para $k = 0, 1, \cdots$ **hacer**

$$\boldsymbol{\omega}^{[k]} := Q^{-1} \sum_{i=1}^{l} \boldsymbol{L}_i^\top (\boldsymbol{v}_i^{[k]} - \boldsymbol{\varrho}_i^{[k]})$$

para $i = 1, \cdots, l$ **hacer**

$$\begin{aligned} \boldsymbol{\vartheta}_i^{[k]} &:= \boldsymbol{L}_i \boldsymbol{\omega}^{[k]} \\ \boldsymbol{v}_i^{[k+1]} &:= \underset{\gamma g_i}{\mathbf{prox}}(\boldsymbol{\vartheta}_i^{[k]} + \boldsymbol{\varrho}_i^{[k]}) \\ \boldsymbol{\varrho}_i^{[k+1]} &:= \boldsymbol{\varrho}_i^{[k]} + \boldsymbol{\vartheta}_i^{[k]} - \boldsymbol{v}_i^{[k+1]} \end{aligned}$$

fin
fin

asume que $g_1(\boldsymbol{L}_1\boldsymbol{\omega}) + \cdots + g_l(\boldsymbol{L}_l\boldsymbol{\omega}) \to +\infty$ *cuando* $\|\boldsymbol{\omega}\| \to \infty$ *y que* $\boldsymbol{Q} := \sum_{i=1}^{l} \boldsymbol{L}_i^\top \boldsymbol{L}_i$ *es invertible.*

La solución a (6.5) se obtiene como el límite de la sucesión $\{\boldsymbol{\omega}^{[k]}\}_{k\in\mathbb{N}}$ construida por el Algoritmo 3 (Combettes y Pesquet, 2011). Notar en dicho algoritmo que los operadores proximales y los vectores auxiliares pueden ser calculados simultáneamente en cada iteración, permitiendo su implementación de manera paralela.

6.2.2 *Algoritmos proximales de primer orden y segundo orden*

Gran cantidad de problemas de estimación regularizados pueden escribirse en la siguiente forma compuesta:

Problema 6.2 (Problema compuesto)**.**

$$\min_{\boldsymbol{\omega}\in\mathbb{R}^m} g(\boldsymbol{\omega}) + h(\boldsymbol{\omega}), \tag{6.6}$$

donde g *es convexa con derivada continua y* $h : \mathbb{R}^m \to \mathbb{R}$ *es convexa pero no necesariamente diferenciable.*

Gradiente proximal para modelos pGM

Particularizando el Algoritmo 2 al Problema 6.2, y considerando una aproximación de Taylor de primer orden en el operador proximal de g, obtenemos el Algoritmo 4 de gradiente proximal, donde $\lambda^{[k]}$ controla la velocidad de convergencia. La convergencia es lineal para g Lipschitz continua de parámetro L_g, eligiendo $\lambda^{[k]} \in (0, 1/L_g]$.

En particular, el Algoritmo 4 permite minimizar la función de costo dada en (5.3) sobre el espacio de parámetros $\boldsymbol{\Omega}$, ya que el problema es separable y se conoce sus operadores proximales en forma cerrada:

Observación 6.1. *El minimizante de* (5.3) *puede encontrarse aplicando el Algoritmo* 4 *donde* $g(\boldsymbol{\omega}) = \hat{ES}(X, f(X \mid Y, \boldsymbol{\omega}))$ *es diferenciable y*

Algoritmo 4: Gradiente proximal (Parikh, Boyd et al., 2014)

Datos: g, h, $\boldsymbol{\omega}^{[0]}$, $\lambda > 0$
Resultado: $\boldsymbol{\omega}^{[k]}$
para $k = 0, 1, \cdots$ **hacer**

$$\boldsymbol{\omega}^{[k+1]} := \underset{\lambda^k h}{\mathbf{prox}} \left(\boldsymbol{\omega}^{[k]} - \lambda^{[k]} \frac{\partial}{\partial \boldsymbol{\omega}} g(\boldsymbol{\omega}^{[k]}) \right)$$

fin

los operadores proximales de $h(\boldsymbol{\Theta}) = \|\boldsymbol{\Theta}\|_1 + \iota_{\boldsymbol{\Omega}}(\boldsymbol{\Theta})$, *donde* $\iota_{\boldsymbol{\Omega}}(\boldsymbol{\Theta}) = \begin{cases} 0 & si\ \boldsymbol{\omega} \in \boldsymbol{\Omega} \\ +\infty & si\ \boldsymbol{\omega} \notin \boldsymbol{\Omega} \end{cases}$ *es la función indicadora de dominio y* $h(\boldsymbol{\Gamma}) = \sum_{j=1}^{p} \|\boldsymbol{\Gamma}_j\|_2$ *la norma mixta* $\ell_{1,2}$. *Los operadores poroximales correspondientes están disponibles en forma cerrada y se encuentran definidos en el Apéndice E.1*

Sin embargo, experimentos preliminares mostraron que los algoritmos proximales de primer orden no logran un desempeño aceptable en la familia zipGM debido a la diferencia de escala entre los distintos parámetros involucrados. Con ese fin consideremos los algoritmos de segundo orden presentados a continuación.

OPERADOR PROXIMAL DE SEGUNDO ORDEN Una generalización al operador proximal (Definición 6.1) para el Problema 6.2 consiste en considerar un operador proximal escalado, que definimos a continuación.

Definición 6.2 (Operador proximal de segundo orden (Lee, Sun y Saunders, 2012))**.** *El operador proximal de segundo orden* $\boldsymbol{prox}_h^{\boldsymbol{H}}(\boldsymbol{v})$ *de* $h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ *convexa se define como la única solución al siguiente problema fuertemente convexo:*

$$\underset{h}{\overset{\boldsymbol{H}}{\boldsymbol{prox}}}(\boldsymbol{v}) = \underset{\boldsymbol{\omega} \in \mathbb{R}^m}{\arg\min} \left(h(\boldsymbol{\omega}) + \frac{1}{2} \|\boldsymbol{\omega} - \boldsymbol{v}\|_{\boldsymbol{H}}^2 \right), \tag{6.7}$$

donde $\boldsymbol{H}$ *es positiva definida.*

Asumiendo que $\boldsymbol{H}^{[k]} > m\boldsymbol{I}$ para $m > 0$, existen $\{\lambda^{[k]}\}_{k \in \mathbb{N}}$ tales que la sucesión $\{\boldsymbol{\omega}^{[k]}\}_{k \in \mathbb{N}}$ producida por el Algoritmo 5 converge al óptimo del Problema 6.2 (Lee, Sun y Saunders, 2012). Además, una sucesión convergente $\{\lambda^{[k]}\}_{k \in \mathbb{N}}$ puede obtenerse en cada paso realizando una búsqueda lineal hasta satisfacer $g(\boldsymbol{\omega}^{[k+1]}) \leq g(\boldsymbol{\omega}^{[k]}) + \alpha\lambda\Delta$, con $\Delta = (\Delta\boldsymbol{\omega}^{[k]})^\top \frac{\partial}{\partial \boldsymbol{\omega}} g(\boldsymbol{\omega}^{[k]}) + h(\boldsymbol{\omega}^{[k]} + \Delta\boldsymbol{\omega}^{[k]}) - h(\boldsymbol{\omega}^{[k]})$, para algún $\alpha \in (0, 1/2)$. Si $H^{[k]}$ es el Hessiano de g en $\boldsymbol{\omega}^{[k]}$, la convergencia es q-cuadrática, similar al algoritmo de Newton.

6.2.2.1 *Algoritmo proximal para minimizar* (5.6)

Contribución original: solución al problema de optimización inducido por una penalización jerárquica pesada

Para minimizar (5.6) implementamos el Algoritmo 5 para el Problema 6.2 donde $g(\boldsymbol{\omega}) := \hat{ES}(\boldsymbol{X}, f(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{\omega}))$ con S :

Algoritmo 5: Newton proximal (Lee, Sun y Saunders, 2012)

Datos: g, h, $\boldsymbol{\omega}^{[0]}$, $\lambda^{[k]} > 0$
Resultado: $\boldsymbol{\omega}^{[k]}$
para $k = 0, 1, \cdots$ **hacer**

$$\Delta\boldsymbol{\omega}^{[k]} := \operatorname*{prox}_{h}^{\boldsymbol{H}^{[k]}}\left(\boldsymbol{\omega}^{[k]} - (\boldsymbol{H}^{[k]})^{-1}\frac{\partial}{\partial\boldsymbol{\omega}}g(\boldsymbol{\omega}^{[k]})\right)$$
$$\boldsymbol{\omega}^{[k+1]} := \boldsymbol{\omega}^{[k]} + \lambda^{[k]}\Delta\boldsymbol{\omega}^{[k]}$$

fin

$\mathcal{X} \times \mathcal{P} \to \mathbb{R}$ un score propio y donde $f(\boldsymbol{X} \mid Y, \boldsymbol{\omega})$ es una modelo gráfico condicional lineal de segundo orden (Definición 2.12) de la familia zipGM (Definición 2.10) junto con (3.10) y donde $h(\boldsymbol{\omega}) := \lambda_{\mathcal{R}} \sum_{j=1}^{k} \|(\boldsymbol{\Gamma}_j; \boldsymbol{\Psi}_j; (\Theta_{jl}, \Phi_{lj}; \Phi_{jl}; \Lambda_{jl})_{l\neq j})\|_{\mathcal{R}_j} + \lambda_{\mathcal{C}} \sum_{j=1}^{k}\sum_{l\neq j} \|(\Theta_{jl}, \Phi_{lj}; \Phi_{jl}; \Lambda_{jl})\|_{\mathcal{C}_{jl}}$ es la penalización jerárquica con pesos óptimos dados por la matriz de información de Fisher bajo el modelo independiente $\mathcal{I}(\boldsymbol{\omega}^0)$.

Para ello, consideramos el operador proximal de segundo orden (Defición 6.2) con matriz de curvatura $\boldsymbol{H}^{[k]} := \mathcal{I}(\boldsymbol{\omega}^{[k]}) =: \mathcal{I}$ (Martens, 2020), dando lugar al siguiente subproblema dado $t > 0$:

$$\begin{aligned}
&\operatorname*{prox}_{th}^{\mathcal{I}}\left(\boldsymbol{\omega}^{[k]} - t\mathcal{I}^{-1}\frac{\partial}{\partial\boldsymbol{\omega}}\hat{E}S(\boldsymbol{\omega}^{[k]})\right) = \\
&= \operatorname*{arg\,min}_{\boldsymbol{\omega}\in\mathbb{R}^m} \frac{1}{2t}\left\|\boldsymbol{\omega} - \left(\boldsymbol{\omega}^{[k]} - t\mathcal{I}^{-1}\frac{\partial}{\partial\boldsymbol{\omega}}\hat{E}S(\boldsymbol{\omega}^{[k]})\right)\right\|_{\mathcal{I}}^2 + h(\boldsymbol{\omega}) \\
&= \operatorname*{arg\,min}_{\boldsymbol{\omega}\in\mathbb{R}^m} \tilde{g}_{\ell}(\mathcal{I}^{1/2}\boldsymbol{\omega}) + \bar{g}_{\mathcal{R}}(\mathcal{I}(\boldsymbol{\omega}^0)^{1/2}\boldsymbol{\omega}) \qquad (6.8) \\
&\qquad\qquad + \bar{g}_{\mathcal{C}}(\mathcal{I}(\boldsymbol{\omega}^0)^{1/2}\boldsymbol{\omega}) + \iota_{\boldsymbol{\Omega}}(\boldsymbol{\omega}),
\end{aligned}$$

donde definimos

$$\tilde{g}_{\ell}(\tilde{\boldsymbol{\omega}}) = \frac{1}{2t}\|\boldsymbol{\omega} - \left(\boldsymbol{\omega}^{[k]} - t\mathcal{I}^{-1}\frac{\partial}{\partial\boldsymbol{\omega}}\hat{E}S(\boldsymbol{\omega}^{[k]})\right)\|_2^2, \qquad (6.9)$$

$$\bar{g}_{\mathcal{R}}(\bar{\boldsymbol{\omega}}) = \lambda_{\mathcal{R}} \sum_{j=1}^{k} \|(\overline{\boldsymbol{\Gamma}}_j; \overline{\boldsymbol{\Psi}}_j; (\overline{\Theta}_{jl}, \overline{\Phi}_{lj}; \overline{\Phi}_{jl}; \overline{\Lambda}_{jl})_{l\neq j})\|_2, \qquad (6.10)$$

$$\bar{g}_{\mathcal{C}}(\bar{\boldsymbol{\omega}}) = \lambda_{\mathcal{C}} \sum_{j=1}^{k}\sum_{l\neq j} \|(\overline{\Theta}_{jl}, \overline{\Phi}_{lj}; \overline{\Phi}_{jl}; \overline{\Lambda}_{jl})\|_2, \qquad (6.11)$$

$$\iota_{\boldsymbol{\Omega}}(\bar{\boldsymbol{\omega}}) = \begin{cases} 0 & \text{if } \bar{\boldsymbol{\omega}} \in \boldsymbol{\Omega} \\ \infty & \text{otherwise} \end{cases}, \qquad (6.12)$$

con $\tilde{\boldsymbol{\omega}} = \mathcal{I}^{1/2}\boldsymbol{\omega}$ y $\bar{\boldsymbol{\omega}} = \mathcal{I}(\boldsymbol{\omega}^0)^{1/2}\boldsymbol{\omega}$ y sus correspondientes bloques forman los parámetros transformados $\overline{\boldsymbol{\Gamma}}_j, \overline{\boldsymbol{\Psi}}_j, \overline{\Theta}_{jl}, \overline{\Phi}_{lj}, \overline{\Phi}_{jl}, \overline{\Lambda}_{jl}$.

Observación 6.2. *La transformación $\bar{\boldsymbol{\omega}} = \mathcal{I}(\boldsymbol{\omega}^0)^{1/2}\boldsymbol{\omega}$ tiene dos efectos:*

- *Define una penalización del tipo LASSO por grupos usual en (6.10), (6.11).*

- (6.9) *corresponde a un método precondicionado de primer otden en* $\boldsymbol{\omega}$ *(Yang et al., 2016). Esta formulación mejora el condicionamiento del sub-problema, resultando en convergencia más rápida.*

A continuación, describimos el uso del Algoritmo 3 (SDMM) para evaluar el operador proximal escalado (6.8).

SDMM COMO SUBPROGRAMA PARA OPTIMIZAR (6.8) Para optimizar (6.8) aplicamos el Algoritmo 3 considerando el correspondiente problema en su forma (6.2). Los operadores proximales (Definición 6.1) correspondientes a las funciones convexas $\tilde{g}_\ell, \tilde{g}_\mathcal{R}$, $\tilde{g}_\mathcal{C}$ y ι_Ω dadas en (6.9)-(6.12) se definen en forma cerrada en el Apéndice E.2.1.

RESTRICCIONES IMPUESTAS POR LOS MODELOS PGM Y ZIPGM Las restricciones impuestas por el espacio paramétrico $\boldsymbol{\Omega}$ de cada modelo se incorporan en el problema de optimización mediante la función indicadora (6.12), y cuyo operador proximal corresponde a la proyección ortogonal al espacio de parámetros. Los ejemplos analizados presentan las siguientes restricciones:

- $\boldsymbol{\Theta}$ es SDP para los modelos Normal-pGM y zipGM. En este caso, $\mathbf{prox}_{\iota_\Omega}(\boldsymbol{\Theta}) = U(A)_+U^\top$, donde UAU es la descomposición en valores singulares de $\boldsymbol{\Theta}$ y donde $(\cdot)_+$ devuelve la parte no negativa.
- $\Theta_{jl} \leq 0$ para todo $i \neq j$ y $\Theta_{jj} = 0$ para todo j en el modelo Poisson-pGM y zipGM. Por lo tanto, $\mathbf{prox}_{\iota_\Omega}(\boldsymbol{\Theta}) = -(-\boldsymbol{\Theta})_+$.
- $\Theta_{jj} = 0$ para todo j en los modelos Ising-pGM, TPoisson-pGM y zipGM, resultando en $\mathbf{prox}_{\iota_\Omega}(\operatorname{diag}(\boldsymbol{\Theta})) = \mathbf{0}$.

6.3 ALGORITMO PROPUESTO PARA EL APRENDIZAJE EN ALTA DIMENSIÓN DE MODELOS ZIPGM

Asumiendo que se cuenta con conjuntos independientes de entrenamiento y validación, la estimación y selección del modelo sigue los pasos:

1. Obtener el MLE $\hat{\boldsymbol{\omega}}^0$ iterando con el gradiente natural realizando la iteración $\boldsymbol{\omega}^{[k+1]} = \boldsymbol{\omega}^{[k]} - \tilde{\nabla}\hat{ES}(\boldsymbol{X}, f(\boldsymbol{X} \mid Y, \boldsymbol{\omega}^{[k]})$ para el modelo independiente (5.7) usando los datos de entrenamiento.
2. Calcular la matriz de información de Fisher $\hat{\mathcal{I}}^0 = \mathcal{I}(\hat{\boldsymbol{\omega}}^0)$ usando (4.9), donde la esperanza $E_{\hat{\boldsymbol{\omega}}^0}$ es aproximada generando muestras de la distribución (5.7). Los bloques de $\hat{\mathcal{I}}^0$ definen las normas $\|.\|_{\mathcal{R}_j}$ y $\|.\|_{\mathcal{C}_j}$ del problema regularizado (5.6).
3. Definir las cotas superiores $\lambda_{\mathcal{R},\text{máx}} \propto \text{máx}_j\, J_j^T\mathcal{R}_jJ_j$ y $\lambda_{\mathcal{C},\text{máx}} \propto \text{máx}_{jl}\, J_{jl}\mathcal{C}_{jl}J_{jl}$, donde J_j, J_{jl} son los correspondientes bloques del Jacobiano de la función objetivo dada por el primer término de (5.6) evaluada en $\hat{\boldsymbol{\omega}}^0$. Luego, para cada par $(\lambda_\mathcal{R}, \lambda_\mathcal{C})$ tomando valores en una grilla definida en $[0, \lambda_{\mathcal{R},\text{máx}}] \times [0, \lambda_{\mathcal{C},\text{máx}}]$:

a) obtener $\hat{\boldsymbol{\omega}}$ minimizando (5.6) usando el algoritmo proximal de segundo orden (Sección 6.2.2.1) que construye en cada iteración la matriz $\mathcal{I}(\hat{\boldsymbol{\omega}}^{(k)})$ usando (4.43) y donde la esperanza $E_{\boldsymbol{X}|Y}$ es aproximada numéricamente a partir de muestras generadas por Gibbs sampling (Algoritmo 1) a partir de las condicionales (2.18). Martens (2020) demostró que la convergencia lineal está ganartizada en un entorno del óptimo y que al considerar la matriz regularizada $\mathcal{I}(\boldsymbol{\omega}) + \epsilon I$, $\epsilon > 0$, se asegura convergencia global.

b) Optimizar el modelo de regresión $\boldsymbol{X} \mid Y$ zipGM en (2.16) junto con (3.10) usando solamente los parámetros seleccionados en el paso 3*a*, de dónde se obtienen estimadores insesgados $\widehat{\boldsymbol{\Gamma}}$ y $\widehat{\boldsymbol{\Psi}}$ que permiten evaluar la reducción $R(\boldsymbol{X})$ en (3.11) o (3.12).

c) Calcular una medida de desempeño predictiva para los datos de validación basada en $R(\boldsymbol{X})$ obtenida en el paso 3*b* (ver tabla 5.1).

4. Usando los parámetros óptimos de regularización $(\lambda^*_{\mathcal{R}}, \lambda^*_{\mathcal{C}})$ los cuales presentan la mejor desempeño en predicción en el paso 3*c*, repetir los pasos 3*a* y 3*b* combinando los conjuntos de datos de entrenamiento y validación para obtener el modelo final.

Contribución original: software escalable

Una implementación en *Pytorch* del algoritmo propuesto para los modelos zipGM presentados en los Ejemplos 2.7-2.9 con el score definido por la pseudolikelihood (Definición 4.6) y penalización jerárquica (Sección 5.4), conformando la función de costo compuesta para el problema de regresión $\boldsymbol{X} \mid Y$ (5.6) se encuentra disponible en `https://github.com/ekoplin/SDR-zipGM`. En el Apéndice E.2.2 se detalla la estructura asumida en el cálculo de la matriz de información así como las derivadas de la función de costo que permiten definir el operador proximal de segundo orden (6.8).

6.4 CONTRIBUCIÓN

En este capítulo encontramos una solución algorítmica para optimizar la función de costo asociada a los modelos zipGM (5.6) compuesta por la esperanza muestral de un score y de una penalización jerárquica pesada. Si bien la solución encontrada combina algoritmos estudiados previamente como el Algoritmo 5 y el Algoritmo 3, hasta donde se conoce este tipo de problemas de optimización no se encuentran estudiados ni existen herramientas disponibles para su optimización. Además, debido a la escala del problema, se brindó una implementación que aprovecha la separación que brindan los operadores proximales para evaluar de manera paralela las operaciones involucradas.

Parte IV

RESULTADOS NUMÉRICOS

En los siguientes capítulos estudiamos el desempeño de los modelos y estimadores propuestos en los capítulos precedentes. La evaluación se realiza en el contexto de RSD en alta dimensión, aunque los modelos propuestos podrían encontrar aplicaciones adicionales más allá de la reducción dimensional. También pondremos énfasis en la estimación utilizando penalización jerárquica (Parte III), lo que nos permite ayudar a la interpretación de las reducciones identificando las variables más relevantes. En el Capítulo 7 estudiamos el comportamiento de los algoritmos propuestos con datos simulados, mientras que en el Capítulo 8 mostraremos su aplicación a datos composicionales reales de microbioma.

7

SIMULACIONES

En este capítulo estudiaremos el comportamiento de los estimadores de los modelos de la familia zipGM, presentados en los Ejemplos 2.7-2.9, usando el Algoritmo propuesto en Sección 6.3 que utiliza pseudolikelihood (Definición 4.6) y penalización jerárquica (Sección 5.4). Simulaciones adicionales destinadas a evaluar los estimadores de modelos pGM obtenidos mediante scores propios se presentan en el Apéndice F.1.

En todos los casos presentados en este capítulo estudiamos el poder predictivo de los modelos estimados y su habilidad para seleccionar correctamente las variables relacionadas con la respuesta. La intención es estudiar cuánto las reducciones obtenidas preservan la información discriminante, ayudando además a revelar las relaciones clave entre los predictores individuales y la respuesta. También estudiaremos las interacciones presentes en el modelo gráfico, que en aplicaciones con datos reales ayudará a comprender relaciones entre los distintos predictores.

En la Sección 7.1 estudiamos el impacto de la proporción de ceros en los datos sobre los estimadores de los modelos gráficos no adaptados a estos escenarios (Ejemplos 2.1-2.4). Estas pruebas nos permitirán ilustrar la utilidad de los modelos propuestos, que sí están adaptados a escenarios de alta proporción de ceros. También evaluamos cuánto se deteriora el desempeño cuando el modelo zipGM elegido para el análisis no corresponde a la distribución real de los datos. En la Sección 7.2 mostramos el impacto de las interacciones entre los predictores sobre los estimadores obtenidos. En la sección 7.3 estudiamos la capacidad de los modelos propuestos frente a datos realistas simulados mediante inteligencia artificial generativa entrenada con datos reales de microbioma.

ESCENARIO GENERAL. A lo largo del capítulo mantendremos fijas algunas condiciones experimentales. En particular, consideramos $p = 100$ y $n \in \{200, 500, 1000\}$. En cada experimento generamos conjuntos de entrenamiento de n muestras y conjuntos independientes de validación y de test. Los datos de validación se componen de 1000 muestras y se utilizan para seleccionar la combinación más adecuada de los parámetros de regularización de los estimadores. Los modelos elegidos se evalúan luego sobre el conjunto de test, que consiste en 10000 muestras. Los conjuntos de entrenamiento y de validación se generaron de forma independiente para cada repetición del experimento, mientras que el conjunto de test se mantuvo fijo durante la experimentación. Las medidas de desempeño reportadas se obtuvieron sobre 100 repeticiones del experimento en cada condición evaluada.

DESEMPEÑO	FPR	FNR
VS	$\#(\hat{S}_{\text{VS}} \cap \bar{S}^*_{\text{VS}})/\#\bar{S}^*_{\text{VS}}$	$\#(\hat{\bar{S}}_{\text{VS}} \cap S^*_{\text{VS}})/\#S^*_{\text{VS}}$
CI	$\#(\hat{S}_{\text{CI}} \cap \bar{S}^*_{\text{CI}})/\#\bar{S}^*_{\text{CI}}$	$\#(\hat{\bar{S}}_{\text{CI}} \cap S^*_{\text{CI}})/\#S^*_{\text{CI}}$

Tabla 7.1: Medidas de desempeño en selección definidas en base a los conjuntos complementarios de selección de variables $\bar{S}_{\text{VS}} = \{j : \boldsymbol{\Gamma}_j = \boldsymbol{\Psi}_j = \mathbf{0}\}$ donde $\boldsymbol{\Gamma}_j$y $\boldsymbol{\Psi}_j$ corresponden a la j-ésima fila de la matriz correspondiente y los conjuntos complementarios de selección de interacciones $\bar{S}_{\text{CI}} = \{(i,j) : \Theta_{ij} = \Phi_{ji} = \Phi_{ij} = \Lambda_{ij} = 0\}$. $\bar{S}^*_{(\cdot)}$ define los conjuntos poblacionales, mientras que $\hat{\bar{S}}_{(\cdot)}$ los conjuntos estimados.

MEDIDAS DE PERFORMANCE. Cuando la respuesta Y es binaria, la reducción $R(\boldsymbol{X})$ dada en el Corolario 3.2 es un escalar y usamos la AUC sobre los datos de test para medir la performance. Cuando la respuesta Y es continua (ver Sección 7.3), usamos el MSE entre la respuesta $Y \in \mathbb{R}$ y la predicción $\hat{Y}$ basada en la reducción $\boldsymbol{R}(\boldsymbol{X})$, medido también sobre el conjunto de test. La predicción se obtiene a partir de un estimador no paramétrico basado en un núcleo Gaussiano, cuyo ancho de banda está determinado por la mediana de las distancias entre pares de observaciones de entrenamiento (Adragni y Cook, 2009).

También es de interés evaluar la capacidad del modelo estimado para seleccionar las variables vinculadas con la respuesta Y, como también los conjuntos de independencia condicional entre dichas variables. Esta propiedad se encuentra codificada en los ceros de las interacciones: como consecuencia del Teorema 2.1, para $\boldsymbol{X}$ siguiendo un modelo zipGM (Definición 2.10), se tiene que $X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{\setminus\{i,j\}}$ si y solo si $\Theta_{ij} = \Phi_{ji} = \Phi_{ij} = \Lambda_{ij} = 0$. Dichos conjuntos son inducidos por el tercer término en (5.6), mientras que la selección de variables se debe al efecto del segundo término.

Para cuantificar la performance en selección de variables (VS, por sus siglas en inglés) y en selección de conjuntos de independencia condicional (CI), utilizamos la tasa de falsos positivos (FPR) y de falsos negativos (FNR) como se detalla en la Tabla 7.1. En particular, FPR(VS) reporta la proporción de variables que son seleccionadas incorrectamente como asociadas con la respuesta. Equivalentemente, FPR(CI) reporta la proporción de pares de variables que no son seleccionadas como condicionalmente independientes cuando en verdad lo son; es decir, cuenta la proporción de las aristas estimadas en el grafo de independencia condicional que no existen en la versión poblacional. De manera complementaria, la FNR considera la proporción de componentes que se encuentran presentes en el modelo poblacional pero que no son seleccionadas por el estimador considerado.

En general, un FNR alto indica modelos muy sesgados, mientras que un FPR alto indica modelos con mucha varianza. El caso de selección ideal se da cuando ambos valores son nulos. Con el objetivo de aislar efectos, debido a que la performance en selección se encuentra vinculada a la predicción a través del criterio predictivo seleccionado

en el Algoritmo propuesto en la Sección 6.3, adoptamos una medida de "oráculo" definida en McDavid et al. (2019), que selecciona las variables que minimizan la FPR(VS) sujeto a FNR(VS)< 0.1.

7.1 PROPORCIÓN DE CEROS Y MODELOS MAL ESPECIFICADOS

Consideramos que el modelo poblacional para $\boldsymbol{X} \mid Y$ es Normal-zipGM, Poisson-zipGM o TPoisson-zipGM con interacciones únicamente entre variables vinculadas a la respuesta Y a través de (3.10). Por simplicidad analizamos el caso de respuesta binaria balanceada $Y \in \{0, 1\}$ y la AUC como medida predictiva.

7.1.1 *Parámetros poblacionales*

ESTRUCTURA FUERTEMENTE JERÁRQUICA. Consideramos que $X_1, \cdots, X_{15}$ están asociadas con Y a través de (3.10) y a su vez, estas variables se encuentran correlacionadas entre ellas a través de interacciones no nulas $\Theta_{jl}, \Lambda_{jl}, \Phi_{jl}$ y Φ_{lj}, $j, l \in \{1, \cdots, 15\}$. Los demás elementos de $\boldsymbol{\Gamma}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, así como los demás elementos fuera de la diagonal de $\boldsymbol{\Theta}$, son considerados nulos.

MODELO INDEPENDIENTE. Definimos los parámetros del modelo independiente (5.7) que nos permitirán modelar la esperanza condicional de $X_j \mid \mu(X_j)$. Para el modelo Normal-zipGM, generamos η_{0j} y Θ_{jj} como muestras independientes de una $\mathcal{N}(0, 0.01)$ y $\mathcal{N}(-1, 0.01)$ respectivamente, para cada $j \in \{1, \cdots, 100\}$. De este modo, cada variable condicional es aproximadamente una normal estándar. Para los modelos Poisson y TPoisson-zipGM, consideramos $\eta_{0j} \sim \mathcal{N}(\log(3), 1)$, por lo que cada variable condicional corresponde aproximadamente a una Poisson truncada en cero con media $E[\boldsymbol{X}_j \mid \nu(X_j) = 1] = \exp\{\eta_j\} \exp \exp\{\eta_j\} / [\exp\{\eta_j\} - 1] \approx 3.15$.

PROPORCIÓN DE CEROS. La proporción de ceros en el modelo independiente (5.7) está controlada en cada variable por el parámetro ξ_j correspondiente. Para variar la proporción de ceros en los datos consideramos $\xi_{0j} = -A^+(\eta_{0j}, \Theta_{jj}) + \Delta_\xi$ con $\Delta_\xi \in \{-2, 0, 2\}$ y $A^+(\cdot)$ definida en el Teorema 2.4. De esta manera logramos que la proporción de ceros esté entre el $87 - 90\,\%$ para $\Delta_\xi = -2$, $45 - 65\,\%$ para $\Delta_\xi = 0$ y $10 - 28\,\%$ para $\Delta_\xi = 2$ al incorporar las interacciones.

INTERACCIONES. Para generar las interacciones no nulas $(\Theta_{jl}, \Phi_{lj}, \Phi_{jl}, \Lambda_{jl})$, $j, l = 1, \ldots, 15$, primero generamos muestras independientes $(\tilde{\Theta}_{jl}, \tilde{\Phi}_{lj}, \tilde{\Phi}_{jl}, \tilde{\Lambda}_{jl})$ de la mezcla, $0.5\mathcal{N}(-3, 1) + 0.5\mathcal{N}(3, 1)$.

Luego acomodamos por un factor de escala tendiendo en cuenta el modelo independiente sobre el que actúan: para el modelo Normal-zipGM, $\Theta_{jl} = c\tilde{\Theta}_{jl} / (\Theta_{jj}\Theta_{ll})^{1/2}$ y $\Phi_{lj} = c\tilde{\Phi}_{lj} / (-\Theta_{jj})^{1/2}$ con $c = 0.01$. Mientras que para los modelos Poisson y TPoisson-zipGM, consideramos $\Theta_{jl} = c - \tilde{\Theta}_{jl} \exp\{(\eta_{0j} + \eta_{0l})/2\}$ y $\Phi_{lj} = c\tilde{\Phi}_{lj} \exp(\eta_{0l}/2)$ con $c = 0.005$. De esta manera balanceamos las contribuciones entre las

interacciones y a su vez, la constante c fue elegida de manera de fijar un nivel de señal predictiva caracterizada por el span que definimos a continuación.

DIRECCIONES DEL SUBESPACIO DE REDUCCIÓN DE DIMENSIONES. Para definir el span de RSD caracterizado en el Corolario 3.11, debemos definir las matrices $\mathbf{\Gamma}$ y $\mathbf{\Psi}$. Notemos que en el modelo Possion, tanto la proporción de ceros como la media se encuentran ligadas por el parámetro $\boldsymbol{\eta}_j$. Para crear un escenario desfavorable al modelo Poisson, consideramos un span que haga aumentar tanto la proporción de ceros en los datos como el promedio de las cuentas no nulas. En la Sección 7.1.3 analizaremos este comportamiento.

De manera similar a las interacciones, definimos los valores auxiliares $\bar{\Psi}_j = 1$ y $\bar{\Gamma}_j = -\exp(-\eta_{0j}/2)$ para los modelos Poisson y TPoisson-zipGM y $\bar{\Gamma}_j = -(-\Theta_{jj})^{1/2}$ para el modelo Normal-zipGM, para cada $j = 1, \ldots, 15$. Fijamos $\mathbf{\Gamma} = a_1\bar{\mathbf{\Gamma}}$ en (3.10a) y $\mathbf{\Psi} = a_2\bar{\mathbf{\Psi}}$ en (3.10b) de manera que las constante a_1 y a_2 controlan la fuerza de la asociación entre $\boldsymbol{X}$ e Y.

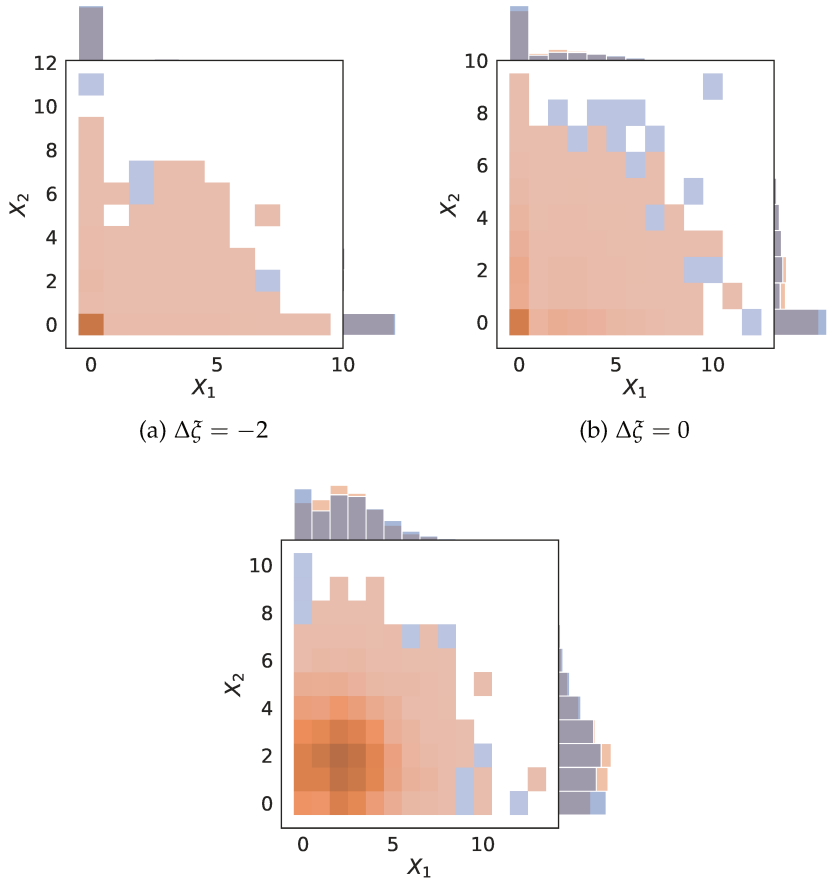
Estas constantes fueron elegidas de manera de balancear las contribuciones predictivas de cada dimensión de la reducción dada en el Corolario 3.3. La constante a_1 se eligió de manera de obtener una AUC≈ 0.65 en predicción a partir de $\boldsymbol{\alpha}^T\boldsymbol{T}(\boldsymbol{X})$, con $\boldsymbol{\alpha} = \text{span}\{\mathbf{\Gamma}\}$ con el modelo definido anteriormente, pero considerando $\Delta_\xi = 5$, lo que resulta en datos con muy pocos ceros. La constante a_2 fue elegida de manera tal de obtener AUC≈ 0.65 en predicción a partir de $\boldsymbol{\zeta}^T\boldsymbol{\nu}(\boldsymbol{X})$ con $\boldsymbol{\zeta} = \text{span}\{\mathbf{\Psi}\}$ cuando $\Delta_\xi = 0$.

En la Figura 7.1 se muestra el histograma conjunto de las dos primeras variables generadas por el modelo Poisson-zipGM en función de la respuesta Y a medida que disminuye la proporción de ceros de acuerdo con $\Delta\xi \in \{-2, 0, 2\}$. En la Figura 7.2 se muestra el efecto de la proporción de ceros en las reducciones poblacionales para el caso Poisson-zipGM. Se puede observar que la parte binaria de la reducción separada (Proposición 3.3) es más informativa a medida que su variabilidad aumenta ($\Delta\xi = 0$), mientras que las cuentas aportan más información a medida que la proporción de ceros decrece ($\Delta\xi = 2$).

7.1.2 *Resultados*

NORMAL-ZIPGM. En la Figura 7.3 se muestran los resultados obtenidos cuando los modelos Normal-pGM, Normal-zipGM y el Ising son entrenados con datos generados por el Normal-zipGM definido en la Sección 7.1.1 para $\Delta_\xi \in \{-2, 0, 2\}$, correspondientes a distintas proporciones de ceros. Observamos que para todo Δ_ξ el modelo Normal-zipGM logra la mayor AUC, siendo mayor la diferencia a medida que n y la proporción de ceros disminuye.

Para $\Delta_\xi = -2$, i. e. cuando la proporión de ceros es mayor al 80 %, la AUC del modelo Normal-zipGM es similar a la del Ising, aunque mucho mayor a la AUC obtenida por el modelo Normal-pGM. Para



(a) $\Delta\xi = -2$

(b) $\Delta\xi = 0$

(c) $\Delta\xi = 2$

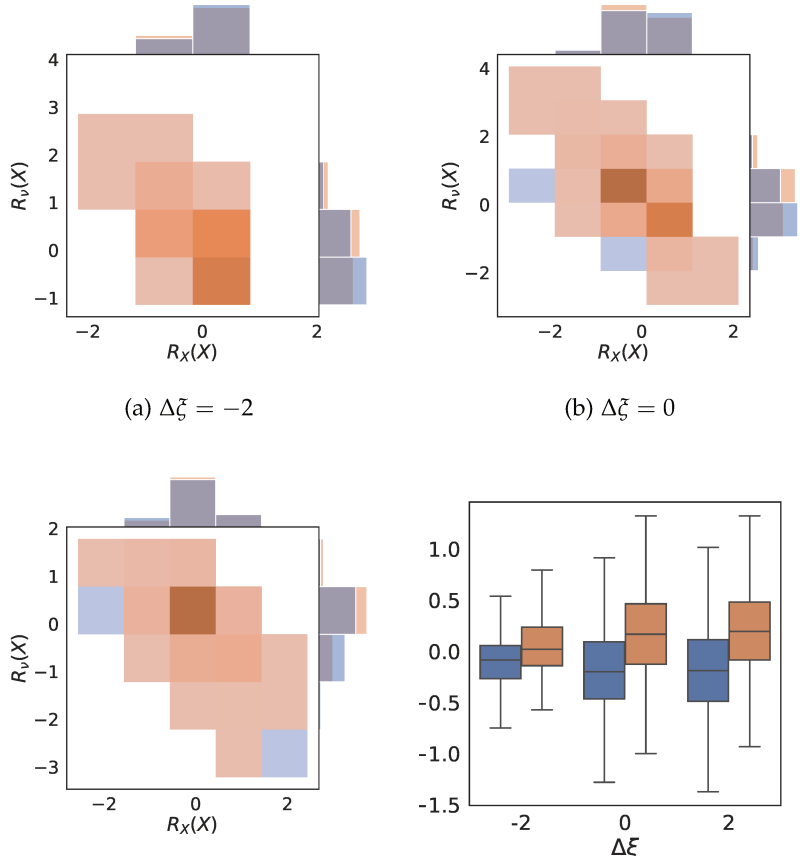
Figura 7.1: Interacción de las primeras dos variables generadas por el modelo Poisson-zipGM en relación con la respuesta Y (color) a medida que disminuye la proporción de ceros.

$\Delta_\xi = 2$ ($10 - 30\,\%$ de ceros en los datos), el modelo Normal-pGM muestra mejor performance en predicción que el modelo Ising.

En este caso, como la estructura inducida por la penalización es correcta, todos los modelos presentaron muy baja FPR(VS), FNR(VS) y FPR(CI) para todo Δ_ξ. De todos modos, el modelo Ising consistentemente presenta mayor FNR(CI) debido al efecto de la mala especificación del modelo.

POISSON-ZIPGM. En la Figura 7.4, se muestran los resultados para distintas proporciones de ceros (valores de Δ_ξ) cuando se entrenan los modelos Poisson-pGM, Poisson-zipGM e Ising con datos generados por el modelo Poisson-zipGM definido en la Sección 7.1.1.

El modelo Poisson-zipGM presenta la mayor AUC en predicción, seguido por el modelo Ising, mientras que el Poisson-pGM presenta una performance en predicción mucho más baja. Notar que la diferencia entre el Poisson-pGM y el Poisson-zipGM fue mayor que para los correspondientes modelos Normales debido a la mala especificación del primero causada por la selección del span. Analizamos este comportamiento en la Sección 7.1.3.



(a) $\Delta\xi = -2$ (b) $\Delta\xi = 0$

(c) $\Delta\xi = 2$ (d) Reducción conjunta $R(\boldsymbol{X})$

Figura 7.2: Reducciones obtenidas por el modelo poblacional Poisson-zipGM en relación con la respuesta Y (color) a medida que disminuye la proporción de ceros. En las figuras (a)-(c) se muestran las interacciones entre las componentes de la RSD separada (Proposición 3.3) mientras que en (d) se detalla la reducción conjunta (Proposición 3.2).

Para $\Delta_\xi = 2$ ($10 - 28\,\%$ de ceros), la AUC promedio en predicción fue menor a 0.55 para todo n, mientras que fue de alrededor de 0.7 para el modelo Poisson-zipGM. En este escenario, el modelo Poisson-pGM obtuvo mucho mayor FPR(VS) que los otros modelos, mientras que cuando la proporción de ceros en los datos es extremadamente alta ($\Delta_\xi = -2$), muestra muy alta FNR(VS). Los resultados sugieren que este modelo no resulta confiable para seleccionar variables.

El FPR(CI) fue bajo para todos los modelos cuando $\Delta_\xi = -2$, mientras que el FNR(CI) decrece para los modelos Poisson-zipGM y Poisson-pGM a medida que baja la proporción de ceros en los datos.

RESULTADOS ADICIONALES. Los resultados obtenidos cuando generamos muestras a partir del modelo TPoisson-zipGM son similares a los obtenidos con el modelo Poisson-zipGM y se detallan en el Apéndice F.2.1. Además, en el Apéndice F.2.1 se muestran los resultados obtenidos utilizando el criterio de selección "oráculo". Estos resultados resaltan que la penalización adoptada induce de manera robusta

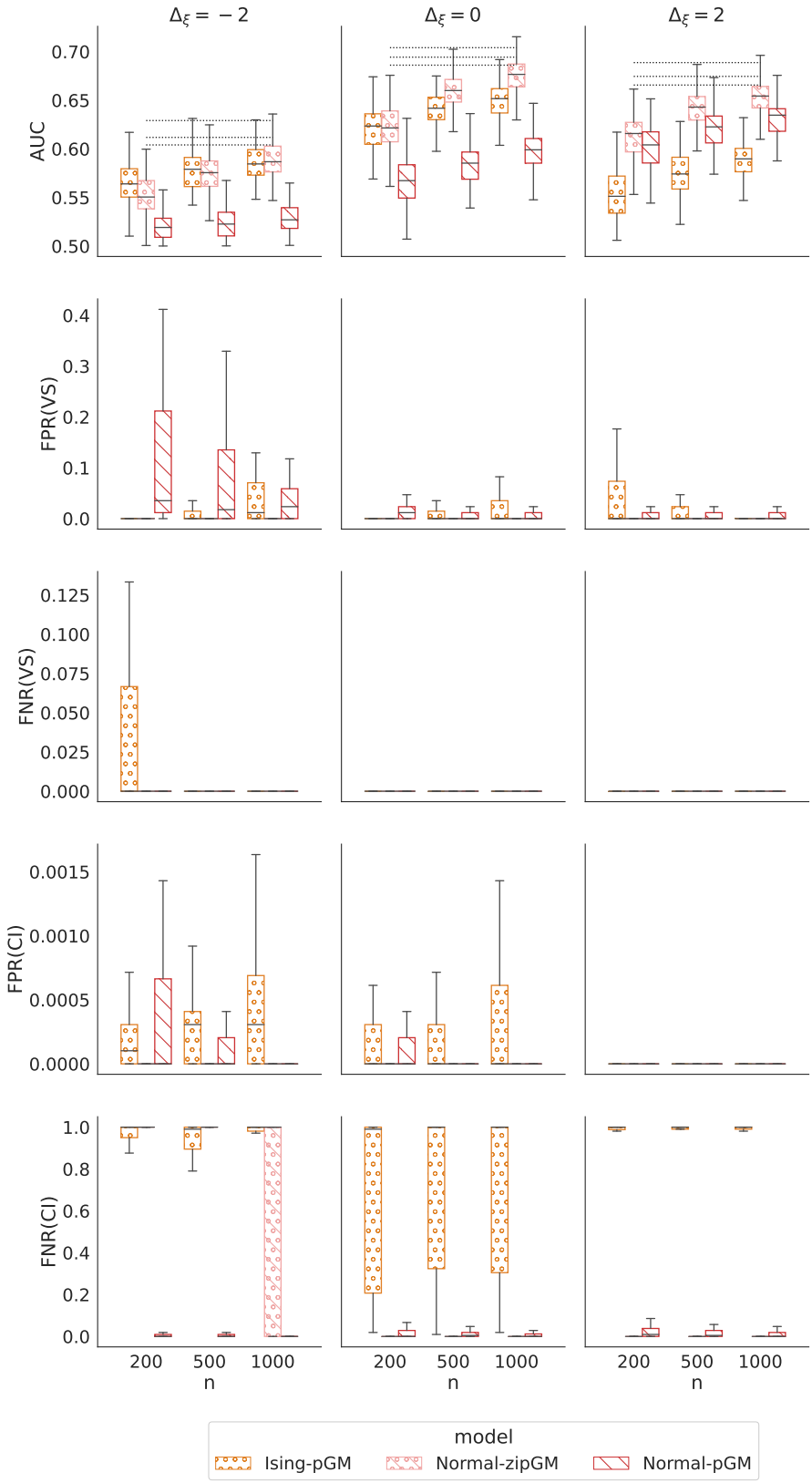


Figura 7.3: Medidas de performance en predicción y selección (filas) para los modelos Ising, Normal-pGM y Normal-zipGM con datos generados por el modelo Normal-zipGM detallado en la Sección 7.1.1 para $n \in \{200, 500, 1000\}$ para distintas proporciones de ceros en los datos (columnas).

