



UNIVERSIDAD NACIONAL DE ROSARIO

FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

CARRERA DE POSTGRADO

DOCTORADO EN ESTADÍSTICA

“Modelización de Mapeo Asociativo en situaciones de variable fenotípica multidimensional de carácter mixto”

Autora: Verónica Patricia Lac Prugent

Director: Dr. Sergio Bramardi

Co-Director: Dr. Ramiro Curti

Rosario, 6 de marzo de 2026

Índice general

Resumen	VI
Abstract	IX
Publicaciones	XI
Dedicatorias	XIII
Agradecimientos	XIV
1. Introducción	1
1.1. Presentación del Problema	5
1.2. Objetivos e Hipótesis	8
1.3. Descripción del trabajo	9
1.4. Alcance del Trabajo	9
2. Aspectos Biológicos y Agronómicos	11
2.1. Introducción	11
2.2. Bancos de Germoplasma	12
2.3. Caracterización de recursos genéticos	14
2.3.1. Caracterización morfológica y agronómica	14
2.3.2. Caracterización citogenética	15
2.3.3. Caracterización bioquímica	16
2.3.4. Caracterización molecular. ADN	17
2.3.4.1. Marcadores moleculares	18
2.3.4.2. Marcadores SNP (Single Nucleotide Polymorphisms)	23
2.3.4.3. Secuenciación de ADN	23
2.4. Desequilibrio de ligamiento	24

2.4.1. Medidas de desequilibrio de ligamiento	26
2.4.2. Aplicaciones en mapeo asociativo	27
2.5. QTL, Mapeo Asociativo y GWAS	28
3. Aspectos Estadísticos	31
3.1. Análisis Multivariado.	31
3.1.1. Test de esfericidad de Barlett	35
3.2. Análisis Multivariado a dos vías.	36
3.2.1. Componentes Principales	36
3.3. Análisis Multivariado a tres vías.	37
3.3.1. Análisis Factorial Multiple (AFM)	37
3.4. Modelos Lineales.	44
3.4.1. Modelos de Regresión	45
3.4.1.1. Regresión lineal simple	46
3.4.1.2. Regresión lineal múltiple	47
3.4.1.3. Regresión lineal múltiple multivariada	48
3.4.2. Selección de variables	52
3.4.3. Modelos de regresión penalizados	53
3.4.3.1. Modelo de Regresión Ridge	54
3.4.3.2. Modelo de Regresión LASSO	58
3.4.3.3. Modelos de Regresión Red Elástica (Elastic Net)	60
3.4.3.4. Métodos de Optimización	63
3.4.4. Modelos de Regresión de mínimos cuadrados parciales (PLS)	69
3.4.5. Modelos Lineales Generalizados (MLG)	74
3.4.5.1. Familia exponencial y modelos asociados	75
3.4.5.2. Modelo Bernoulli multivariado condicionado (Modelo logístico multivariado conjunto para respuestas binarias)	82
3.4.5.3. Inferencia en MLG	84
3.5. Métodos de Remuestreo	87

3.5.0.1. Validación cruzada	88
3.5.0.2. Bootstrap	90
3.6. Análisis de Redundancia	91
3.6.1. Fundamentos metodológicos	92
3.6.2. Visualización mediante triplot	94
4. Mapeo Asociativo	96
4.1. Modelo sin corrección por estructura (<i>Naive</i>)	96
4.2. Modelo por estructura genética poblacional (EGP) P y Q	97
4.3. Modelo kinship (K)	98
4.4. Métodos de corrección por multiplicidad	99
5. Materiales	101
5.1. Base de datos de Quinoa	101
5.2. Base de datos de Arroz	104
5.3. Entorno de trabajo y Herramientas Computacionales	106
6. Estudio comparativo de alternativas metodológicas disponibles para evaluar asociación de caracteres fenotípicos con marcadores moleculares	109
6.1. Análisis preliminar	110
6.1.1. Desequilibrio de ligamiento	110
6.1.2. Análisis Multivariado: Factorial Múltiple (AFM)	110
6.2. Mapeo Variables Cuantitativas	112
6.2.1. Coeficiente de Escoufier	112
6.2.2. Modelos univariados <i>Naive</i>	113
6.2.3. Modelos univariados P	115
6.2.4. Modelo de Regresión Elastic Net Múltiple Multivariado	118
6.2.5. Modelo Multivariado de Regresión de Mínimos Cuadrados Parciales (PLS)	121
6.2.6. Comparación de los modelos	123