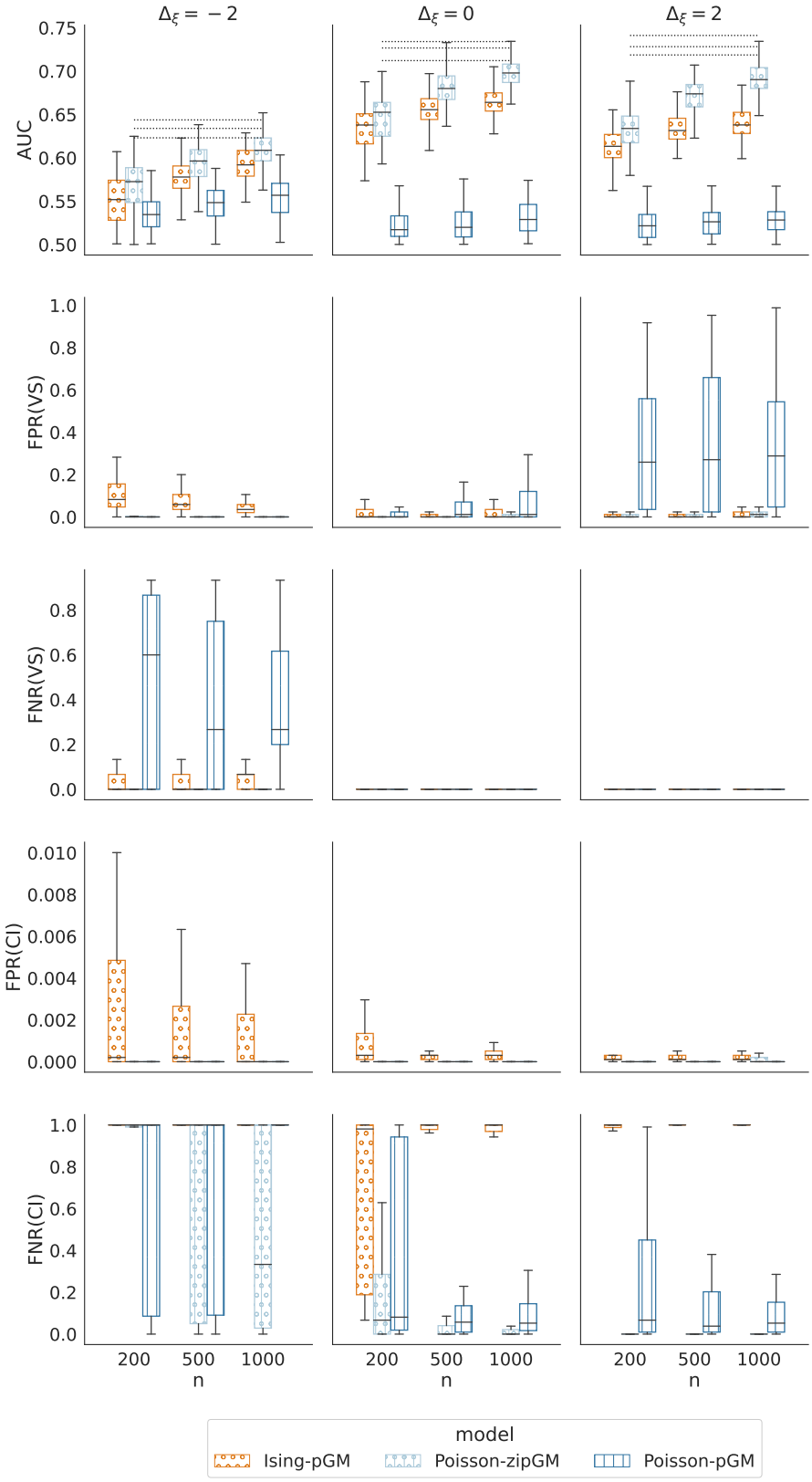


Figura 7.4: Medidas de performance en predicción y selección (filas) para los modelos Ising, Poisson-pGM y Poisson-zipGM con datos generados por el modelo Poisson-zipGM detallado en la Sección 7.1.1 para $n \in \{200, 500, 1000\}$ para distintas proporciiones de ceros en los datos (columnas).

la correcta selección de variables incluso cuando el modelo está mal especificado, como en el caso de los pGMs analizados. Cuando se adopta un criterio de selección basado en predicción, la selección tiende a incorporar algunas variables que no se encontraban vinculadas con la respuesta en el modelo poblacional.

7.1.3 *Influencia de las direcciones del span en modelos pGM*

Por simplicidad, en esta sección únicamente consideramos el caso de respuesta Y binaria.

POISSON-PGM. En la parte (a) de la Figura 7.5 se muestra el efecto de la mala especificación de un modelo Poisson-pGM cuando es entrenado a partir de muestras generadas por un modelo Poisson-zipGM, donde la proporción de ceros, así como la media de las cuentas condicionales de $X \mid \nu(X) = 1$ aumentan o disminuyen simultáneamente como función de la respuesta Y. Es decir, que en el modelo poblacional Poisson-zipGM, las componentes de span $\boldsymbol{\kappa}$ correspondientes a $\boldsymbol{\Gamma}$ y a $\boldsymbol{\Psi}$ tienen signos opuestos. Además consideramos el caso extremo en que $E[X|Y = 1] \approx E[X|Y = 0]$.

En este caso se observa que, como en el modelo Poisson-pGM la probabilidad de observar un cero está vinculada a la distribución de las muestras positivas, si condicional en Y la probabilidad de observar ceros aumenta pero la media de los datos que no son ceros también aumenta, la distribución Poisson pierde capacidad predictiva respecto a la respuesta Y. Este efecto ilustra en (a), donde las líneas punteadas solapadas en el gráfico indican que los modelos perdieron toda capacidad predictiva sobre Y.

NORMAL-PGM En este caso, como se ilustra en la parte (b) de la Figura 7.5, el modelo Normal-pGM es mucho más robusto en cuanto a aprender la distribución de los datos generados por el modelo Normal-zipGM con $\boldsymbol{\eta}_0 \approx \mathbf{0}$; i. e. cuando los datos están centrados condicionalmente a ser no nulos, es decir, $EX \mid (\nu(X) = 1) = 0$. Por otro lado, cuando los datos no están condicionalmente centrados, el modelo Normal-pGM podría caer en el comportamiento de pérdida de señal como en el caso del modelo Poisson-pGM mostrado anteriormente.

Estos resultados sugieren centrar los datos en forma condicional, $(\boldsymbol{X}|\nu(\boldsymbol{X}) = 1)$, para aprender un modelo Normal-pGM en condiciones de exceso de ceros. Esta normalización condicional hace que el parámetro $\boldsymbol{\eta}_0$ en el modelo Normal-zipGM equivalente sea aproximadamente nulo. Formalizamos este argumento mediante la Proposición F.1 en el Apéndice F.2.2, que muestra que cuando $\boldsymbol{\eta}_0$ es no nulo, existen direcciones del span en los modelos condicionales lineales en $T(\boldsymbol{X})$ (Definición 2.12) de la familia zipGM (Definición 2.10 junto con (3.10)) que deja a la RSD (Proposición 3.1) obtenida con los modelos pGM correspondientes (Definición 2.5 junto con (3.10a)) sin capacidad predictiva respecto a la respuesta Y. Por otro lado, cuando $\boldsymbol{\eta}_0 = \mathbf{0}$ en el modelo Normal-zipGM, decimos que el modelo está centrado condi-

cionalmente y el modelo Normal-pGM conserva capacidad predictiva.

7.2 EFECTO DE LAS INTERACCIONES EN MODELOS ZIPGM CON PENALIZACIÓN JERÁRQUICA

Nuevamente consideramos que el modelo poblacional para $\boldsymbol{X} \mid Y$ es Normal-zipGM, Poisson-zipGM o TPoisson-zipGM, pero permitimos interacciones en posiciones aleatorias entre todas las variables. También definimos conjuntos de variables que están relacionadas con la respuesta Y a través de (3.10a), a través de (3.10b) o a través de ambas. Por simplicidad consideramos el caso de respuesta binaria balanceada $Y \in \{0, 1\}$ y la AUC como medida predictiva.

7.2.1 *Parámetros poblacionales*

ESTRUCTURA. Consideramos que $X_1, \ldots, X_5$ están asociadas con Y sólamente a través de $\boldsymbol{\eta}_y$, $X_6, \ldots, X_{10}$, por medio de ambas $\boldsymbol{\eta}_y$ y $\boldsymbol{\xi}_y$ y $X_{11}, \ldots, X_{15}$ únicamente a través de $\boldsymbol{\xi}_y$.

A fin de comprender mejor el efecto de estructuras particulares de las matrices de interacción $\boldsymbol{\Theta}, \boldsymbol{\Lambda}$, y $\boldsymbol{\Psi}$, condieramos tres patrones o bloques típicos: un bloque fuertemente jerárquico (FJ) implica interacciones entre las variables relacionadas con la respuesta, i. e. $X_1, \ldots, X_{15}$, un bloque anti-jerárquico (AJ) considera interacciones entre las variables que no están vinculadas con la respuesta, i. e. $X_{16}, \ldots, X_{100}$; y el bloque débilmente jerárquico (DJ) considera interacciones entre una variable vinculada a la respuesta ($X_1, \ldots, X_{15}$) y una que no lo está ($X_{16}, \ldots, X_{100}$).

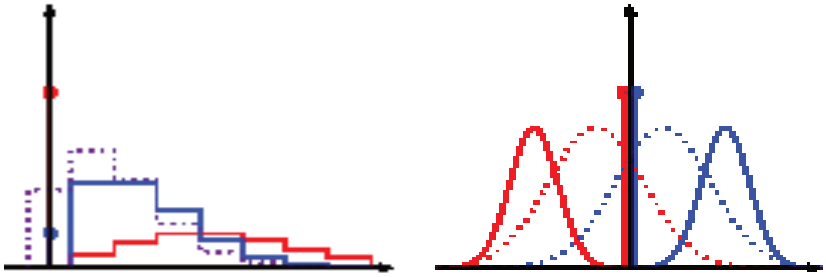
A continuación estudiaremos las siguientes estructuras en las matrices de interacciones:

(S1) Independencia: no contiene ninguna interacción.

(S2) Estructura fuertemente jerárquica: consideramos 30 interacciones en (FJ) formando 3 bloques densos de 5 variables cada uno. Las demás interacciones son nulas.

(S3) Estructura general: Consideramos la misma estructura que en el caso anterior para el bloque (FJ), pero incorporamos 15 interacciones en posiciones aleatorias tanto de (AJ) como de (DJ), definiendo 60 interacciones no nulas en total.

Las interacciones son generadas como en la Sección 7.1.1, aunque permitimos variar la constante de escala c para evaluar el efecto del aumento de interacciones. En los gráficos codificamos la estructura y la constante de escala de las interacciones de acuerdo al siguiente patrón: XX-YY-XX(c), donde XX,YY y ZZ indican la estructura adoptada para el bloque (FJ), (DJ) y (AJ) respectivamente, pudiendo ser `indep`, `blockdiag` o `random`, mientras que el subfijo c indica el valor de la constante de escala en las interacciones c.



(a) Datos generados (líneas sólidas) con un Poisson-zipGM con $\text{sign}(\Gamma) \neq \text{sign}(\Psi)$ y ajustados (líneas punteadas) por el modelo Poisson-pGM condicional a $Y \in \{0, 1\}$ (color). Las distribuciones ajustadas coinciden, indicando la pérdida de información de la respuesta debido a la mala especificación del modelo.

(b) Datos generados (líneas sólidas) con un Normal-zipGM con $\eta_0 = 0$ y ajustados (líneas punteadas) por el modelo Normal-pGM condicional a $Y \in \{0, 1\}$ (color). Como los datos se encuentran condicionalmente centrados, el modelo Normal-pGM no pierde toda la información sobre la respuesta, aunque a mayor proporción de ceros, más solapamiento y pérdida de información.

Figura 7.5: Las líneas muestran la función de probabilidad puntual o densidad condicional a $Y \in \{0, 1\}$, la probabilidad puntual en el origen se especifica con un punto coloreado. En (a) se muestra el caso en que los datos fueron generados a partir de un modelo Poisson-zipGM con esperanza condicional $E[X \mid \nu(X) = 1]$ y proporción de ceros elevada para el caso $Y = 1$ (rojo), mientras que para $Y = 0$ tiene baja esperanza condicional y proporción de ceros (azul), de manera que $E[X|Y = 0] = E[X|Y = 1]$. Las líneas punteadas muestran cómo el modelo Poisson-pGM ajusta a los datos condicionales a la respuesta Y. En particular se observa que el modelo no puede diferenciar las distintas poblaciones condicionales a Y (líneas punteadas solapadas). En (b) consideramos que los datos fueron generados por un modelo Normal-zipGM centrados condicionalmente a que $\nu(X) = 1$, i. e. $\boldsymbol{\eta}_0 = 0$ en (3.10a). Las líneas punteadas indican el ajuste del modelo Normal-pGM, mostrando que existe separación entre las distintas poblaciones, aunque dicha separación se hace menor a medida que aumenta la proporción de ceros en los datos.

MODELO INDEPENDIENTE. Para el modelo Normal-zipGM, generamos η_{0j} y Θ_{jj} como en la Sección 7.1.1, mientras que para los modelos Poisson-zipGM y TPoisson-zipGM consideramos $\eta_{0j} \sim \mathcal{N}(\log(100), 1)$ para todo j, por lo que la esperanza de las cuentas no nulas es de 100. En todos los casos consideramos $\xi_{0j} \sim N(-A^+(\eta_{0j}, \Theta_{jj}), 1)$, por los que los datos obtenidos de los modelos independientes (5.7) contienen aproximadamente un 50 % de ceros en cada variable.

REGRESIÓN. En todos los casos, generamos los valores auxiliares $\bar{\Gamma}_j$ y $\bar{\Psi}_j$ a partir de la mezcla $0.5N(-3, 1) + 0.5N(3, 1)$ y escalamos cada $\bar{\Gamma}_j$, dividiéndo por el desvío estándar de la variable correspondiente X_j bajo el modelo independiente (5.7).

Finalmente, definimos los parámetros de regresión $\Gamma_j = a_1\bar{\Gamma}_j$ y $\Psi_j = a_2\bar{\Psi}_j$, donde las constantes a_1 y a_2 se eligen de manera tal que la AUC en predicción sea aproximadamente 0.65 cuando se utiliza cada componente de la reducción $\boldsymbol{R}(\boldsymbol{X})$ definida en el Corolario 3.3, de esta manera logramos balancear la señal proveniente de la componente $\boldsymbol{T}(\boldsymbol{X})$ del estadístico suficiente del modelo Hurdle y de la componente binaria $\boldsymbol{\nu}(\boldsymbol{x})$.

7.2.2 *Resultados*

NORMAL-ZIPGM. En la Figura 7.6 mostramos para distintos tamaños muestrales los resultados correspondientes a $\boldsymbol{X} \mid Y$ con distribución Normal-zipGM. En la primera columna se muestra el caso independiente (S1), mientras que la segunda y tercera columna corresponden a la estructura general (S3) a medida que aumenta la fuerza de las interacciones.

Cuando la reducción $R(\boldsymbol{X})$ es estimada con el modelo Normal-zipGM, la AUC en predicción es aproximadamente 0.7, alcanzando el AUC de la predicción del modelo poblacional (líneas punteadas), mientras que cuando la reducción $R(\boldsymbol{X})$ es estimada por los modelos Poisson-zipGM and TPoisson-zipGM, la AUC en predicción es significativamente menor, del orden de 0.6 en todos los casos.

En cuanto a las medidas de selección, encontramos que la FPR(VS) es mayor para el modelo Normal-zipGM, mientras que para interacciones fuertes y $n = 1000$, el modelo TPoisson-zipGM también alcanza valores elevados de FPR(VS). Por otro lado, los modelos Poisson-zipGM y TPoisson-zipGM presentan mayor FNR(VS), perdiendo variables asociadas a la respuesta. Por otro lado, la FPR(CI) fue baja (< 0.2) para todos los modelos, excepto por el TPoisson-zipGM $n = 1000$ e interacciones fuertes. Para interacciones moderadas, obtuvimos muy alta FNR(CI), indicando que los modelos no fueron capaces de estimar esas interacciones, mientras que cuando las interacciones fueron más fuertes, la FNR(CI) fue menor en todos los modelos y en particular para tamaños muestrales $n \geq 500$, el modelo TPoisson-zipGM obtuvo la menor FNR(CI), confirmando que la penalización jerárquica redujo tanto la FNR(VS) como la FNR(CI) a medida que las interacciones se hicieron más fuertes, incluso en presencia de mala especificación.

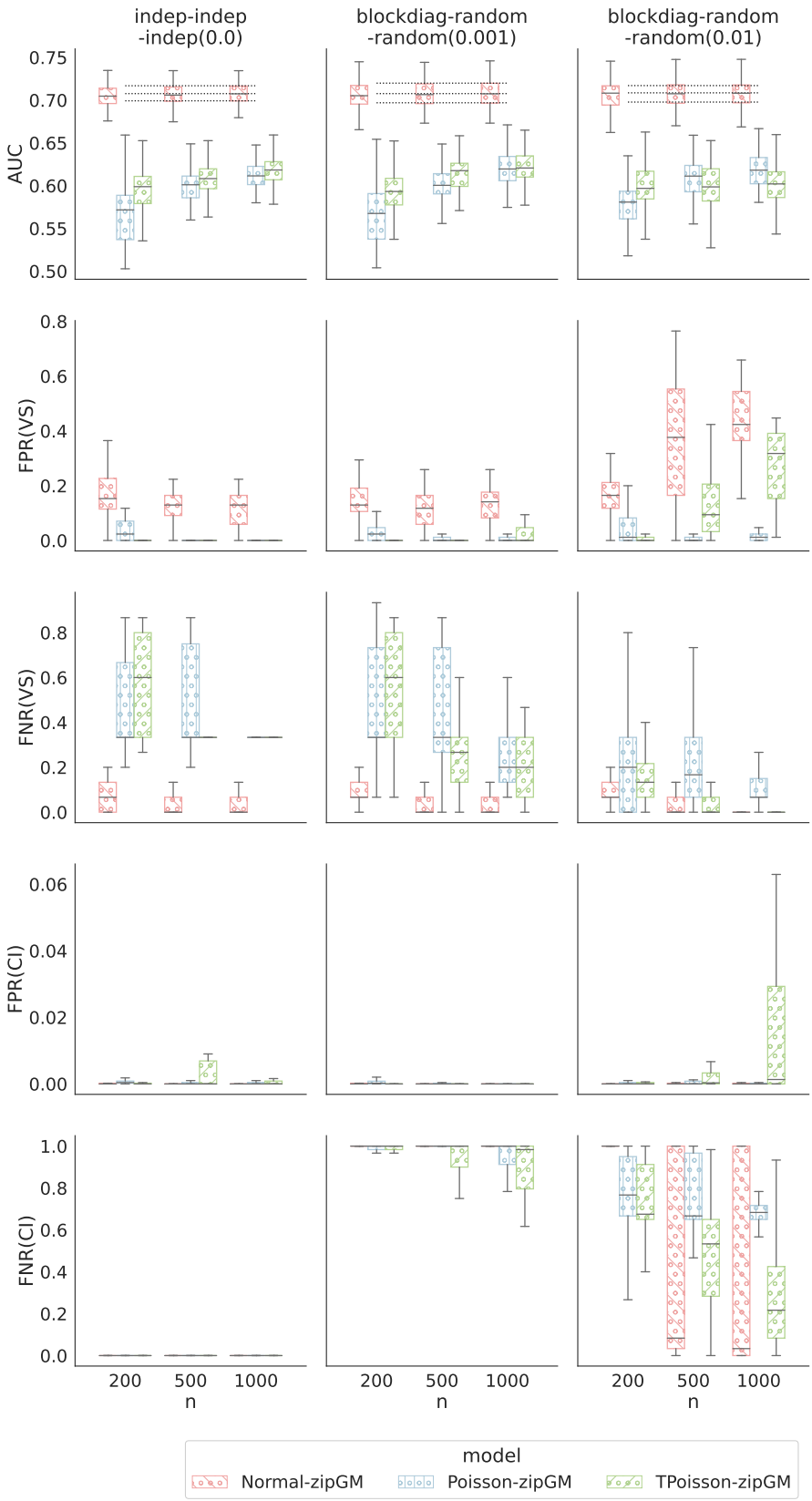


Figura 7.6: Medidas de performance en predicción y selección (filas) para los modelos zipGM propuestos con datos generados por el modelo Normal-zipGM detallado en la Sección 7.2.1 para $n \in \{200, 500, 1000\}$ a medida que la fuerza de las interacciones crece (columnas).

POISSON-ZIPGM. Cuando $\boldsymbol{X} \mid Y$ se distribuye de acuerdo con el modelo Poisson-zipGM (Figura 7.7) bajo independencia (S1), o bajo una estructura general de interacciones (S3) con intensidad creciente (columnas 1, 2 y 3 respectivamente), la AUC en predicción fue similar para los modelos Poisson-zipGM y TPoisson-zipGM, aunque menor para el modelo Normal-zipGM, el cual es más afectado por la mala especificación.

En el caso de interacciones fuertes (tercera columna), todos los modelos convergen, a medida que aumenta el tamaño muestral, a seleccionar un conjunto de variables que no se encontraban vinculadas con la respuesta en el modelo poblacional a través de (3.10). Si bien cuentan como falsos positivos, estas variables interaccionan con las variables vinculadas con la respuesta $(X_1, \ldots, X_{15})$, a través de la estructura (DJ) asumida en (S3). En este caso, la penalización jerárquica incorporó dichas variables, que capturan información de segundo orden sobre la respuesta Y, mejorando en el caso de los modelos Poisson-zipGM y TPoisson-zipGM la predicción dada por el modelo poblacional (líneas punteadas). Bajo independencia o interacciones débiles, la FPR(VS) obtenida por el modelo Normal-zipGM fue mucho mayor que para el Poisson o TPoisson-zipGM.

En todos los casos, la FPR(CI) fue muy baja, mientras que la FNR(CI) fue muy elevada para interacciones débiles, especialmente para tamaños muestrales chicos, pero parece converger a cero lentamente a medida que aumenta el tamaño muestral. En presencia de interacciones fuertes, todos los modelos presentan baja FNR(CI), por lo que logran capturar la estructura del grafo de independencia condicional.

RESULTADOS ADICIONALES. En el Apéndice F.3 mostramos los resultados para el criterio de selección "oráculo", observándose que la penalización jerárquica también es capaz de capturar las variables vinculadas con la respuesta a través de (3.10), pero el criterio de selección basado en predicción tiende a incorporar variables vinculadas con la respuesta a través de las interacciones únicamente. También en el Apéndice F.3 se muestran los resultados correspondientes a la estructura (S2), mostrando que el modelo Normal-zipGM generalmente logra mejor performance en presencia de interacciones fuertes.

7.3 DATOS SIMULADOS TIPO MICROBIOMA

A diferencia de las simulaciones mostradas en las Secciones 7.1 y 7.2 donde generábamos muestras a partir del modelo de regresión inversa $\boldsymbol{X} \mid Y$ perteneciente a la familia zipGM, en este caso consideramos un modelo directo $Y|\boldsymbol{X}$. Usamos el modelo generativo `MB-GAN` (Rong et al., 2021) para generar muestras de $\{\boldsymbol{x}^{(s)}\}$ (cuentas) similares a datos de microbioma y consideramos un modelo de regresión logística para definir una respuesta binaria $Y \in \{0, 1\}$ y un modelo lineal para definir una respuesta continua $Y \in \mathbb{R}$. En ambos casos consideramos la respuesta tanto como función de las cuentas normalizadas $Y \mid \boldsymbol{X}$ como de las indicadoras $Y \mid \boldsymbol{\nu}(\boldsymbol{X})$.

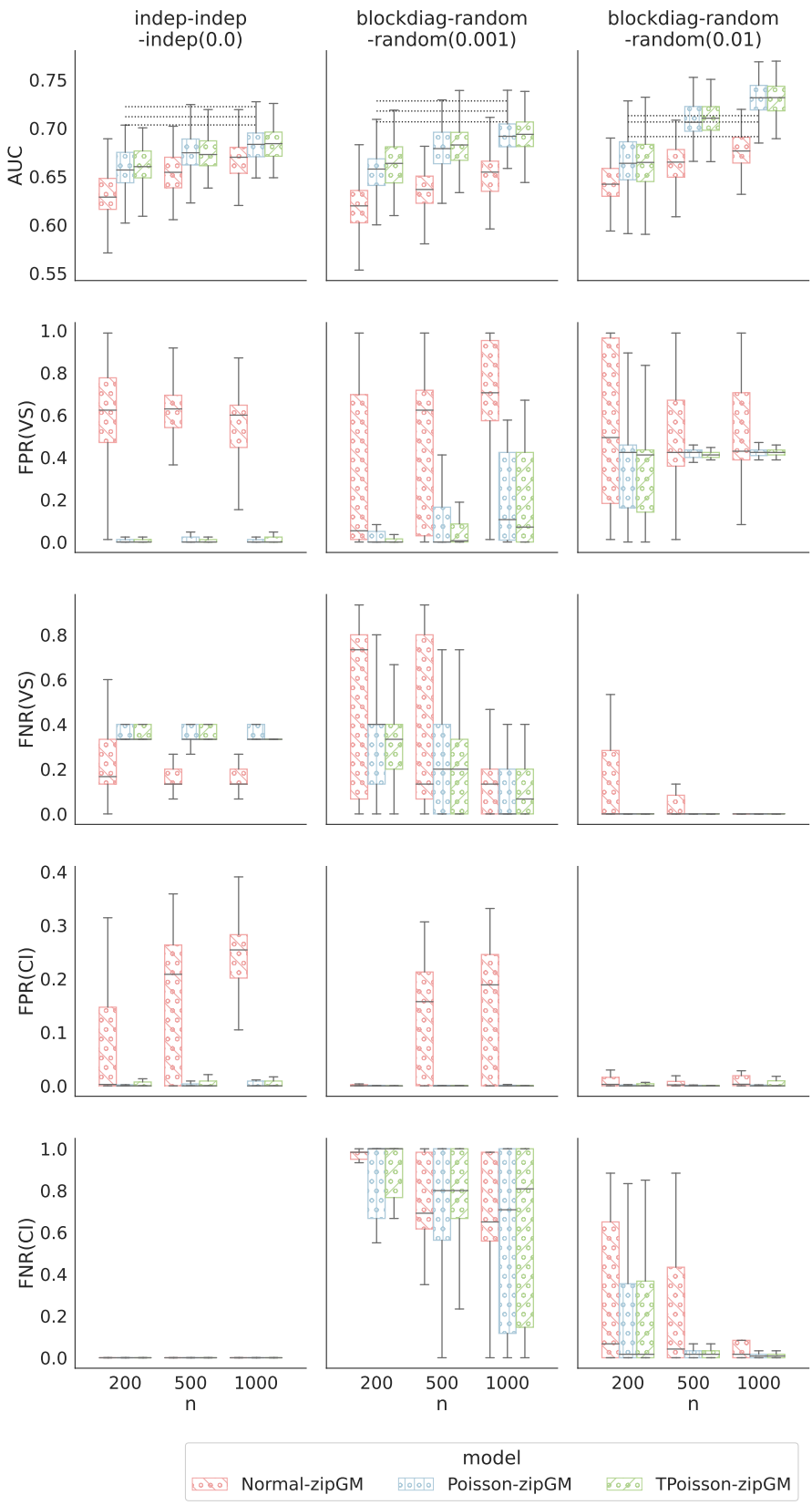


Figura 7.7: Medidas de performance en predicción y selección (filas) para los modelos zipGM propuestos con datos generados por el modelo Poisson-zipGM detallado en la Sección 7.2.1 para $n \in \{200, 500, 1000\}$ a medida que la fuerza de las interacciones crece (columnas).

Para cada muestra $\boldsymbol{x}^{(s)}$, generamos una respuesta $y^{(s)}$ continua o binaria como función de $m_{10}^{(s)}$, la suma de las 10 componentes más abundantes de $\boldsymbol{X}$ normalizada por las cuentas totales $\sum_{j=1}^{p} x_j^s$ en el caso $Y \mid \boldsymbol{X}$, y de la suma de las 10 componentes con mayor entropía de $\boldsymbol{\nu}(\boldsymbol{X})$ en el caso $Y \mid \boldsymbol{\nu}(\boldsymbol{X})$.

CASO RESPUESTA CONTINUA. En cada caso obtuvimos la respuesta $Y \mid \boldsymbol{X}$ o $Y \mid \boldsymbol{\nu}(\boldsymbol{X})$ continua estandarizando la función de los correspondientes predictores $m_{10}^{(s)}$. Comparamos la performance en predicción como en selección de variables contra el método Sparse Partial Least Squares (SPLS), disponible en el paquete `mixOmics` (Rohart et al., 2017).

CASO RESPUESTA BINARIA. Generamos una respuesta binaria Y a partir de los predictores $\boldsymbol{X}$ o $\boldsymbol{\nu}(\boldsymbol{X})$ definiendo el modelo logístico $\text{logit}(p^{(s)}) = \tilde{m}_{10}^{(s)}$, donde $\tilde{m}_{10}^{(s)}$ es la función de los predictores correspondientes definida anteriormente para cada caso. Luego generamos $y^{(s)} \sim \text{Bernoulli}(p^{(s)})$. Comparamos la performance tanto en predicción como en selección de variables contra el método Sparse Partial Least Squares Discriminant Analysis (SPLS-DA), también disponible en el paquete `mixOmics` (Rohart et al., 2017).

7.3.1 *Resultados*

En todos los casos estimamos la RSD, $R(\boldsymbol{X}) \in \mathbb{R}$, dada en el Corolario 3.2 con $f_y = y$ para los modelos zipGM.

CASO RESPUESTA CONTINUA. La primera columna de la Figura 7.8 muestra que cuando la respuesta Y es generada utilizando únicamente $\boldsymbol{\nu}(\boldsymbol{X})$ como predictor, el método SPLS (mixOmics) presenta un MSE mucho mayor que los modelos zipGM, los cuales logran una performance similar para todos los tamaños muestrales considerados. Cuando Y depende de las cuentas $\boldsymbol{X}$ (segunda columna de la Figura 7.8), el modelo TPoisson-zipGM junto con el estimador SPLS logran el menor MSE en predicción, mientras que el modelo Normal-zipGM presenta una baja performance, lo cual se puede adjudicar a la mala especificación del modelo. En todo caso, el estimador SPLS no logra estimar correctamente las variables asociadas con la respuesta, presentando un FNR(VS) muy alto para todos los tamaños muestrales y configuraciones.

CASO RESPUESTA BINARIA. En la Figura 7.9 se muestran los resultados correspondientes al caso de respuesta Y binaria. Para el modelo $Y \mid \boldsymbol{\nu}(\boldsymbol{X})$ (primera columna), todos los métodos logran una predicción cercana al modelo poblacional. Sin embargo, el modelo SPLS-DA (mixOmics) presenta una FNR(VS) muy alta, mientras que la FPR(VS) fue similar en todos los métodos. Por otro lado, cuando consideramos el modelo $Y \mid \boldsymbol{X}$ (segunda columna), los modelos Poisson-zipGM y TPoisson-zipGM lograron superar por poco al estimador SPLS-DA de mixOmics en predicción, mientras que el modelo Normal-zipGM pre-

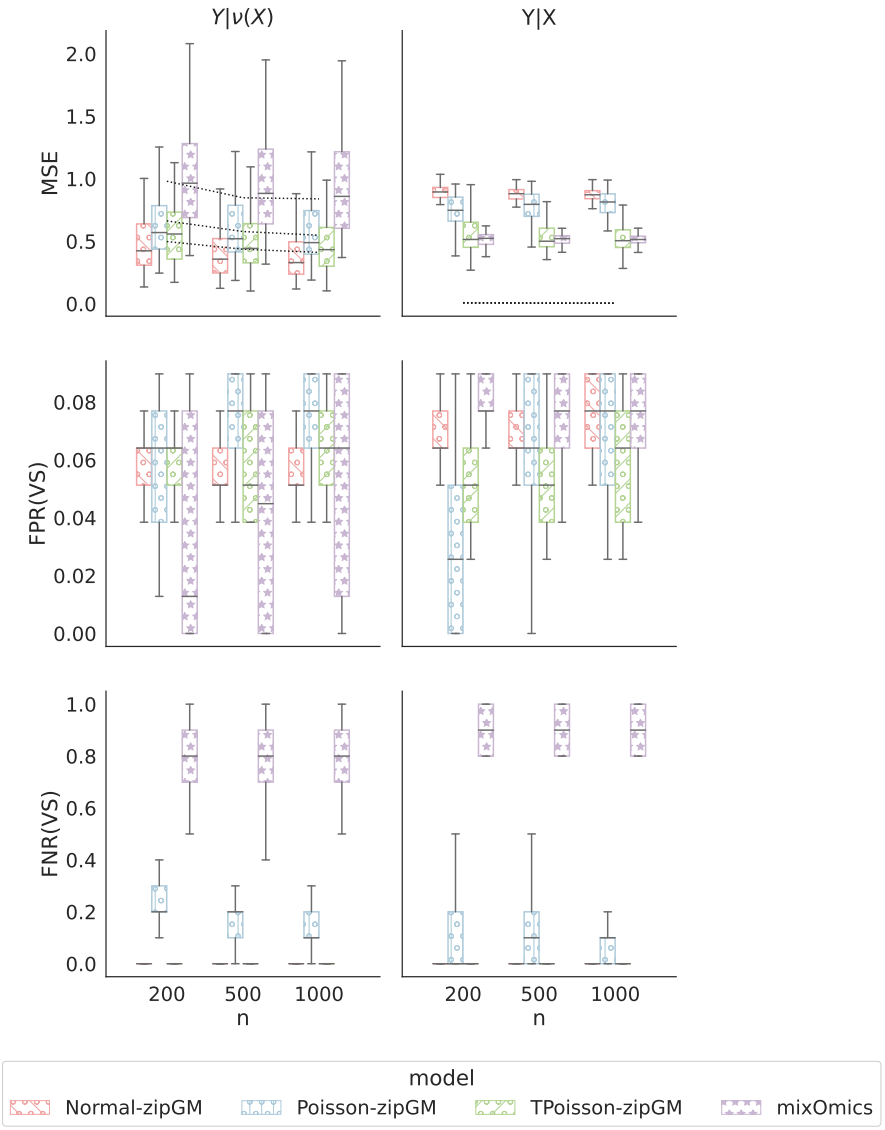


Figura 7.8: Medidas de performance en predicción y selección de variables (filas) para los modelos zipGM propuestos y el modelo SPLS del paquete `mixOmics`, cuando los datos son generados por un modelo directo $Y|\boldsymbol{\nu}(\boldsymbol{X})$ (columna 1) or $Y|\boldsymbol{X}$ (columna 2) para respuesta continua Y basados en muestras generadas por el paquete `MB-GAN` para $n \in \{200, 500, 1000\}$.

sentó una pérdida importante de información debido a la mala especificación, causada principalmente porque la transformación logarítmica (Aitchison y Bacon-Shon, 1984) aplicada a los predictores distorsiona los datos debido a que las cuentas son pequeñas ($E\boldsymbol{X} \mid \boldsymbol{\nu}(\boldsymbol{x}) = \mathbf{1}$). Al igual que en el caso de respuesta continua, el estimador SPLS-DA de mixOmics tiene la FPR(VS) más alta, mientras que los modelos zipGM presentan mucha variabilidad. En todos los casos la FNR(VS) fue muy baja.

RESULTADOS ADICIONALES. Los resultados correspondientes al criterio de selección "oráculo" se encuentran en el Apéndice F.4, logrando excelentes resultados en selección de variables, demostrando nuevamente la robustez de la penalización adoptada.

7.4 COMENTARIOS DE CIERRE DEL CAPÍTULO

Las pruebas con datos simulados presentadas en este capítulo ilustran las ventajas aportadas por los procedimientos propuestos de RSD para datos composicionales de alta dimensión que tienen en cuenta la alta proporción de ceros, como así también las potenciales limitaciones. La primera serie de simulaciones, con datos generados bajo modelos zipGM, muestran las ventajas de considerar los patrones de ceros al estimar la reducción. En particular, cuando la proporción de ceros aumenta considerablemente, el desempeño de los modelos no adaptados al exceso de ceros se deteriora rápidamente, tanto que suele resultar conveniente modelar únicamente los patrones de ceros mediante un modelo Ising, sin agregar información cuantitativa. Los diversos escenarios estudiados también nos muestran que el desempeño de los métodos propuestos, con muestras finitas, depende de interrelaciones complejas entre la respuesta y los predictores y patrones de ceros, como así también de la propia estructura de correlación presente en los predictores, los patrones de ceros y entre ambos. De modo general, los estimadores propuestos con penalización jerárquica logran capturar correctamente los patrones de dependencia con la respuesta cuando las correlaciones son suficientemente fuertes.

La segunda tanda de experimentos usando datos que semejan notablemente los datos reales de microbioma permiten una evaluación en un escenario de aplicación más realista. Estos resultados muestran que los procedimientos propuestos permiten obtener reducciones más útiles que las herramientas disponibles para interpretar la relación subyacente entre datos composicionales y la respuesta de interés, sin deteriorar comparativamente la capacidad predictiva. Esta ventaja se observa principalmente en el desempeño en la identificación de variables relevantes. La observación es importante porque el escenario analizado supone una condición ideal para la identificación de variables relevantes mediante el uso de herramientas como SPLS y SPLS-DA, en el sentido que las reducciones son intrínsecamente unidimensionales. Para reducciones de mayor dimensión, estos pro-

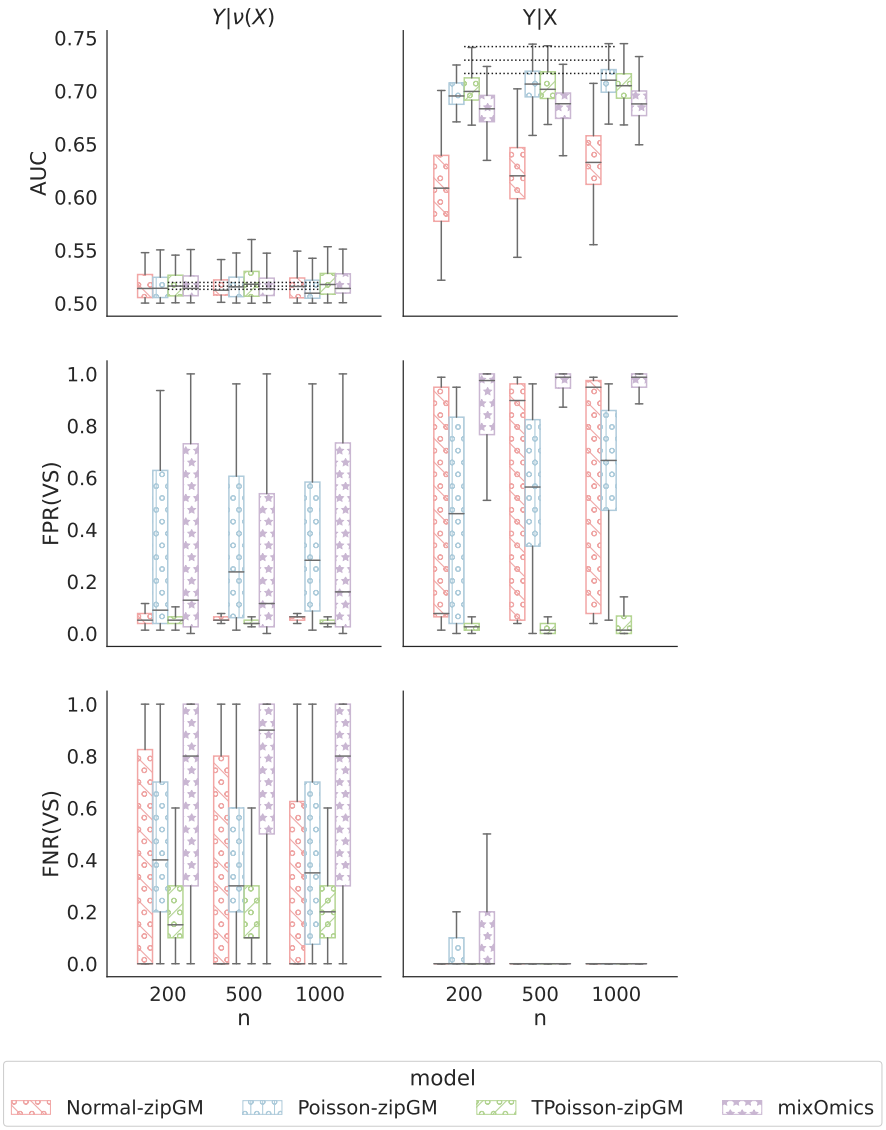


Figura 7.9: Medidas de performance en predicción y selección de variables (filas) para los modelos zipGM propuestos y el modelo SPLSDA del paquete `mixOmics`, cuando los datos son generados por un modelo directo $Y|\boldsymbol{\nu}(\boldsymbol{X})$ (columna 1) or $Y|\boldsymbol{X}$ (columna 2) para respuesta binaria Y basados en muestras generadas por el paquete `MB-GAN` para $n \in \{200, 500, 1000\}$.

cedimientos ofrecen selección de variables por coordenadas, no de forma integral como sí lo hacen los métodos propuestos en esta tesis.

8

APLICACIÓN A DATOS REALES

El avance tecnológico en la secuenciación genética permiten actualmente obtener mediciones precisas del perfil genético del microbioma. A pesar de ello, la gran complejidad de estos datos, sumado a ciertas características particulares como la gran proporción de ceros a niveles taxonómicos inferiores (ya que sólo algunas especies se encuentran en cada muestra) y el hecho de ser datos composicionales definidos sobre los enteros positivos, establecen desafíos importantes para encontrar asociaciones entre dichos perfiles microbianos y las variables de interés clínico consideradas como respuesta.

En la Sección 8.1 mostramos la aplicación de los modelos pGM al análisis de datos de microbioma en relación al nivel taxonómico de las variables predictivas. Para ello consideramos los mismos datos en dos niveles taxonómicos distintos: uno que resulta de agregar información de los niveles inferiores, dando lugar a un escenario con pocas variables donde los modelos clásicos funcionan relativamente bien; otro correspondiente a un nivel taxonómico inferior, que resulta en un problema de alta dimensión y gran proporción de ceros. La motivación principal de este ejemplo es mostrar que los estimadores de la RSD basados en scores propios logran escalar satisfactoriamente permitiendo analizar datos de microbioma a niveles taxonómicos inferiores (más detallados), aspecto difícil de lograr con los estimadores de máxima verosimilitud propuestos originalmente en Tomassi et al. (2019). A su vez, ilustraremos las reducciones obtenidas por los distintos modelos, y observaremos que los modelos zipGM obtienen mejor separación entre clases que los modelos pGM correspondientes.

En la Sección 8.2, mostraremos la aplicación de los nuevos métodos propuestos para datos de microbioma, que agregan el modelado explícito de los patrones de ceros en los datos. En este caso, consideramos como respuesta el sexo de la persona de quien se le tomó la muestra. Notamos que no estamos interesados en predecir el sexo de un individuo en función de la composición de su microbiota intestinal, sino que existe evidencia suficiente sobre dicha asociación. La variable respuesta elegida nos permitirá una comparación directa con métodos de reducción de dimensiones usados frecuentemente con datos de microbioma que sólo identifican los componentes relevantes en el caso de respuestas binarias. El objetivo de este análisis deberá entenderse desde un punto de vista exploratorio y estará centrado principalmente en determinar qué variables (especies) están asociadas a la respuesta, como así también buscaremos investigar las asociaciones entre las especies involucradas.

8.1 HUMAN MICROBIOME PROJECT

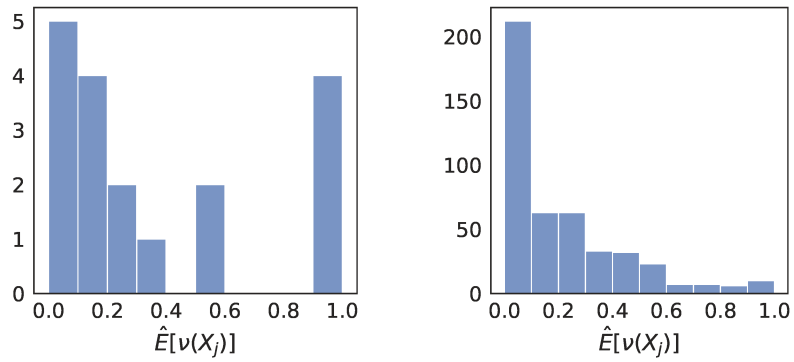
Tomassi et al. (2019) propusieron RSDs basadas en modelos para datos de conteo para visualizar posibles asociaciones globales entre la composición de la microbiota y una variable de interés, pudiendo también identificar los principales componentes responsables de tal asociación. No obstante, sus estimadores basados en máxima verosimilitud, no escalan adecuadamente con el número de variables. En esta sección mostramos que los estimadores basados en scores propios propuestos en esta tesis (Sección 4.3) ofrecen una solución a tales limitaciones.

Tomando el ejemplo presentado en Tomassi et al. (2019), analizamos los datos del Proyecto Microbioma Humano (HMP, por sus siglas en Inglés), disponibles en `https://commonfund.nih.gov/hmp`, en el cual se describen las comunidades de microbioma en en relación con el lugar del cuerpo de donde se obtuvieron las muestras. En particular, se consideraron como respuesta Y las categorías `Anterior_nares`, `Left_Anticubital_fossa`, `Posterior_fornix`, `saliva` y `stool`, correspondientes a muestras obtenidas mediante hisopado de la nariz, piel, vagina, saliva y materia fecal, respectivamente.

Indexando cada respuesta por un entero, $Y \in \{0, \cdots, 4\}$, definimos la función $f(y)$ correspondiente al caso discreto tratado en la Sección 3.2.1. Con esto logramos evitar restricciones adicionales o problemas de identificabilidad en la matriz de regresión Γ del modelo gráfico de segundo orden condicional lineal en $T(\boldsymbol{X})$ (Definición 2.12) asociada a los Ejemplos 2.1-2.6.

Consideramos los niveles taxonómicos L2 (*Phylum*) y L6 (*Genus*). En el primer caso, luego de eliminar variables con menos de 5 mediciones a lo lardo de todo el conjunto de datos, el vector de predictores composicionales $\boldsymbol{X}$ contiene $p = 23$ componentes y se dispone de $n = 681$ muestras. En cambio, luego de filtrar las variables menos expresadas en el nivel taxonómico 6, se obtuvieron $p = 456$ componentes, con una gran proporción de ceros. En la Figura 8.1 se muestra la cantidad de variables con una proporción de ceros dada. Si bien en el nivel L2 tenemos variables con proporción de ceros mayor al 90 %, en L6 esto se acentúa aún más, ya que casi la mitad de las variables predictivas cuentan con una proporción de ceros mayor al 90 %.

MODELOS CONSIDERADOS. En ambos casos consideramos los modelos Ising-pGM, Normal-pGM, Poisson-pGM, FPoisson-pGM y sqPoisson-pGM dados en los Ejemplos 2.1-2.6 optimizando la función de costo penalizada (5.3) mediente el Algoritmo 4. Los scores considerados en cada caso se encuentran resumidos en la Tabla 8.1. En el caso del nivel taxonómico L6 consideramos además los modelos Normal-zipGM y Poisson-zipGM presentados en los Ejemplos 2.7 y 2.8 respectivamente, empleando la función de costo penalizada (5.6) definida por la pseudolikelihood y empleando el algoritmo dado en la Sección 6.3.



(a) L2 (b) L6

Figura 8.1: Proporción de ceros en los datos HMP para los niveles taxonómicos L2 y L6.

MODELO	SCORE
Normal-pGM	score-matching (4.47)
Ising-pGM	ratio-matching (4.49)
Poisson-pGM	score propio de composición (4.51)
FPoisson-pGM	pseudolikelihood (4.58)
sqPoisson-pGM	score propio de composición (4.62)

Tabla 8.1: Scores considerados para cada modelo.

8.1.1 *Resultados*

En la Tabla 8.2 se resumen los resultados en predicción para ambos niveles taxonómicos considerados. Los resultados obtenidos usando los scores (Tabla 8.1) fueron similares a los publicados por Tomassi et al. (2019).

Observamos que al considerar los predictores a un nivel taxonómico mayor (L6) permite que el modelo Ising-pGM gane poder predictivo, mientras que los modelos Poisson-pGM y Normal-pGM lo pierden producto del efecto de la mala especificación causada por una proporción de ceros excesiva para estos modelos. Esto se debe, en parte, a que aumentar la variabilidad de los predictores binarios en el nivel L6 respecto al L2. Esta tendencia también se observa en las Figuras 8.2 y 8.3, donde se muestran las reducciones obtenidas por los modelos Ising-pGM y sqPoisson-pGM, respectivamente, para el nivel taxonómico L2 y L6.

Por otro lado, en la Figura 8.4 se muestran las reducciones obtenidas en el nivel taxonómico L6 por los modelos Normal-pGM y Normal-zipGM, donde se observa que el modelo zipGM presenta mayor separación entre clases que su contraparte pGM.

En el Apéndice G.1 se muestran las reducciones obtenidas con los modelos Normal-pGM, Poisson-pGM y FPoisson-pGM y Poisson-zipGM. Además se detallan distintas medidas de predicción, así como las interacciones seleccionadas por los distintos modelos en los niveles

taxonómicos L2 y L6. En relación con la capacidad predictiva de los modelos, se observó que las clases con menor cantidad de muestras (`Left_Anticubital_fossa` y `Posterior_fornix`) son más desfavorecidas en el nivel L2, mientras que su predicción es mejor en el nivel L6, dando cuenta de la pérdida de información en la composición de los predictores. Por otro lado, las interacciones estimadas fueron más sparse en el modelo sqPoisson-pGM, seguido por el modelo Normal-pGM, el cual comparte en gran parte su estructura con los modelos Poisson-pGM y Ising-pGM, aunque estos las estiman en mayor número. Las interacciones estimadas por los modelos zipGM presentan un patrón similar a las de los correspondientes pGM.

NIVEL TAXONÓMICO	MODELO	PROMEDIO	STD
L2	Ising-pGM	0.74	0.05
	Normal-pGM	0.91	0.05
	Poisson-pGM	0.89	0.04
	FPoisson-pGM	0.89	0.09
	sqPoisson-pGM	0.88	0.08
L6	Ising-pGM	0.9	0.01
	Normal-pGM	0.86	0.09
	Poisson-pGM	0.88	0.06
	FPoisson-pGM	0.91	0.07
	sqPoisson-pGM	0.93	0.04
	Normal-zipGM	0.95	0.03
	Poisson-zipGM	0.93	0.04

Tabla 8.2: Error de predicción (accuracy) de la respuesta Y basándonos en distintas RSDs de los datos de microbioma HMP en niveles taxonómicos L2 y L6 sobre 10 particiones independientes de test mediante validación cruzada.

8.2 AMERICAN GUT PROJECT

En esta sección estudiamos la asociación entre datos composicionales de microbioma de intestino con el sexo de la persona. Para ello utilizamos los datos del Proyecto Intestino Americano (AGP, por sus iniciales en Inglés), analizados a nivel de Genus (L6). Los datos se encuentran disponibles en `http://americangut.org`. La elección del escenario experimental intenta facilitar la comparación con otros métodos analíticos usados frecuentemente en el análisis de datos de microbioma, que no son aplicables a respuestas continuas o con más categorías.

PREPROCESAMIENTO. Las muestras de individuos de edad menor a 50 años con al menos 1000 cuentas fueron normalizadas a 1000 cuentas mediante el proceso denominado rarefacción con el objetivo de mejorar las características estadísticas de los datos (Hong et al., 2022).

Luego de excluir las variables que sólo se encuentran en menos del 10 % de las muestras, nos quedamos con $p = 71$ variables descriptivas en 937 mediciones, de las cuales 488 son mujeres y 449 hombres.

MODELOS CONSIDERADOS. Ajustamos nuestros modelos zipGM, así como los pGM usando el Algoritmo propuesto en la Sección 6.3 con el score definido por la pseudolikelihood (Definición 4.6) y la AUC como medida predictiva. También incluimos el modelo gráfico diferencial de segundo orden (Definición 4.12) junto con la función de costo (4.34) penalizada por la norma ℓ_1. Comparamos los resultados obtenidos tanto en predicción como en selección de variables con los métodos SPLS-DA del paquete `mixOmics` (Le Cao, Boitard y Besse, 2011; Rohart et al., 2017), *selbal* (Rivera-Pinto et al., 2018) y *CoDA-lasso*, implementado en el paquete `coda4microbiome` (Calle, Pujolassos y Susin, 2023). *Selbal* y *CoDA-lasso* consideran todas las transformaciones logarítmicas de las abundancias relativas $\log(X_i/X_j)$ para todo $i \neq j$ y realizan la selección de las variables asociadas a una respuesta binaria.

Como *mixOmics*, *selbal* y *coda4microbiome* no brindan una estimación interpretable de las interacciones, estimamos modelos gráficos usando los paquetes especializados para datos de microbioma: *Spring* (Yoon, Gaynanova y Müller, 2019) y *SpiecEasi* (Kurtz et al., 2015). Como estos modelos no ajustan a una covariable o repuesta, también entrenamos nuestros modelos con los datos sin condicionar en la respuesta de manera de poder compararlos fácilmente. *Spring* construye un modelo cópula gaussiano truncado, estimando las correlaciones entre variables latentes mediante un estadístico de rango. El modelo se aplica soble los datos pretransformados usando una transformación logarítmica normalizada por la media geométrica de las cuentas no nulas (transformación CLR, por Centered Log-Ratio), conservando los ceros en los datos. Por su parte, *SpiecEasi* considera la misma transformación logarítmica y entrena un modelo gráfico Normal-pGM. Ambos métodos inducen conjuntos de independencia condicional entre las variables penalizando las correspondientes entradas de las matrices de interacciones.

VALIDACIÓN CRUZADA. Para seleccionar los modelos, consideramos validación cruzada (Sección 5.5) con 5 particiones estratificadas por sexo para reportar resultados insesgados en predicción. Para aplicar el algoritmo propuesto en la Sección 6.3, definimos dentro de cada partición de entrenamiento 5 nuevas particiones y consideramos el promedio del error de validación obtenido sobre cada una de estas particiones para seleccionar el valor más adecuado de los parámetros de regularización (i. e. selección del mejor modelo). Para simplificar la interpretación y comparación entre los distintos métodos, en todos los casos consideramos los modelos que seleccionan hasta 15 variables en cada partición.

8.2.1 *Resultados*

PREDICCIÓN Y SELECCIÓN DE VARIABLES. En la Figura 8.5 se muestra la cantidad de veces que cada variable es seleccionada como asociada a la respuesta y en la Figura 8.6 la AUC en predicción obtenida por cada modelo a partir de las 5 particiones consideradas de entrenamiento.

Los modelos Poisson-zipGM y TPoisson-zipGM seleccionaron las mismas variables en cada partición, mientras que los modelos pGM estándard así como los modelos de referencia *mixOmics*, *coda4microbiome* y *selbal* muestran una selección con mucha variabilidad, ya que seleccionaron diferentes conjuntos de variables en cada partición de los datos. El modelo diferencial de segundo orden comparte principalmente las variables seleccionadas con los modelos Poisson y TPoisson-zipGM, aunque su selección es poco estable y su capacidad predictiva más baja y comparable con los modelos pGM. Si consideramos el conjunto total de variables seleccionadas a partir de los 5 conjuntos de entrenamiento, observamos que *mixOmics* y *coda4microbiome* requirieron el mayor número de variables: 23 y 24 respectivamente, y obtuvieron un AUC en predicción de 0.65 en ambos casos, mientras que los modelos Poisson-zipGM y TPoisson-zipGM usaron sólamente 13 variables para predecir la respuesta, resultando en una AUC en predicción de 0.63 en ambos casos, performance similar a la obtenida por *selbal*, pero con una representación más compacta en términos de la cantidad de variables seleccionadas. La AUC en predicción para el modelo Normal-zipGM fue ligeramente menor, pero la performance en predicción de los modelos pGM estándard considerados fue mucho más baja. En particular observamos que el modelo Ising obtuvo una AUC en predicción de 0.57.

Estos resultados sugieren que el modelado conjunto del patrón de ceros en los datos mejora la performance tanto en predicción como en selección de variables respecto de los pGMs. También muestran que modelar solamente el patrón de ceros de los datos, como en el caso del modelo Ising, pierde información importante sobre la respuesta considerada.

Por otro lado, comparamos la capacidad predictiva contenida en las variables seleccionadas de forma estable, es decir, aquellas que fueron seleccionadas en cada una de las particiones de los datos de entrenamiento. En la Tabla 8.4 mostramos el error en predicción estimado en una nueva validación cruzada de 10 particiones usando dichas variables estables para predecir el sexo utilizando distintos modelos estadísticos: regresión logística (LR, paquete `glm`), analisis discriminante cuadrático (qda, paquete `MASS`), random forest (RF, paquete `randomForest`), classificadores de vector soporte con kernel polinómico (svm-poly, paquete `kernlab`).

Los modelos predictivos LR y qda, logran el menor error en predicción (accuracy) cuando utilizan las variables estables estimadas por *mixOmics*, 0.37 y 0.38 respectivamente, cercano al error obtenido al emplear las variables estables estimadas por los modelos Poisson-zipGM y TPoisson-zipGM, logrando un error en predicción de 0.39

en ambos casos. Cuando empleamos svm como modelo predictivo, el error en predicción obtenido al emplear las variables estables estimadas por los modelos Poisson-zipGM y TPoisson-zipGM como predictores es de 0.34, notablemente menor al obtenido utilizando las variables seleccionadas por los otros modelos. El error en predicción al utilizar las variables seleccionadas por los modelos Normal-pGM, Normal-zipGM, Poisson-PGM y *selbal* fue substancialmente mayor para todos los modelos predictivos.

Si bien no hay un acuerdo amplio en la literatura sobre posibles diferencias típicas en la composición de la microbiota intestinal en función del sexo, las variables identificadas por nuestros modelos son coherentes con los resultados de Kim et al. (2019), que reportan una asociación entre *Bacteroides-Prevotella* y sexo.

INDEPENDENCIA CONDICIONAL. Los correspondientes grafos de interacciones estimados a partir de las variables descriptivas $\boldsymbol{X}$ se muestran en gris en la Figura 8.7 para los modelos propuestos y los algoritmos especializados *SpiecEasi* y *Spring*. Además, para los modelos pGM y zipGM propuestos, se muestra en azul la selección de variables e interacciones obtenidas al considerar el modelo condicional $X \mid Y$ junto con la penalización jerárquica (5.6).

Se puede observar que los modelos zipGM son más robustos seleccionando interacciones que los correspondientes pGMs, mostrando mayor intersección tanto entre ellos como con los algoritmos *SpiecEasi* y *Spring*. Este comportamiento se puede observar sobre el cluster formado por los nodos $\{6, \cdots, 10\}$, los cuales no fueron seleccionados ni por el modelo Ising ni por los modelos Poisson y TPoisson-pGM, pero sí por todos los zipGMs propuestos, así como por los algoritmos especializados.

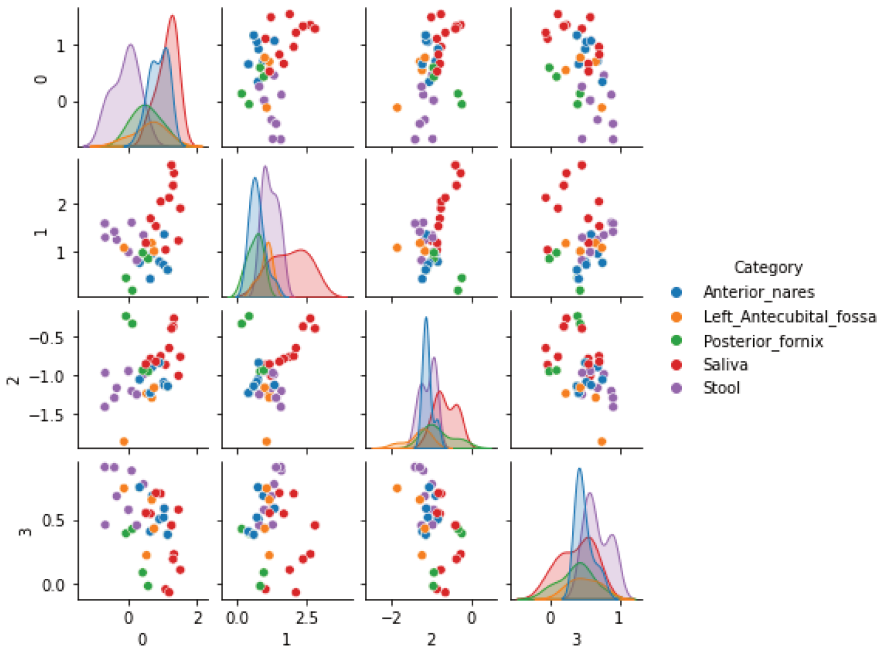
RESULTADOS ADICIONALES. En el Apéndice G.2 se muestra la selección de interacciones (aristas) y de variables (nodos) inducida por la penalización jerárquica en el modelo Normal-pGM para una grilla regular de los parámetros de regularización $\lambda_{\mathcal{R}}, \lambda_{\mathcal{C}}$, caracterizando su comportamiento.

En conjunto, con los resultados obtenidos utilizando el criterio "oráculo" en el Capítulo 7, se observa que la estructura inducida por la penalización es suficientemente flexible para detectar distintos patrones tanto en las variables asociadas con la respuesta como en las interacciones entre ellas. La principal limitación a la hora de controlar el nivel de falsos positivos en selección es el modelo predictivo. Esto sugiere el estudio de otros criterios de selección de modelos o la aplicación de métodos de submuestreo para disminuir la varianza en selección.

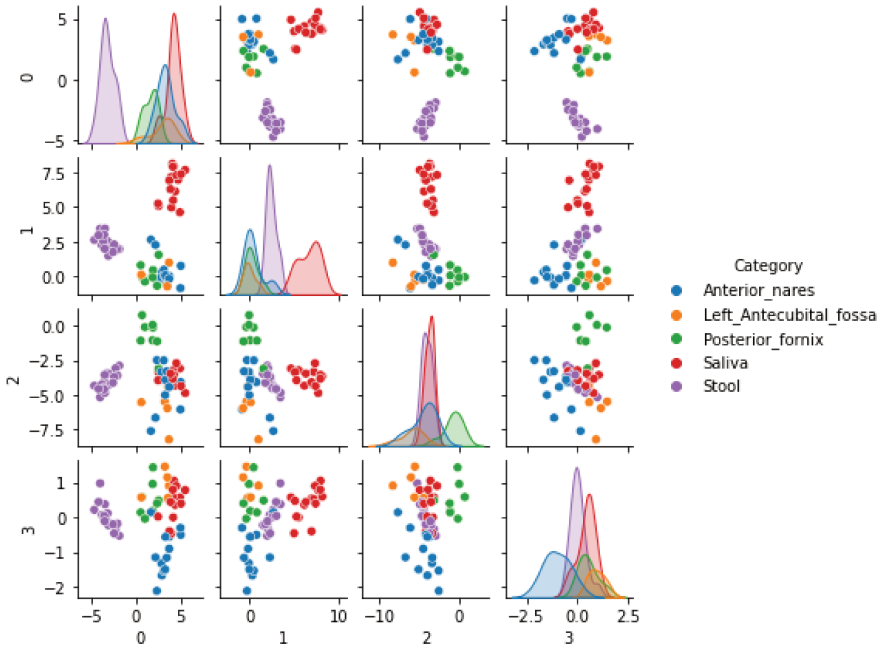
8.3 COMENTARIOS DE CIERRE DEL CAPÍTULO

Los ejemplos presentados en este capítulo confirman en primer lugar la utilidad de estimadores de la RSD basados en criterios de

optimización alternativos a la estimación de máxima verosimilitud, que presenten ventajas computacionales importantes sin degradar de forma significativa las propiedades estadísticas. Los resultados obtendidos con modelos pGM muestran también que estas ventajas computacionales no son suficientes para abordar problemas complejos con datos de microbioma al nivel de detalle requerido por aplicaciones de valor práctico. En línea con los resultados obtenidos con datos simulados, el análisis de datos reales de microbioma revela la importancia de modelar adecuadamente los patrones de ceros presentes en los datos. En ausencia de esta capacidad, cuando la proporción de ceros es elevada, los modelos tipo Ising que describen patrones de co-ocurrencia de predictores mostraron mejor desempeño que el modelado más fino de abundancias. Al mismo tiempo, el modelado conjunto de patrones de ceros junto con la información cuantitativa dio consistentemente mejores resultados en el ejemplo con datos del American Gut Project que el uso de modelos tipo Ising sin información cuantitativa adicional. Los resultados referidos a la selección de variables relevantes también muestran que los métodos propuestos de RSD representan una alternativa interesante a los métodos establecidos de representación de datos de microbioma en baja dimensión, ya que logran identificar de forma más estable las variables más importantes que describen la relación funcional entre la composición de la microbiota y la variable respuesta de interés, sin deteriorar significativamente la capacidad de estimar esa relación.

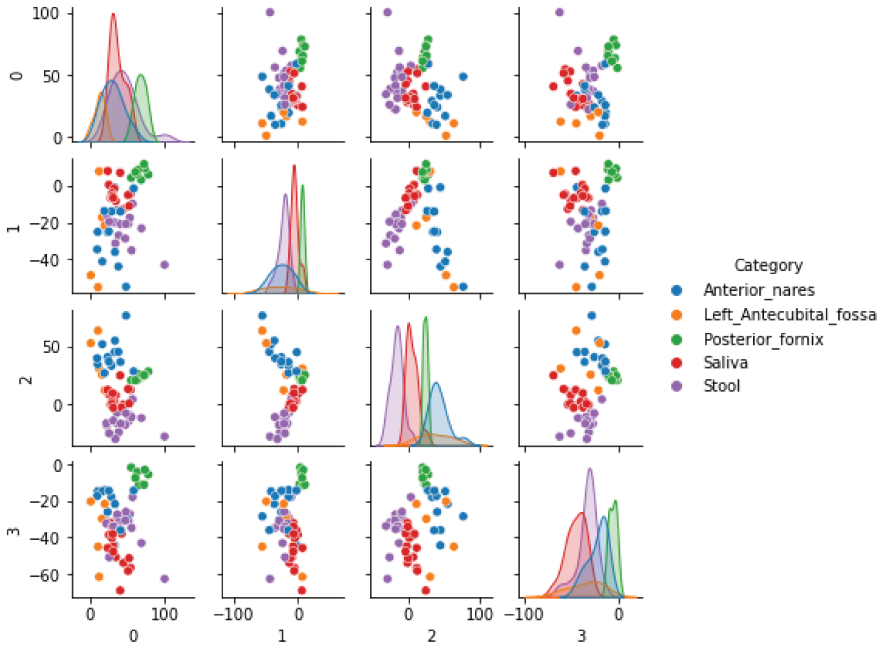


(a) L2

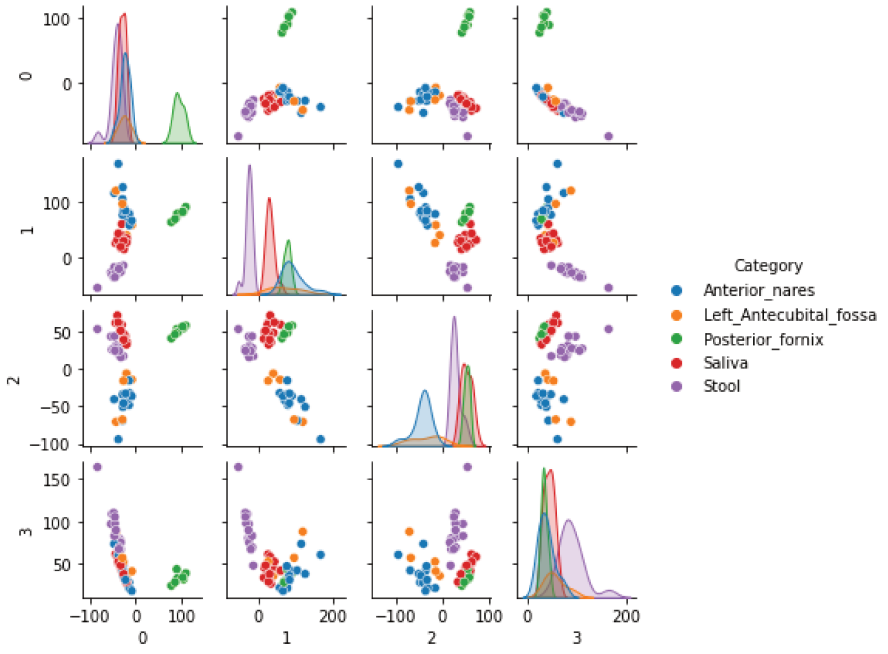


(b) L6

Figura 8.2: Reducción en predicción de los datos de microbioma HMP aprendida por el modelo Ising-pGM en los niveles taxonómicos L2 y L6 en relación con el lugar de donde se extrajeron las muestras.

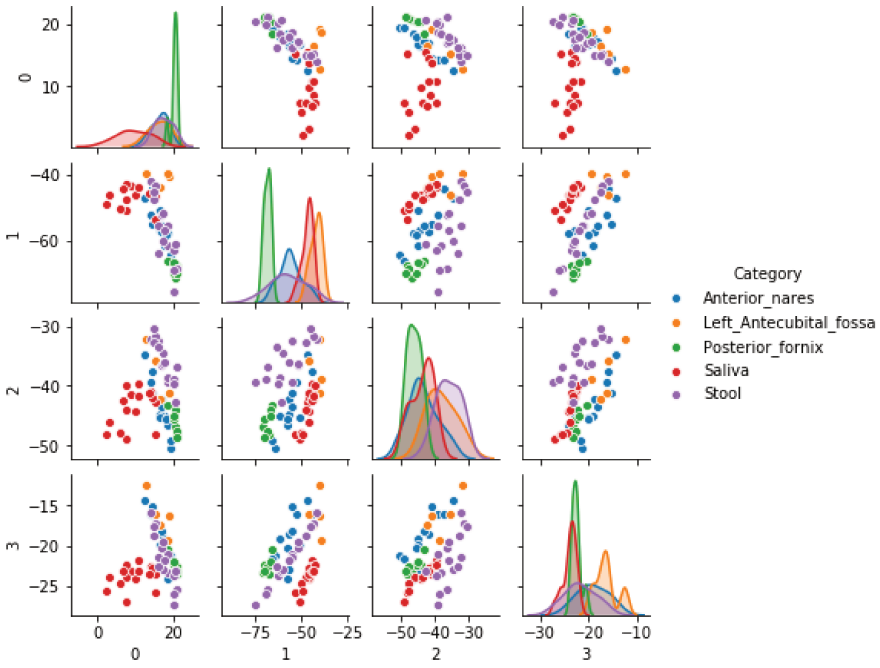


(a) L2

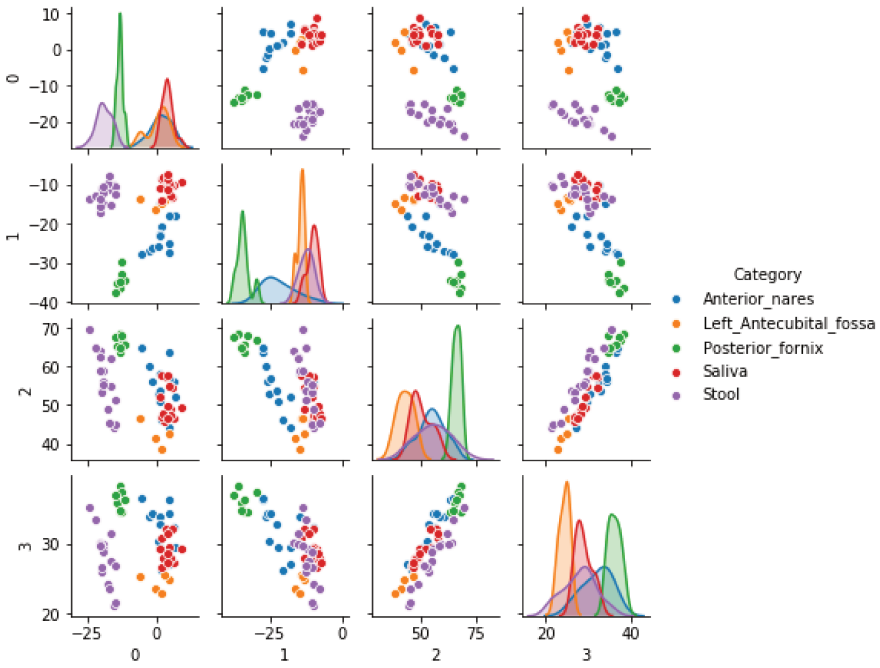


(b) L6

Figura 8.3: Reducción en predicción de los datos de microbioma HMP aprendida por el modelo sqPoisson-pGM en los niveles taxonómicos L2 y L6 en relación con el lugar de donde se extrajeron las muestras.



(a) Normal-pGM



(b) Normal-zipGM

Figura 8.4: Reducción en predicción de los datos de microbioma HMP aprendida por los modelos Normal-pGM y Normal-zipGM. En ambos casos consideramos el nivel taxonómicos L6.

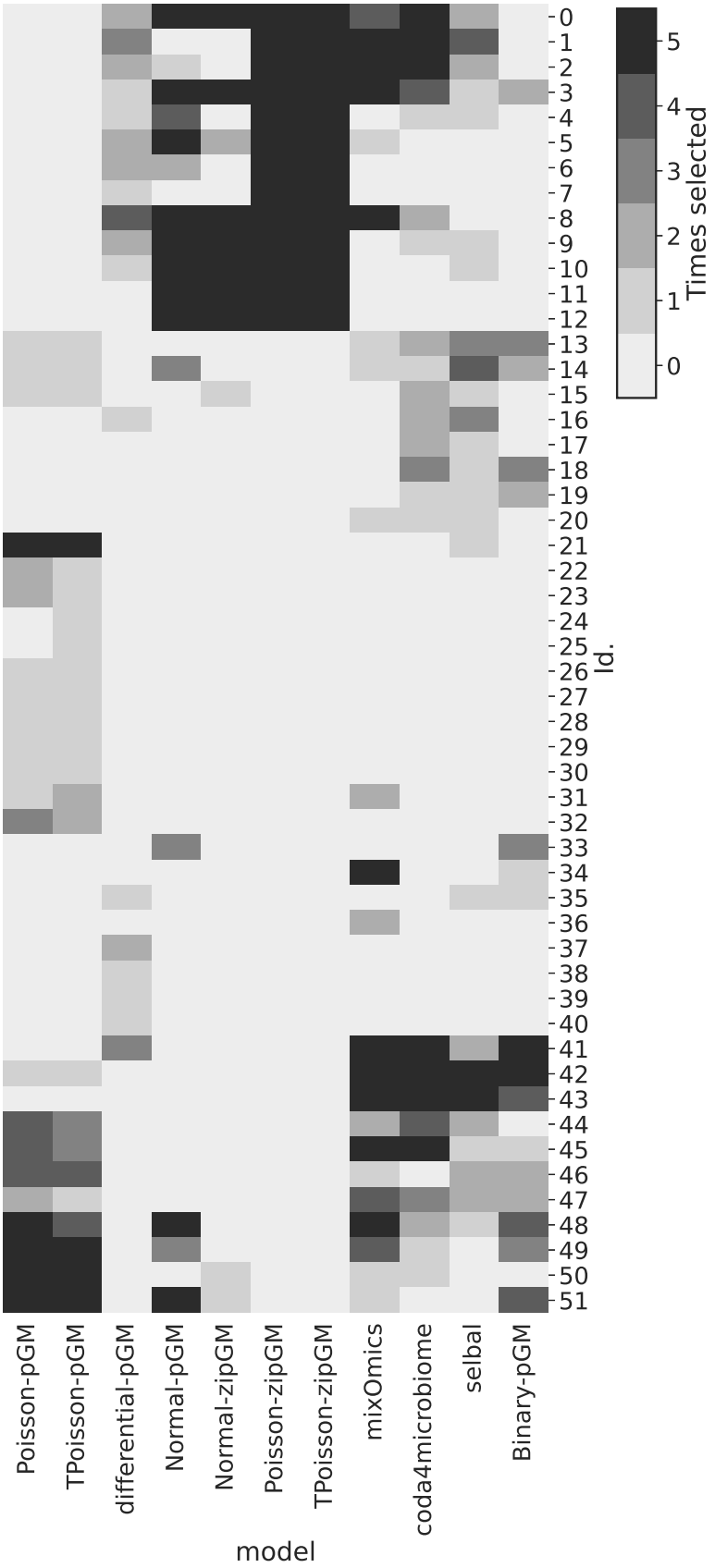


Figura 8.5: Análisis de asociación entre datos de microbioma de intestino y sexo. Se muestra el número de veces que cada modelo selecciona cada variable en base a 5 particiones independientes del conjunto de datos. Sólo se muestran las variables seleccionadas por algún modelo. El id. de dichas variables se encuentra en la Tabla 8.3.

ID.	PHILUM	CLASS	ORDER	FAMILY	GENUS
0	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinomyces
1	Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	Corynebacterium
2	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium
3	Actinobacteria	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	
4	Actinobacteria	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	Collinsella
5	Actinobacteria	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	Eggerthella
6	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
7	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Parabacteroides
8	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella
9	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	
10	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes
11	Bacteroidetes	Bacteroidia	Bacteroidales	[Odoribacteraceae]	Odoribacter
12	Cyanobacteria	4Cod-2	YS2		
13	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus
14	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus
15	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
16	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
17	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Lactococcus
18	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
19	Firmicutes	Bacilli	Turicibacterales	Turicibacteraceae	Turicibacter
20	Firmicutes	Clostridia	Clostridiales		
21	Firmicutes	Clostridia	Clostridiales	Christensenellaceae	
22	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	
23	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	SMB53
24	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	
25	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Blautia
26	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus
27	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Dorea
28	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnobacterium
29	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Roseburia
30	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	[Ruminococcus]
31	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	
32	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	
33	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Anaerotruncus
34	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
35	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillospira
36	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus
37	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
38	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Veillonella
39	Firmicutes	Clostridia	Clostridiales	[Tissierellaceae]	1-68
40	Firmicutes	Clostridia	Clostridiales	[Tissierellaceae]	Anaerococcus
41	Firmicutes	Clostridia	Clostridiales	[Tissierellaceae]	Finegoldia
42	Firmicutes	Clostridia	Clostridiales	[Tissierellaceae]	Peptoniphilus
43	Firmicutes	Clostridia	Clostridiales	[Tissierellaceae]	WAL_1855D
44	Firmicutes	Erysipelotrichi	Erysipelotrichales	Erysipelotrichaceae	Coprobacillus
45	Firmicutes	Erysipelotrichi	Erysipelotrichales	Erysipelotrichaceae	[Eubacterium]
46	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	Oxalobacter
47	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	Bilophila
48	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	
49	Tenericutes	Mollicutes	RF39		
50	Tenericutes	RF3	ML615J-28		
51	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Verrucomicrobiaceae	Akkermansia

Tabla 8.3: Especies seleccionadas por los modelos considerados junto con su identificador. En todos los casos las especies pertenecen a la familia Bacteria.

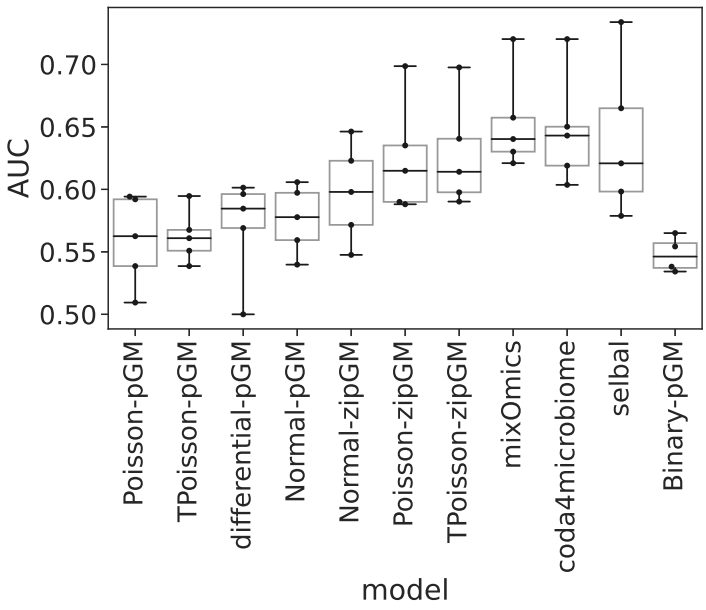


Figura 8.6: Análisis de asociación entre datos de microbioma de intestino y sexo. Se muestra el AUC en predicción en base a 5 particiones independientes del conjunto de datos.

MODELO	N° VARS	LR	QDA	SVM-POLY	RF
coda4microbiome	7	0.39	0.42	0.39	0.00
mixOmics	10	0.37	0.38	0.37	0.00
selbal	2	0.43	0.45	0.44	0.13
Ising-pGM	2	0.45	0.47	0.44	0.14
Normal-pGM	10	0.40	0.45	0.38	0.00
Normal-zipGM	7	0.40	0.41	0.40	0.00
Poisson-pGM	5	0.45	0.46	0.46	0.33
Poisson-zipGM	13	0.39	0.39	0.34	0.00
TPoisson-pGM	4	0.45	0.45	0.46	0.35
TPoisson-zipGM	13	0.39	0.39	0.34	0.00

Tabla 8.4: Error de predicción (accuracy) obtenido al emplear las variables estables estimadas con cada modelo y distintos modelos predictivos sobre 10 particiones independientes de test. Los modelos predictivos consideramos fueron: LR (regresión logística), qda (análisis discriminante cuadrático), svm-poly (clasificador de vector soporte con kernel polinómico), RF (random forest).

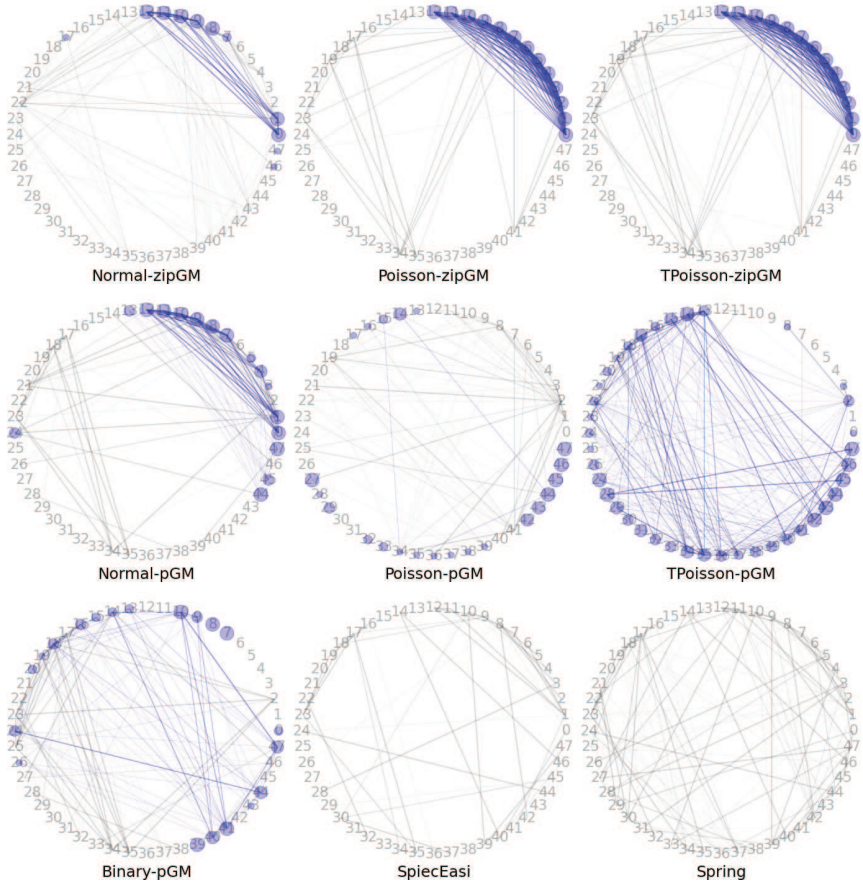


Figura 8.7: En gris se muestran los grafos de interacciones entre las varibales descriptivas obtenidos por los modelos zipGM propuestos, los pGM y los algoritmos *SpiecEasi* y *Spring*. En azul se muestran las variables e interacciones seleccionadas por la penalización jerárquica propuesta para los modelos zipGM y pGM al considerar el modelo condicional $\boldsymbol{X} \mid Y$. Los números en cada nodo corresponden a las especies detalladas en la Tabla 8.3, mientras que el tamaño de cada nodo indica el número de veces que la variable fue seleccionada dadas las 5 particiones de entrenamiento.

Parte V

CONCLUSIONES

En el siguiente capítulo resumiremos los resultados obtenidos, las dificultades encontradas y propondremos algunas alternativas a explorar en trabajos futuros.

9

CONCLUSIONES

En esta tesis caracterizamos nuevas familias paramétricas para el modelado de distribuciones de probabilidad multivariadas en la familia exponencial, que permiten modelar variables aleatorias que tomen el valor cero con probabilidad arbitraria. Estos modelos permiten una mejor descripción de datos composicionales de gran dimensión, que típicamente presentan una gran cantidad de ceros ya que muchos componentes se observan solamente en unas pocas muestras. Los modelos presentados corresponden a la subfamilia de modelos gráficos y están definidos a partir de un estadístico suficiente que se compone del vector de cuentas $\boldsymbol{X}$, que caracterizan la abundancia de cada componente, y de la indicadora $\boldsymbol{\nu}(\boldsymbol{X})$, que da cuenta de la presencia de cada uno en una comunidad dada.

Las interacciones definidas en los modelos gráficos propuestos (zipGM) caracterizan tres tipos de patrones:

1. coabundancia: la parte del estadístico suficiente $\boldsymbol{x}\boldsymbol{x}^\top$ cuenta las interacciones entre los niveles composicionales de las distintas variables. Este tipo de interacciones es modelada por ejemplo por los modelos gráficos pGM como el Poisson-pGM definido en el Ejemplo 2.3, aunque dichos modelos no se adecúan a datos con exceso de ceros.

2. coexistencia: define las interacciones entre las especies presentes y las ausentes en la muestra. El modelo gráfico Ising definido en el Ejemplo 2.2 modela particularmente este tipo de interacciones.

3. mixta: las interacciones entre los niveles de cada especie en la muestra y la presencia o ausencia de una determinada especie. Este tipo de interacciones es propia del modelo gráfico zipGM.

Los modelos gráficos zipGM incorporan así nuevos tipos de interacciones a las existentes en los modelos gráficos pGM, que resultan mejor adaptados a problemas con predictores con una gran cantidad de ceros. Suponiendo que los parámetros naturales de estos modelos dependen linealmente de una función fija de la respuesta, obtuvimos reducciones dimensionales suficientes para preservar la información predictiva siguiendo el enfoque general para la familia exponencial introducido en Bura, Duarte y Forzani (2016).

Además de modelos adecuados, la solución de problemas de escala real exige disponer de algoritmos adecuados para su estimación. Advertidos de la dificultad de adaptar enfoques de máxima verosimilitud, gran parte del trabajo de tesis lo dedicamos a explorar estimadores alternativos basados en funciones de scores. Estos enfoques mostraron ser útiles con modelos pGM, como así también con extensiones que modelan los patrones de abundancia condicionados a observar los componentes. Este tipo de modelos, que son un paso intermedio

entre los modelos pGM y los modelos zipGM, no han sido discutidos en este documento, principalmente porque no conducen a reducciones suficientes. No obstante, podrían encontrar aplicación en otros contextos.

La RSD, por sí sola, permite explorar asociaciones globales entre los datos composicionales y otra variable de interés. Buena parte del esfuerzo de esta tesis estuvo puesto en conseguir RSDs que logren también identificar las componentes predictoras que son las verdaderas responsables de ese efecto observado a nivel global. Por ser lineales en el estadístico suficiente de la familia exponencial elegida, esta tarea puede encararse promoviendo estimadores regularizados que induzcan la anulación de todos los coeficientes asociados a variables irrelevantes. Aunque las estrategias generales para conseguir este efecto son bien conocidas, la obtención de selección de variables en simultáneo con la estimación de la reducción exigió recurrir a penalizantes estructurados con un alto grado de complejidad y actuando sobre variables de distinto tipo, lo que implicó el planteo de algoritmos especialmente adaptados.

En particular, debido a la gran dispersión entre la parte del estadístico suficiente vinculada a la función indicadora $\boldsymbol{\nu}(\boldsymbol{x})$ y la parte vinculada a las cuentas $\boldsymbol{x}$, los parámetros naturales se encuentran en distintas escalas, provocando que la penalización dada por la norma ℓ_1 esté fuertemente descalibrada, redundando en mayor error en selección, aunque algunas simulaciones mostraron que dicha penalización logra buen nivel de generalización en predicción. Este comportamiento negativo en selección puede ser mitigado considerando una norma pesada. En nuestro trabajo consideramos los pesos dados por la matriz de información de Fisher evaluada en el modelo independiente. McDavid et al. (2019, Proposition 1) demostró que esta penalización aproxima un test estadístico para determinar si un parámetro es nulo.

Por otro lado, debido al gran número de parámetros involucrados en problemas de algunos cientos de variables, fue necesario asumir una estructura extra en la penalización a fin de simplificar el problema de optimización asociado. Se consideró una penalización jerárquica de manera tal que se modelen interacciones únicamente entre las variables asociadas a la respuesta. La penalización jerárquica cuenta con dos estructuras superpuestas:

- una penalización sobre todos los parámetros vinculados a una determinada variable.
- una penalización que induce variables independientes de a pares.

Esta penalización estructurada permite lograr modelos compactos facilitando el problema de optimización asociado.

Las pruebas experimentales conducidas nos permitieron evaluar el desempeño de las distintas opciones de modelado y estimación exploradas a lo largo de la tesis, y su comparación con metodologías de uso frecuente en el análisis de datos de microbioma. Comenzando por validar la mayor escalabilidad de los enfoques basados en scores

propios por sobre los enfoques de máxima verosimilitud previamente propuestos en (Tomassi et al., 2019), observamos de forma general que los modelos y estimadores propuestos presentan ventajas de interpretabilidad de las soluciones obtenidas, sin un deterioro significativo de la capacidad predictiva de esas reducciones. Su valor sobre la comprensión de las relaciones subyacentes entre predictores y respuesta resulta de una identificación más estable ante el remuestreo de los datos.

Un repaso más detallado de los resultados obtenidos nos muestra que cuando la proporción de ceros en los datos crece por encima de un 30 %, situación muy común en datos de microbioma reales, el modelo Ising ofrece mejor desempeño en la obtención de reducciones suficientes que los otros modelos pGM no adaptados al exceso de ceros. Por otra parte, cuando se modelan explícitamente los patrones de ceros, los modelos Poisson-zipGM brindan una selección de variables más estable que métodos de uso frecuente como los disponibles en (Rohart et al., 2017), con un desempeño predictivo similar. En este sentido, los resultados también sugieren que los modelos Normal-zipGM a menudo logran mejor capacidad predictiva que sus análogos tipo Poisson, probablemente gracias a una mayor flexibilidad en el modelado de correlaciones entre predictores.

Tanto los modelos zipGM como los pGM comparados fueron penalizados jerárquicamente empleando los pesos dados por la matriz de información de Fisher. Adoptando distintas estructuras en las matrices de interacciones, se observó que dicha penalización induce soluciones que agregan variables no asociadas linealmente con la respuesta cuando las mismas interactúan con otras variables que sí están asocidas. En este sentido, esta estructura aumenta la FPR mientras que mantiene acotada la FNR. Esta observación sugiere que podría mejorarse el desempeño de estos modelos para seleccionar las variables realmente asociadas con la respuesta introduciendo estrategias de remuestreo para evaluar la estabilidad de las interacciones identificadas.

PROPUESTAS PARA EXPLORAR

En esta tesis se optó por la minimización de una divergencia entre distribuciones multivariadas para el aprendizaje de los modelos. El estimador de máxima verosimilitud, aunque es el más eficiente desde el punto de vista estadístico, presenta dificultades para su cómputo y alta sensibilidad a *outliers*. La divergencia inducida por pseudo-verosimilitud presenta ventajas de cómputo y relaja la hipótesis de modelado al no incorporar la restricción de orden del modelo gráfico. En este sentido, es posible interpretar su minimización como la proyección a modelos gráficos de orden 2 (para los casos mostrados en los ejemplos) de una densidad sustituta de orden superior que es entrenada con los datos. Otra solución adoptada es considerar regresiones lineales generalizadas separadas. En este caso se pierden las restricciones impuestas por el espacio de parámetros como la simetría de las interacciones, pero guarda resultados similares al estimador de

máxima pseudo-verosimilitud. Si bien estas tres alternativas fueron comparadas con modelos mixtos en (Lee y Hastie, 2015), sólo fue analizado el caso del modelo conjunto Normal-Ising. Allí se reportó una importante mejora en eficiencia al considerar la pseudo-verosimilitud en lugar de las regresiones separadas.

Durante el trabajo de tesis se exploró la construcción de divergencias que no dependen de las constantes de normalización. Estas divergencias presentan mayor flexibilidad para incorporar conocimientos previos propios del campo de la aplicación. Gneiting y Raftery (2007) muestran algunos ejemplos de ello aplicados a economía. En este sentido, sería valioso contar con simulaciones extensivas que permitan definir la pérdida de eficiencia al considerar diversos estimadores, así también como su estabilidad frente a outliers.

Por otra parte, en la construcción del modelo Hurdle presentada en el Capítulo 2, se recorta el dominio de la distribución general adoptada excluyendo el cero para luego incorporar en su lugar una probabilidad puntual. Esto permite definir el modelo más simple para datos con exceso de ceros. Otras alternativas recientes consideran un *momentum* entre la proporción de ceros y un entorno de la distribución general (Haslett, Parnell y Sweeney, 2018). La definición de distribuciones multivariadas que consideren estos modelos y su potencial encuadre en la familia exponencial es un problema abierto.

Por último, nuevos trabajos en el aprendizaje de modelos gráficos incorporan la eliminación del sesgo causado por la penalización durante el proceso de aprendizaje (Kim, Liu y Kolar, 2021). Esto permite la definición de tests sobre los parámetros, pudiendo controlar la tasa de falsos positivos. Aunque los trabajos que incorporan esta metodología no logran escalar con la dimensión de los problemas vinculados a la aplicación, presentan un avance conceptual que sería interesante extender a los modelos gráficos propuestos.

Parte VI

ANEXOS

A

ANEXO AL CAPÍTULO 2

A.1 MODELOS GRÁFICOS

Demostración de la Proposición 2.2. Asumiendo que P es un modelo gráfico positivo y está definido sobre un dominio numerable, consideramos $C_{k,i}$ la i-ésima configuración (enumeración del espacio producto de $\boldsymbol{X}_k$) del clique k y definimos los parámetros naturales $\omega_{k,i} = \log \phi_k(C_{k,i})$ y los estadísticos suficientes que indican la configuración del clique, i. e.

$$f_{k,i}(\boldsymbol{x}_k) = \begin{cases} 1 & \text{si } \boldsymbol{x}_k = C_{k,i} \\ 0 & \text{en otro caso} \end{cases}.$$

Con esto,

$$P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\omega}) = \exp\{\sum_k \boldsymbol{\omega}_k^\top \boldsymbol{f}_k(x_k) - Z(\boldsymbol{\omega})\} = \exp\{\langle \boldsymbol{\omega}, \boldsymbol{f} \rangle - Z(\boldsymbol{\omega})\}.$$

□

A.2 NORMAL-PGM

Considerando el cambio de variables que nos lleva a los parámetros canónicos a los naturales en la distribución Normal multivariada: $\boldsymbol{\Theta} = -\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\eta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, luego

$$\begin{aligned} P(\boldsymbol{x}) &= \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\} \\ &= \frac{1}{(2\pi)^{p/2}|-\boldsymbol{\Theta}^{-1}|^{1/2}} \exp\{\frac{1}{2}(\boldsymbol{x}+\boldsymbol{\Theta}^{-1}\boldsymbol{\eta})^T\boldsymbol{\Theta}(\boldsymbol{x}+\boldsymbol{\Theta}^{-1}\boldsymbol{\eta})\} \\ &= \exp\{\boldsymbol{\eta}^T\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Theta}\boldsymbol{x} - Z(\boldsymbol{\eta},\boldsymbol{\Theta})\}, \\ Z(\boldsymbol{\eta},\boldsymbol{\Theta}) &= -\frac{1}{2}\boldsymbol{\eta}^T\boldsymbol{\Theta}^{-1}\boldsymbol{\eta} - \frac{1}{2}\log(|-\boldsymbol{\Theta}|) + \frac{p}{2}\log(2\pi). \end{aligned}$$

Las distribuciones condicionales están dadas por

$$\begin{aligned} P(x_j \mid x_{\setminus j}) &= \frac{P(x)}{P(x_{\setminus j})} = \frac{\exp\{\boldsymbol{\eta}^T x + \frac{1}{2}x^T\boldsymbol{\Theta}x\}}{\int_\infty^\infty \exp\{\boldsymbol{\eta}^T x + \frac{1}{2}x^T\boldsymbol{\Theta}x\}dx_j} \\ &= \exp\{\eta_{j|\setminus j}x_j + \frac{1}{2}\Theta_{jj}x_j^2 - Z_{j|\setminus j}\}, \\ \eta_{j|\setminus j} &= \eta_j + \sum_{i\neq j}\theta_{ji}x_i, \\ Z_{j|\setminus j} &= -\frac{\eta_{j|\setminus j}^2}{2\Theta_{jj}} - \frac{1}{2}\log(-\Theta_{jj}) + \frac{1}{2}\log(2\pi) \end{aligned}$$

A.3 FIXED-LENGTH POISSON-PGM

La siguiente proposición permite transformar la distribución con restricciones de igualdad en la suma de las variables dada en la Definición 2.8 en la distribución (Definición 2.5) con restricciones de desigualdad en la suma de todas las variables menos una. Observar que la nueva familia incorpora parametros no nulos en la diagonal de las interacciones.

Proposición A.1. *Si $\boldsymbol{X}$ distribuye de acuerdo con (2.11) con parámetros $m_{\boldsymbol{x}}, \boldsymbol{\eta}$ y $\boldsymbol{\Theta}$, el vector reducido $X_{\setminus k} \mid m_{\boldsymbol{x}}$ distribuye de acuerdo con (2.12), i. e. son FPoisson-pGM$_{\setminus k}(m_{\boldsymbol{x}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\Theta}})$, donde*

$$\begin{aligned} \tilde{\boldsymbol{\Theta}} &= \boldsymbol{\Theta}_{\setminus k,\setminus k} - \boldsymbol{\theta}_{k,\setminus k}\mathbf{1}_{k-1}^T \\ \tilde{\boldsymbol{\eta}} + \mathit{diag}(\tilde{\boldsymbol{\Theta}})m_{\boldsymbol{x}} &= \boldsymbol{\eta}_{\setminus k} - \eta_k \mathbf{1}_{k-1}, \end{aligned}$$

donde $\boldsymbol{\Theta}_{\setminus k,\setminus k}$ representa la submatriz principal obtenida después de remover el índice k de $\boldsymbol{\Theta}$ y $\boldsymbol{\theta}_{k,\setminus k}$ es la k-ésima columna de $\boldsymbol{\Theta}$ después de remover el k-ésimo elemento.

Demostración. El resultado se obtiene al reemplazar X_k en (2.11) por $m_{\boldsymbol{x}} - \mathbf{1}^\top \boldsymbol{x}_{\setminus k}$. □

El corolario que se enuncia a continuación considera un modelo de regresión lineal en la familia FPoisson-pGM (Definición 2.8) y encuentra el modelo equivalente en la familia reducida FPoisson-pGM$_{\setminus k}(m_{\boldsymbol{x}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\Theta}})$ dada en la Definición 2.5.

Corolario A.1 (Modelo de regresión inversa con restricción de suma no constante)**.** *Cuando $\boldsymbol{X} \mid Y$ distribuye de acuerdo con (2.11) con parámetros $m_{\boldsymbol{x}}, \boldsymbol{\eta}_y = \boldsymbol{\eta}_0 + \Gamma \boldsymbol{f}_y$ y $\boldsymbol{\Theta}$, el vector reducido $\boldsymbol{X}_{\setminus k} \mid Y, m_{\boldsymbol{x}}$ distribuye de acuerdo con FPoisson-pGM$_{\setminus k}(m_{\boldsymbol{x}}, \tilde{\boldsymbol{\eta}}_{y,m_{\boldsymbol{x}}}, \tilde{\boldsymbol{\Theta}})$, con*

$$\tilde{\boldsymbol{\eta}}_{y,m_{\boldsymbol{x}}} = \tilde{\boldsymbol{\eta}}_0 - \mathit{diag}(\tilde{\boldsymbol{\Theta}})m_{\boldsymbol{x}} + \tilde{\boldsymbol{\Gamma}}\boldsymbol{f}_y, \tag{A.1}$$

donde

$$\begin{aligned} \tilde{\boldsymbol{\eta}}_0 &= \boldsymbol{\eta}_{0,\setminus k} - \eta_{0,k}\mathbf{1}_{k-1} \in \mathbb{R}^{k-1} \\ \tilde{\boldsymbol{\Gamma}} &= \boldsymbol{\Gamma}_{\setminus k} - \boldsymbol{\gamma}_k^T \mathbf{1}_{k-1} \in \mathbb{R}^{k-1\times r} \\ \tilde{\boldsymbol{\Theta}} &= \boldsymbol{\Theta}_{\setminus k} - \boldsymbol{\theta}_{k,\setminus k}\mathbf{1}_{k-1}^T \in \mathbb{R}^{k-1\times k-1} \text{ simétrica}, \end{aligned}$$

donde $\boldsymbol{\Gamma}_{\setminus k}$ se obtiene removiendo la k-ésima fila $\boldsymbol{\gamma}_k$ de $\boldsymbol{\Gamma}$.

Demostración. El resultado se obtiene al considerar el modelo de regresión inversa $\boldsymbol{X} \mid Y$ definiendo $\boldsymbol{\eta} = \boldsymbol{\eta}_0 + \boldsymbol{\Gamma}\boldsymbol{f}(y)$ en (2.11) y aplicar la Proposición A.1. □

La familia FPoisson-pGM$_{\setminus k}(m_{\boldsymbol{x}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\Theta}})$ dada en la Definición 2.5 tiene condicionales en la misma familia:

Proposición A.2 (Condicionales)**.** *Cuando $X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_p$ distribuyen conjuntamente FPoisson-pGM$_{\setminus k}(m_{\boldsymbol{x}}, \tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\Theta}})$, las condicionales*

$X_j|\boldsymbol{X}_{\setminus j}$ distribuyen de acuerdo a una FPoisson-pGM$_{\setminus k}(m_{\boldsymbol{x}|\boldsymbol{x}_{\setminus j}}, \tilde{\eta}_{j|\setminus j}, \tilde{\Theta}_{jj})$, donde

$$m_{\boldsymbol{x}|\boldsymbol{x}_{\setminus j}} = m_{\boldsymbol{x}} - \sum_{l \neq j,k} x_l,$$
$$\tilde{\eta}_{j|\setminus j} = \tilde{\eta}_j + \sum_{l \neq j} \tilde{\Theta}_{jl} x_l.$$

A.4 ZERO-INFLATED PGM

Demostración de la Proposición 2.4. Usando que $Z(\tau) = \log\{1 + \exp(\tau)\}$ y $Z(\tau, \tau_0) = Z(\tau) + \tau_0$,

$$\begin{aligned} P(\nu(X) = 0) &= P(X = 0) = \exp\{\tau_0 - Z(\tau, \tau_0)\} \\ &= \exp\{-Z(\tau)\} = (1 + \exp\{\tau\})^{-1}, \\ P(\nu(X) = 1) &= 1 - P(\nu(X) = 0) = (1 + \exp\{-\tau\})^{-1}, \end{aligned}$$

donde usamos la definición de $Z^+(\boldsymbol{\theta})$. Además,

$$\begin{aligned} P(X = x \mid \nu(X) = \nu) &= \frac{P(X = x)}{P(\nu(X) = \nu(x))} \\ &= \frac{\exp\{\nu(x)[\tau + \tau_0 - Z^+(\boldsymbol{\theta})] + \boldsymbol{\theta}^T \boldsymbol{t}(x) + h(x) - Z(\tau, \tau_0)\}}{\exp\{\tau\nu(x) - Z(\tau)\}} \\ &= \exp\{\nu(x)[\tau_0 - Z^+(\boldsymbol{\theta})] + \boldsymbol{\theta}^T \boldsymbol{t}(x) + h(x) - \tau_0\}, \end{aligned}$$

entonces,

$$\begin{aligned} P(X = 0 \mid \nu(X) = 0) &= 1, \\ P(X = x \mid \nu(X) = 1) &= \exp\{\boldsymbol{\theta}^T \boldsymbol{t}(x) + h(x) - Z^+(\boldsymbol{\theta})\}. \end{aligned}$$

□

Demostración del Teorema 2.4. ($\Rightarrow$) Partiendo de (2.16), la probabilidad condicional de X_j dadas las demás variables $\boldsymbol{X}_{\setminus j}$ resulta

$$P(X_j = x_j \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{x}_{\setminus j}) = \frac{P(\boldsymbol{X} = (\boldsymbol{x}_{\setminus j}; x_j))}{\sum_{m \in \mathcal{X} \cup \{0\}} P(\boldsymbol{X} = (\boldsymbol{x}_{\setminus j}; m))}, \tag{A.2}$$

donde

$$\begin{aligned} P(\boldsymbol{X} = (\boldsymbol{x}_{\setminus j}; x_j)) = \exp\{&\boldsymbol{\xi}^T(\boldsymbol{\nu}_{\setminus j}; \nu_j) + \boldsymbol{\eta}^T(\boldsymbol{T}(\boldsymbol{x}_{\setminus j}); T(x_j)) + \\ &\frac{1}{2}(\boldsymbol{\nu}_{\setminus j}; \nu_j)^T \boldsymbol{\Lambda}(\boldsymbol{\nu}_{\setminus j}; \nu_j) + (\boldsymbol{\nu}_{\setminus j}; \nu_j)^T \boldsymbol{\Phi}(\boldsymbol{T}(\boldsymbol{x}_{\setminus j}); T(x_j)) + \\ &\frac{1}{2}(\boldsymbol{T}(\boldsymbol{x}_{\setminus j}); T(x_j))^T \boldsymbol{\Theta}(\boldsymbol{T}(\boldsymbol{x}_{\setminus j}); T(x_j)) + \mathbf{1}^T \boldsymbol{h}((\boldsymbol{x}_{\setminus j}; x_j))\} \end{aligned} \tag{A.3}$$

y

$$\begin{aligned} P(\boldsymbol{X} = (\boldsymbol{x}_{\setminus j}; m)) = \exp\{&\boldsymbol{\xi}^T(\boldsymbol{\nu}_{\setminus j}; \nu(m)) + \boldsymbol{\eta}^T(\boldsymbol{T}(\boldsymbol{x}_{\setminus j}); T(m)) + \\ &\frac{1}{2}(\boldsymbol{\nu}_{\setminus j}; \nu(m))^T \boldsymbol{\Lambda}(\boldsymbol{\nu}_{\setminus j}; \nu(m)) + (\boldsymbol{\nu}_{\setminus j}; \nu(m))^T \boldsymbol{\Phi}(\boldsymbol{T}(\boldsymbol{x}_{\setminus j}); T(m)) + \\ &\frac{1}{2}(\boldsymbol{T}(\boldsymbol{x}_{\setminus j}); T(m))^T \boldsymbol{\Theta}(\boldsymbol{T}(\boldsymbol{x}_{\setminus j}); T(m)) + \mathbf{1}^T \boldsymbol{h}((\boldsymbol{x}_{\setminus j}; m))\}, \end{aligned} \tag{A.4}$$

donde $\mathcal{X} \subsetneq \mathbb{R}$ es el soporte de la variable aleatoria $X_j \mid \nu(X_j) = 1$. En lo que sigue definimos $\boldsymbol{\nu} := \boldsymbol{\nu}(\boldsymbol{x})$. Considerando la estructura en bloques

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{\setminus j,\setminus j} & \boldsymbol{\Lambda}_{\setminus j,j} \\ \boldsymbol{\Lambda}_{j,\setminus j} & 0 \end{pmatrix} \quad \boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\Phi}_{\setminus j,\setminus j} & \boldsymbol{\Phi}_{\setminus j,j} \\ \boldsymbol{\Phi}_{j,\setminus j} & 0 \end{pmatrix} \quad \boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\Theta}_{\setminus j,\setminus j} & \boldsymbol{\Theta}_{\setminus j,j} \\ \boldsymbol{\Theta}_{j,\setminus j} & \Theta_{jj} \end{pmatrix},$$

por simetría, $\boldsymbol{\Lambda}_{\setminus j,j} = \boldsymbol{\Lambda}_{j,\setminus j}^T$ y $\boldsymbol{\Theta}_{\setminus j,j} = \boldsymbol{\Theta}_{j,\setminus j}^T$. Luego de algunas simplificaciones,

$$P(X_j = x_j \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{x}_{\setminus j}) = \frac{\exp\{\xi_{j|\setminus j}\nu_j + \eta_{j|\setminus j}T(x_j) + \frac{\Theta_{jj}}{2}T(x_j)^2 + B(x_j)\}}{\sum_{m\in\mathcal{X}\cup\{0\}} \exp\{\xi_{j|\setminus j}\nu(m) + \eta_{j|\setminus j}T(m) + \frac{\Theta_{jj}}{2}T(m)^2 + B(m)\}}, \tag{A.5}$$

donde $\xi_{j|\setminus j} = \xi_j + \boldsymbol{\Lambda}_{j,\setminus j}\boldsymbol{\nu}_{\setminus j} + \boldsymbol{\Phi}_{j,\setminus j}\boldsymbol{T}(\boldsymbol{x}_{\setminus j})$ and $\eta_{j|\setminus j} = \eta_j + \boldsymbol{\nu}_{\setminus j}^T\boldsymbol{\Phi}_{\setminus j,j} + \boldsymbol{\Theta}_{j,\setminus j}\boldsymbol{T}(\boldsymbol{x}_{\setminus j})$ que resultan (2.20) y (2.21) respectivamente.

($\Leftarrow$) Sea

$$Q(\boldsymbol{X} \mid \boldsymbol{\omega}) := \log \frac{P(\boldsymbol{X} \mid \boldsymbol{\omega})}{P(\boldsymbol{0} \mid \boldsymbol{\omega})}, \tag{A.6}$$

para cualquier $\boldsymbol{X}$ en el espacio muestral $(\mathcal{X} \cup \{0\})^k$, donde $P(\boldsymbol{0} \mid \boldsymbol{\omega})$ denota la probabilidad de que todas las variables tomen el valor 0, con $P(\boldsymbol{0} \mid \boldsymbol{\omega}) > 0$. Por hipótesis, $P(\boldsymbol{X} \mid \boldsymbol{\omega})$ es el producto de factores de a lo sumo dos componentes, y como Q es homogéneo, podemos escribirlo como

$$Q(\boldsymbol{X} \mid \boldsymbol{\omega}) = \sum_{j=1}^{k} X_j g_j(X_j) + \frac{1}{2}\sum_j \sum_{k\neq j} X_j X_k g_{jk}(X_j, X_k) \tag{A.7}$$

donde $g_j(\cdot)$, $g_{jk}(\cdot) = g_{kj}(\cdot)$ están definidas salvo una constante, por lo que $\log G_{jk}(X_j, X_k) = X_j g_j(X_j) + X_k g_k(X_k) + X_j X_k[g_{jk}(X_j, X_k) + g_{kj}(X_k, X_j)]/2$ también lo está para $j > k$. De Besag (1974),

$$Q(\boldsymbol{X} \mid \boldsymbol{\omega}) - Q(\boldsymbol{X}^{\setminus j} \mid \boldsymbol{\omega}) = \log \frac{P(X_j \mid \boldsymbol{X}_{\setminus j}, \boldsymbol{\omega})}{P(0 \mid \boldsymbol{X}_{\setminus j}, \boldsymbol{\omega})}, \tag{A.8}$$

donde $\boldsymbol{X}^{\setminus j} = (X_1, \cdots, X_{j-1}, 0, X_{j+1}, \cdots, X_k)$. Usando (A.7) en (A.8) resulta

$$Q(\boldsymbol{X} \mid \boldsymbol{\omega}) - Q(\boldsymbol{X}^{\setminus j} \mid \boldsymbol{\omega}) = X_j g_j(X_j) + \frac{1}{2}\sum_{k\neq j} X_j X_k g_{jk}(X_j, X_k). \tag{A.9}$$

Por (2.18), el lado derecho de (A.8) se simplifica en

$$\log \frac{P(X_j \mid \boldsymbol{X}_{\setminus j}, \boldsymbol{\omega})}{P(0 \mid \boldsymbol{X}_{\setminus j}, \boldsymbol{\omega})} = \frac{1}{2}\Theta_{jj}[T^2(X_j) - T^2(0)] + \eta_{j|\setminus j}(\boldsymbol{X}_{\setminus j})[T(X_j) - T(0)] + \xi_{j|\setminus j}(\boldsymbol{X}_{\setminus j})\nu_j + [B(X_j) - B(0)], \tag{A.10}$$

donde $\nu_j = \nu(X_j)$, y hacemos evidente la dependencia de $\eta_{j|\setminus j}$ y $\xi_{j|\setminus j}$ en las variables $\boldsymbol{X}_{\setminus j}$. Tomando $X_k = 0$ para todo $k \neq j$, (A.9) y (A.10) con (A.8) implican

$$X_j g_j(X_j) = \frac{1}{2}\Theta_{jj}[T^2(X_j) - T^2(0)] + \eta_{j|\setminus j}(\mathbf{0})[T(X_j) - T(0)] + \xi_{j|\setminus j}(\mathbf{0})\nu_j + [B(X_j) - B(0)], \quad \text{(A.11)}$$

y evaluando $X_k = 0$ para todo $k \notin \{j, l\}$,

$$X_j g_j(X_j) + X_j X_l g_{jl}(X_j, X_l) = \frac{1}{2}\Theta_{jj}[T^2(X_j) - T^2(0)] + \eta_{j|\setminus j}(\cdots, 0, X_l, 0, \cdots)[T(X_j) - T(0)] + \xi_{j|\setminus j}(\cdots, 0, X_l, 0, \cdots)\nu_j + [B(X_j) - B(0)]. \quad \text{(A.12)}$$

Combinando (A.11) y (A.12), tenemos

$$X_j g_j(X_j) = \frac{1}{2}\Theta_{jj}[T^2(X_j) - T^2(0)] + \eta_{j|\setminus j}(\mathbf{0})[T(X_j) - T(0)] + \xi_{j|\setminus j}(\mathbf{0})\nu_j + [B(X_j) - B(0)], \quad \text{(A.13)}$$

$$X_j X_l g_{jl}(X_j, X_l) = [\eta_{j|\setminus j}(\cdots, 0, X_l, 0, \cdots) - \eta_{j|\setminus j}(\mathbf{0})][T(X_j) - T(0)] + [\xi_{j|\setminus j}(\cdots, 0, X_l, 0, \cdots) - \xi_{j|\setminus j}(\mathbf{0})]\nu_j.$$

Usando (2.20) y (2.21),

$$\eta_{j|\setminus j}(\mathbf{0}) = \eta_j + \sum_{l \neq j}\Theta_{jl}T(0), \quad \text{(A.14)}$$

$$\xi_{j|\setminus j}(\mathbf{0}) = \xi_j + \sum_{l \neq j}\Phi_{jl}T(0),$$

$$\eta_{j|\setminus j}(\cdots, 0, X_l, 0, \cdots) - \eta_{j|\setminus j}(\mathbf{0}) = \Theta_{jl}[T(X_l) - T(0)] + \Phi_{lj}\nu_l,$$

$$\xi_{j|\setminus j}(\cdots, 0, X_l, 0, \cdots) - \xi_{j|\setminus j}(\mathbf{0}) = \Phi_{jl}[T(X_l) - T(0)] + \Lambda_{jl}\nu_l,$$

donde η_j, ξ_j, Θ_{jl}, Φ_{lj}, Φ_{jl} y Λ_{jl} son constantes. Reemplazando (A.14) en (A.13), y reacomodando términos, la distribución conjunta (A.7) se simplifica en

$$
\begin{aligned}
&Q(\boldsymbol{X} \mid \boldsymbol{\omega}) = \\
&= \sum_{j=1}^{k} \Bigg(\frac{1}{2}\Theta_{jj}[T^2(X_j) - T^2(0)] + [\eta_j + \sum_{l \neq j} \Theta_{jl} T(0)][T(X_j) - T(0)] + \\
&\quad [\xi_j + \sum_{l \neq j} \Phi_{jl} T(0)]\nu_j + [B(X_j) - B(0)] \Bigg) + \frac{1}{2} \sum_{j=1}^{k} \sum_{l \neq j} \Big(\\
&\quad \Theta_{jl}[T(X_j) - T(0)][T(X_l) - T(0)] + \Phi_{lj}[T(X_j) - T(0)]\nu_l + \\
&\quad \Phi_{jl}\nu_j[T(X_l) - T(0)] + \Lambda_{jl}\nu_j\nu_l \Big) \\
&= \sum_{j=1}^{k} \Bigg(\eta_j[T(X_j) - T(0)] + \xi_j\nu_j + \frac{1}{2}\Theta_{jj}[T^2(X_j) - T^2(0)] + \\
&\quad [B(X_j) - B(0)] \Bigg) - \sum_{j=1}^{k} \sum_{l \neq j} \Theta_{jl} T(0)^2 + \sum_{j=1}^{k} \sum_{l \neq j} \Theta_{jl} T(X_j) T(0) + \\
&\quad \sum_{j=1}^{k} \sum_{l \neq j} \Phi_{jl}\nu_j T(0) + \frac{1}{2} \sum_{j=1}^{k} \sum_{l \neq j} \left(\Theta_{jl} T(X_j) T(X_l) + \Lambda_{jl}\nu_j\nu_l \right) + \\
&\quad \frac{1}{2} \sum_{j=1}^{k} \sum_{l \neq j} \Theta_{jl} T(0)^2 - \frac{1}{2} \sum_{j=1}^{k} \sum_{l \neq j} \left(\Theta_{jl} T(X_j) T(0) + \Theta_{jl} T(0) T(X_l) \right) + \\
&\quad \frac{1}{2} \sum_{j=1}^{k} \sum_{l \neq j} \left(\Phi_{lj}\nu_l T(X_j) + \Phi_{jl}\nu_j T(X_l) \right) - \\
&\quad \frac{1}{2} \sum_{j=1}^{k} \sum_{l \neq j} \left(\Phi_{lj}\nu_l T(0) + \Phi_{jl}\nu_j T(0) \right), \\
&= \boldsymbol{\eta}^T[\boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{T}(\boldsymbol{0})] + \boldsymbol{\xi}^T\boldsymbol{\nu}(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{T}(\boldsymbol{x})^T\boldsymbol{\Theta}\boldsymbol{T}(\boldsymbol{x}) - \frac{1}{2}\boldsymbol{T}(\boldsymbol{0})^T\boldsymbol{\Theta}\boldsymbol{T}(\boldsymbol{0}) + \\
&\quad \frac{1}{2}\boldsymbol{\nu}(\boldsymbol{x})^T\boldsymbol{\Lambda}\boldsymbol{\nu}(\boldsymbol{x}) + \boldsymbol{\nu}(\boldsymbol{x})^T\boldsymbol{\Phi}\boldsymbol{T}(\boldsymbol{x}) + \mathbf{1}^T[\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{h}(\boldsymbol{0})].
\end{aligned}
$$

donde usamos en la última igualdad que $\Theta_{jl} = \Theta_{lj}$ y $\sum_{j=1}^{k} \sum_{l \neq j} \Phi_{lj}\nu_l T(0) = \sum_{j=1}^{k} \sum_{l \neq j} \Phi_{jl}\nu_j T(0)$. Definiendo los vectores o matrices

$$
[\boldsymbol{\eta}]_j = \eta_j, \quad [\boldsymbol{\xi}]_j = \xi_j, \qquad [\boldsymbol{\Theta}]_{jl} = \begin{cases} \Theta_{jj} & \text{if } j = l \\ \Theta_{jl} & \text{otherwise} \end{cases},
$$

$$
[\boldsymbol{\Phi}]_{jl} = \begin{cases} 0 & \text{if } j = l \\ \Phi_{jl} & \text{otherwise} \end{cases}, \qquad [\boldsymbol{\Lambda}]_{jl} = \begin{cases} 0 & \text{if } j = l \\ \Lambda_{jl} & \text{otherwise} \end{cases}.
$$

A partir de (A.6), y como $\boldsymbol{\nu}(\mathbf{0}) = \mathbf{0}$,

$$
\begin{aligned}
\log P(\boldsymbol{X} \mid \boldsymbol{\omega}) = \boldsymbol{\eta}^T\boldsymbol{T}(\boldsymbol{x}) + \boldsymbol{\xi}^T\boldsymbol{\nu}(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{T}(\boldsymbol{x})^T\boldsymbol{\Theta}\boldsymbol{T}(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{\nu}(\boldsymbol{x})^T\boldsymbol{\Lambda}\boldsymbol{\nu}(\boldsymbol{x}) + \\
\boldsymbol{\nu}(\boldsymbol{x})^T\boldsymbol{\Phi}\boldsymbol{T}(\boldsymbol{x}) + \mathbf{1}^T B(\boldsymbol{x}) - A(\boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}, \boldsymbol{\Phi}),
\end{aligned}
$$

resultando en (2.16), cuyo logaritmo de su función de partición está dado por

$$A(\boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}, \boldsymbol{\Phi}) = \log \left(\sum_{\boldsymbol{x} \in (\mathcal{X} \cup \{0\})^k} \exp\{\boldsymbol{\eta}^T \boldsymbol{T}(\boldsymbol{x}) + \boldsymbol{\xi}^T \boldsymbol{\nu}(\boldsymbol{x}) + \right.$$
$$\left. \frac{1}{2} \boldsymbol{T}(\boldsymbol{x})^T \boldsymbol{\Theta} \boldsymbol{T}(\boldsymbol{x}) + \frac{1}{2} \boldsymbol{\nu}(\boldsymbol{x})^T \boldsymbol{\Lambda} \boldsymbol{\nu}(\boldsymbol{x}) + \boldsymbol{\nu}(\boldsymbol{x})^T \boldsymbol{\Phi} \boldsymbol{T}(\boldsymbol{x}) + \mathbf{1}^T B(\boldsymbol{x})\} \right).$$

□

B

ANEXO AL CAPÍTULO 3

Demostración de la Proposición 3.1. De la definición 2.12, podemos factorizar la distribución condicional como :

$$\begin{aligned} P(\boldsymbol{X}=\boldsymbol{x} \mid Y=y) &= \exp\left\{(\boldsymbol{\eta}_0+\boldsymbol{\Gamma}\boldsymbol{f}_y)^\top \boldsymbol{T}(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{T}(\boldsymbol{x})^\top \boldsymbol{\Theta}\boldsymbol{T}(\boldsymbol{x}) + \right. \\ &\qquad \left. \sum_{j=1}^{k} h(x_j) - Z(\boldsymbol{\eta},\boldsymbol{\Theta})\right\} \\ &= g(\boldsymbol{R}(\boldsymbol{T}(\boldsymbol{x})),Y)c(\boldsymbol{T}(\boldsymbol{x}))l(y), \end{aligned}$$

donde

$$\begin{aligned} l(y) &= 1 \\ c(\boldsymbol{T}(\boldsymbol{x})) &= \exp\left(\boldsymbol{\eta}_0^\top \boldsymbol{T}(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{T}(\boldsymbol{x})^\top \boldsymbol{\Theta}\boldsymbol{T}(\boldsymbol{x}) + \sum_{j=1}^{k} h(x_j) \right. \\ &\qquad \left. - Z(\boldsymbol{\eta},\boldsymbol{\Theta})\right) \\ g(\boldsymbol{R}(\boldsymbol{T}(\boldsymbol{x})),Y) &= \exp\left(\boldsymbol{f}_y^\top \boldsymbol{\Gamma}^\top \boldsymbol{T}(\boldsymbol{x})\right) = \exp\left((\boldsymbol{\beta}\boldsymbol{f}_y)^\top(\boldsymbol{\alpha}^\top \boldsymbol{T}(\boldsymbol{x}))\right), \end{aligned}$$

donde usamos $\boldsymbol{\Gamma} = \boldsymbol{\alpha}\boldsymbol{\beta}$. Aplicando el Teorema 3.2 obtenemos el resultado. □

Demostración del Corolario 3.1. El corolario es consecuencia de la inclusión de los modelos gráficos en la familia exponencial enunciada en la Proposición 2.2. □

Demostración del Corolario 3.2. La distribución condicional de $\boldsymbol{X} \mid Y$ dado en (2.16), junto con (3.10a) y (3.10b) resulta

$$\begin{aligned} P(\boldsymbol{X}=\boldsymbol{x} \mid Y=y) = \exp\{&(\boldsymbol{\eta}_0+\boldsymbol{\Gamma}\boldsymbol{f}(y))^\top \boldsymbol{T}(\boldsymbol{x}) + (\boldsymbol{\xi}+\boldsymbol{\Psi}\boldsymbol{f}(y))^\top \boldsymbol{\nu}(\boldsymbol{x}) \\ &+ \frac{1}{2}\boldsymbol{T}(\boldsymbol{x})^\top \boldsymbol{\Theta}\boldsymbol{T}(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{\nu}(\boldsymbol{x})^\top \boldsymbol{\Lambda}\boldsymbol{\nu}(\boldsymbol{x}) \\ &+ \boldsymbol{\nu}(\boldsymbol{x})^\top \boldsymbol{\Phi}\boldsymbol{T}(\boldsymbol{x}) + \mathbf{1}^\top \boldsymbol{h}(\boldsymbol{x}) \\ &- Z(\boldsymbol{\xi}_y,\boldsymbol{\eta}_y,\boldsymbol{\Theta},\boldsymbol{\Phi},\boldsymbol{\Lambda})\} \end{aligned} \tag{B.1}$$

donde $Z(\boldsymbol{\xi}_y,\boldsymbol{\eta}_y,\boldsymbol{\Theta},\boldsymbol{\Phi},\boldsymbol{\Lambda})$ es el logaritmo de la función de partición que normaliza la distribución condicional (B.1) y depende de todos los parámetros en el modelo. Esta distribución condicional admite la siguiente factorización:

$$P(\boldsymbol{X}=\boldsymbol{x} \mid Y=y) = g(\boldsymbol{R}(\boldsymbol{x}),Y)c(\boldsymbol{x})l(y), \tag{B.2}$$

donde

$$
\begin{aligned}
l(y) &= \exp\{-Z(\boldsymbol{\xi}_y, \boldsymbol{\eta}_y, \boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Lambda})\}, \\
c(\boldsymbol{x}) &= \exp\Bigg\{\boldsymbol{\eta}_0^\top \boldsymbol{T}(\boldsymbol{x}) + \boldsymbol{\xi}_0^\top \boldsymbol{\nu}(\boldsymbol{x}) \\
&\quad + \frac{1}{2}\boldsymbol{T}(\boldsymbol{x})^\top \boldsymbol{\Theta}\boldsymbol{T}(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{\nu}(\boldsymbol{x})^\top \boldsymbol{\Lambda}\boldsymbol{\nu}(\boldsymbol{x}) \\
&\quad + \boldsymbol{\nu}(\boldsymbol{x})^\top \boldsymbol{\Phi}\boldsymbol{T}(\boldsymbol{x}) + \mathbf{1}^\top \boldsymbol{h}(\boldsymbol{x})\Bigg\}, \\
g(\boldsymbol{R}(\boldsymbol{x}), y) &= \exp\{(\boldsymbol{\Psi}\boldsymbol{f}(y))^\top \boldsymbol{\nu}(\boldsymbol{x}) + (\boldsymbol{\Gamma}\boldsymbol{f}(y))^\top \boldsymbol{T}(\boldsymbol{x})\} \\
&= \exp\left(\boldsymbol{f}_y^\top \begin{pmatrix}\boldsymbol{\Gamma}\\ \boldsymbol{\Psi}\end{pmatrix}^\top \begin{pmatrix}\boldsymbol{T}(\boldsymbol{x})\\ \boldsymbol{\nu}(\boldsymbol{x})\end{pmatrix}\right) \\
&= \exp\left((\boldsymbol{\beta}\boldsymbol{f}_y)^\top \boldsymbol{\kappa}^\top \begin{pmatrix}\boldsymbol{T}(\boldsymbol{x})\\ \boldsymbol{\nu}(\boldsymbol{x})\end{pmatrix}\right) \\
&= \exp\left((\boldsymbol{\beta}\boldsymbol{f}_y)^\top \boldsymbol{R}(\boldsymbol{X})\right),
\end{aligned}
$$

donde usamos $\begin{pmatrix}\boldsymbol{\Gamma}\\ \boldsymbol{\Psi}\end{pmatrix} = \boldsymbol{\kappa}\boldsymbol{\beta}$, con $\boldsymbol{\kappa} \in \mathbb{R}^{2p\times d}$ y $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$, es la representación en rango reducido de la matriz. Se concluye como consecuencia del Teorema 3.2. □

Demostración del Corolario 3.3. Partiendo de la demostración del Corolario 3.2, considerando

$$
\begin{aligned}
g(R(\boldsymbol{x}), y) &= \exp\left\{\boldsymbol{f}_y^\top \boldsymbol{\Gamma}^\top \boldsymbol{T}(\boldsymbol{x}) + \boldsymbol{f}_y^\top \boldsymbol{\Psi}^\top \boldsymbol{\nu}(\boldsymbol{x})\right\} \\
&= \exp\left(\boldsymbol{f}_y^\top \begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\tau}\end{pmatrix}^\top \begin{pmatrix}\boldsymbol{\alpha}^\top \boldsymbol{T}(\boldsymbol{x})\\ \boldsymbol{\zeta}^\top \boldsymbol{\nu}(\boldsymbol{x})\end{pmatrix}\right) \\
&= \exp\left(\boldsymbol{f}_y^\top \begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\tau}\end{pmatrix}^\top R_s(\boldsymbol{X})\right),
\end{aligned}
$$

donde $\boldsymbol{\Gamma} = \boldsymbol{\alpha}\boldsymbol{\beta}$ y $\boldsymbol{\Psi} = \boldsymbol{\zeta}\boldsymbol{\tau}$ son representaciones de rango reducido, con rango $d, d_0 \leq \text{mín}\{k, r\}$ respectivamente. Se concluye como consecuencia del Teorema 3.2. □

C

ANEXO AL CAPÍTULO 4

C.1 EQUIVALENCIA DIVERGENCIA-SCORE

La siguiente proposición establece la equivalencia (4.3) para la divergencia PKL definida en (4.15), por la cual la divergencia entre la distribución epírica de los datos y el modelo es proporcional a la esperanza empírica del score inducido.

Proposición C.1. *Sea $\hat{Q}_{\boldsymbol{X},Y}$ la distribución empírica de $Q_{\boldsymbol{X},Y}$, definida a partir del conjunto de datos $(\boldsymbol{X}^s, Y^s), s = 1, \dots, n$. Considerando que ambas distribuciones tienen como marginal a $\hat{Q}_Y$, la divergencia entre modelos condicionales (4.33) definida a partir de la divergencia $PKL(\hat{Q}_{\boldsymbol{X}|Y} \| P_{\boldsymbol{X}|Y})$ definida en (4.15) es proporcional a la esperanza empírica del score $\sum_{j=1}^{k} \log P(X_j = x_j \mid \boldsymbol{X}_{\setminus j} = \boldsymbol{x}_{\setminus j}, Y = y, \boldsymbol{\omega})$.*

Demostración. Consideremos las distribuciones empíricas:

$$\hat{Q}(\boldsymbol{X}, Y) = \frac{1}{n} \sum_{s=1}^{n} \delta(\boldsymbol{X} = \boldsymbol{X}^s)\delta(Y = Y^s) \tag{C.1}$$

$$\hat{Q}(Y) = \frac{1}{n} \sum_{s=1}^{n} \delta(Y = Y^s) \tag{C.2}$$

$$\hat{Q}(\boldsymbol{X} \mid Y) = \begin{cases} \frac{\sum_{s=1}^{n} \delta(\boldsymbol{X}=\boldsymbol{X}^s)\delta(Y=Y^s)}{\sum_{s=1}^{n} \delta(Y=Y^s)} & \text{if } Y \in \{Y^s\}_{s=1}^{n} \\ C_1 & \text{otherwise} \end{cases} \tag{C.3}$$

$$\hat{Q}(X_j \mid \boldsymbol{X}_{\setminus j}, Y) = \begin{cases} \frac{\sum_{s=1}^{n} \delta(\boldsymbol{X}=\boldsymbol{X}^s)\delta(Y=Y^s)}{\sum_{s=1}^{n} \delta(\boldsymbol{X}_{\setminus j}=\boldsymbol{X}^s_{\setminus j})\delta(Y=Y^s)} & \text{if } (\boldsymbol{X}_{\setminus j}, Y) \in \{(\boldsymbol{X}^s_{\setminus j}, Y^s)\}_{s=1}^{n} \\ C_2 & \text{otherwise} \end{cases} \tag{C.4}$$

con $\boldsymbol{X} \in \mathcal{X}^k$ y $Y \in \mathcal{Y}$ y donde C_1, C_2 son constantes. Luego,

$$\mathrm{PKL}(\hat{Q}_{\boldsymbol{X},Y} \| P_{\boldsymbol{X},Y}(\boldsymbol{\omega})) = \sum_{Y \in \mathcal{Y}} \hat{Q}(Y) \sum_{j=1}^{k} \sum_{\boldsymbol{X} \in \mathcal{X}^k} \hat{Q}(\boldsymbol{X} \mid Y) \Big[\log \hat{Q}(X_j \mid \boldsymbol{X}_{\setminus j}, Y) - \log P(X_j \mid \boldsymbol{X}_{\setminus j}, Y, \boldsymbol{\omega}) \Big].$$

Usando (C.2), (C.3) y (C.4), $\mathrm{PKL}(\hat{Q}_{\boldsymbol{X},Y} \| P_{\boldsymbol{X},Y}(\boldsymbol{\omega}))$ resulta

$$\sum_{Y \in \mathcal{Y}} \sum_{s=1}^{n} \frac{1}{n} \delta(Y = Y^s) \sum_{j=1}^{k} \sum_{\boldsymbol{X} \in \mathcal{X}^k} \frac{\sum_{s=1}^{n} \delta(\boldsymbol{X} = \boldsymbol{X}^s)\delta(Y = Y^s)}{\sum_{s=1}^{n} \delta(Y = Y^s)} \Bigg[\log \frac{\sum_{s=1}^{n} \delta(\boldsymbol{X} = \boldsymbol{X}^s)\delta(Y = Y^s)}{\sum_{s=1}^{n} \delta(\boldsymbol{X}_{\setminus j}\boldsymbol{X}^s_{\setminus j})\delta(Y = Y^s)} - \log P(X_j \mid \boldsymbol{X}_{\setminus j}, Y, \boldsymbol{\omega}) \Bigg],$$

luego de reacomodar los términos,

$$\begin{aligned}&\mathrm{PKL}(\hat{Q}_{X,Y}\|P_{X,Y}(\boldsymbol{\omega})) = \\ &\frac{1}{n}\sum_{j=1}^{k}\sum_{Y\in\mathcal{Y}}\sum_{\boldsymbol{X}\in\mathcal{X}^k}\sum_{s=1}^{n}\delta(\boldsymbol{X}=\boldsymbol{X}^s)\delta(Y=Y^s)\Bigg[\\ &\qquad \log\frac{\sum_{s=1}^{n}\delta(\boldsymbol{X}=\boldsymbol{X}^s)\delta(Y=Y^s)}{\sum_{s=1}^{n}\delta(\boldsymbol{X}_{\setminus j}=\boldsymbol{X}_{\setminus j}^s)\delta(Y=Y^s)} - \log P(X_j \mid \boldsymbol{X}_{\setminus j}, Y, \boldsymbol{\omega})\Bigg].\end{aligned}$$

Definiendo la cardinalidad de los eventos $(\boldsymbol{X}, Y)$ y $(\boldsymbol{X}_{\setminus j}, Y)$ en la muestra por $|(\boldsymbol{X}, Y)|_S = \sum_{s=1}^{n}\delta(\boldsymbol{X}=\boldsymbol{X}^s)\delta(Y=Y^s)$ y $|(\boldsymbol{X}_{j\setminus}, Y)|_S = \sum_{s=1}^{n}\delta(\boldsymbol{X}_{\setminus j}=\boldsymbol{X}_{\setminus j}^s)\delta(Y=Y^s)$ respectivamente, obtenemos

$$\begin{aligned}\mathrm{PKL}(\hat{Q}_{X,Y}\|P_{X,Y}(\boldsymbol{\omega})) = \frac{1}{n}\sum_{j=1}^{k}\sum_{Y\in\mathcal{Y}}\sum_{\boldsymbol{X}\in\mathcal{X}^k}|(\boldsymbol{X},Y)|_S\Bigg[\\ \log\frac{|(\boldsymbol{X},Y)|_S}{|(\boldsymbol{X}_{\setminus j},Y)|_S} - \log P(X_j \mid \boldsymbol{X}_{\setminus j}, Y, \boldsymbol{\omega})\Bigg].\end{aligned}$$

Finalmente,

$$\begin{aligned}\mathrm{PKL}(\hat{Q}_{X,Y}\|P_{X,Y}(\boldsymbol{\omega})) &= \frac{1}{n}\sum_{j=1}^{k}\sum_{Y\in\mathcal{Y}}\sum_{\boldsymbol{X}\in\mathcal{X}^k}|(\boldsymbol{X},Y)|_S\log\frac{|(\boldsymbol{X},Y)|_S}{|(\boldsymbol{X}_{\setminus j},Y)|_S} - \\ &\quad \frac{1}{n}\sum_{j=1}^{k}\sum_{s=1}^{n}\log P(X_j^s \mid \boldsymbol{X}_{\setminus j}^s, Y^s, \boldsymbol{\omega}) \\ &= C - \frac{1}{n}\sum_{s=1}^{n}\sum_{j=1}^{k}\log P(X_j^s \mid \boldsymbol{X}_{\setminus j}^s, Y^s, \boldsymbol{\omega}) \\ &= C - \hat{E}\sum_{j=1}^{k}\ell_{j,\setminus j}(\boldsymbol{\omega}),\end{aligned}$$

donde C es una constante independiente de $\boldsymbol{\omega}$.

□

C.2 SCORE PROPIO PARA DISTRIBUCIONES EN EL DOMINIO DE LOS ENTEROS, EJEMPLO 4.1

Dado el conjunto de cliques C_x y la entropía local $H_x(p_x, p_{x+1})$ definidas en el ejemplo 4.1, como consecuencia del Teorema 4.1,

$$\begin{aligned}S(x,P) &= \frac{\partial}{\partial\alpha_x}\sum_{C_x:x\in\mathbb{N}\cup\{0\}}H_x(\boldsymbol{\alpha}_{C_x}) \\ &= \frac{\partial}{\partial\alpha_x}\left(H_{x-1}(\alpha_{x-1},\alpha_x) + H_x(\alpha_x,\alpha_{x+1})\right) \\ &= \frac{\partial}{\partial\alpha_x}\left(\alpha_{x-1}G_{x-1}(\alpha_x/\alpha_{x-1}) + \alpha_x G_x(\alpha_{x+1}/\alpha_x)\right) \\ &= G'_{x-1}(\nu_{x-1}) + G_x(\nu_x) - \nu_x G'_x(\nu_x) \quad (x=0,1,\cdots),\end{aligned}$$

resulta un score propio local, donde $\nu_x = p_{x+1}/p_x$ y con el primer termino nulo para $x = 0$.

Con esto tenemos que

$$S(x,\mu) = G'_{x-1}\left(\frac{\mu}{x}\right) + G_x\left(\frac{\mu}{x+1}\right) - \frac{\mu}{x+1}G'_x\left(\frac{\mu}{x+1}\right),$$
$$\frac{dS(x,\mu)}{d\mu} = \frac{1}{x}G''_{x-1}\left(\frac{\mu}{x}\right) - \frac{\mu}{(x+1)^2}G''_x\left(\frac{\mu}{x+1}\right).$$

Considerando una función convexa dada en el Ejemplo 4.1 y sus derivadas,

$$G_x(v) = -(x+1)^a\frac{v^m}{m(m-1)} \quad \text{for } m \neq 0,1$$
$$G'_x(v) = -(x+1)^a\frac{v^{m-1}}{m-1}$$
$$G''_x(v) = -(x+1)^a v^{m-2},$$

el score local inducido puede escribirse como

$$S(x,\mu) = -x^{a-m+1}\frac{\mu^{m-1}}{m-1} - (x+1)^{a-m}\frac{\mu^m}{m(m-1)} + (x+1)^{a-m}\frac{\mu^m}{m-1} \tag{C.5}$$
$$= -x^{a-m+1}\frac{\mu^{m-1}}{m-1} + (x+1)^{a-m}\frac{\mu^m}{m},$$
$$\frac{dS(x,\mu)}{d\mu} = -x^{a-m+1}\mu^{m-2} + (x+1)^{a-m}\mu^{m-1}, \tag{C.6}$$

donde el primer término es nulo para $x = 0$. Notar que con $m \leq a$, esta condición es satisfecha.

C.3 NORMAL-PGM, EJEMPLO 2.1

En esta sección derivamos algunos scores para la distribución Normal-pGM. Veremos también que algunos scores nos llevan a estimadores en forma cerrada:

C.3.1 *MLE*

Es posible obtener el MLE (Definición 4.3) en forma cerrada para el modelo Normal-pGM, para ello escribimos las ecuaciones de estimación:

$$\left.\begin{aligned}\frac{\partial \hat{E}S_{\text{SME}}(\boldsymbol{x},(\boldsymbol{\eta};\boldsymbol{\Theta}))}{\partial \boldsymbol{\eta}} &= \mathbf{0}\\ \frac{\partial \hat{E}S_{\text{SME}}(\boldsymbol{x},(\boldsymbol{\eta};\boldsymbol{\Theta}))}{\partial \boldsymbol{\Theta}} &= \mathbf{0}\end{aligned}\right\}$$
$$\left.\begin{aligned}\hat{E}\boldsymbol{x} + \boldsymbol{\Theta}^{-1}\boldsymbol{\eta} &= \mathbf{0}\\ \hat{E}\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{\Theta}^{-1}\boldsymbol{\eta}\boldsymbol{\eta}^\top\boldsymbol{\Theta}^{-1} + \boldsymbol{\Theta}^{-1} &= \mathbf{0}\end{aligned}\right\}$$
$$\left.\begin{aligned}\hat{E}\boldsymbol{x} + \boldsymbol{\Theta}^{-1}\boldsymbol{\eta} &= \mathbf{0}\\ \hat{E}\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{\Theta}^{-1}\boldsymbol{\Theta}\hat{E}\boldsymbol{x}\hat{E}\boldsymbol{x}^\top\boldsymbol{\Theta}\boldsymbol{\Theta}^{-1} + \boldsymbol{\Theta}^{-1} &= \mathbf{0}\end{aligned}\right\}$$
$$\left.\begin{aligned}\hat{E}\boldsymbol{x} + \boldsymbol{\Theta}^{-1}\boldsymbol{\eta} &= \mathbf{0}\\ \hat{\boldsymbol{S}} + \boldsymbol{\Theta}^{-1} &= \mathbf{0}\end{aligned}\right\}$$

de donde resulta $\hat{\boldsymbol{\Theta}} = -\hat{\boldsymbol{S}}^{-1}$ y $\hat{\boldsymbol{\eta}} = -\hat{\boldsymbol{\Theta}}\hat{E}\boldsymbol{x}$, donde definimos la matriz de covarianza muestral $\hat{\boldsymbol{S}} = \hat{E}\boldsymbol{x}\boldsymbol{x}^\top - \hat{E}\boldsymbol{x}\hat{E}\boldsymbol{x}^\top$.

C.3.2 *SME*

El score S_{SM} dado en (4.20) se especializa para el caso del modelo Normal-pGM como

$$\begin{aligned} S_{\text{SME}}(\boldsymbol{\eta}, \boldsymbol{\Theta} \mid \boldsymbol{x}) &= \sum_{j=1}^{k} \frac{\partial^2}{\partial x_j^2}\ell + \frac{1}{2}\left(\frac{\partial}{\partial x_j}\ell\right)^2 \\ &= \sum_{j=1}^{k} \Theta_{jj} + \frac{1}{2}\left(\boldsymbol{\theta}_j^T\boldsymbol{x} + \boldsymbol{\eta}_j\right)^2 \\ &= \operatorname{tr}(\boldsymbol{\Theta}) + \frac{1}{2}\|\boldsymbol{\Theta}\boldsymbol{x} + \boldsymbol{\eta}\|_2^2. \end{aligned} \tag{C.7}$$

Es posible obtener el SME (Proposición 4.2) en forma cerrada para el modelo Normal-pGM, para ello escribimos las ecuaciones de estimación:

$$\left.\begin{aligned} \frac{\partial \hat{E} S_{\text{SME}}(\boldsymbol{x}, (\boldsymbol{\eta}; \boldsymbol{\Theta}))}{\partial \boldsymbol{\eta}} &= \boldsymbol{0} \\ \frac{\partial \hat{E} S_{\text{SME}}(\boldsymbol{x}, (\boldsymbol{\eta}; \boldsymbol{\Theta}))}{\partial \boldsymbol{\Theta}} &= \boldsymbol{0} \end{aligned}\right\}$$

$$\left.\begin{aligned} \boldsymbol{\eta} + \boldsymbol{\Theta}\hat{E}\boldsymbol{x} &= \boldsymbol{0} \\ \boldsymbol{I} + \hat{E}(\boldsymbol{\eta} + \boldsymbol{\Theta}\boldsymbol{x})\boldsymbol{x}^\top + \hat{E}\boldsymbol{x}(\boldsymbol{\eta} + \boldsymbol{\Theta}\boldsymbol{x})^\top - \hat{E}\ \operatorname{diag}(\boldsymbol{\Theta}\boldsymbol{x}\boldsymbol{x}^\top + \eta\boldsymbol{x}^\top) &= \boldsymbol{0} \end{aligned}\right\}$$

$$\left.\begin{aligned} \boldsymbol{\eta} + \boldsymbol{\Theta}\hat{E}\boldsymbol{x} &= \boldsymbol{0} \\ \boldsymbol{I} + \boldsymbol{\Theta}\hat{\boldsymbol{S}} + \hat{\boldsymbol{S}}\boldsymbol{\Theta} - \operatorname{diag}(\boldsymbol{\Theta}\hat{\boldsymbol{S}}) &= \boldsymbol{0} \end{aligned}\right\}$$

de donde resulta $\hat{\boldsymbol{\theta}} = -\hat{\boldsymbol{S}}^{-1}$ y $\hat{\boldsymbol{\eta}} = -\hat{\boldsymbol{\Theta}}\hat{E}\boldsymbol{x}$.

De manera equivalente, cuando consideramos el modelo gráfico de segundo orden condicional lineal en $\boldsymbol{X}$ (Definición 2.12), donde definimos $\boldsymbol{\eta}_y = \boldsymbol{\eta}_0 + \boldsymbol{\Gamma}\boldsymbol{f}_y$, con $\boldsymbol{\Gamma}$ una matriz de dimensión $p \times r$ y $\boldsymbol{f}_y$ una función de la respuesta Y que devuelve un vector de dimensión r centrado en la muestra i. e. $\hat{E}\boldsymbol{f}_y = 0_p$. Considerando el score $S_{\text{SME}}(\boldsymbol{\eta}_0, \boldsymbol{\Gamma}, \boldsymbol{\Theta} \mid \boldsymbol{x}, y) = S_{\text{SME}}(\boldsymbol{\eta}_0 + \boldsymbol{\Gamma}\boldsymbol{f}_y, \boldsymbol{\Theta} \mid \boldsymbol{x})$ definido en (C.7), el SME para este modelo resulta:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\eta}_0}\hat{E} S_{\text{SME}}(\boldsymbol{\eta}_0, \boldsymbol{\Gamma}, \boldsymbol{\Theta} \mid \boldsymbol{x}, y) &= 0 \\ \hat{E}\frac{\partial}{\partial \boldsymbol{\eta}_y} S_{\text{SME}}(\boldsymbol{\eta}_y, \boldsymbol{\Theta} \mid \boldsymbol{x})\frac{\partial}{\partial \boldsymbol{\eta}_0}\boldsymbol{\eta}_y &= 0 \\ \boldsymbol{\Theta}\hat{E}\boldsymbol{x} + \hat{E}\boldsymbol{\eta}_y &= 0 \\ \boldsymbol{\Theta}\hat{E}\boldsymbol{x} + \boldsymbol{\eta}_0 + \beta\hat{E}\boldsymbol{f}_y &= 0 \\ \boldsymbol{\Theta}\hat{E}\boldsymbol{x} + \boldsymbol{\eta}_0 &= 0 \qquad \text{(C.8)} \\ \hat{\boldsymbol{\eta}}_0 &= -\boldsymbol{\Theta}\hat{E}\boldsymbol{x}, \qquad \text{(C.9)} \end{aligned}$$

donde (C.8) resulta de la hipótesis de f_y centrada.

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Gamma}} \hat{E} S_{\text{SME}}(\boldsymbol{\eta}_0, \boldsymbol{\Gamma}, \boldsymbol{\Theta} \mid \boldsymbol{x}, y) &= 0 \\
\hat{E} \frac{\partial}{\partial \boldsymbol{\eta}_y} S_{\text{SME}}(\boldsymbol{\eta}_y, \boldsymbol{\Theta} \mid \boldsymbol{x}) \frac{\partial}{\partial \boldsymbol{\Gamma}} \boldsymbol{\eta}_y &= 0 \\
\boldsymbol{\Theta} \hat{E} \boldsymbol{x} \boldsymbol{f}_y^T + \hat{E}(\boldsymbol{\eta}_y \boldsymbol{f}_y^T) &= 0 \\
\boldsymbol{\Theta} \hat{E} \boldsymbol{x} \boldsymbol{f}_y^T + \boldsymbol{\eta}_0 \hat{E} \boldsymbol{f}_y^T + \boldsymbol{\Gamma} \hat{E} \boldsymbol{f}_y \boldsymbol{f}_y^T &= 0 \\
\boldsymbol{\Theta} \hat{E} \boldsymbol{x} \boldsymbol{f}_y^T + \boldsymbol{\Gamma} \hat{E} \boldsymbol{f}_y \boldsymbol{f}_y^T &= 0 \\
\hat{\boldsymbol{\Gamma}} &= -\boldsymbol{\Theta} \hat{\boldsymbol{S}}_{\boldsymbol{x} \boldsymbol{f}_y} \hat{\boldsymbol{S}}_{\boldsymbol{f}_y}^{-1}
\end{aligned} \tag{C.10}$$

donde nuevamente se usó que f_y está centrada. Finalmente,

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Theta}} \hat{E} S_{\text{SME}}(\boldsymbol{\eta}_0, \boldsymbol{\Gamma}, \boldsymbol{\Theta} \mid \boldsymbol{x}, y) &= 0 \\
\hat{E} \frac{\partial}{\partial \boldsymbol{\Theta}} S_{\text{SME}}(\boldsymbol{\eta}_0 + \boldsymbol{\Gamma} \boldsymbol{f}_y, \boldsymbol{\Theta} \mid \boldsymbol{x}) &= 0 \\
I_k + \hat{E}(\boldsymbol{\Theta} \boldsymbol{x} + \boldsymbol{\eta}_y) \boldsymbol{x}^T + \hat{E} \boldsymbol{x} (\boldsymbol{\Theta} \boldsymbol{x} + \boldsymbol{\eta}_y)^T - \hat{E} \operatorname{diag}(\boldsymbol{\Theta} \boldsymbol{x} \boldsymbol{x}^T + \boldsymbol{\eta}_y \boldsymbol{x}^T) &= 0 \\
I_k + \boldsymbol{\Theta} S_{\boldsymbol{x}|\boldsymbol{f}_y} + S_{\boldsymbol{x}|\boldsymbol{f}_y} \boldsymbol{\Theta} - \operatorname{diag}(\boldsymbol{\Theta} S_{\boldsymbol{x}|\boldsymbol{f}_y}) &= 0 \\
\hat{\boldsymbol{\Theta}} &= -\hat{\boldsymbol{S}}_{\boldsymbol{x}|\boldsymbol{f}_y}^{-1}.
\end{aligned} \tag{C.11}$$

donde se define la covarianza condicional muestral $\hat{\boldsymbol{S}}_{\boldsymbol{x}|\boldsymbol{f}_y} = \hat{\boldsymbol{S}}_{\boldsymbol{x}} - \hat{\boldsymbol{S}}_{\boldsymbol{x}\boldsymbol{f}_y} \hat{\boldsymbol{S}}_{\boldsymbol{f}_y}^{-1} \hat{\boldsymbol{S}}_{\boldsymbol{f}_y \boldsymbol{x}}$ donde $\hat{\boldsymbol{S}}_{\boldsymbol{x}\boldsymbol{f}_y} = \hat{E}(\boldsymbol{x} - \hat{E})(\boldsymbol{f}_y - \hat{E}\boldsymbol{f}_y)^T$, $\hat{\boldsymbol{S}}_{\boldsymbol{f}_y} = \hat{E}(\boldsymbol{f}_y - \hat{E}\boldsymbol{f}_y)(\boldsymbol{f}_y - \hat{E}\boldsymbol{f}_y)^T$ y $\hat{\boldsymbol{S}}_{\boldsymbol{x}} = \hat{E}\boldsymbol{x}\boldsymbol{x}^\top - \hat{E}\boldsymbol{x}\hat{E}\boldsymbol{x}^\top$. La solución es unica como consecuencia de la condición de Sylvester. Luego, el SME coincide con el MLE, y está dado por las ecuaciones (C.11), (C.10) y (C.9).

C.4 POISSON-PGM, EJEMPLO 2.3

Partiendo del Ejemplo 4.1, donde encontramos un score propio para la distribución Poisson de media μ y considerando $m = a = 2$ en (4.23), el score y sus derivadas se simplifican en

$$\begin{aligned}
S(x, \mu) &= \left(\frac{\mu}{2} - x \right) \mu, \\
\frac{dS(x, \mu)}{d\mu} &= \mu - x.
\end{aligned} \tag{C.12}$$

Construimos un score propio para la distribución multivariada Poisson-pGM (Ejemplo 2.3) a partir del Teorema 4.2. En particular consideramos (4.25) con cada score condicional S_j igual a S en (C.12).

D

ANEXO AL CAPÍTULO 5

D.1 CONDICIONES DE OPTIMALIDAD PARA PROBLEMAS CONVEXOS

En esta sección definimos los problemas convexos y resumimos algunos resultados de optimalidad. En la Sección D.1.1 se presentan las condiciones de optimalidad KKT para problemas convexos diferenciables, las cuales se amplían al caso no diferenciable en la Sección D.1.2. Esto nos permitirá estudiar dichas condiciones en el contexto de aprendizaje estadístico en el Apéndice D.2.

Definición D.1 (Conjunto convexo)**.** *Un conjunto $\mathcal{B} \subseteqq \mathbb{R}^m$ es convexo si para cada $\omega, \omega' \in \mathcal{B}$,*

$$s\omega + (1-s)\omega' \in \mathcal{B}, \qquad \forall s \in (0,1)$$

Definición D.2 (función convexa)**.** *Una función $f : \mathbb{R}^m \to \mathbb{R}$ es convexa si para cada ω, ω' en el dominio de f,*

$$f(s\omega + (1-s)\omega') \leq sf(\omega) + (1-s)f(\omega'), \qquad \forall s \in (0,1)$$

Proposición D.1. *Si $f : \mathbb{R}^m \to \mathbb{R}$ es convexa, los mínimos locales de f son mínimos globales.*

D.1.1 *Problemas diferenciables*

Consideremos el siguiente problema de optimización con restricciones de dominio:

Problema D.1 (Problema convexo con restricciones de dominio)**.**

$$\min_{\omega \in \mathcal{B}} f(\omega) \tag{D.1}$$

donde $f : \mathbb{R}^m \to \mathbb{R}$ es convexa y diferenciable con continuidad y $\mathcal{B} \subseteqq \mathbb{R}^m$ es un conjunto convexo.

Proposición D.2 (Optimalidad para problemas diferenciables)**.** *Si f es diferenciable, $\omega^* \in \mathcal{B}$ optimiza el Problema D.1 si y solo si*

$$\langle \frac{\partial}{\partial \omega} f(\omega^*), \omega - \omega^* \rangle \geq 0 \quad \forall \omega \in \mathcal{B}. \tag{D.2}$$

Demostración. Para cualquier $\omega \in \mathcal{B}$, la convexidad de f implica para todo $s \in (0,1)$

$$\begin{aligned} f(\omega^* + s(\omega - \omega^*)) &\leq f(\omega^*) + s(f(\omega) - f(\omega^*)) \\ f(\omega) - f(\omega^*) &\geq \frac{f(\omega^* + s(\omega - \omega^*)) - f(\omega^*)}{s}, \end{aligned} \tag{D.3}$$

donde usamos la convexidad de $\mathcal{B}$ para asegurar que el argumento de f en el lado izquierdo de (D.3) se encuentra dentro del mismo conjunto. Para $s \to 0$, tenemos

$$f(\boldsymbol{\omega}) - f(\boldsymbol{\omega}^*) \geq \langle \frac{\partial}{\partial \boldsymbol{\omega}} f(\boldsymbol{\omega}^*), \boldsymbol{\omega} - \boldsymbol{\omega}^* \rangle,$$

de donde resulta

$$f(\boldsymbol{\omega}) \geq f(\boldsymbol{\omega}^*) + \langle \frac{\partial}{\partial \boldsymbol{\omega}} f(\boldsymbol{\omega}^*), \boldsymbol{\omega} - \boldsymbol{\omega}^* \rangle \geq f(\boldsymbol{\omega}^*),$$

donde usamos la condición de optimalidad. □

Cuando $\mathcal{B} = \mathbb{R}^m$, la condición de optimalidad (D.2) se reduce en $\frac{\partial}{\partial \boldsymbol{\omega}} f(\boldsymbol{\omega}^*) = 0$. Cuando la restricción $\mathcal{B}$ puede escribirse como la intersección de conjuntos de subnivel de funciones convexas $g_i : \mathbb{R}^m \to \mathbb{R}$, el problema D.1 resulta

Problema D.2 (Problema convexo compuesto)**.**

$$\min_{\boldsymbol{\omega} \in \mathbb{R}^m} f(\boldsymbol{\omega}) \text{ tal que } g_i(\boldsymbol{\omega}) \leq 0 \, i = 1, \cdots, q \tag{D.4}$$

donde g_i, $i = 1, \cdots, q$ y f son funciones convexas.

Definición D.3 (Lagrangiano)**.** *Definimos la función $L : \mathbb{R}^{m+q} \to \mathbb{R}$ asociada al Problema D.2 como*

$$L(\boldsymbol{\omega}, \lambda_1, \cdots, \lambda_q) = f(\boldsymbol{\omega}) + \sum_{i=1}^{q} \lambda_i g_i(\boldsymbol{\omega}),$$

donde $\lambda_i \geq 0$ para todo $i = 1, \cdots, q$ son llamados multiplicadores de Lagrange.

Notar que cada término $\lambda_i g_i(\boldsymbol{\omega})$ impone una penalización cuando no se satisface la restricción $g_i(\boldsymbol{\omega}) \leq 0$.

Las Condiciones de Karush-Kuhn-Tucker (KKT) relacionan el óptimo en los multiplicadores de lagrange $\lambda^{*q}_{i\,i=1}$, llamdo *óptimo dual*, con el *óptimo primal* $\boldsymbol{\omega}^* \in \mathbb{R}^m$.

KKT

Teorema D.1 (KKT problemas convexos compuestos diferenciables Bertsekas, 2016)**.** *Las condiciones que enumeramos debajo son necesarias y suficientes para que $\boldsymbol{\omega}^*$ sea el óptimo global cuando el Problema D.2 es diferenciable y satisface las condiciones de regularidad llamadas dualidad fuerte por la cual existen $\boldsymbol{\omega}^* \in \mathbb{R}^m$ y $\lambda_i^* > 0$, $i = 1, \cdots, q$ tales que*

$$L(\boldsymbol{\omega}^*, \lambda_1, \cdots, \lambda_q) \leq L(\boldsymbol{\omega}^*, \lambda_1^*, \cdots, \lambda_q^*) \leq L(\boldsymbol{\omega}, \lambda_1^*, \cdots, \lambda_q^*),$$

para todo $\boldsymbol{\omega} \in \mathbb{R}^m$ y $\lambda_i > 0$, $i = 1, \cdots, q$.

1. *Factibilidad primal: $g_i(\boldsymbol{\omega}^*) \leq 0$ para todo $i = 1, \cdots, q$.*
2. *Holgura complementaria: $\lambda_I^* g_i(\boldsymbol{\omega}^*) = 0$ para todo $i = 1, \cdots, q$.*
3. *Optimalidad del Lagrangiano:*

$$0 = \frac{\partial}{\partial \boldsymbol{\omega}} L(\boldsymbol{\omega}, \lambda_1, \cdots, \lambda_q) = \frac{\partial}{\partial \boldsymbol{\omega}} f(\boldsymbol{\omega}^*) + \sum_{i=1}^{q} \lambda_i \frac{\partial}{\partial \boldsymbol{\omega}} g_i(\boldsymbol{\omega}^*). \tag{D.5}$$

Notar que la condición de holgura complementaria no dice que $\lambda_i^* = 0$ si la restricción está inactiva en el óptimo i. e. $g_i(\boldsymbol{\omega}^)) < 0$. Es decir, bajo homgura complementaria, la condición 3 garantiza que el vector normal $-\frac{\partial}{\partial \boldsymbol{\omega}} f(\boldsymbol{\omega}^*)$ viva en el span positivo generado por los vectores $\{\frac{\partial}{\partial \boldsymbol{\omega}} g_i(\boldsymbol{\omega}^*) \mid \lambda_i^* > 0\}$.

D.1.2 *Problemas no diferenciables*

Para funciones convexas no diferenciables, consideramos una generalización del gradiente llamado subgradiente. Esta generaliza la propiedad de las funciones convexas diferenciables que asegura que la aproximación de primer orden dada por la tangente provee una cota inferior de la funcion.

Definición D.4 (subgradiente)**.** *Dada una función convexa $f : \mathbb{R}^m \to \mathbb{R}$, decimos que $\boldsymbol{z} \in \mathbb{R}^m$ es un subdiferencial de f en $\boldsymbol{\omega}$ si*

$$f(\boldsymbol{\omega}') \geq f(\boldsymbol{\omega}) + \langle \boldsymbol{z}, \boldsymbol{\omega}' - \boldsymbol{\omega} \rangle, \qquad \forall \boldsymbol{\omega}' \in \mathbb{R}^m.$$

Definiendo al subgradiente como el vector normal a un hiperplano que soporta el epígrafo de f dado por $\{(\boldsymbol{\omega}, r) \in \mathbb{R}^m \times \mathbb{R} : r \geq f(\boldsymbol{\omega})\}$.

Definición D.5 (subdiferencial)**.** *El subdiferencial $\partial f(\boldsymbol{\omega})$ de f en $\boldsymbol{\omega}$ es el conjunto de todos los subgradientes de f en $\boldsymbol{\omega}$.*

Si f es diferenciable en $\boldsymbol{\omega}$, el subdiferencial se reduce al gradiente de f en $\boldsymbol{\omega}$. En puntos donde f no es diferenciable, el subdiferencial es el conjunto convexo que contiene a todos los subgradientes.

PROPIEDADES

Proposición D.3 (Subdiferencial de una suma Bertsekas, 2016)**.** *Dadas funciones convexas continuas f y g, el subdiferencial de $f + g$ resulta*

$$\partial(g + g)(\boldsymbol{\omega}) = \boldsymbol{conv}\left(\partial f(\boldsymbol{\omega}) + \partial g(\boldsymbol{\omega})\right),$$

donde consideramos la suma de Minkowski entre conjuntos definida como $A + B = \{a + b \mid a \in A,\, b \in B\}$.

Un corolario del teorema Bertsekas-Danskin Bertsekas, 2016 permite obtener subdiferenciales de máximos puntuales:

Proposición D.4 (Bertsekas-Danskin)**.** *Supongamos que $f_1, \cdots, f_m$ son funciones convexas y sea $f(\boldsymbol{\omega}) = \max_{i=1,\cdots,m} f_i(\boldsymbol{\omega})$. El subdiferencial de f está dado por*

$$\partial f(\boldsymbol{\omega}) = \boldsymbol{conv}\left(\cup_{i=1}^m \{\partial f_i(\boldsymbol{\omega}) \mid f_i(\boldsymbol{\omega}) = f(\boldsymbol{\omega})\}\right),$$

donde $\boldsymbol{conv}(\cdot)$ indica la envolvente convexa y está dada por $\boldsymbol{conv}(\{x_i\}_{i=1}^m) = \{\sum_{i=1}^m \alpha_i x_i \mid \alpha_i \geq 0,\, i = 1, \cdots, m,\, \sum_{i=1}^m \alpha_i = 1\}$.

Las condiciones de optimalidad pueden ser generalizadas a problemas no diferenciables considerando la condición sobre el subdiferencial del lagrangeano, como enunciamos a continuación:

Teorema D.2 (Optimalidad para problemas convexos no diferenciables Bertsekas, 2016)**.** *Bajo condiciones de regularidad de las funciones involucradas, el vector $\boldsymbol{\omega}^*$ el el óptimo del Problema D.2 si y solo si*

$$\mathbf{0} \in \partial f(\boldsymbol{\omega}^*) + \sum_{j=1}^{m} \lambda_j^* \partial g_j(\boldsymbol{\omega}^*) \tag{D.6}$$

Como el subdiferencial es un conjunto, la condición (D.6) significa que el origen pertenece al conjunto resultante de la suma de Minkowski de los subdiferenciales.

KKT

En particular, las condiciones de optimalidad KKT pueden aplicarse en el caso de considerar problemas no diferenciables considerando (D.6) en lugar de (D.5) en la condición 3.

D.2 OPTIMALIDAD EN PROBLEMAS COMPUESTOS EN ESTADÍSTICA

En lo que sigue analizamos las condiciones de optimalidad de algunos problemas de estimación penalizados en estadística. Primero presentamos algunos resultados generales sobre los subgradientes involucrados necesarios para definir las condiciones de optimalidad del estimador LASSO, y LASSO por grupos presentados en las proposiciones D.8 y D.9. Dichas condiciones permiten caracterizar la penalización jerárquica dada en (5.4a) y (5.4b), además del resultado dado en (5.5).

Subgradiente de las normas ℓ_1 y $\ell_{2,1}$

Proposición D.5 (Subdiferencial del valor absoluto)**.** *El subdiferencial del valor absoluto está dado por*

$$\partial|\omega| = \begin{cases} \{+1\} & si\ \omega > 0 \\ \{-1\} & si\ \omega < 0 \\ [-1,1] & si\ \omega = 0 \end{cases}.$$

Demostración. Considerando que $|\omega| = \max\{z\omega \mid |z| \leq 1\}$ y por la Proposición D.4,

$$\begin{aligned} \partial|\omega| &= \mathbf{conv}\left(\frac{\partial}{\partial\omega} z\omega \mid z \in [-1,1],\, z\omega = |\omega|\right) \\ &= \mathbf{conv}\left(z \in [-1,1] \mid z\omega = |\omega|\right) \\ &= \begin{cases} [-1,1] & \text{si } \omega = 0 \\ \omega/|\omega|] & \text{en otro caso} \end{cases}. \end{aligned}$$

□

Proposición D.6 (Subdiferencial de la norma ℓ_1)**.** *El subdiferencial de la norma ℓ_1 está dado por*

$$\partial\|\boldsymbol{\omega}\|_1 = \partial|\omega_1| \times \cdots \times |\omega_m|,$$

donde $\times$ indica el producto cartesiano y $\partial|\omega_j|$ es el subdiferencial del valor absoluto dado en la Proposición D.5.

Demostración. Considerando que $\|\boldsymbol{\omega}\|_1 = \sum_{j=1}^{m} |\omega_j|$, por la Proposición D.3, tenemos que

$$\partial\|\boldsymbol{\omega}\|_1 = \partial|\omega_1| \times \cdots \times |\omega_m|.$$

□

La siguiente Proposición generaliza este resultado a normas generales:

Proposición D.7 (Subdiferencial de una norma general)**.** *El subdiferencial de la norma $\|\cdot\|$ está dado por*

$$\partial\|\boldsymbol{\omega}\| = \{v \in \mathbb{R}^m \mid \langle v, \boldsymbol{\omega}\rangle = \|\boldsymbol{\omega}\|, \|\boldsymbol{v}\|_* \leq 1\},$$

donde $\|\cdot\|_*$ indica la norma dual de $\|\cdot\|$ definida como

$$\|\boldsymbol{v}\|_* = \sup_{\|\boldsymbol{u}\|\leq 1} \langle \boldsymbol{v}, \boldsymbol{u}\rangle.$$

Corolario D.1 (Subdiferencial de una norma pesada)**.** *El subdiferencial de una norma ℓ_2 pesada $\|\cdot\|_{\mathcal{I}}$ definida como $\|\boldsymbol{\beta}\|_{\mathcal{I}} = \sqrt{\{\boldsymbol{\beta}^T\mathcal{I}\boldsymbol{\beta}\}}$ donde $\mathcal{I}$ es una matriz definida positiva, está dado por*

$$\partial\|\boldsymbol{\omega}\|_{\mathcal{I}} = \{v \in \mathbb{R}^m \mid \langle v, \boldsymbol{\omega}\rangle = \|\boldsymbol{\omega}\|_{\mathcal{I}}, \|\boldsymbol{v}\|_{\mathcal{I}^{-1}} \leq 1\},$$

Proposición D.8 (Optimalidad LASSO)**.** *El problema de optimización inducido por LASSO*

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \mathbb{J}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1,$$

donde $\mathbb{J}(\boldsymbol{\beta})$ es una función de costo convexa diferenciable tiene como condición necesaria de optimalidad

$$\begin{cases} \dfrac{\partial}{\partial\beta_j}\mathbb{J}(\boldsymbol{\beta}) = -\lambda\,\mathrm{sign}(\boldsymbol{\beta}), & \text{si } \beta_j \neq 0 \\ \left|\dfrac{\partial}{\partial\beta_j}\mathbb{J}(\boldsymbol{\beta})\right| \leq \lambda, & \text{si } \beta_j = 0. \end{cases}$$

Demostración. Las condiciones de optimalidad resultan de aplicar el Teorema D.2, dando lugar a la condición

$$\mathbf{0} \in \frac{\partial}{\partial\beta_j}\mathbb{J}(\boldsymbol{\beta}) + \partial|\beta_j|, \quad j = 1, \cdots, p,$$

donde $\partial|\beta_j|$ es el subgradiente del valor absoluto dado en la Proposición D.5. □

Se considera una norma pesada general en cada grupo

Proposición D.9 (Optimalidad LASSO por grupos)**.** *El problema de optimización inducido por LASSO por grupos dada por la partición $\{\boldsymbol{\beta}_j\}_{j=1}^{g}$ del vector $\boldsymbol{\beta}$ resulta*

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \mathbb{J}(\boldsymbol{\beta}) + \lambda\sum_{j=1}^{g}\|\boldsymbol{\beta}_j\|_{\mathcal{I}_j},$$

donde $\mathbb{J}(\boldsymbol{\beta})$ es una función de costo convexa diferenciable y donde $\|\cdot\|_{\mathcal{I}}$ es una norma ℓ_2 pesada por la matriz definida positiva $\mathcal{I}$, tiene como condición necesaria de optimalidad

$$\begin{cases} \dfrac{\partial}{\partial \boldsymbol{\beta}_j}\mathbb{J}(\boldsymbol{\beta}) = -\lambda \mathcal{I}_j \dfrac{\boldsymbol{\beta}_j}{\|\boldsymbol{\beta}_j\|_{\mathcal{I}_j}}, & si\ \|\boldsymbol{\beta}_j\|_2 \neq 0 \\ \left\| \dfrac{\partial}{\partial \boldsymbol{\beta}_j}\mathbb{J}(\boldsymbol{\beta}) \right\|_{\mathcal{I}_j^{-1}} \leq \lambda, & si\ \|\boldsymbol{\beta}_j\|_2 = 0. \end{cases}$$

Demostración. las condiciones de optimalidad (Teorema D.2) se reducen a

$$\mathbf{0} \in \frac{\partial}{\partial \boldsymbol{\beta}_j}\mathbb{J}(\boldsymbol{\beta}) + \partial\|\boldsymbol{\beta}_j\|_{\mathcal{I}_j}, \quad j = 1, \cdots, g,$$

Usando el Corolario D.1, se encuentra el resultado. □

E

ANEXO AL CAPÍTULO 6

E.1 OPTIMIZACIÓN DEL PROBLEMA PENALIZADO (5.3) PARA MODELOS PGM

En esta sección definimos los operadores proximales necesarios para la aplicación del Algoritmo 4 (gradiente proximal) vinculados al problema (5.3).

La siguiente proposición permite incorporar la restricción de simetría en las interacciones $\mathbf{\Theta}$ de los modelos pGM (Definición 2.5).

Proposición E.1. *$\mathbf{prox}_{\lambda\tilde{g}}(v_{ij}, v_{ji}) = S_\lambda((v_{ij}+v_{ji})/2)$, where $S_\lambda(\cdot)$ is the soft-thresholding operator and $\tilde{g}(\theta_{ij}, \theta_{ji}) = \|\theta_{ij}\|_1 + \|\theta_{ji}\|_1 + \iota_{\mathbf{\Omega}}$, where $\tilde{\mathcal{C}} = \{\theta_{ij} : \theta_{ij} = \theta_{ji}\}$*

Demostración.

$$
\begin{aligned}
\underset{\lambda\tilde{g}}{\mathbf{prox}}(v_{ij}, v_{ji}) &= \underset{\theta_{ij},\theta_{ji}}{\arg\min}\, \tilde{g}(\theta_{ij}, \theta_{ji}) + 1/(2\lambda)\left((\theta_{ij}-v_{ij})^2 + (\theta_{ji}-v_{ji})^2\right) \\
&= \underset{\theta_{ij}=\theta_{ji}}{\arg\min}\, \|\theta_{ij}\|_1 + \|\theta_{ji}\|_1 + 1/(2\lambda)\left((\theta_{ij}-v_{ij})^2 + (\theta_{ji}-v_{ji})^2\right) \\
&= \underset{\theta_{ij}}{\arg\min}\, 2\|\theta_{ij}\|_1 + 2/(2\lambda)(\theta_{ij} - (v_{ij}+v_{ji})/2)^2 + \text{cte} \\
&= \underset{\theta_{ij}}{\arg\min}\, \|\theta_{ij}\|_1 + 1/(2\lambda)(\theta_{ij} - (v_{ij}+v_{ji})/2)^2 \\
&= S_\lambda((v_ij + v_{ji})/2).
\end{aligned}
$$

□

La siguiente proposición define el operador proximal vinculado a la norma ℓ_1 de $\mathbf{\Theta}$ para modelos Poisson-pGM que incorporan la restricción $\Theta_{j,l} \leq 0, \quad j \neq l = 1, \cdots, p$. Dicha restricción se incorpora mediante la función indicadora $\iota_{\mathcal{C}^-}$ con $\mathcal{C}^- = \{x | x \leq 0\}$, esta función toma el valor cero cuando el argumento es no positivo e infinito cuando es positivo.

Proposición E.2. *$\mathbf{prox}_{\lambda|\cdot|+\iota_{\mathcal{C}^-}}(v) = -(-v-\lambda)_+$ donde $\mathcal{C}^- = \{x | x \leq 0\}$.*

Demostración.

$$
\begin{aligned}
\underset{\lambda|\cdot|+\iota_{\mathcal{C}^-}}{\mathbf{prox}}(v) &= \underset{\theta}{\arg\min}\, \lambda|\theta| + \iota_{\mathcal{C}^-} + 1/2(\theta - v)^2 \\
&= \underset{\theta\leq 0}{\arg\min}\, g_{\lambda,v}(\theta),
\end{aligned}
$$

donde $g_{\lambda,v}(\theta) = \lambda|\theta| + 1/2(\theta - v)^2$ es convexa y su subdiferencial (Definición D.5) resulta

$$
\begin{aligned}
\partial g(\bar{\theta}) &= \bar{\theta} - s - \lambda, \text{if } \bar{\theta} < 0 \\
\partial g(0) &= [-s-\lambda, 0]
\end{aligned}
$$

$0 \in \partial g(0) \Leftrightarrow s > -\lambda$ and if $0 > -\lambda > s$, $0 \in \partial g(\hat{\theta}_\lambda) \Leftrightarrow \hat{\theta}_\lambda = s + \lambda$. Resultando en $\hat{\theta}_\lambda = -(-s - \lambda)_+$, donde $(\cdot)_+$ devuelve la parte no negativa del argumento. □

En relación con la Observación 6.1: Notemos primero que la Proposición E.1 nos dice que la restricción de simetría impuesta en $\imath_{\mathbf{\Omega}}(\cdot)$ puede combinarse con el operador proximal de soft-thresholding; i. e. $\mathbf{prox}_{\lambda h}(v)$ es separable sobre los pares simétricos $(\theta_{ij}, \theta_{ji})$, con lo cual, el operador proximal está dado por $\mathbf{prox}_{\lambda_1 g_{ij}}((v_{ij}, v_{ji})) = -\left(-(v_{ij} + v_{ji})/2 - \lambda_1\right)_+$, donde $(\cdot)_+$ es el operador que devuelve la parte no negativa. Esto resuelve la condición de simetría en las interacciones Θ de cada modelo. En el caso del modelo Poisson-pGM donde $\mathbf{\Omega}$ incluye además la restricción de no-positividad de las interacciones debemos considerar la Proposición E.2 que define al operador de soft-thresholding restringido a los valores no positivos. Por otro lado, el operador proximal de $h(\mathbf{\Gamma})$ es el operador de LASSO por grupos (Parikh y Boyd, 2014, 6.5.4 Sum of norms).

E.2 OPTIMIZACIÓN DEL PROBLEMA PENALIZADO (5.6) PARA MODELOS ZIPGM

Para optimizar el problema (5.6), en la Sección 6.2.2.1 propusimos una solución que combina el Algoritmo 5 con el Algoritmo 3. En la Sección E.2.1 de este apéndice se detallan los operadores proximales involucrados permitiendo la aplicación de la solución propuesta. Seguidamente, en la Sección E.2.2, detallamos primero la estructura que se asume para las métricas $\mathcal{I}$ e $\mathcal{I}(\boldsymbol{\omega}_0)$ y luego la forma que toma el Jacobiano de la función de costo inducida por pseudolikelihood.

E.2.1 *Operadores proximales asociados al problema de optimización de modelos zipGM*

Enunciamos a continuación los operadores proximales (Definición 6.1) correspondientes a las funciones convexas $\tilde{g}_\ell, \tilde{g}_\mathcal{R}$, $\tilde{g}_\mathcal{C}$ y $\imath_{\mathbf{\Omega}}$ dadas en (6.9)-(6.12) necesarias para optimizar (6.8) usando el Algoritmo 3. Dichos operadores resultan de la especialización de operadores usuales estudiados por Parikh, Boyd et al. (2014).

$$\begin{aligned}
\underset{\gamma\tilde{g}_\ell}{\mathbf{prox}}(\tilde{\boldsymbol{\omega}}_0) &:= \arg\min_{\tilde{\boldsymbol{\omega}}} \tilde{g}_\ell(\tilde{\boldsymbol{\omega}}) + \frac{1}{2\gamma}\|\tilde{\boldsymbol{\omega}} - \tilde{\boldsymbol{\omega}}_0\|_2^2 && \text{(E.1)}\\
&= \arg\min_{\bar{\boldsymbol{\omega}}} \frac{1}{2t}\|\boldsymbol{\omega} - \left(\boldsymbol{\omega}^{[k]} - t\mathcal{I}^{-1}\frac{\partial}{\partial\boldsymbol{\omega}}\hat{E}S(\boldsymbol{\omega}^{[k]})\right)\|_2^2 \\
&\qquad + \frac{1}{2\gamma}\|\bar{\boldsymbol{\omega}} - \bar{\boldsymbol{\omega}}_0\|_2^2 \\
&= \left(\frac{1}{t} + \frac{1}{\gamma}\right)^{-1}\left(\frac{1}{t}\tilde{\boldsymbol{\omega}}^{[k]} + \frac{1}{\gamma}\tilde{\boldsymbol{\omega}}_0 - \mathcal{I}^{-1/2}\frac{\partial}{\partial\boldsymbol{\omega}}\hat{E}S(\boldsymbol{\omega}^{[k]})\right),
\end{aligned}$$

$$\left[\underset{\gamma\bar{g}_\mathcal{R}}{\mathbf{prox}}(\bar{\boldsymbol{\omega}}_0)\right]_{\mathcal{R}_j} := \arg\min_{[\bar{\boldsymbol{\omega}}]_{\mathcal{R}_j}} g_\mathcal{R}([\bar{\boldsymbol{\omega}}]_{\mathcal{R}_j}) + \frac{1}{2\gamma}\|\,[\bar{\boldsymbol{\omega}}]_{\mathcal{R}_j} - [\bar{\boldsymbol{\omega}}_0]_{\mathcal{R}_j}\,\|_2^2 \qquad \text{(E.2)}$$

$$= \left(1 - \frac{\lambda_\mathcal{R}\gamma}{\sum_{j=1}^k \|\,[\bar{\boldsymbol{\omega}}_0]_{\mathcal{R}_j}\,\|_2}\right)_+ [\bar{\boldsymbol{\omega}}_0]_{\mathcal{R}_j},$$

$$\left[\underset{\gamma\bar{g}_\mathcal{C}}{\mathbf{prox}}(\bar{\boldsymbol{\omega}}_0)\right]_{\mathcal{C}_{jl}} := \arg\min_{[\bar{\boldsymbol{\omega}}]_{\mathcal{C}_{jl}}} g_\mathcal{C}([\bar{\boldsymbol{\omega}}]_{\mathcal{C}_{jl}}) + \frac{1}{2\gamma}\|\,[\bar{\boldsymbol{\omega}}]_{\mathcal{C}_{jl}} - [\bar{\boldsymbol{\omega}}_0]_{\mathcal{C}_{jl}}\,\|_2^2 \qquad \text{(E.3)}$$

$$= \left(1 - \frac{\lambda_\mathcal{C}\gamma}{\|\,[\bar{\boldsymbol{\omega}}_0]_{\mathcal{C}_{jl}}\,\|_2}\right)_+ [\bar{\boldsymbol{\omega}}_0]_{\mathcal{C}_{jl}},$$

$$\begin{aligned}
\underset{\iota_\mathbf{\Omega}}{\mathbf{prox}}(\bar{\boldsymbol{\omega}}_0) &:= \arg\min_{\bar{\boldsymbol{\omega}}} \iota_\mathbf{\Omega}(\bar{\boldsymbol{\omega}}) + \frac{1}{2\gamma}\|\bar{\boldsymbol{\omega}} - \bar{\boldsymbol{\omega}}_0\|_2^2 && \text{(E.4)}\\
&= \arg\min_{\bar{\boldsymbol{\omega}}\in\mathbf{\Omega}} \|\bar{\boldsymbol{\omega}} - \bar{\boldsymbol{\omega}}_0\|_2,
\end{aligned}$$

donde $[\cdot]_{\mathcal{R}_j}$ y $[\cdot]_{\mathcal{C}_{jl}}$ considera los bloques correspondientes a los parámetros $(\mathbf{\Gamma}_j; \mathbf{\Psi}_j; (\Theta_{jl}; \Phi_{lj}; \Phi_{jl}; \Lambda_{jl}))$ y $(\Theta_{jl}; \Phi_{lj}; \Phi_{jl}; \Lambda_{jl})$ respectivamente.

Además, (E.2) y (E.3) son operadores llamados de *block soft thresholding* en inglés, donde $(\cdot)_+$ devuelve la parte positiva del argumento. Dada la estructura de la penalidad en (6.8), los operadores proximales de las funciones $\bar{g}_\mathcal{R}$ y $\bar{g}_\mathcal{C}$ pueden separarse de acuerdo al índice j y j, l respectivamente.

La estructura asumida para el cómputo (4.43) de las métricas $\mathcal{I}$ e $\mathcal{I}(\boldsymbol{\omega}_0)$ fue presentada en la Sección E.2.2.1. Los gradientes $\frac{\partial}{\partial\boldsymbol{\omega}}\hat{E}S(\boldsymbol{\omega}^{[k]})$ involucrados correspondientes a los ejemplos 2.7-2.9 fueron presentados en la Sección E.2.2.2.

E.2.2 *Jacobiano y Hessiano de la pseudolikelihood para modelos zipGM*

En esta sección se presenta la estructura asumida para el Hessiano de la función de costo inducida por pseudolikelihood (Definición 4.6) para la familia zipGM, lo cuál nos permite estimar la métrica local $\mathcal{I}$ (4.43) que define la penalización y se emplea para encontrar direcciones descendentienes en el espacio de parámetros junto con los gradientes involucrados.

E.2.2.1 *Hessiano diagonal por bloques*

Para el cálculo de las matrices de información inducidas por la pseudolikelihhod, y que en el caso independiente coincide con la matriz de información de Fisher, asumimos la estructura diagonal por bloques definidas por los grupos para cada $j \neq l \in \{1, \ldots, k\}$,

$$\boldsymbol{b}_j = (\eta_j; \xi_j; \Theta_{jj}), \quad \boldsymbol{r}_j = (\Gamma_j; \Psi_j), \quad \boldsymbol{w}_{jl} = (\Theta_{jl}; \Phi_{lj}; \Phi_{jl}; \Lambda_j). \quad \text{(E.5)}$$

Usando (E.5), las matrices que definen las normas pesadas en (5.6) resultan los bloques $\mathcal{R}_{jl} = \operatorname{diag}(\mathcal{I}_{\boldsymbol{r}_j}, (\mathcal{I}_{\boldsymbol{w}_{jl}})_{l \neq j}$ y $\mathcal{C}_{jl} = \mathcal{I}_{\boldsymbol{w}_{jl}}$ de la matriz de información de Fisher, respectivamente.

Usando la estructura diagonal por bloques (E.5), $\mathcal{I}$ puede escribirse en forma cerrada como la esperanza del Hessiano. Detallamos debajo cada bloque de esta matriz.

El bloque correspondiente a los parámetros independientes $\boldsymbol{b}_j = (\eta_j; \xi_j; \Theta_{jj})$ resulta

$$\begin{aligned}\mathcal{I}_{\boldsymbol{b}_j} &= E \frac{\partial^2}{\partial \boldsymbol{b}_j \partial \boldsymbol{b}_j^T} \sum_{j'=1}^{k} \ell_{j'|\setminus j'} = E \frac{\partial^2 \ell_{j|\setminus j}}{\partial(\eta_{j\setminus j}; \xi_{j\setminus j}; \Theta_{jj}) \partial(\eta_{j\setminus j}; \xi_{j\setminus j}; \Theta_{jj})^T} \\ &= E \frac{\partial^2 A(\eta_{j|\setminus j}, \xi_{j|\setminus j}, \Theta_{jj})}{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j}, \Theta_{jj}) \partial(\eta_{j|\setminus j}; \xi_{j|\setminus j}, \Theta_{jj})^T}.\end{aligned}$$

Para los parámetros de regresión $\boldsymbol{r}_j = (\Gamma_j, \Psi_j)$ obtenemos

$$\begin{aligned}\mathcal{I}_{\boldsymbol{r}_j} &= E \frac{\partial^2}{\partial \boldsymbol{r}_j \partial \boldsymbol{r}_j^T} \sum_{j'=1}^{k} \ell_{j'|\setminus j'} \\ &= E \frac{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j})}{\partial \boldsymbol{r}_j^T} \frac{\partial^2 \ell_{j|\setminus j}}{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j}) \partial(\eta_{j|\setminus j}; \xi_{j|\setminus j})^T} \frac{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j})}{\partial \boldsymbol{r}_j},\end{aligned}$$

donde

$$\begin{aligned}\frac{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j})}{\partial \boldsymbol{r}_j^T} &= I_2 \otimes f_y^T, \\ \frac{\partial^2 \ell_{j|\setminus j}}{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j}) \partial(\eta_{j|\setminus j}; \xi_{j|\setminus j})^T} &= \frac{\partial^2 A(\eta_{j|\setminus j}, \xi_{j|\setminus j}, \Theta_{jj})}{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j}) \partial(\eta_{j|\setminus j}; \xi_{j|\setminus j})^T},\end{aligned}$$

donde I_2 es la matriz identidad de dimensión 2. Finalmente, para las interacciones $\boldsymbol{w}_{jl} = (\Theta_{jl}, \Phi_{lj}, \Phi_{jl}, \Lambda_j)$, resulta

$$\begin{aligned}\mathcal{I}_{\boldsymbol{w}_{jl}} &= E \frac{\partial^2}{\partial \boldsymbol{w}_{jl} \partial \boldsymbol{w}_{jl}^T} \sum_{j'=1}^{k} \ell_{j'|\setminus j'} \\ &= E \frac{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j})}{\partial \boldsymbol{w}_{jl}^T} \frac{\partial^2 \ell_{j|\setminus j}}{\partial(\eta_j; \xi_j) \partial(\eta_j; \xi_j)^T} \frac{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j})}{\partial \boldsymbol{w}_{jl}} + \\ &\quad E \frac{\partial(\eta_{l|\setminus l}; \xi_{l|\setminus l})}{\partial \boldsymbol{w}_{jl}^T} \frac{\partial^2 \ell_{l|\setminus l}}{\partial(\eta_l; \xi_l) \partial(\eta_l; \xi_l)^T} \frac{\partial(\eta_{l|\setminus l}; \xi_{l|\setminus l})}{\partial \boldsymbol{w}_{jl}} \\ &= E \frac{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j})}{\partial \boldsymbol{w}_{jl}^T} \frac{\partial^2 A(\eta_{j|\setminus j}, \xi_{j|\setminus j}, \Theta_{jj})}{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j}) \partial(\eta_{j|\setminus j}; \xi_{j|\setminus j})^T} \frac{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j})}{\partial \boldsymbol{w}_{jl}} + \\ &\quad E \frac{\partial(\eta_{l|\setminus l}; \xi_{l|\setminus l})}{\partial \boldsymbol{w}_{jl}^T} \frac{\partial^2 A(\eta_{l|\setminus l}, \xi_{l|\setminus l}, \Theta_{ll})}{\partial(\eta_{l|l}; \xi_{l|l}) \partial(\eta_{l|l}; \xi_{l|l})^T} \frac{\partial(\eta_{l|\setminus l}; \xi_{l|\setminus l})}{\partial \boldsymbol{w}_{jl}},\end{aligned}$$

donde

$$\frac{\partial(\eta_j;\xi_j)}{\partial \boldsymbol{w}_{jl}} = I_2 \otimes \begin{bmatrix} X_l \\ \nu(X_l) \end{bmatrix}^T, \frac{\partial(\eta_l;\xi_l)}{\partial \boldsymbol{w}_{jl}} = \begin{bmatrix} X_j \\ \nu(X_j) \end{bmatrix}^T \otimes I_2.$$

CÁLCULO DEL HESSIANO DE $A(\eta_{j|\setminus j}, \xi_{j|\setminus j}, \Theta_{jj})$ Usando (2.19) y considerando $T(0) = 0$, tenemos

$$A(\xi_{j|\setminus j}, \eta_{j|\setminus j}, \Theta_{jj}) = \log\left(1 + \exp\left\{\xi_{j|\setminus j} + A^+(\eta_{j|\setminus j}, \Theta_{jj})\right\}\right). \quad \text{(E.6)}$$

Las derivadas de primer orden resultan

$$\frac{\partial A(\eta_{j|\setminus j}, \xi_{j|\setminus j}, \Theta_{jj})}{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j}; \Theta_{jj})} = g\left(\xi_{j|\setminus j} + A^+(\eta_{j|\setminus j}, \Theta_{jj})\right) \begin{bmatrix} \frac{\partial A^+(\eta_{j|\setminus j}, \Theta_{jj})}{\partial \eta_{j|\setminus j}} \\ 1 \\ \frac{\partial A^+(\eta_{j|\setminus j}, \Theta_{jj})}{\partial \Theta_{jj}} \end{bmatrix},$$

donde $g(x) = (1 + e^{-x})^{-1}$ es la función sigmoidea. Las derivadas de segundo orden resultan

$$\begin{aligned}
&\frac{\partial^2 A(\eta_{j|\setminus j}, \xi_{j|\setminus j}, \Theta_{jj})}{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j}; \Theta_{jj})\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j}; \Theta_{jj})^T} = \\
&= g\left(\xi_{j|\setminus j} + A^+(\eta_{j|\setminus j}, \Theta_{jj})\right)\left(1 - \right. \\
&\qquad \left. g\left(\xi_{j|\setminus j} + A^+(\eta_{j|\setminus j}, \Theta_{jj})\right)\right) \begin{bmatrix} \frac{\partial A^+(\eta_{j|\setminus j}, \Theta_{jj})}{\partial \eta_{j|\setminus j}} \\ 1 \\ \frac{\partial A^+(\eta_{j|\setminus j}, \Theta_{jj})}{\partial \Theta_{jj}} \end{bmatrix} \begin{bmatrix} \frac{\partial A^+(\eta_{j|\setminus j}, \Theta_{jj})}{\partial \eta_{j|\setminus j}} \\ 1 \\ \frac{\partial A^+(\eta_{j|\setminus j}, \Theta_{jj})}{\partial \Theta_{jj}} \end{bmatrix}^T + \\
&g\left(\xi_{j|\setminus j} + A^+(\eta_{j|\setminus j}, \Theta_{jj})\right) \begin{bmatrix} \frac{\partial^2 A^+(\eta_{j|\setminus j}, \Theta_{jj})}{\partial \eta_{j|\setminus j}^2} & 0 & \frac{\partial^2 A^+(\eta_{j|\setminus j}, \Theta_{jj})}{\partial \eta_{j|\setminus j} \partial \Theta_{jj}} \\ 0 & 0 & 0 \\ \frac{\partial^2 A^+(\eta_{j|\setminus j}, \Theta_{jj})}{\partial \Theta_{jj} \partial \eta_{j|\setminus j}} & 0 & \frac{\partial^2 A^+(\eta_{j|\setminus j}, \Theta_{jj})}{\partial \Theta_{jj}^2} \end{bmatrix},
\end{aligned}$$

ya que $g'(x) = g(x)\{1 - g(x)\}$.

E.2.2.2 *Cálculo del Jacobiano*

Computamos el jacobiano $J_{\boldsymbol{b}_j}, J_{\boldsymbol{r}_j}, J_{\boldsymbol{w}_{jl}}$ de cada bloque $\boldsymbol{\omega}_{\boldsymbol{b}_j}, \boldsymbol{\omega}_{\boldsymbol{r}_j}$ y $\boldsymbol{\omega}_{\boldsymbol{w}_{lj}}$ mediante

$$\begin{aligned}
J_{\boldsymbol{b}_j} &= E\frac{\partial \ell_{j|\setminus j}}{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j}; \Theta_{jj})} = E\frac{\partial A(\eta_{j|\setminus j}, \xi_{j|\setminus j}, \Theta_{jj})}{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j}; \Theta_{jj})} - E(T(X_j); \nu(X_j); T(X_j)^2/2), \\
J_{\boldsymbol{r}_j} &= E\frac{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j})}{\partial \boldsymbol{r}_j}\frac{\partial \ell_{j|\setminus j}}{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j})}, \\
J_{\boldsymbol{w}_{jl}} &= E\frac{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j})}{\partial \boldsymbol{w}_{jl}}\frac{\partial \ell_{j|\setminus j}}{\partial(\eta_{j|\setminus j}; \xi_{j|\setminus j})} + \frac{\partial(\eta_{l|\setminus l}; \xi_{l|\setminus l})}{\partial \boldsymbol{w}_{jl}}\frac{\partial \ell_{l|\setminus l}}{\partial(\eta_{l|\setminus l}; \xi_{l|\setminus l})}.
\end{aligned}$$

A continuación definimos las derivadas de $A^+(\eta_{j|\setminus j}, \Theta_{jj})$ para las distribuciones definidas en los Ejemplos 2.7-2.9.

NORMAL-ZIPGM A partir de (2.24),

$$A^+(\eta_{j|\setminus j}, \Theta_{jj}) = -\frac{\eta_{j|\setminus j}^2}{2\Theta_{jj}} - \frac{1}{2}\log(-\Theta_{jj}) + \frac{1}{2}\log(2\pi).$$

Luego,

$$\frac{\partial A^+(\eta_{j|\setminus j}, \Theta_{jj})}{\partial(\eta_{j|\setminus j}; \Theta_{jj})} = \begin{bmatrix} -\frac{\eta_{j|\setminus j}}{\Theta_{jj}}, \\ \frac{\eta_{j|\setminus j}^2}{2\Theta_{jj}^2} - \frac{1}{2\Theta_{jj}} \end{bmatrix}$$

$$\frac{\partial^2 A^+(\eta_{j|\setminus j}, \Theta_{jj})}{\partial(\eta_{j|\setminus j}; \Theta_{jj})\partial(\eta_{j|\setminus j}; \Theta_{jj})^T} = \begin{bmatrix} -\frac{1}{\Theta_{jj}} & \frac{\eta_{j|\setminus j}}{\Theta_{jj}^2} \\ \frac{\eta_{j|\setminus j}}{\Theta_{jj}^2} & -\frac{\eta_{j|\setminus j}^2}{\Theta_{jj}^3} + \frac{1}{2\Theta_{jj}^2} \end{bmatrix}.$$

POISSON-ZIPGM A partir de (2.26),$A^+(\eta_{j|\setminus j}) = \log(\exp(\exp(\eta_{j|\setminus j})) - 1)$. Definiendo la variable auxiliar $\widetilde{X}_{j|\setminus j} \sim$Poisson$(\exp(\eta_{j|\setminus j}))$ y llamando $\beta_{j|\setminus j} = \frac{P(\widetilde{X}_{j|\setminus j}>1)}{P(\widetilde{X}_{j|\setminus j}=1)}$, tenemos:

$$\begin{aligned} \frac{\partial A^+(\eta_{j|\setminus j})}{\partial \eta_{j|\setminus j}} &= \exp(\eta_{j|\setminus j}) \frac{P(\widetilde{X}_{j|\setminus j} = 0) + P(\widetilde{X}_{j|\setminus j} > 0)}{P(\widetilde{X}_{j|\setminus j} > 0)} \\ &= \exp(\eta_{j|\setminus j}) \left[1 + \frac{P(\widetilde{X}_{j|\setminus j} = 0)}{P(\widetilde{X}_{j|\setminus j} > 0)}\right] = \exp(\eta_{j|\setminus j}) + \frac{1}{1+\beta_{j|\setminus j}}, \\ \frac{\partial^2 A^+(\eta_{j|\setminus j})}{\partial \eta_{j|\setminus j}^2} &= \exp(\eta_{j|\setminus j}) \left\{1 - \frac{1}{1+\beta_{j|\setminus j}} \left[1 - \frac{1}{\exp(\eta_{j|\setminus j})} \frac{\beta_{j|\setminus j}}{1+\beta_{j|\setminus j}}\right]\right\}. \end{aligned}$$

Cuando $\eta_{j|\setminus j} \to \infty$, luego $\beta_{j|\setminus j} \to \infty$ y $\frac{\partial A^+(\eta_{j|\setminus j})}{\partial \eta_{j|\setminus j}} \to \exp(\eta_{j|\setminus j})$, o cuando $\eta_{j|\setminus j} \to -\infty$, resulta $\beta_{j|\setminus j} \to 0$ y $\frac{\partial A^+(\eta_{j|\setminus j})}{\partial \eta_{j|\setminus j}} \to \exp(\eta_{j|\setminus j}) + 1$. Además,

$$\begin{aligned} \frac{\partial A^+(\eta_{j|\setminus j}, \Theta_{jj})}{\partial(\eta_{j|\setminus j}; \Theta_{jj})} &= \begin{bmatrix} \frac{\exp(\eta_{j|\setminus j})}{1-\exp(-\exp(\eta_{j|\setminus j}))} \\ 0 \end{bmatrix} \approx \begin{bmatrix} \exp(\eta_{j|\setminus j}) \\ 0 \end{bmatrix}, \\ \frac{\partial^2 A^+(\eta_{j|\setminus j}, \Theta_{jj})}{\partial(\eta_{j|\setminus j}; \Theta_{jj})\partial(\eta_{j|\setminus j}; \Theta_{jj})^T} &= \begin{bmatrix} \frac{\exp(\eta_{j|\setminus j}+\exp(\eta_{j|\setminus j}))(\exp\exp(\eta_{j|\setminus j})-\exp(\eta_{j|\setminus j})-1)}{(\exp\exp(\eta_{j|\setminus j})-1)^2} & 0 \\ 0 & 0 \end{bmatrix} \\ &\approx \begin{bmatrix} \exp(\eta_{j|\setminus j}) & 0 \\ 0 & 0 \end{bmatrix}, \end{aligned}$$

y la aproximación es ajustada cuando $\exp(\eta_{j|\setminus j}) \gg 1$.

TPOISSON-ZIPGM A partir de (2.27),

$$A^+(\eta_{j|\setminus j}) = \log\left(\exp(\exp(\eta_{j|\setminus j})) - 1 - \sum_{i=T^*+1} \exp(i\eta_{j|\setminus j})/i!\right).$$

Then,

$$\begin{aligned} \frac{\partial A^+(\eta_{j|\setminus j})}{\partial \eta_{j|\setminus j}} &= \frac{\exp(\eta_{j|\setminus j})\exp(\exp(\eta_{j|\setminus j})) - \sum_{i=T^*+1}\exp(i\eta_{j|\setminus j})/(i-1)!}{\exp(\exp(\eta_{j|\setminus j})) - 1 - \sum_{i=T^*+1}\exp(i\eta_{j|\setminus j})/i!} \\ &= \left[1 + \beta_{j|\setminus j}\right]^{-1} + \left[\lambda^{-1} + \frac{(T^*)^{-1}}{1+\alpha_{j|\setminus j}}\right]^{-1}, \end{aligned}$$

con

$$\beta_{j|\setminus j} = \frac{P(2 \leq \widetilde{X}_{j|\setminus j} \leq T^*)}{P(\widetilde{X}_{j|\setminus j} = 1)}, \qquad \alpha_{j|\setminus j} = \frac{P(1 \leq \widetilde{X}_{j|\setminus j} \leq T^* - 2)}{P(\widetilde{X}_{j|\setminus j} = T^* - 1)},$$

donde definimos la variable auxiliar $\widetilde{X}_{j|\setminus j} \sim$Poisson$(\lambda_{j|\setminus j})$ con media $\lambda_{j|\setminus j} = \exp(\eta_{j|\setminus j})$. Además,

$$\begin{aligned}
\frac{1}{\lambda_{j|\setminus j}} \frac{\partial^2 A^+(\eta_{j|\setminus j})}{\partial \eta_{j|\setminus j}^2} &= -\left(1 + \beta_{j|\setminus j}\right)^{-2} \frac{d\beta_{j|\setminus j}}{d\lambda_{j|\setminus j}} - \left[\lambda_{j|\setminus j}^{-1} + \frac{(T^*)^{-1}}{1 + \alpha_{j|\setminus j}}\right]^{-2} \Bigg(\\
&\qquad - \lambda_{j|\setminus j}^{-2} - (T^*)^{-1}(1 + \alpha_{j|\setminus j})^{-2} \frac{d\alpha_{j|\setminus j}}{d\lambda_{j|\setminus j}} \Bigg) \\
\frac{d\beta_{j|\setminus j}}{d\lambda_{j|\setminus j}} &= \frac{P(1 \leq \widetilde{X}_{j|\setminus j} \leq T^* - 1)}{P(\widetilde{X}_{j|\setminus j} = 1)} - \\
&\qquad \frac{P(2 \leq \widetilde{X}_{j|\setminus j} \leq T^*)}{P^2(\widetilde{X}_{j|\setminus j} = 1)} P(\widetilde{X}_{j|\setminus j} = 0) \\
&= 1 - \frac{\lambda_{j|\setminus j}^{T^*-1}}{T^*!} + \left(1 - \frac{1}{\lambda_{j|\setminus j}}\right) \beta_{j|\setminus j}, \\
\frac{d\alpha_{j|\setminus j}}{d\lambda_{j|\setminus j}} &= \frac{P(0 \leq \widetilde{X}_{j|\setminus j} \leq T^* - 3)}{P(\widetilde{X}_{j|\setminus j} = T^* - 1)} - \\
&\qquad \frac{P(1 \leq \widetilde{X}_{j|\setminus j} \leq T^* - 2)}{P^2(\widetilde{X}_{j|\setminus j} = T^* - 1)} P(\widetilde{X}_{j|\setminus j} = T^* - 2) \\
&= \frac{(T^* - 1)!}{\lambda_{j|\setminus j}^{T^*-1}} - \frac{T^* - 1}{\lambda_{j|\setminus j}} + \left(1 - \frac{T^* - 1}{\lambda_{j|\setminus j}}\right) \alpha_{j|\setminus j}.
\end{aligned}$$

Cuando $\eta_{j|\setminus j} \to -\infty$, entonces $\beta_{j|\setminus j} \to 0$, $\beta'_{j|\setminus j} \to 1/2$, y $\alpha_{j|\setminus j} \to \infty$, $\alpha'_{j|\setminus j} \to -\infty$. Cuando $\eta_{j|\setminus j} \to \infty$, resulta $\beta_{j|\setminus j} \to \infty$, $\beta'_{j|\setminus j} \to \infty$ y $\alpha_{j|\setminus j} \to 0$, $\alpha'_{j|\setminus j} \to 0$.

F

ANEXO AL CAPÍTULO 7

F.1 ESTIMACIÓN DE MODELOS PGM MEDIANTE SCORES PROPIOS

Con el objetivo de evaluar el comportamiento de scores propios penalizados, consideramos el aprendizaje de pGMs mediante distintos scores propios presentados en la Sección 4.8. En todos los casos consideramos el problema de minimización de la esperanza empírica del score, penalizado por la norma ℓ_1 de las interacciones y la norma compuesta $\ell_{1,2}$ sobre cada fila de la matriz de regresión $\mathbf{\Gamma}$. La función de costo penalizada resultante (5.3) se optimizó usando el Algoritmo 4.

F.1.1 *Parámetros poblacionales*

Los datos se generaron siguiendo un modelo gráfico de segundo orden condicional lineal en $T(\boldsymbol{X})$ (Definición 2.12), con $\boldsymbol{\eta} = a\mathbf{1}_p$. Las filas no nulas de $\mathbf{\Gamma}$ fueron generadas a partir de p realizaciones Bernoulli(0.6) independientes y sus entradas no nulas fueron generadas al azar a partir de una distribución uniforme en el intervalo $(0, 1)$. La matriz de interacciones $\mathbf{\Theta}$ fue definida de manera tal de asociar variables consecutivas, esto es

$$\Theta_{jl} = \begin{cases} b & \text{si } |j - l| = 1 \\ 0 & \text{en otro caso} \end{cases}$$

En la Tabla F.1 se detallan los valores de las constantes a y b que definen los modelos poblacionales correspondientes. También se presenta el score considerado en cada caso para su estimación.

Modelo	score	a	b
Ising-pGM	ratio-matching (4.49)	0.5	0.1
Poisson-pGM	score propio de composición (4.51)	4	-0.001
Normal-pGM	score-matching (4.47)	4	0.1

Tabla F.1: Parámetros poblacionales y score considerado para cada familia de distribuciones considerada en la simulación.

F.1.2 *Resultados*

Consideramos problemas con una cantidad de predictores $p \in \{10, 20\}$ y tamaños muestrales $n \in \{20, 50, 100, 200, 500, 1000\}$. La Figura F.1 muestra distintas medidas de calidad de las estimaciones obtenidas para el caso del modelo condicional Ising-pGM. La Figura F.2

muestra los resultados correspondientes sin considerar covariables para el modelo Poisson-PGM, mientras que en la Figura F.3 se muestran los resultados correspondientes al modelo Normal-pGM.

Observamos que para los parámetros $\boldsymbol{\eta}$ y $\boldsymbol{\Gamma}$, el error de estimación converge rápidamente para tamaños muestrales que superan la cantidad de variables p en cada caso. En el caso de las interacciones, la convergencia de dicho error es más lenta a medida que aumentamos la cantidad de variables involucradas, debido a que el número de estos parámetros crece con el cuadrado de la dimensión p. Por otro lado, la estimación de las interacciones es especialmente difícil en los modelos tipo Ising cuando las interacciones son fuertes (Montanari y Pereira, 2009).

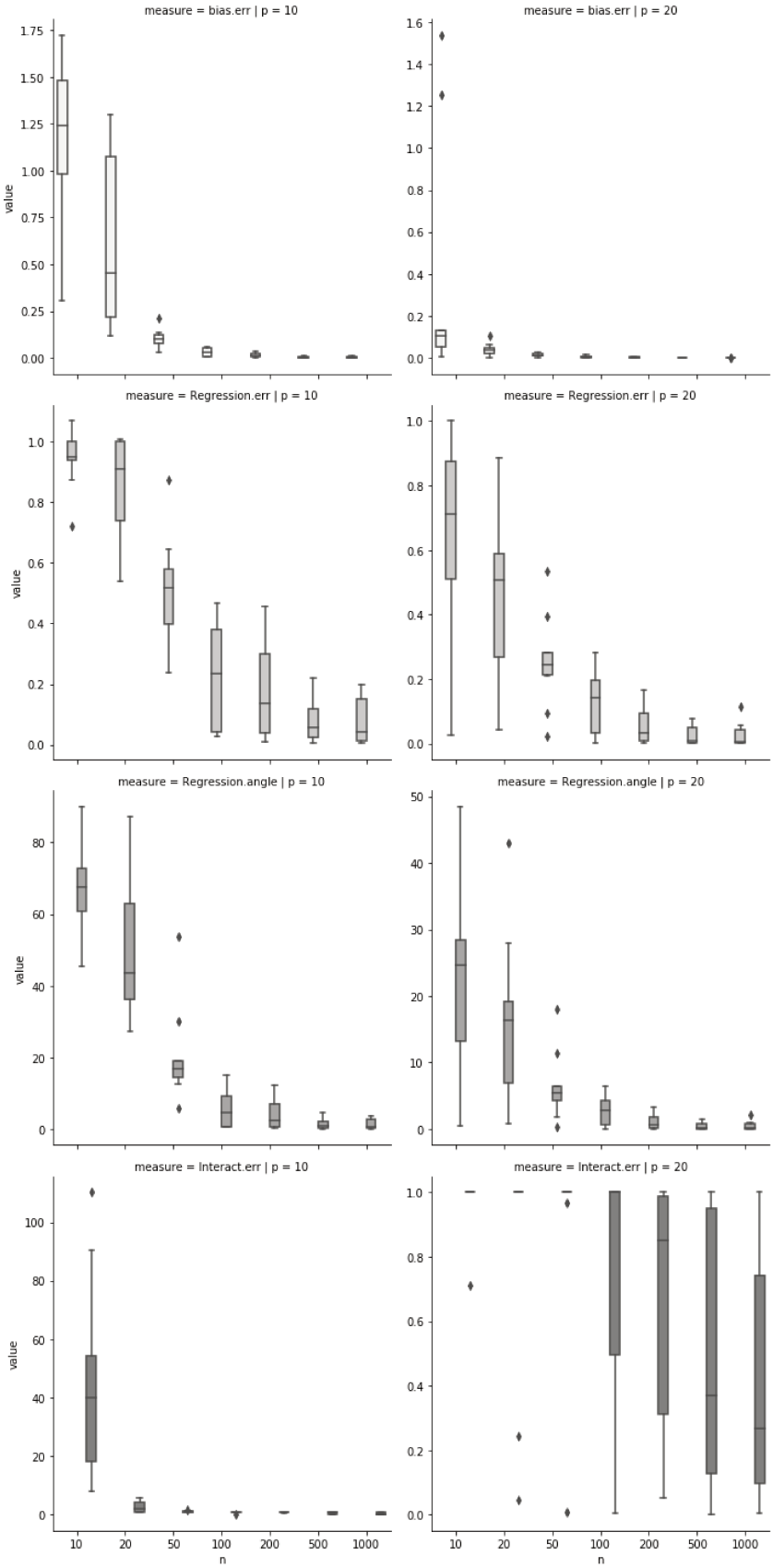


Figura F.1: Medidas de performance en estimación del modelo Ising-pGM mediante la función de costo dada por el score de ratio matching penalizada. Cada columna corresponde a $p \in \{10, 20\}$ variables, mientras que cada fila corresponde al error relativo de estimación del parámetro $\boldsymbol{\eta}$, $\boldsymbol{\Gamma}$, el ángulo entre los subespacios generados por las columnas de $\boldsymbol{\Gamma}$ y $\hat{\boldsymbol{\Gamma}}$ y el error relativo de estimación de $\boldsymbol{\Theta}$ resp.

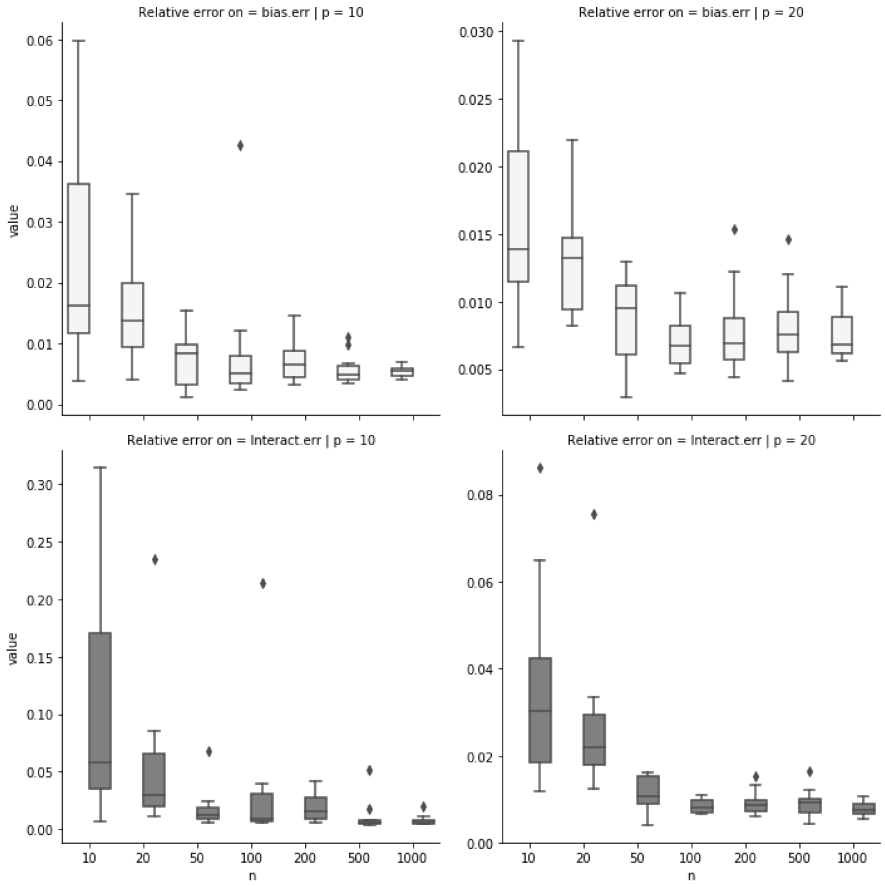


Figura F.2: Medidas de performance en estimación del modelo Poisson-pGM mediante la función de costo dada por el score de ratio matching penalizado.

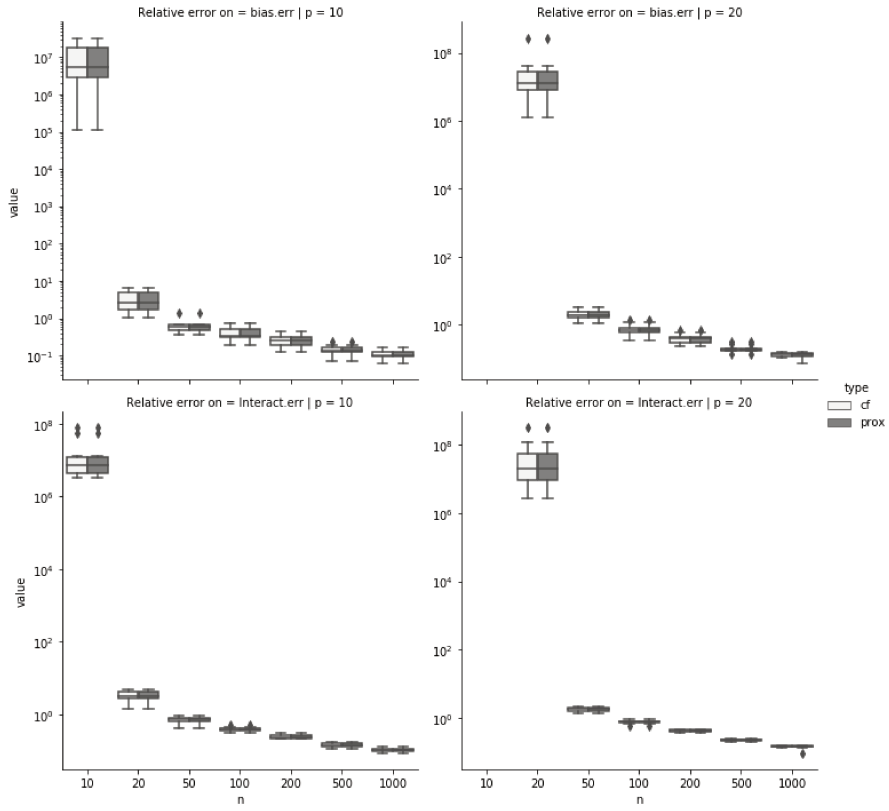


Figura F.3: Medidas de performance en estimación del modelo Normal-pGM mediante la función de costo dada por el score de ratio matching penalizado.

F.2 PROPORCIÓN DE CEROS EN MODELOS MAL ESPECIFICADOS

F.2.1 *Resultados adicionales*

La Figura F.4 muestra los resultados correspondientes a la simulación detallada en la Sección 7.1 cuando los datos son generados a partir del modelo TPoisson-zipGM y el criterio de selección es AUC, mientras que en la Figura F.5 se muestran los resultados al emplear el criterio de selección "oráculo". Se concluye de manera similar a al caso Poisson-zipGM presentado en la Figura 7.4.

En las Figuras F.6 y F.7 se muestran los resultados correspondientes al criterio de selección "oráculo", cuando los datos son generados a partir de un modelo Normal-zipGM o Poisson-zipGM respectivamente. Estos resultados señalan que la penalización es capaz de recuperar perfectamente las variables asociadas con la respuesta de manera robusta, incluso en casos de mala especificación del modelo.

F.2.2 *Influencia de las direcciones del span en modelos pGM*

La siguiente proposición tiene por objeto el análisis de pérdida de señal en modelos pGM:

Proposición F.1. *Si $\eta_0 \geq \Gamma > 0$, para $Y \sim$ Bernoulli(p) existe un Ψ tal que para $X \mid Y \sim$ escalar que se distribuye de acuerdo con un modelo zipGM con parámetros $(\eta_0, \Gamma, \xi_0, \Psi, \Theta)$, resulta $E[X \mid Y = 0] = E[X \mid Y = 1]$. Cuando $\eta_0 = 0$ (datos condicionalmente centrados), esta condición no se satisface.*

Demostración. Sin pérdida de generalidad, consideremos $Y \sim$ Bernoulli(0.5). Tomando $f_y = 2y - 1$ en (3.10a) y (3.10b). Usando que $\eta_{y=0} = \eta_0 - \Gamma$, $\eta_{y=1} = \eta_0 + \Gamma$, $\xi_{y=0} = \xi_0 - \Psi$ y $\xi_{y=1} = \xi_0 + \Psi$. Como $\eta_0 \geq \Gamma$, llamando $B = \eta_{y=1} - \eta_{y=0} = 2\Gamma$, $A = \eta_{y=0} \exp\{-\xi_{y=0} - A^+(\eta_{y=1}, \Theta)\}$ y $C = \eta_{y=1} \exp\{-\xi_{y=0} - A^+(\eta_{y=0}, \Theta)\}$, obtenemos

$$\begin{aligned}
\frac{B^2 + 4AC}{4} &= \Gamma^2 + (\eta_0 - \Gamma)(\eta_0 + \Gamma) \exp\Big\{ -2\xi_{y=0} - A^+(\eta_{y=0}, \Theta) - \\
&\qquad A^+(\eta_{y=1}, \Theta) \Big\} \\
&= \Gamma^2 + (\eta_0^2 - \Gamma^2) \exp\Big\{ -2\xi_{y=0} - A^+(\eta_{y=0}, \Theta) - \\
&\qquad A^+(\eta_{y=1}, \Theta) \Big\} \\
&\geq \Gamma^2 > 0,
\end{aligned}$$

Además, como $A, B > 0$, podemos definir $\Psi = -\log\left\{\frac{B + \sqrt{B^2 + 4AC}}{2A}\right\}$. A continuación mostramos que para ese Ψ (si $\eta_0 \neq 0$)

$$E[X|Y = 0] = E[X|Y = 1] \tag{F.1}$$

A partir de la definición de Ψ, obtenemos

$$\begin{aligned}
A \exp\{-2\Psi\} - B \exp\{-\Psi\} - C = 0, \\
A \exp\{-\Psi\} - B - C \exp\{\Psi\} = 0,
\end{aligned} \tag{F.2}$$

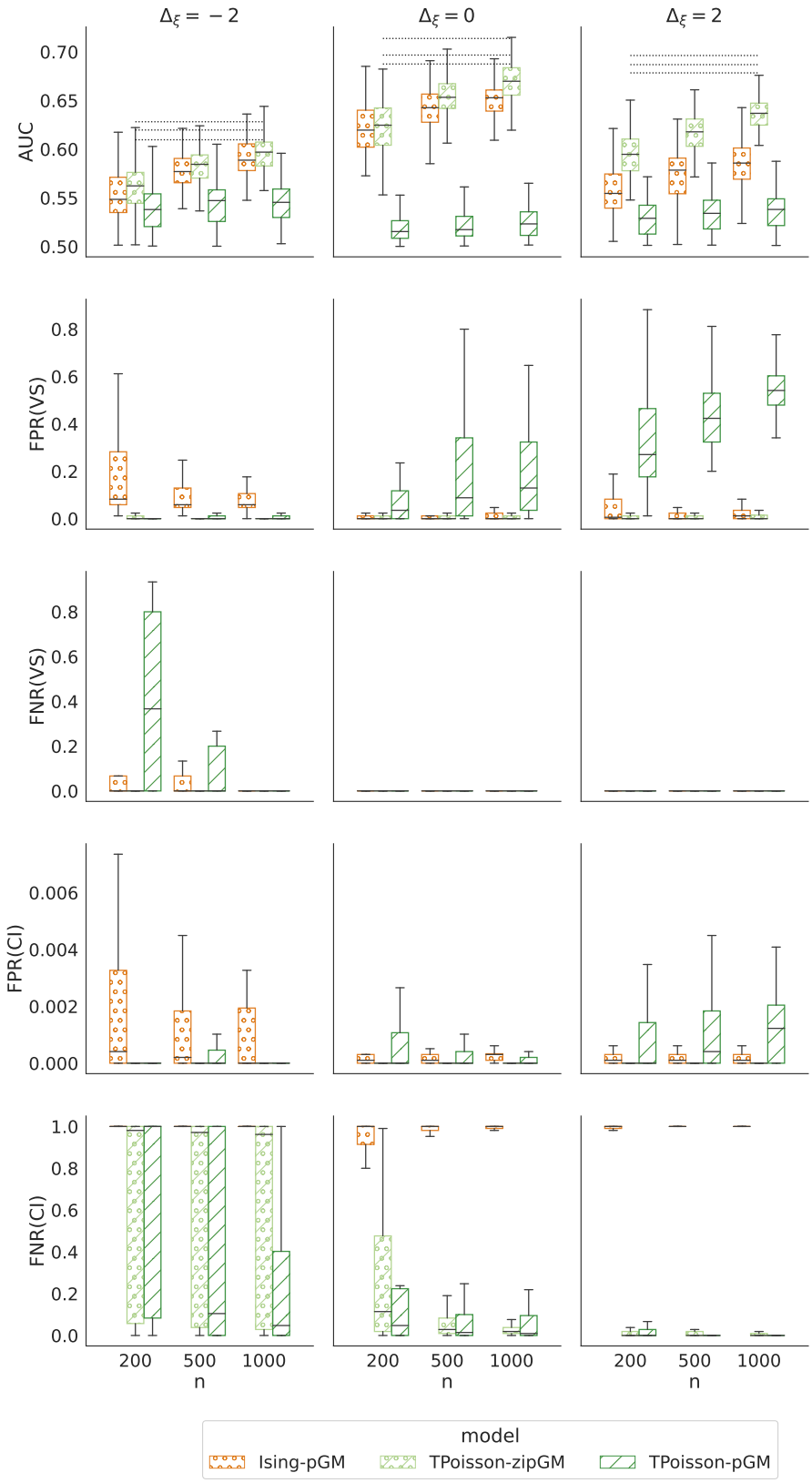


Figura F.4: Resultados obtenidos para el ajuste de los modelos Ising, TPoisson-zipGM, y TPoisson-pGM a datos generados a partir del TPoisson-zipGM definido en Sección 7.1.1 cuando el criterio de selección es la AUC.

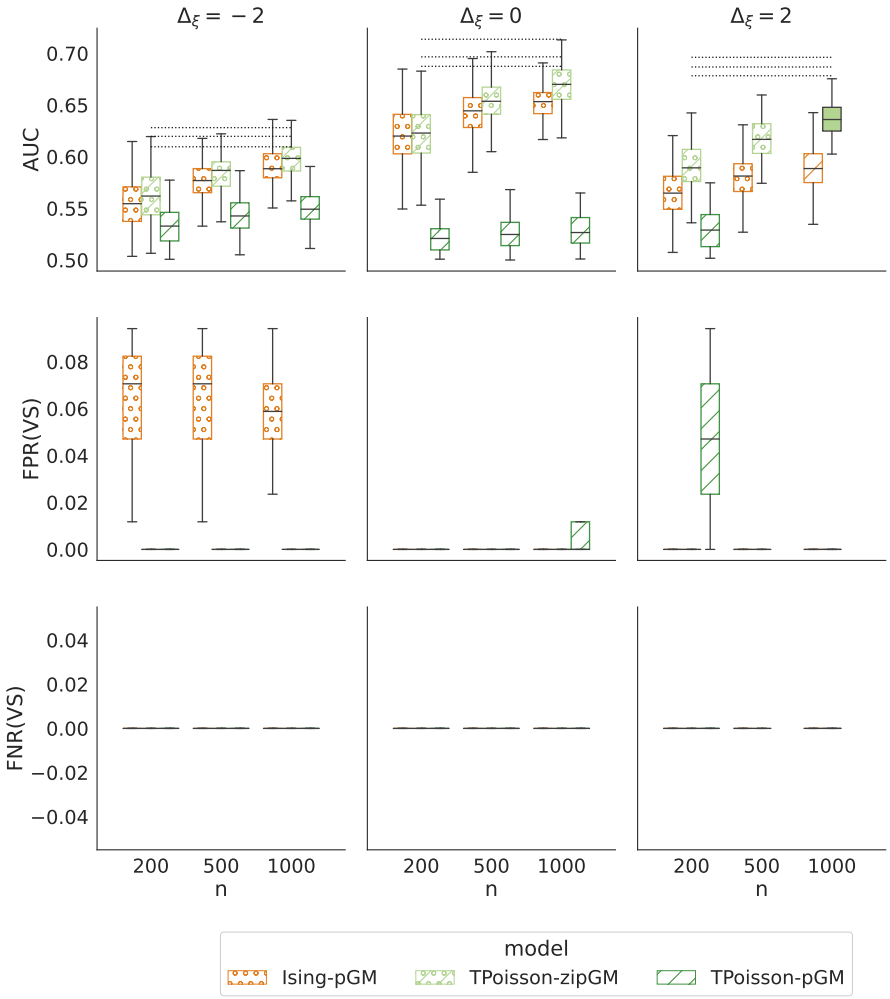


Figura F.5: Resultados obtenidos para el ajuste de los modelos Ising, TPoisson-zipGM, y TPoisson-pGM a datos generados a partir del TPoisson-zipGM definido en Sección 7.1.1 cuando el criterio de selección es el "oráculo".

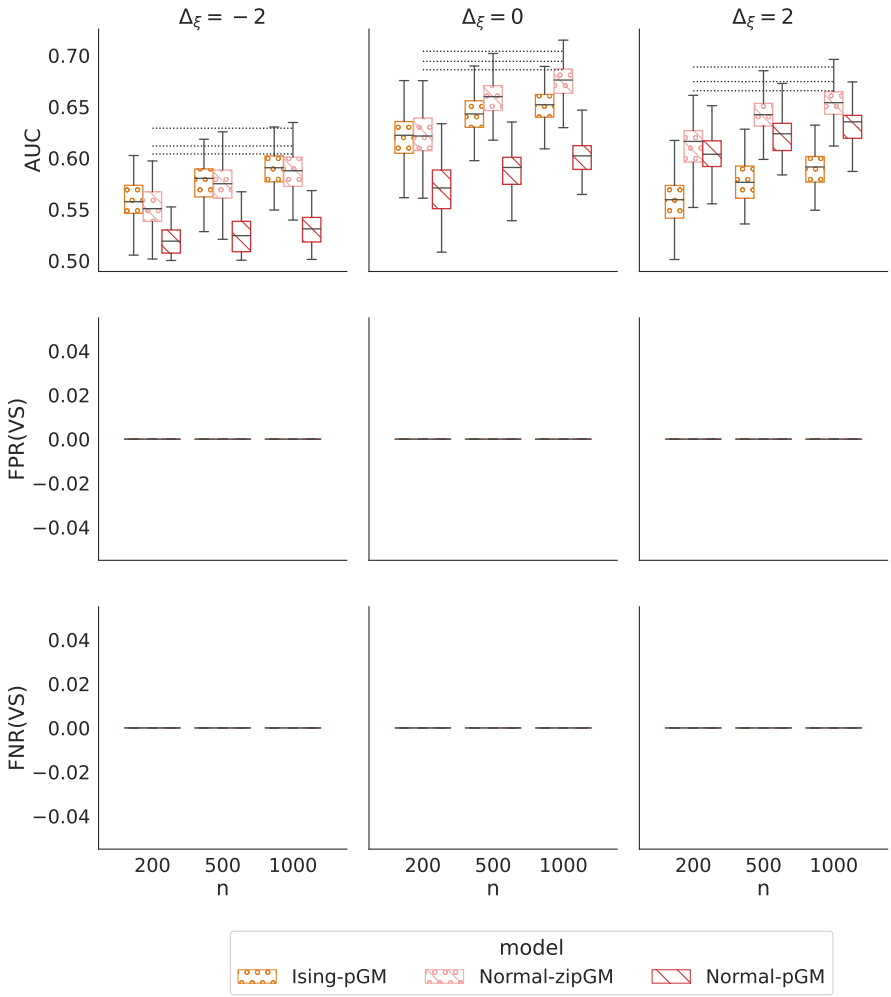


Figura F.6: Resultados obtenidos para el ajuste de los modelos Ising, Normal-zipGM, y Normal-pGM a datos generados a partir del Normal-zipGM definido en Sección 7.1.1 cuando el criterio de selección es el "oráculo".

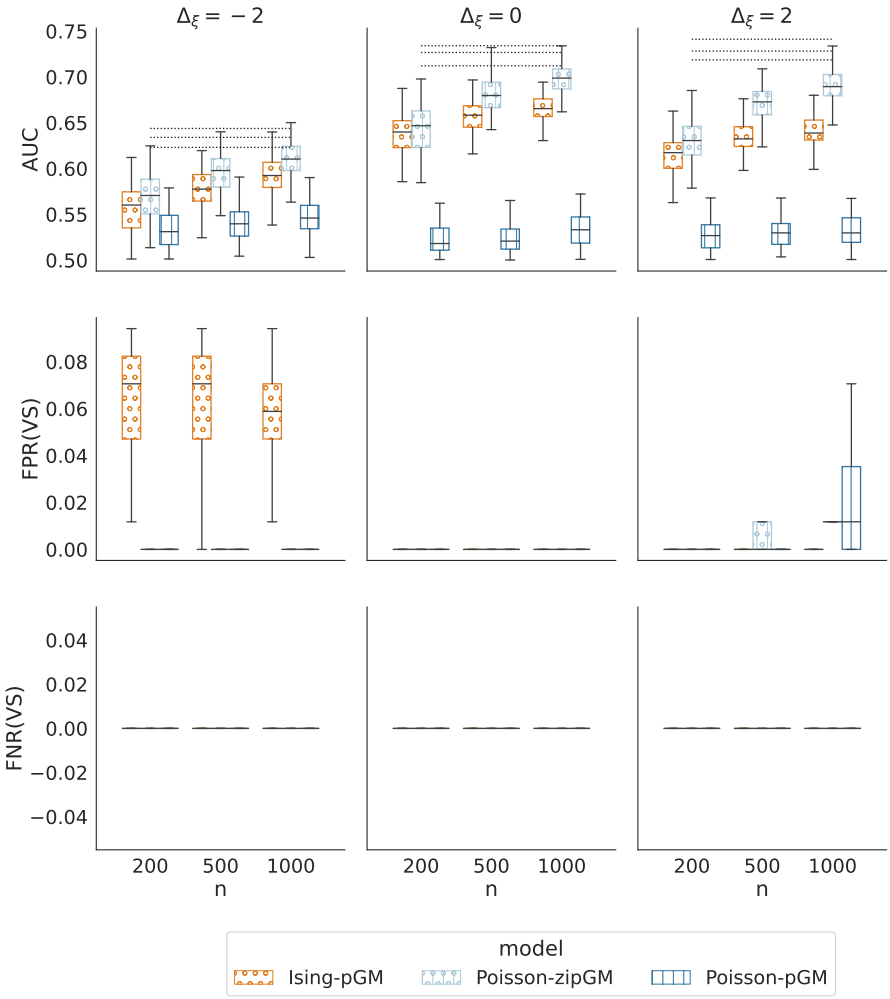


Figura F.7: Resultados obtenidos para el ajuste de los modelos Ising, Poisson-zipGM, y Poisson-pGM a datos generados a partir del Poisson-zipGM definido en Sección 7.1.1 cuando el criterio de selección es el "oráculo".

multiplicando por $\exp(\Psi)$ y usando las definiciones de A, B y C,

$$\eta_{y=0}(1+\exp\{-\xi_0-\Psi-A^+(\eta_{y=1},\Theta)\}) = \eta_{y=1}(1+ \exp\{-\xi_0+\Psi-A^+(\eta_{y=0},\Theta)\}),$$

agrupando

$$\frac{\eta_{y=0}}{1+\exp\{-\xi_{y=0}-A^+(\eta_{y=0},\Theta)\}} = \frac{\eta_{y=1}}{1+\exp\{-\xi_{y=1}-A^+(\eta_{y=1},\Theta)\}} \tag{F.3}$$

Usando (2.22) y (2.23) para el modelo Normal-zipGM y (2.25) para el modelo Poisson-zipGM, el denominador resulta $1/P[\nu(X)=1|Y=y]$, mientras que el numerador es $E[X|\nu(X)=1,Y=y]$ para el modelo Poisson-zipGM y es proporcional a $E[X|\nu=1,Y=y]$ para le modelo Normal-zipGM. Luego, (F.3) resulta equivalente a

$$E[X|\nu(X)=1,Y=0]P[\nu(X)=1|Y=0] = E[X|\nu(X)=1,Y=1] P[\nu(X)=1|Y=1]$$

Ya que $E[X \mid \nu(X)=0,Y=0] = E[X \mid \nu(X)=0,Y=1] = 0$,

$$\begin{aligned} E[X \mid Y=0] &= E[X \mid \nu(X)=0,Y=0]P[\nu(X)=0 \mid Y=0]+ \\ &\quad E[X \mid \nu(X)=1,Y=0]P[\nu(X)=1 \mid Y=0] \\ &= 0 + E[X \mid \nu(X)=1,Y=0]P[\nu(X)=1 \mid Y=0] \\ &= 0 + E[X \mid \nu(X)=1,Y=1]P[\nu(X)=1 \mid Y=1] \\ &= E[X \mid \nu(X)=0,Y=1]P[\nu(X)=0 \mid Y=1]+ \\ &\quad E[X|\nu(X)=1,Y=1]P[\nu(X)=1|Y=1] \\ &= E[X \mid Y=1]. \end{aligned}$$

A continuación mostramos que $\eta_0 \neq 0$ es consecuencia de la condición

$$E[X \mid Y=0] = E[X \mid Y=1]. \tag{F.4}$$

Siguiendo los pasos previos, tenemos que $E[X \mid Y=0] = E[X \mid Y=1]$ resulta equivalente a (F.2), y resolviendo para Ψ, obtenemos $\Psi \in \{\Psi_+, \Psi_-\}$ con

$$\Psi_\pm = -\log\left\{\frac{B \pm \sqrt{B^2+4AC}}{2A}\right\}. \tag{F.5}$$

Asumiendo $\eta_0 = 0$ y que $\Psi \in \mathbb{R}$ en (F.5) existe. Usando las definiciones para A,B and C,

$$\frac{B^2+4AC}{4} = \Gamma^2 - \Gamma^2\exp\{-2\xi_{y=0} - A^+(-\Gamma,\theta) - A^+(\Gamma,\theta)\}, \tag{F.6}$$

y considerando los casos:

(1) $\xi_{y=0} < -\frac{A^+(-\Gamma,\theta)+A^+(\Gamma,\theta)}{2}$: el discriminante $B^2+4AC<0$ in (F.6) and thus $\Psi \notin \mathbb{R}$. Una contradicción.

(2) $\xi_{y=0} \geq -\frac{A^+(-\Gamma,\theta)+A^+(\Gamma,\theta)}{2}$. Luego:

$$\begin{aligned}
&\frac{B-\sqrt{B^2+4AC}}{2A} = -\exp\{\xi_{y=0}+A^+(\Gamma,\Theta)\}\left(1-\right.\\
&\qquad\left.\sqrt{1-\exp\{-2\xi_{y=0}-A^+(-\Gamma,\theta)-A^+(\Gamma,\theta)\}}\right)\\
&\leq 0\\
&\frac{B+\sqrt{B^2+4AC}}{2A} = -\exp\{\xi_{y=0}+A^+(\Gamma,\Theta)\}\left(1+\right.\\
&\qquad\left.\sqrt{1-\exp\{-2\xi_{y=0}-A^+(-\Gamma,\theta)-A^+(\Gamma,\theta)\}}\right)\\
&\leq -\exp\{\xi_{y=0}+A^+(\Gamma,\Theta)\}\\
&\leq 0.
\end{aligned}$$

por lo tanto no existe $\Psi \in \mathbb{R}$ (contradicción), ya que haría negativo al argumento del logaritmo que define $\Psi_{\pm}$.

Notar que la cota superior en el último caso se obtiene de observar que el termino dentro de la raíz cuadrada está siempre entre 0 y 1. □

F.3 EFECTO DE LAS INTERACCIONES EN MODELOS ZIPGM CON PENALIZACIÓN JERÁRQUICA

F.3.1 *Resultados adicionales*

Las figuras F.8 y F.9 contienen los resultados de la simulación presentada en la Sección 7.2 para el criterio de selección "oráculo" cuando los datos son generados a partir del modelo Normal-zipGM o Poisson-zipGM definidos en la Sección 7.2.1 respectivamente. Observamos que la penalización jerárquica permite recuperar las variables asociadas a la respuesta Y incluso cuando la estructura de interacciones de $\boldsymbol{X} \mid Y$ no se ajusta a la estructura inducida por la penalidad (interacciones únicamente entre variables asociadas a la respuesta) También observamos que el modelo TPoisson-zipGM obtiene un especial desempeño en selección de variables, con una FNR comparable al modelo correctamente especificado.

Por otro lado, las Figuras F.10 y F.12 muestran los resultados correspondientes al criterio de selección AUC cuando los datos son generados a partir del los modelos Normal-zipGM o Poisson-zipGM respectivamente y donde se asume la existencia de interacciones únicamente entre las variables asociadas a la respuesta $X_1, \cdots, X_{15}$, acordando con la estructura inducida por la penalización jerárquica y por lo tanto resulta en una configuración favorable para la selección de variables, mostrando excelentes resultados en todos los casos. En las Figuras F.11 y F.13 se muestran los resultados correspondientes cuando el criterio de selección "oráculo" es empleado. En este caso los modelos poblacionales fueron definidos como en la Sección 7.2.1 y las interacciones fueron escaladas por $c = 0.1$ o $c = 0.5$ correspondiente a los casos de interacciones débiles y fuertes respectivamente. Pero los parámetros de regresión fueron escalados por las constantes a_1 y a_2 tales que el AUC en predicción aproxima 0.65 cuando $\alpha^T T(\boldsymbol{X})$ o $\boldsymbol{\zeta} v(\boldsymbol{X})$ son usados como predictores para la respuesta Y en el caso de interacciones débiles ($c = 0.1$). La habilidad predictiva del modelo Poisson-zipGM fue menos sensible a la fuerza de las interacciones que en el caso del modelo Normal-zipGM. Además, los modelos Poisson-zipGM y TPoisson-zipGM presentaron menor FPR y FNR que el modelo Normal-zipGM. También se observó que la performance en selección de variables mejoró con interacciones más fuertes en el caso del modelo Normal-zipGM. Mientras que las interacciones fueron perfectamente recontruídas por los modelos Poisson y Tpoisson zipGMs en el caso de interacciones fuertes. El criterio de selección "oráculo" muestra que una reconstrucción casi perfecta es alcanzable en estos casos.

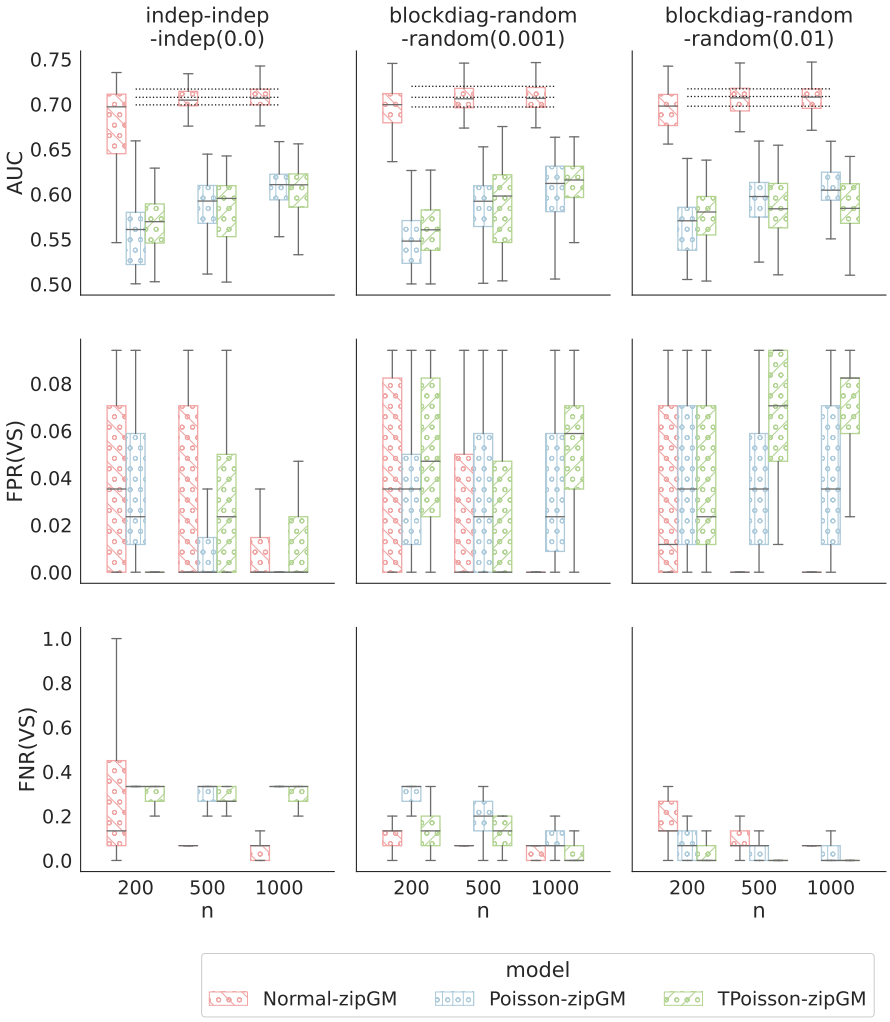


Figura F.8: Resultados obtenidos para el ajuste de los modelos zipGM considerados cuando los datos son generados a partir del modelo Normal-zipGM definido en Sección 7.2.1 cuando el criterio de selección es el "oráculo". Las columnas presentan interacciones con intensidad creciente.

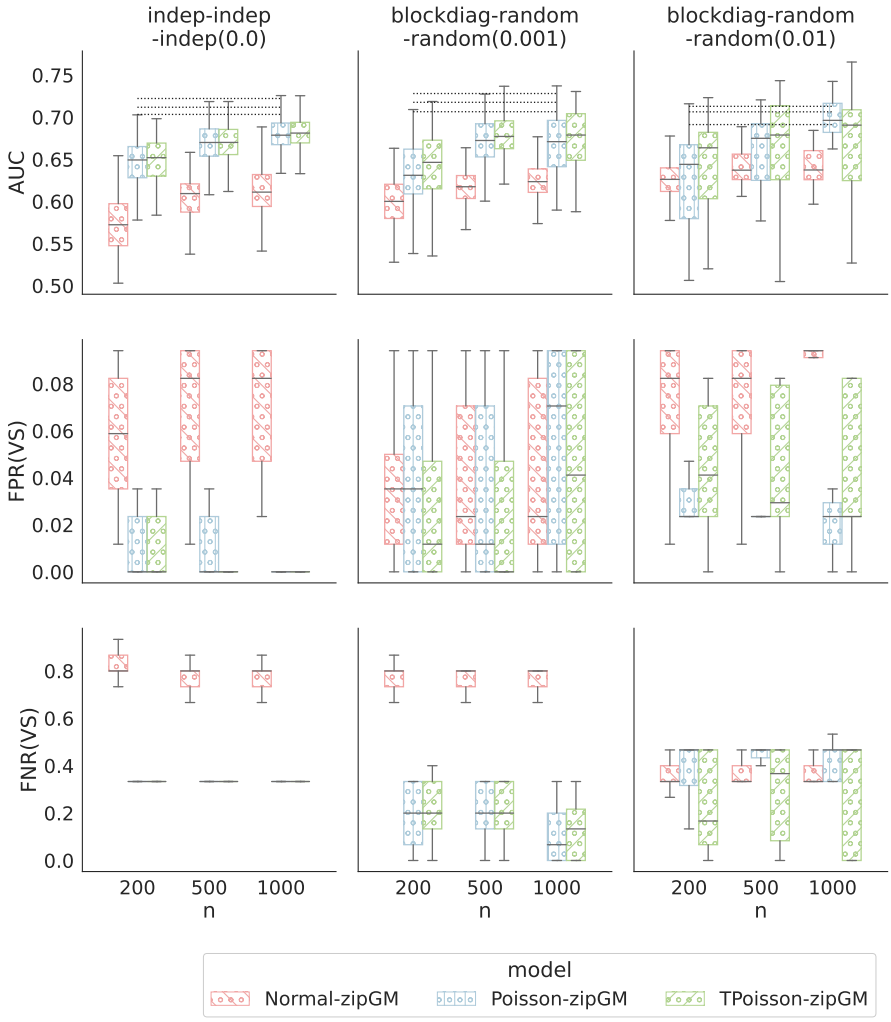


Figura F.9: Resultados obtenidos para el ajuste de los modelos zipGM considerados cuando los datos son generados a partir del modelo Poisson-zipGM definido en Sección 7.2.1 cuando el criterio de selección es el "oráculo". Las columnas presentan interacciones con intensidad creciente.

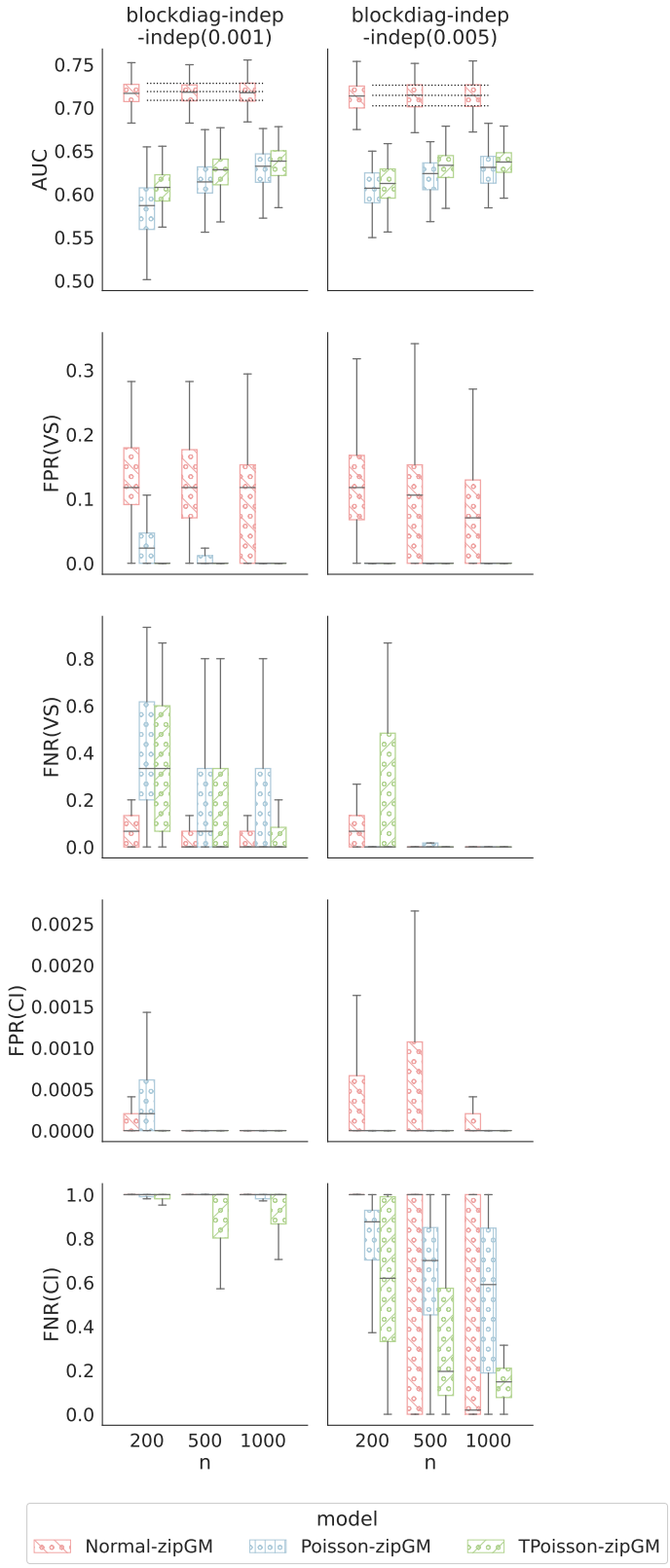


Figura F.10: Resultados obtenidos para el ajuste de los modelos zipGM considerados cuando los datos son generados a partir del modelo Normal-zipGM para interacciones entre variables asociadas a la respuesta cuando el criterio de selección es la AUC. Las columnas presentan interacciones con intensidad creciente.

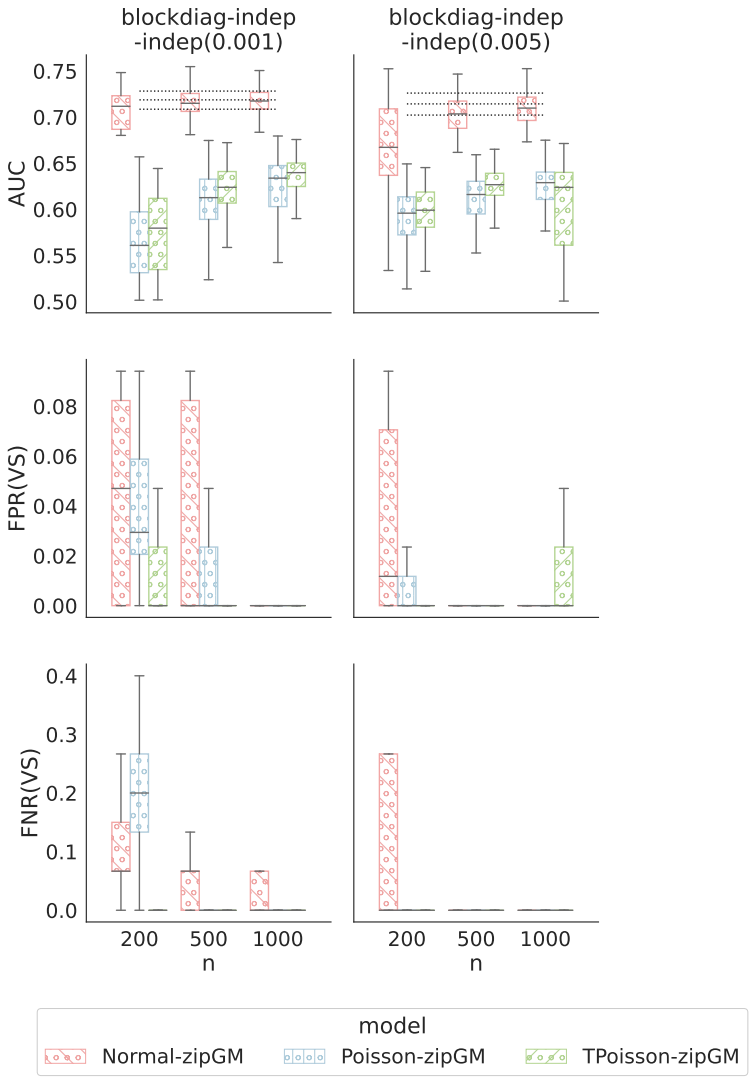


Figura F.11: Resultados obtenidos para el ajuste de los modelos zipGM considerados cuando los datos son generados a partir del modelo Normal-zipGM para interacciones entre variables asociadas a la respuesta cuando el criterio de selección es el "oráculo". Las columnas presentan interacciones con intensidad creciente.

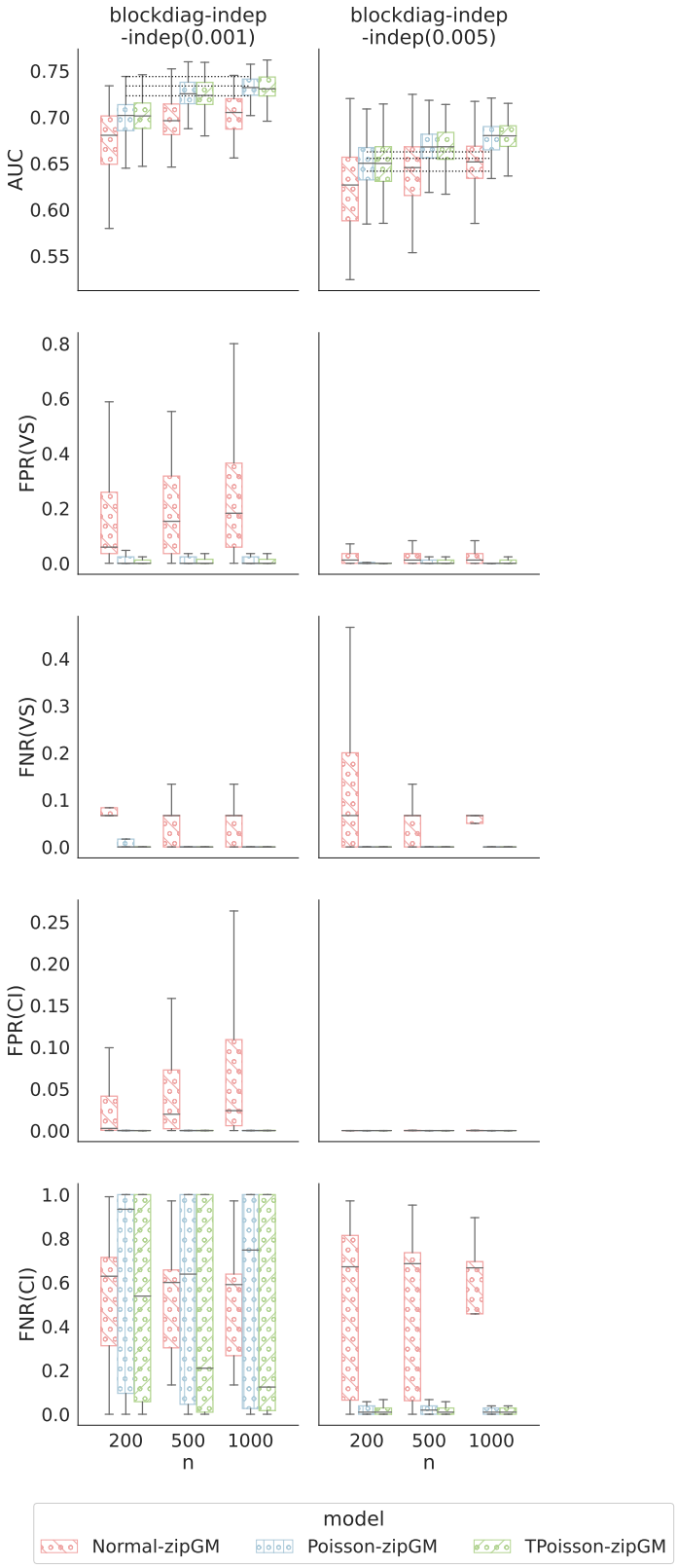


Figura F.12: Resultados obtenidos para el ajuste de los modelos zipGM considerados cuando los datos son generados a partir del modelo Poisson-zipGM para interacciones entre variables asociadas a la respuesta cuando el criterio de selección es la AUC. Las columnas presentan interacciones con intensidad creciente.

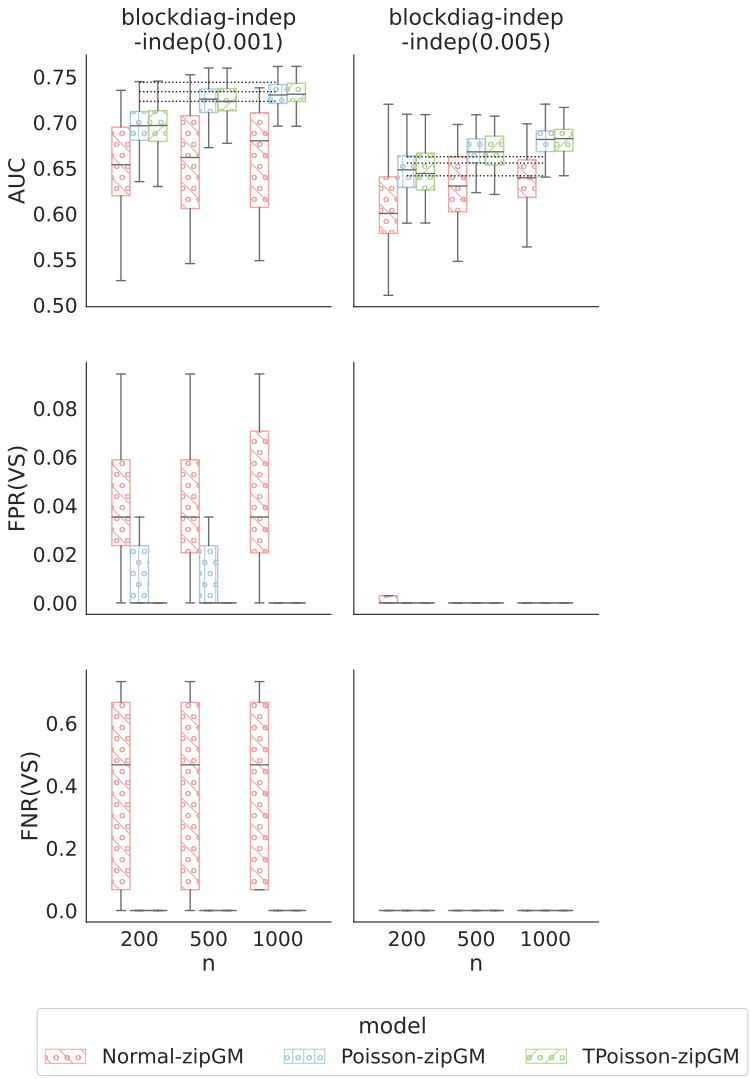


Figura F.13: Resultados obtenidos para el ajuste de los modelos zipGM considerados cuando los datos son generados a partir del modelo Normal-zipGM para interacciones entre variables asociadas a la respuesta cuando el criterio de selección es el "oráculo". Las columnas presentan interacciones con intensidad creciente.

F.4 DATOS SIMULADOS TIPO MICROBIOMA

En las Figuras F.14 y F.15 se detalla el desempeño en predicción de los modelos zipGM considerados, así como también los métodos SPLS y SPLSDA del paquete `mixOmics` cuando empleamos el criterio de selección "oráculo". Se observa que los modelos propuestos logran excelente desempeño en selección de variables, mientras que su competidor del paquete `mixOmics` mantiene una tasa muy alta de falsos negativos, por lo que pierde variables asociadas a la respuesta. En el caso de respuesta binaria se observa que el modelo Poisson-zipGM no logra una proporción de falsos negativos tan baja como los otros modelos de la familia zipGM pero se mantiene muy por debajo de su competidor.

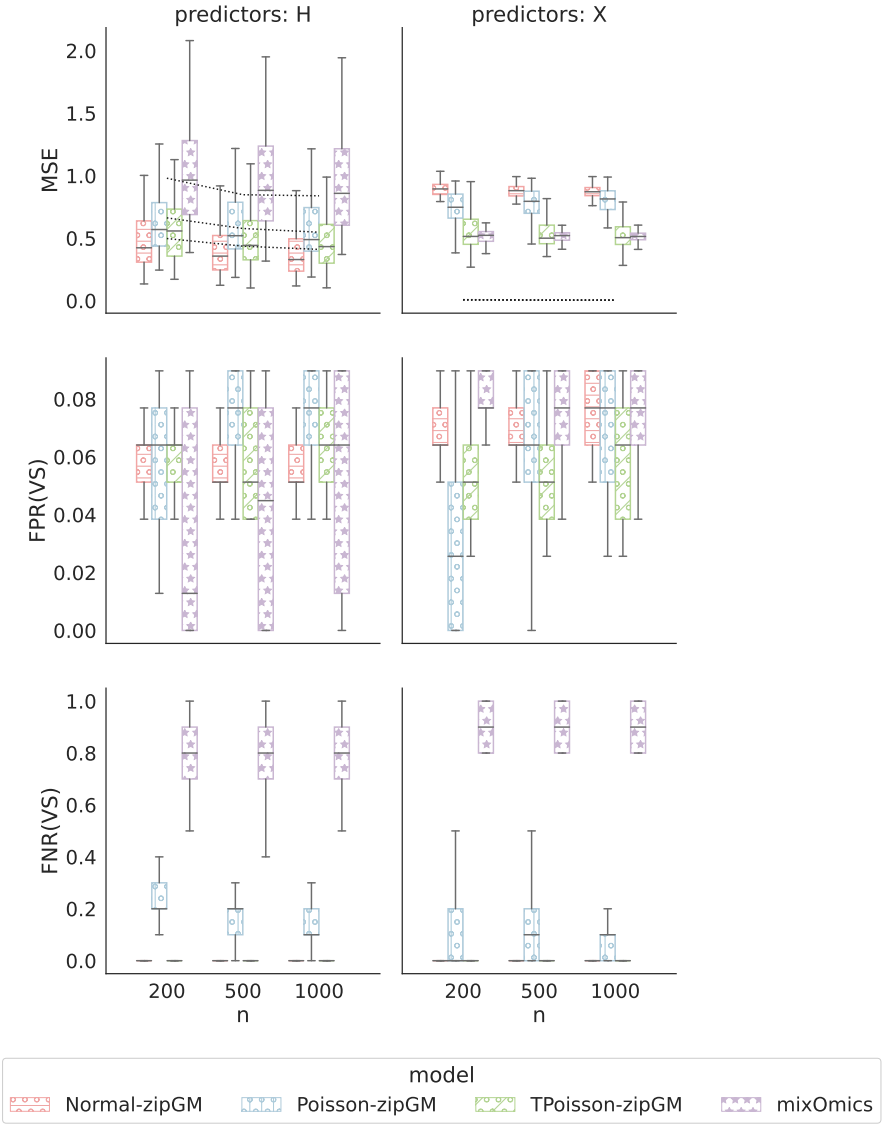


Figura F.14: Medidas de performance en predicción y selección de variables (filas) utilizando el criterio de selección "oráculo" para los modelos zipGM propuestos y el modelo SPLS del paquete `mixOmics`, cuando los datos son generados por un modelo directo $Y|\boldsymbol{\nu}(\boldsymbol{X})$ (columna 1) or $Y|\boldsymbol{X}$ (columna 2) para respuesta continua Y basados en muestras generadas por el paquete `MB-GAN` para $n \in \{200, 500, 1000\}$.

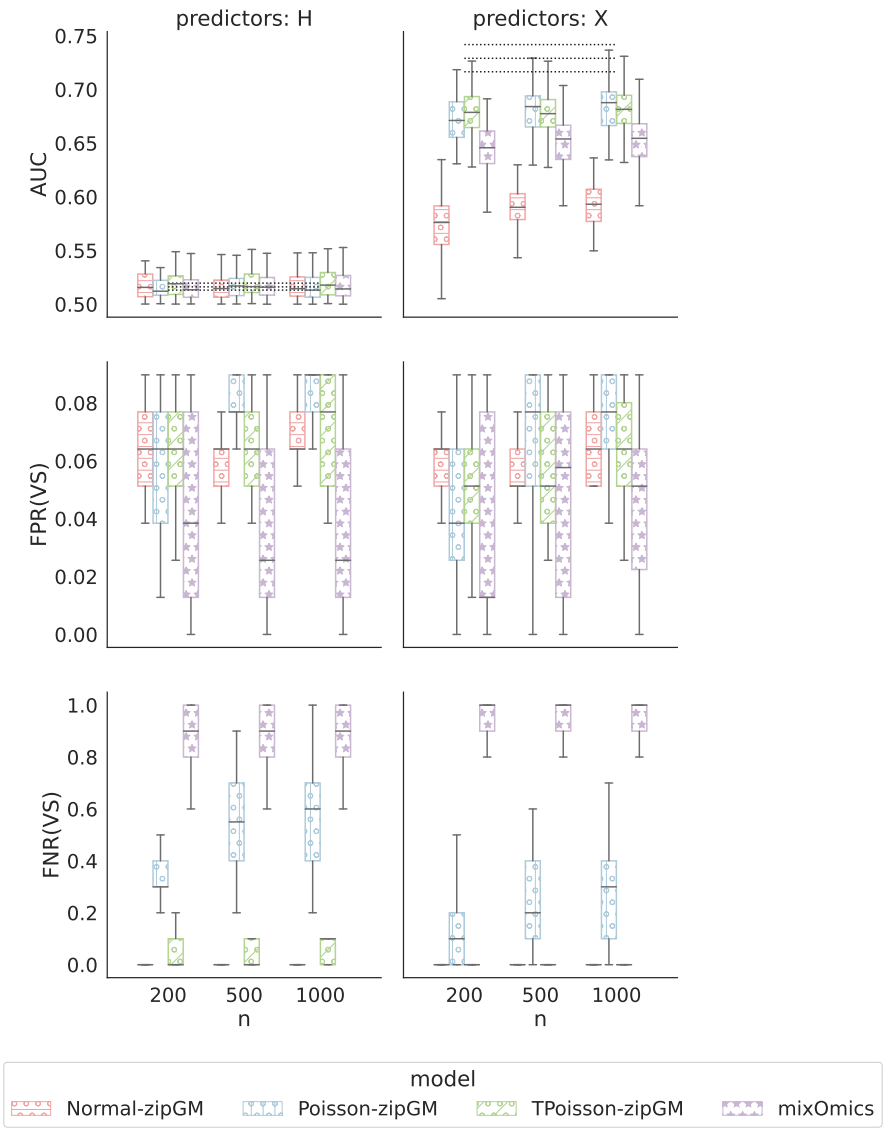


Figura F.15: Medidas de performance en predicción y selección de variables (filas) utilizando el criterio de selección "oráculo" para los modelos zipGM propuestos y el modelo SPLSDA del paquete `mixOmics`, cuando los datos son generados por un modelo directo $Y|\boldsymbol{v}(\boldsymbol{X})$ (columna 1) or $Y|\boldsymbol{X}$ (columna 2) para respuesta binaria Y basados en muestras generadas por el paquete `MB-GAN` para $n \in \{200, 500, 1000\}$.

G

ANEXO AL CAPÍTULO 8

G.1 HUMAN MICROBIOME PROJECT

En las Figuras G.1, G.2 y G.3 se muestran las reducciones obtenidas en los datos HMP por los modelos Normal-pGM, Poisson-pGM y FPoisson-pGM respectivamente en los niveles taxonómicos L2 y L6. Se observa que al aumentar el nivel taxonómico analizado, estos modelos pGM pierden capacidad predictiva respecto de la respuesta. En la Figura G.4 se comparan las reducciones obtenidas en el nivel taxonómico L6 por los modelos Poisson-pGM y zipGM, donde se puede ver que la reducción logra separar mejor las clases al considerar el modelo zipGM.

Además, las Figuras G.5 y G.7 detallan distintas medidas predictivas obtenidas sobre 10 particiones independientes de predicción mediante validación cruzada y la selección de interacciones en el nivel taxonómico L2. En las Figuras G.6 y G.8 se detallan las mismas cantidades cuando se consideró el nivel taxonómico L6.

Se observa que en general el modelo sqPoisson-pGM selecciona menos interacciones que los demás modelos, las cuales también son seleccionadas por el modelo Normal-pGM. Por otro lado, el modelo Poisson-pGM selecciona interacciones más fuertes, algunas de las cuales también son capturadas por el modelo Normal-pGM.

En cuanto a predicción, observamos que todos los modelos logran un desempeño similar, apenas dominado por el modelo Normal-pGM cuando se analiza el nivel taxonómico L2, salvo el modelo Ising-pGM que es desfavorecido por la poca variabilidad de las variables binarias en este nivel. En cambio al analizar el nivel L6, el modelo Ising mejora su desempeño logrando igualar a los demás modelos.

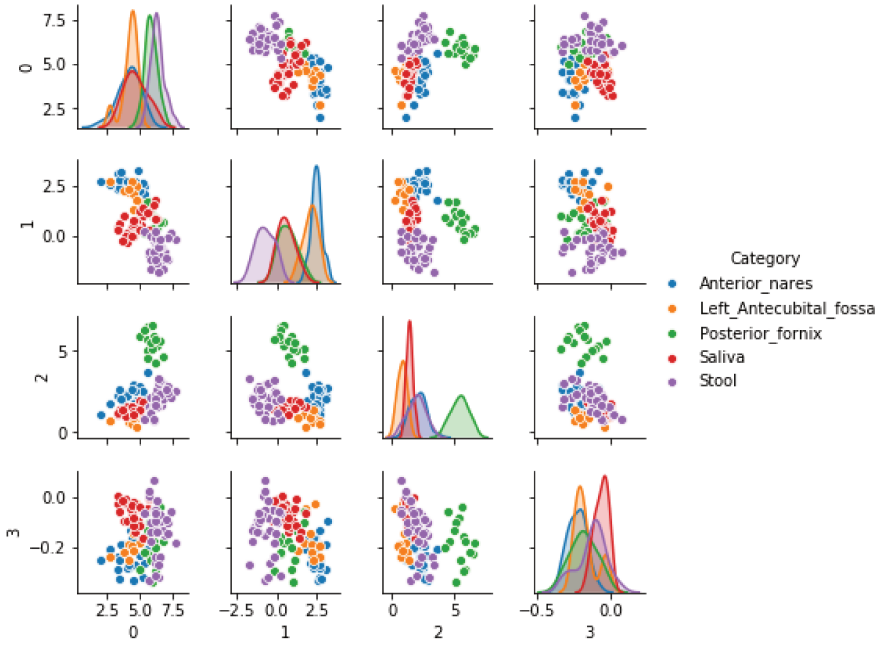
Además, se observa que la categoría `Left_Anticubital_fossa` presenta el mayor error en predicción. Esta clase es afectada por tener el menor número de observaciones (ver Figura G.9), aproximadamente la mitad de muestras que los demás grupos salvo `Posterior_fornix` que tiene un número de muestras intermedio. En L2, únicamente el modelo Normal-pGM es capaz de distinguir dicha clase, pero en L6 la mayoría de los modelos logra mejorar la performance, lo cual muestra que la información respecto de determinadas clases es sensible a la composición de los datos y que podría evaluarse un estimador con pesos de tal manera de considerar dichos desbalances.

G.2 AMERICAN GUT PROJECT

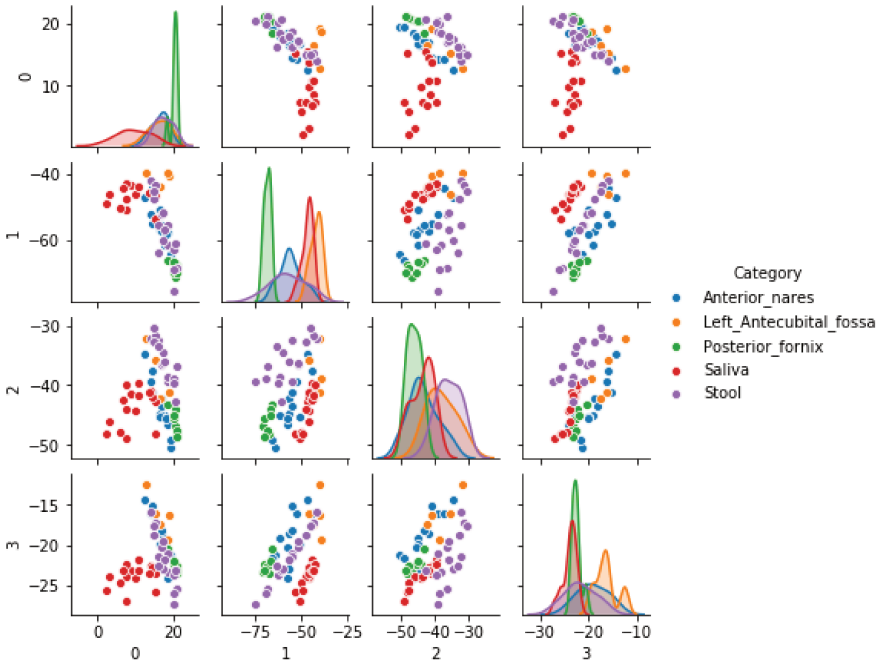
La Figura G.10 resume la estructura de los modelos resultantes de evaluar (5.6) sobre una grilla de parámetros de regularización $\lambda_{\mathcal{C}}, \lambda_{\mathcal{R}} \geq 0$, donde cada fila corresponde a $\lambda_{\mathcal{R}}$ constante, y cada

columna a $\lambda_{\mathcal{C}}$ constante, comenzando con $\lambda_{\mathcal{C}} = \lambda_{\mathcal{R}} = 0$ en la esquina inferior izquierda.

Se observa que a medida que aumenta $\lambda_{\mathcal{R}}$ (ordenadas) se inducen modelos con menos variables indicadas por los nodos azules a la vez que induce interacciones sparse al considerar únicamente las interacciones entre las variables seleccionadas. Por otro lado, al aumentar $\lambda_{\mathcal{C}}$ (abscisas), se inducen interacciones sparse, pero no selecciona variables. La combinación de ambas penalidades resulta en selección de variables y de algunas interacciones entre las variables seleccionadas.

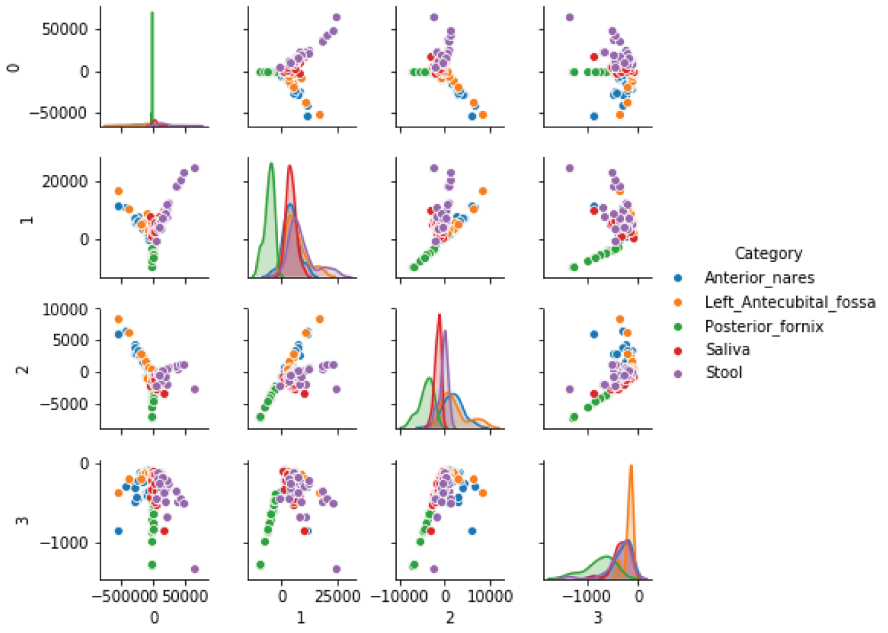


(a) L2

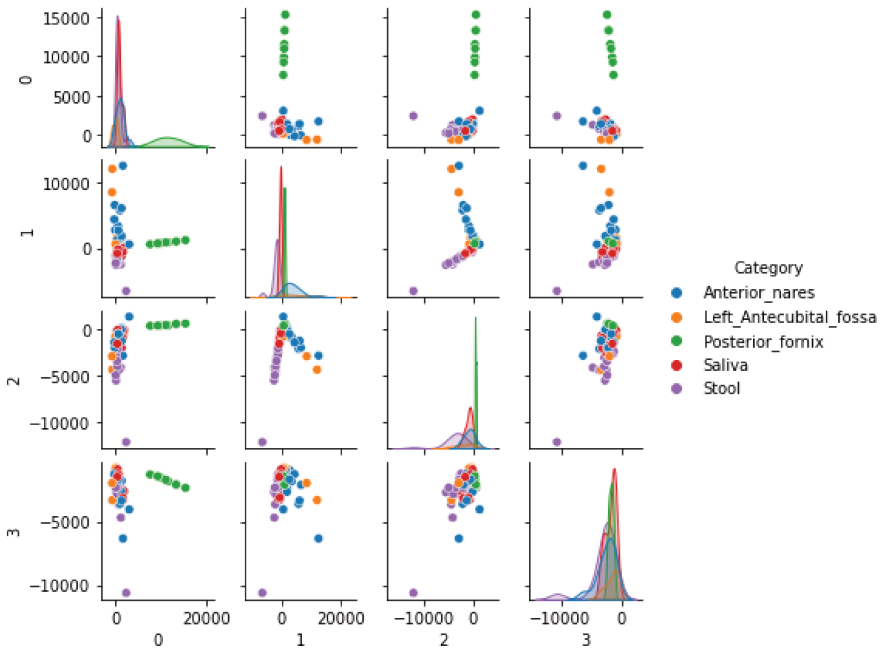


(b) L6

Figura G.1: Reducción en predicción de los datos de microbioma HMP aprendida por el modelo Normal-pGM en los niveles taxonómicos L2 y L6 en relación con el lugar de donde se extrajeron las muestras.

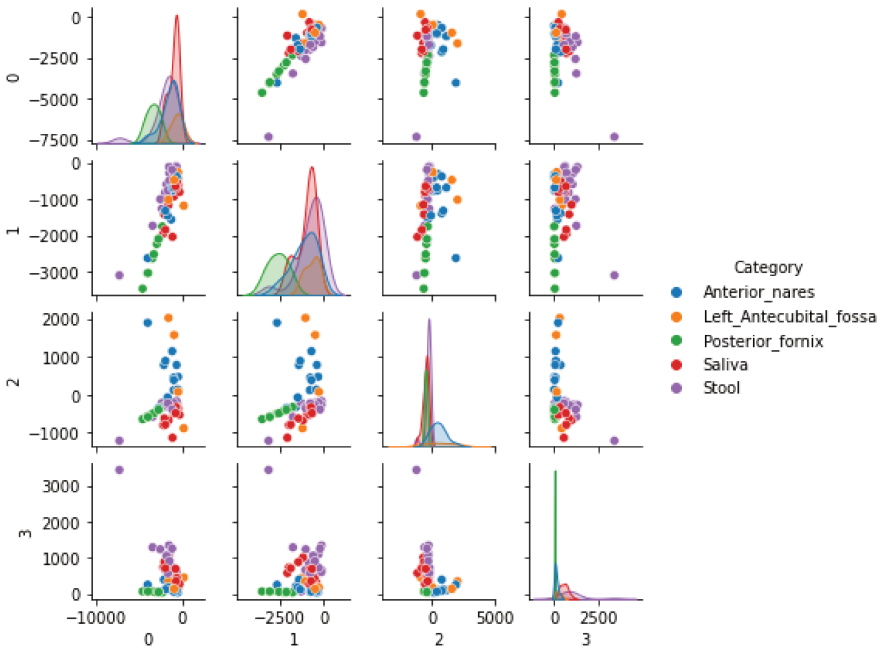


(a) L2

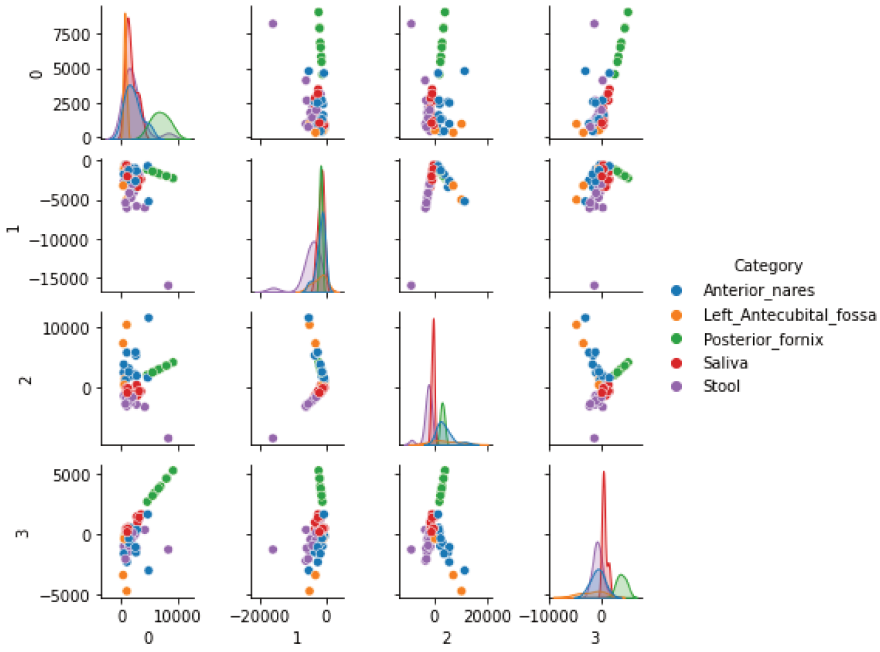


(b) L6

Figura G.2: Reducción en predicción de los datos de microbioma HMP aprendida por el modelo Poisson-pGM en los niveles taxonómicos L2 y L6 en relación con el lugar de donde se extrajeron las muestras.

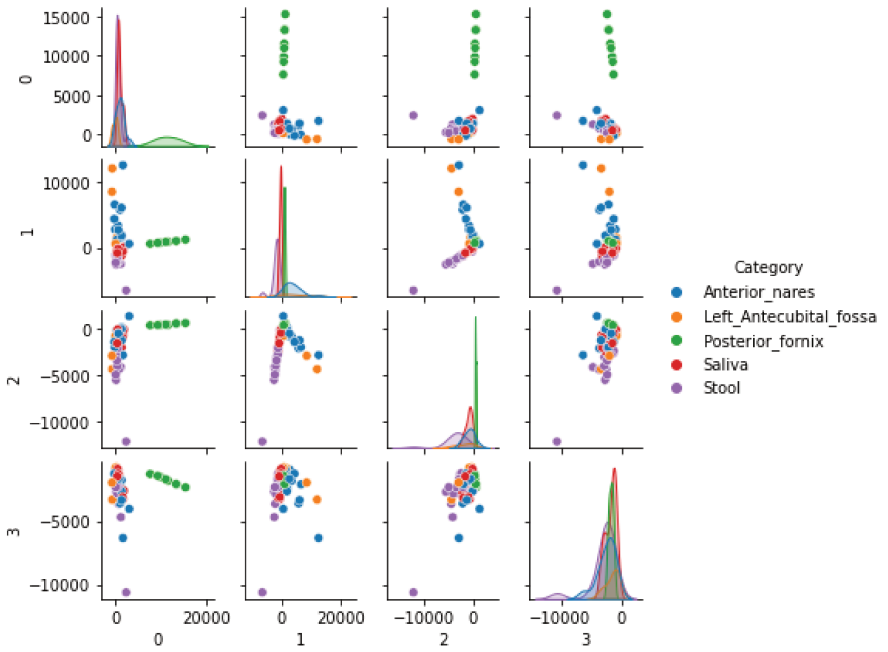


(a) L2

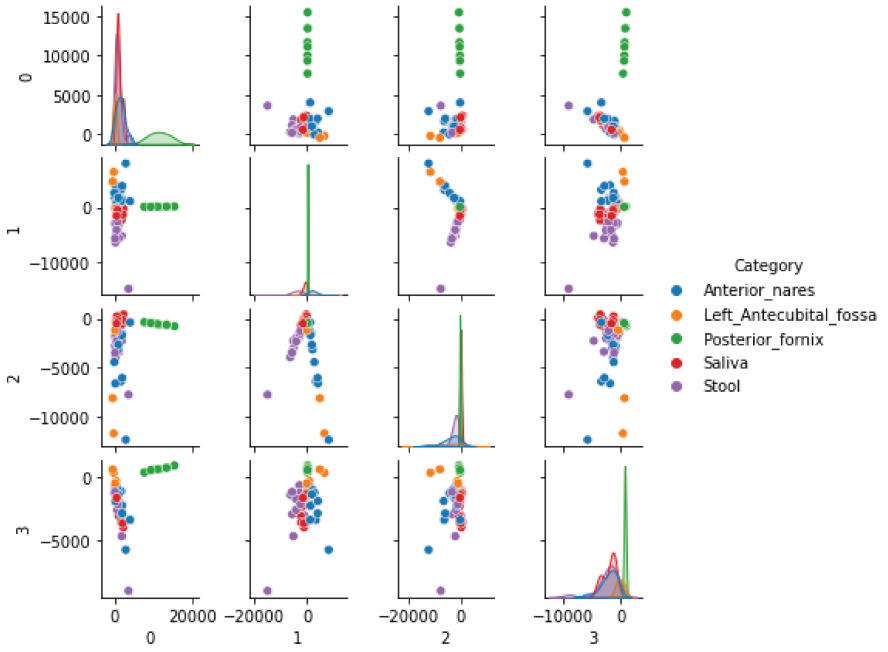


(b) L6

Figura G.3: Reducción en predicción de los datos de microbioma HMP aprendida por el modelo FPoisson-pGM en los niveles taxonómicos L2 y L6 en relación con el lugar de donde se extrajeron las muestras.



(a) Poisson-pGM



(b) Poisson-zipGM

Figura G.4: Reducción en predicción de los datos de microbioma HMP aprendida por el modelo Poisson-pGM y Poisson-zipGM. En ambos casos consideramos el nivel taxonómicos L6.

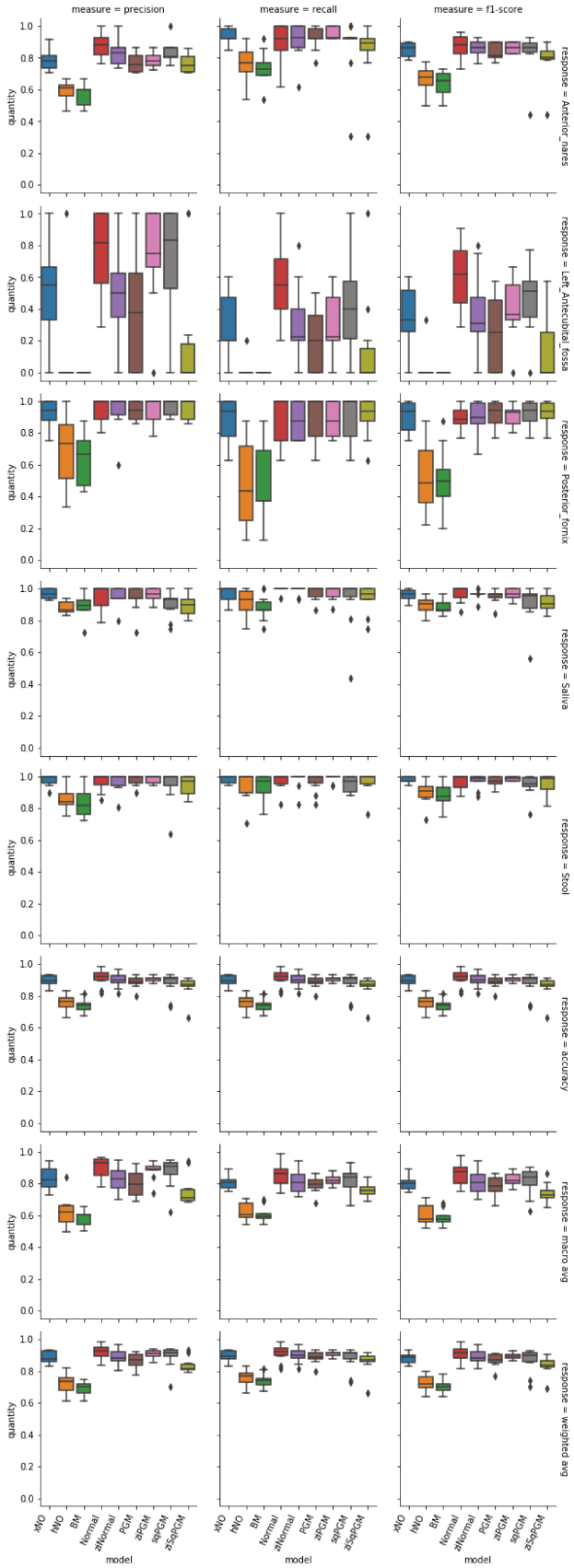


Figura G.5: Medidas de predicción (10 folds) de los datos de microbioma HMP (L2) aprendida por el los distintos pGMs considerados en relación con el lugar de donde se extrajeron las muestras.

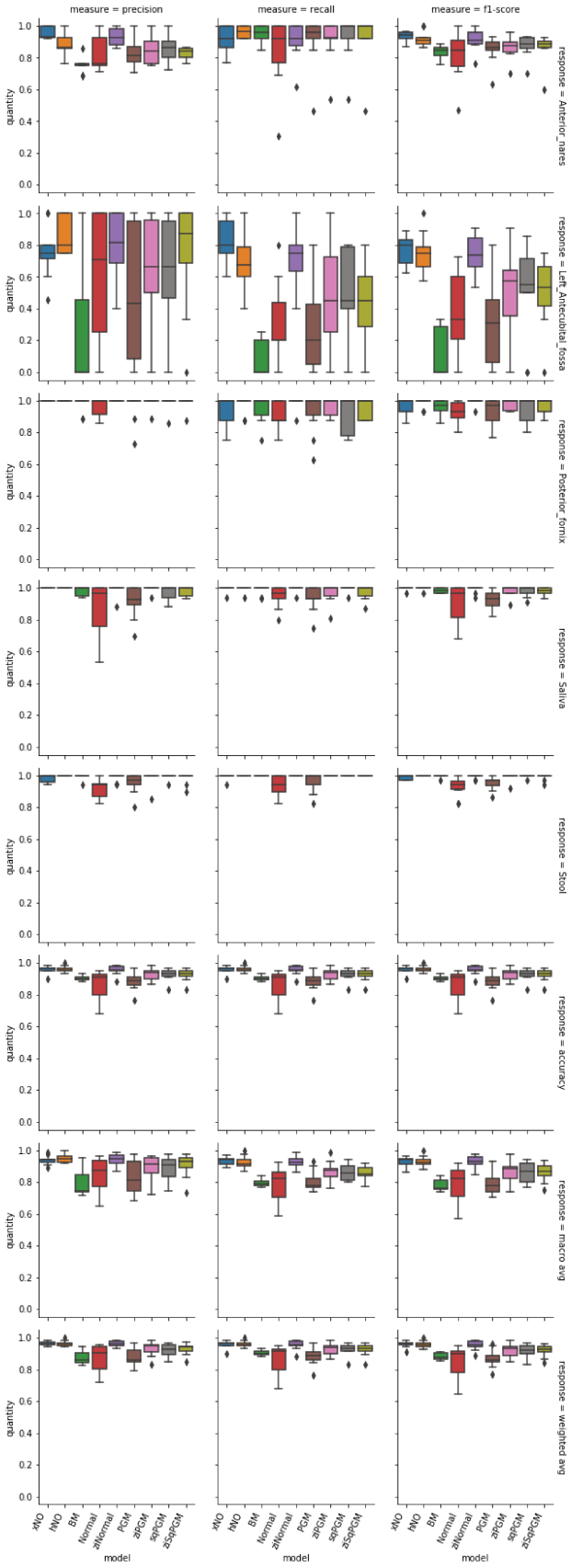


Figura G.6: Medidas de predicción (10 folds) de los datos de microbioma HMP (L6) aprendida por el los distintos pGMs considerados en relación con el lugar de donde se extrajeron las muestras.

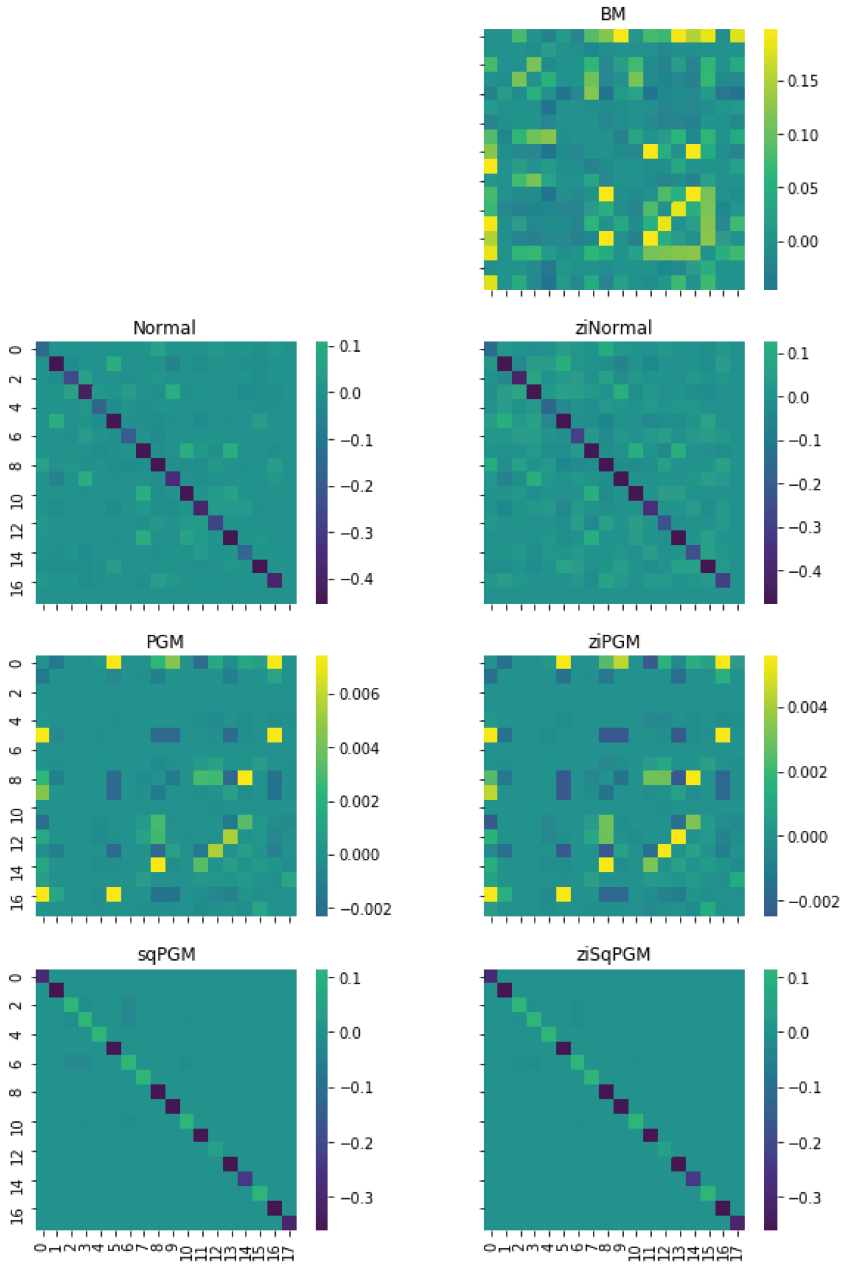


Figura G.7: Interacciones seleccionadas por los distintos pGMs considerados en relación con el lugar de donde se extrajeron las muestras al ser entrenados con datos de microbioma HMP (L2).

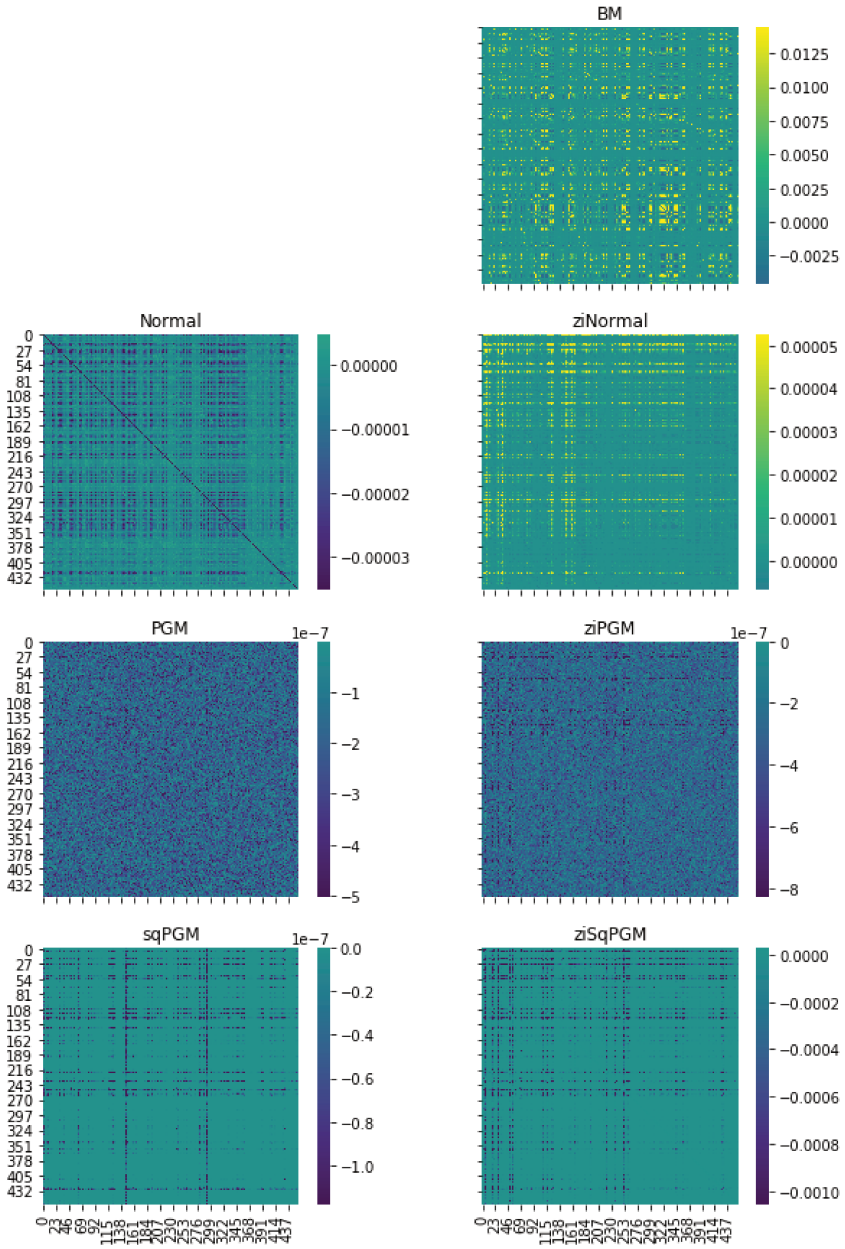


Figura G.8: Interacciones seleccionadas por los distintos pGMs considerados en relación con el lugar de donde se extrajeron las muestras al ser entrenados con datos de microbioma HMP (L6).

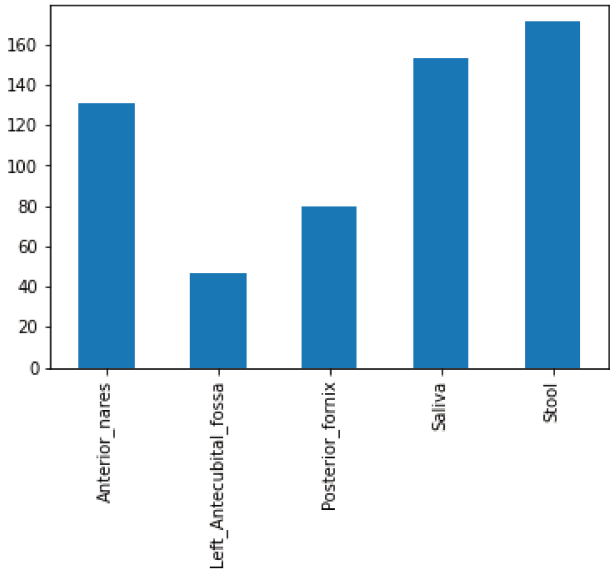


Figura G.9: Cantidad de observaciones por clase en los datos de microbioma HMP.

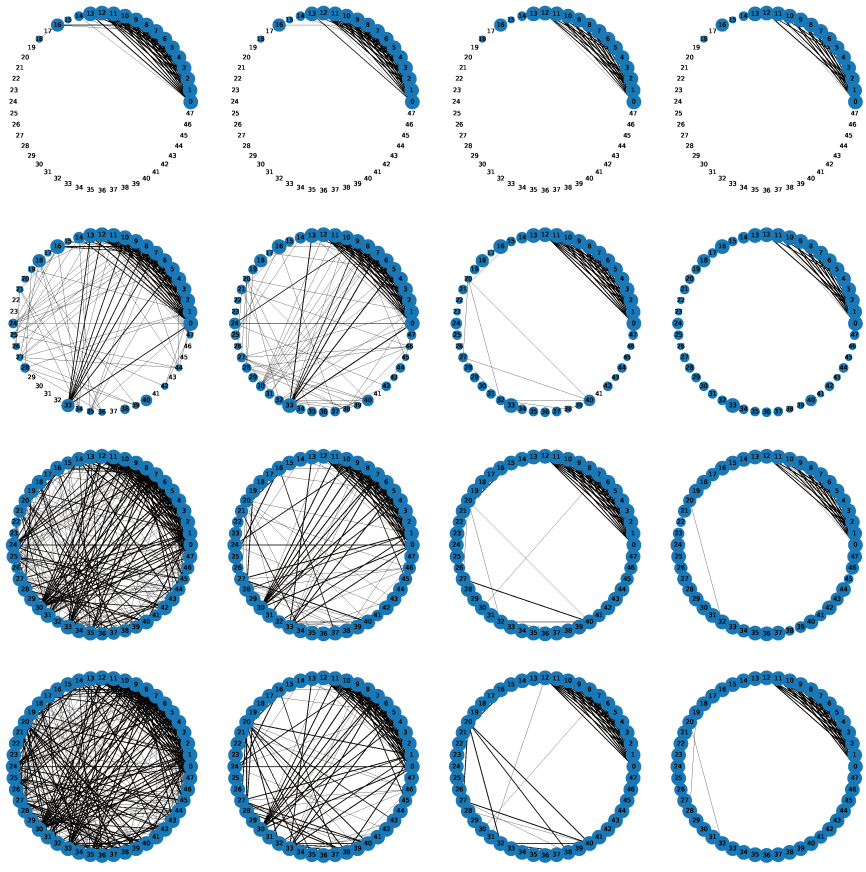


Figura G.10: Grafo de interacciones (aristas) y variables (nodos) seleccionados por la penalización jerárquica (5.6) para el modelo Normal-pGM y pseudolikelihood cuando es evaluada sobre una grilla regular de parámetros de regularización en el cuadrante positivo $\lambda_{\mathcal{C}}, \lambda_{\mathcal{R}} \geq 0$. La primer columna es computada usando $\lambda_{\mathcal{C}} = 0$, mientras que la última fila es computada con $\lambda_{\mathcal{R}} = 0$. Los círculos azules alrededor de los nodos indican las variables seleccionadas.

BIBLIOGRAFÍA

Adragni, Kofi P y R Dennis Cook (2009). «Sufficient dimension reduction and prediction in regression». En: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367.1906, págs. 4385-4405.

Aitchison, J. y J. Bacon-Shon (1984). «Log Contrast Models for Experiments with Mixtures». En: *Biometrika* 71.2, págs. 323-330.

Amari, Shun-ichi (2016). *Information geometry and its applications*. Springer.

Bauschke, HH y PL Combettes (2011). «Convex Analysis and Monotone Operator Theory in Hilbert Spaces». En: *CMS books in mathematics). DOI* 10, págs. 978-1.

Bertsekas, D. (2016). *Nonlinear Programming*. Athena scientific optimization and computation series. Athena Scientific. ISBN: 9781886529052. URL: https://books.google.com.ar/books?id=rC1EEAAAQBAJ.

Besag, J. (1974). «Spatial interaction and the statistical analysis of lattice systems». En: *Journal of the Royal Statistical Society Series B, Statistical Methodology* 36, págs. 192-236.

Besag, Julian (1975). «Statistical analysis of non-lattice data». En: *Journal of the Royal Statistical Society: Series D (The Statistician)* 24.3, págs. 179-195.

Bishop, Christopher M y Nasser M Nasrabadi (2006). *Pattern recognition and machine learning*. Vol. 4. 4. Springer.

Blaser, Martin J, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Iliana Estrada, Zhan Gao y Jose C Clemente (2016). «The Microbiome in Early Life: Implications for Health Outcomes». En: *Nature Medicine* 22.7, págs. 713-722. DOI: 10.1038/nm.4142.

Bura, Efstathia, Sabrina Duarte y Liliana Forzani (2016). «Sufficient reductions in regressions with exponential family inverse predictors». En: *Journal of the American Statistical Association* 111.515, págs. 1313-1329.

Calle, M Luz, Meritxell Pujolassos y Antoni Susin (2023). «coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies». En: *BMC bioinformatics* 24.1, pág. 82.

Cani, Patrice D, Jacques Amar, Miguel A Iglesias, Marjorie Poggi, Claude Knauf, Denis Bastelica, Audrey M Neyrinck, Florence Fava, Kieran M Tuohy, Chantal Chabo et al. (2014). «The Role of the Gut Microbiota in the Development of Obesity and Diabetes». En: *Gastroenterology* 146.6, págs. 1216-1228. DOI: 10.1053/j.gastro.2014.02.008.

Casella, George y R Berger (2002). *Statistical reference*.

Casella, George y Edward I George (1992). «Explaining the Gibbs sampler». En: *The American Statistician* 46.3, págs. 167-174.

Chen, Jun y Hongzhe Li (2013). «Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis». En: *The Annals of Applied Statistics* 7.1, págs. 418 -442. DOI: 10.1214/12-AOAS592. URL: https://doi.org/10.1214/12-AOAS592.

Chiquet, Julien, Mahendra Mariadassou y Stéphane Robin (2021). «The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances». En: *Frontiers in Ecology and Evolution* 9. ISSN: 2296-701X. DOI: 10.3389/fevo.2021.588292. URL: https://www.frontiersin.org/articles/10.3389/fevo.2021.588292.

Combettes, Patrick L y Jean-Christophe Pesquet (2011). «Proximal splitting methods in signal processing». En: *Fixed-point algorithms for inverse problems in science and engineering*. Springer, págs. 185-212.

Cook, R. Dennis (feb. de 2007). «Fisher Lecture: Dimension Reduction in Regression». En: *Statist. Sci.* 22.1, págs. 1-26.

Cook, R.D. (1998). *Regression Graphics: Ideas for studying regressions through graphics*. New York: Wiley.

Dawid, A Philip (1979). «Conditional independence in statistical theory». En: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.1, págs. 1-15.

Dawid, A Philip, Steffen Lauritzen, Matthew Parry et al. (2012). «Proper local scoring rules on discrete sample spaces». En: *Annals of Statistics* 40.1, págs. 593-608.

Díaz-Rizzolo, Marco M., María D. Ricci-Cabello, José M. Kiel-Ma, Carmen Serrano-García, Víctor López-Nicolás, Ángel Gil y María D. Mesa (2018). «Gut Microbiota and Metabolic Health: The Potential Beneficial Effects of a Medium Chain Triglyceride Diet in Obese Individuals». En: *Nutrients* 10.12, pág. 1857. DOI: 10.3390/nu10121857.

Evans, Steven N. y Frederick A. Matsen (2012). «The Phylogenetic Kantorovich–Rubinstein Metric for Environmental Sequence Samples». En: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 74.3, págs. 569-592. DOI: 10.1111/j.1467-9868.2011.01018.x.

Fettweis, Jennifer M., Myrna G. Serrano, J. Paul Brooks et al. (2019). «The Vaginal Microbiome and Preterm Birth». En: *Nature Medicine* 25, págs. 1012-1021. DOI: 10.1038/s41591-019-0450-2.

Gneiting, Tilmann y Adrian E Raftery (2007). «Strictly proper scoring rules, prediction, and estimation». En: *Journal of the American statistical Association* 102.477, págs. 359-378.

Gopalakrishnan, Vancheswaran, Beth A Helmink, Christine N Spencer, Alexandre Reuben y Jennifer A Wargo (2017). «The Microbiome and Cancer». En: *Nature Reviews Cancer* 18, págs. 728-739.

Guarner, Francisco, Juan-R Malagelada et al. (2012). «The Gut Microbiota in Health and in Disease». En: *Cell Host & Microbe* 12, págs. 458-469.

Haslett, John, Andrew Parnell y James Sweeney (2018). «A general framework for modelling zero inflation». En: *arXiv preprint arXiv:1805.00555*.

Hastie, Trevor, Robert Tibshirani y Martin Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Heinemann, Ana B., Vincent A. Martinis, Brandon T. P. Lau, Jennifer A. Charnigo y Steven W. Gallo (2017). «The Vaginal Microbiome: Rethinking Health and Diseases». En: *Microbial Ecology* 73.4, págs. 272-283. DOI: 10.1007/s00248-016-0898-4.

Hong, Johnny, Ulas Karaoz, Perry de Valpine y William Fithian (2022). «To rarefy or not to rarefy: robustness and efficiency trade-offs of rarefying microbiome data». En: *Bioinformatics* 38.9, págs. 2389-2396.

Hyvärinen, Aapo (2005). «Estimation of non-normalized statistical models by score matching». En: *Journal of Machine Learning Research* 6.Apr, págs. 695-709.

– (2007). «Some extensions of score matching». En: *Computational statistics & data analysis* 51.5, págs. 2499-2512.

Inouye, D.I., P. Ravikumar e I.S. Dhillon (2015). «Fixed-length Poisson MRF: adding dependencies to the multinomial». En: *Neural Information Processing Systems* 28.

Inouye, David I, Pradeep Ravikumar e Inderjit S Dhillon (2016). «Square root graphical models: Multivariate generalizations of univariate exponential families that permit positive dependencies». En: *arXiv preprint arXiv:1603.03629*.

Inouye, David I, Eunho Yang, Genevera I Allen y Pradeep Ravikumar (2017). «A review of multivariate distributions for count data derived from the Poisson distribution». En: *Wiley Interdisciplinary Reviews: Computational Statistics* 9.3, e1398.

Jones, Galin L y James P Hobert (2001). «Honest exploration of intractable probability distributions via Markov chain Monte Carlo». En: *Statistical Science*, págs. 312-334.

Khoruts, Alexander y Michael J Sadowsky (2017). «The Gut Microbiota in the Pathogenesis and Therapeutics of Inflammatory Bowel Disease». En: *Cell Host & Microbe* 21, págs. 717-726.

Kim, Byol, Song Liu y Mladen Kolar (2021). «Two-sample inference for high-dimensional markov networks». En: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83.5, págs. 939-962.

Kim, YS, T Unno, BY Kim y MS Park (2019). «Sex Differences in Gut Microbiota». En: *The World Journal of Mens Health* 38, 48–60.

Koller, Daphne y Nir Friedman (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Konishi, Sadanori y Genshiro Kitagawa (2008). *Information criteria and statistical modeling*. Springer Science & Business Media.

Kunstner, Frederik, Philipp Hennig y Lukas Balles (2019). «Limitations of the empirical Fisher approximation for natural gradient descent». En: *Advances in Neural Information Processing systems* 32.

Kurtz, Zachary D, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser y Richard A Bonneau (2015). «Sparse and compositionally robust inference of microbial ecological networks». En: *PLoS Computational Biology* 11.5, e1004226.

Lauritzen, Steffen L (1996). *Graphical models*. Vol. 17. Clarendon Press.

Le Cao, Kim-Anh, Simon Boitard y Philippe Besse (jun. de 2011). «Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems». En: *BMC Bioinformatics* 12.1, pág. 253. ISSN: 1471-2105.

Lee, Jason D y Trevor J Hastie (2015). «Learning the structure of mixed graphical models». En: *Journal of Computational and Graphical Statistics* 24.1, págs. 230-253.

Lee, Jason D, Yuekai Sun y Michael Saunders (2012). «Proximal Newton-type methods for convex optimization». En: *Advances in Neural Information Processing Systems*, págs. 827-835.

Li, Hongzhe (2015). «Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis». En: *Annual Review of Statistics and Its Application* 2.1, págs. 73-94. DOI: 10.1146/annurev-statistics-010814-020351.

Li, Ker-Chau (1991). «Sliced inverse regression for dimension reduction». En: *Journal of the American Statistical Association* 86.414, págs. 316-327.

Lin, Lina, Mathias Drton y Ali Shojaie (2016). «Estimation of high-dimensional graphical models using regularized score matching». En: *Electronic Journal of Statistics* 10.1, pág. 806.

Lindsay, Bruce G, Grace Y Yi y Jianping Sun (2011). «Issues and strategies in the selection of composite likelihoods». En: *Statistica Sinica*, págs. 71-105.

Lyu, Siwei (2012). «Interpretation and generalization of score matching». En: *arXiv preprint arXiv:1205.2629*.

Martens, James (2020). «New insights and perspectives on the natural gradient method». En: *Journal of Machine Learning Research* 21.1, págs. 5776-5851.

McDavid, Andrew, Raphael Gottardo, Noah Simon y Mathias Drton (2019). «Graphical Models for Zero-Inflated Single Cell Gene Expression». En: *Annals of Applied Statistics* 13.2, 848-–873.

McMurdie, Paul J. y Susan Holmes (2014). «Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible». En: *PLOS Computational Biology* 10.4, e1003531. DOI: 10.1371/journal.pcbi.1003531.

Montanari, Andrea y Jose Pereira (2009). «Which graphical models are difficult to learn?» En: *Advances in Neural Information Processing Systems*. Ed. por Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams y A. Culotta. Vol. 22. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2009/file/22fb0cee7e1f3bde58293de743871417-Paper.pdf.

Ovaskainen, Otso y Henrik Johan de Knegt (2017). *Quantitative Ecology and Evolutionary Biology: Integrating Models with Data*. Oxford University Press.

Pang, Daolin, Hongyu Zhao y Tao Wang (2023). «Factor Augmented Inverse Regression and its Application to Microbiome Data Analysis». En: *Journal of the American Statistical Association* 0.0, págs. 1-11. DOI: 10.1080/01621459.2023.2231577. eprint: https://doi.org/10.1080/01621459.2023.2231577. URL: https://doi.org/10.1080/01621459.2023.2231577.

Parikh, Neal y Stephen Boyd (2014). «Proximal algorithms». En: *Foundations and Trends in Optimization* 1.3, págs. 127-239.

Parikh, Neal, Stephen Boyd et al. (2014). «Proximal algorithms». En: *Foundations and Trends® in Optimization* 1.3, págs. 127-239.

Parry, Matthew, A Philip Dawid, Steffen Lauritzen et al. (2012). «Proper local scoring rules». En: *Annals of Statistics* 40.1, págs. 561-592.

Rivera-Pinto, Javier, Juan Jose Egozcue, Vera Pawlowsky-Glahn, Raul Paredes, Marc Noguera-Julian y M Luz Calle (2018). «Balances: a new perspective for microbiome analysis». En: *MSystems* 3.4, e00053-18.

Rohart, Florian, Benoit Gautier, Amrit Singh y Kim-Anh Lê Cao (2017). «mixOmics: An R package for 'omics feature selection and multiple data integration». En: *PLoS Computational Biology* 13.11, e1005752.

Rong, Ruichen, Shuang Jiang, Lin Xu, Guanghua Xiao, Yang Xie, Dajiang J Liu, Qiwei Li y Xiaowei Zhan (2021). «MB-GAN: Microbiome Simulation via Generative Adversarial Network». En: *GigaScience* 10.2, giab005.

Tomassi, Diego, Liliana Forzani, Sabrina Duarte y Ruth M Pfeiffer (dic. de 2019). «Sufficient dimension reduction for compositional data». En: *Biostatistics* 22.4, págs. 687-705. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxz060. eprint: https://academic.oup.com/biostatistics/article-pdf/22/4/687/41781875/kxz060.pdf. URL: https://doi.org/10.1093/biostatistics/kxz060.

Turnbaugh, Peter J, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight y Jeffrey I Gordon (2012a). «Human Microbiome». En: *Nature* 486, págs. 207-214.

Turnbaugh, Peter J, Vanessa K Ridaura, Jeremiah J Faith, Federico E Rey, Rob Knight y Jeffrey I Gordon (2012b). «The Gut Microbiota and Obesity: From Correlation to Causality». En: *Nature Reviews Microbiology* 10, págs. 55-62.

Vaart, Aad W Van der (2000). *Asymptotic statistics*. Vol. 3. Cambridge university press.

Varin, Cristiano, Nancy Reid y David Firth (2011). «An overview of composite likelihood methods». En: *Statistica Sinica* 21, págs. 5-42.

Wainwright, Martin J, Michael I Jordan et al. (2008). «Graphical models, exponential families, and variational inference». En: *Foundations and Trends® in Machine Learning* 1.1–2, págs. 1-305.

Xia, Fan, Jun Chen, Wing Kam Fung y Hongzhe Li (2013). «A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis». En: *Biometrics* 69.4, págs. 1053-1063. DOI: https://doi.org/10.1111/biom.12079. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12079. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12079.

Xu, Huan, Constantine Caramanis y Shie Mannor (2011). «Sparse algorithms are not stable: A no-free-lunch theorem». En: *IEEE transactions on pattern analysis and machine intelligence* 34.1, págs. 187-193.

Yadav, Manoj y Narendra S. Chauhan (2021). «Microbiome Therapeutics: Exploring the Present Scenario and Challenges». En: *Gastroenterology Report* 10, goab046. DOI: 10.1093/gastro/goab046.

Yang, Eunho, Pradeep Ravikumar, Genevera I Allen y Zhandong Liu (2015). «Graphical models via univariate exponential family distributions». En: *Journal of Machine Learning Research* 16.1, págs. 3813-3847.

Yang, Tianbao, Rong Jin, Shenghuo Zhu y Qihang Lin (2016). «On data preconditioning for regularized loss minimization». En: *Machine Learning* 103.1, págs. 57-79.

Yoon, Grace, Irina Gaynanova y Christian L Müller (2019). «Microbial networks in SPRING-Semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data». En: *Frontiers in genetics* 10, pág. 516.

Yuan, Ming y Yi Lin (2006). «Model selection and estimation in regression with grouped variables». En: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1, págs. 49-67.

Zhao, Ni, Jun Chen, Ian M. Carroll, Tamar Ringel-Kulka, Michael P. Epstein, Hua Zhou, Jie J. Zhou, Yehuda Ringel, Hongzhe Li y Michael C. Wu (2015). «Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test». En: *American Journal of Human Genetics* 96.5, págs. 797-807. DOI: 10.1016/j.ajhg.2015.04.003.

Zhao, Peng, Guilherme Rocha y Bin Yu (2009). «The composite absolute penalties family for grouped and hierarchical variable selection». En: *Annals of Statistics* 37.6A, págs. 3468-3497.

Zhao, Sihai Dave, T Tony Cai y Hongzhe Li (2014). «Direct estimation of differential networks». En: *Biometrika* 101.2, págs. 253-268.