6.3. Mapeo Variables Cualitativas	125
6.3.1. Modelos <i>Naive</i> Logísticos	125
6.3.2. Modelos Naive Logit Multinomiales	128
6.4. Consideraciones finales del estudio comparativo	130
7. Propuesta Metodológica	134
7.1. Introducción	134
7.2. Limitaciones y alcances de las metodologías aplicadas en el estudio comparativo	135
7.2.1. Limitaciones en el tratamiento de variables cuantitativas . . .	136
7.2.2. Limitaciones en el tratamiento de variables cualitativas binarias y politómicas	136
7.3. Fundamento de la propuesta	137
7.4. Propuesta <i>RDA-Like</i> : Multivariado para variables cuantitativas . . .	138
7.4.1. Supuestos y condiciones de aplicación	142
7.5. Propuesta <i>RDA-Like</i> : Multivariado para variables dicotómicas	142
7.5.1. Modelo logístico multivariado conjunto para respuestas binarias con penalización <i>Elastic Net</i>	143
7.5.1.1. Aplicación del modelo logístico multivariado con penalización <i>Elastic Net</i>	147
7.5.2. Construcción del análisis <i>RDA-Like</i>	151
7.5.3. Supuestos y condiciones de aplicación	152
7.6. Propuesta <i>RDA-Like</i> : Multivariado mixto	153
7.6.1. Objetivo e idea central	153
7.6.2. Construcción del espacio mixto	154
7.6.3. Proyección de marcadores moleculares y armado del triplot . .	155
7.6.4. Interpretación	155
7.6.5. Implementación y salidas	156
7.6.6. Ventajas y alcances	156
7.6.7. Limitaciones	156

8. Aplicación de la propuesta metodológica de mapeo asociativo para variables mixtas	158
8.1. Datos de quinoa	158
8.1.1. <i>RDA-Like</i> Multivariado para variables cuantitativas	158
8.1.2. <i>RDA-Like</i> multivariado para variables cualitativas dicotómicas	163
8.1.3. <i>RDA-Like</i> multivariado mixto	165
8.2. Datos de arroz	167
8.2.1. <i>RDA-Like</i> multivariado para variables cuantitativas	168
8.2.2. <i>RDA-Like</i> multivariado para variables cualitativas dicotómicas	171
8.2.3. <i>RDA-Like</i> Multivariado mixto	177
9. Discusión	179
10. Conclusiones	186
10.1. Limitaciones y futuras líneas de investigación	191
A. Código en R	193
Referencias	298

Resumen

Los bancos de germoplasma resguardan accesiones que han sido caracterizadas mediante rasgos fenotípicos y marcadores moleculares en distintos programas de evaluación y conservación. Dicha información se encuentra, en gran medida, subutilizada en términos de su potencial para el análisis conjunto. La disponibilidad de estos datos permite su reaprovechamiento mediante estrategias de mapeo asociativo, orientadas a la detección de asociaciones entre fenotipos y marcadores moleculares. Dichas asociaciones constituyen señales candidatas que pueden ser posteriormente profundizadas y validadas en estudios específicos. No obstante, el mapeo asociativo presenta desafíos metodológicos en presencia de alta correlación entre variables, coexistencia de respuestas cuantitativas y cualitativas y alta dimensionalidad. En este contexto, los enfoques univariados y multivariados tradicionales resultan limitados para capturar patrones de asociación conjunta entre marcadores moleculares y rasgos, así como para interpretar posibles señales pleiotrópicas.

En este trabajo se propone una estrategia metodológica de análisis que combina modelos de regresión Elastic Net Multivariados—en su formulación gaussiana para variables cuantitativas y su extensión logística para variables dicotómicas—con un enfoque interpretativo denominado *RDA-like*. Este enfoque retoma la lógica del Análisis de Redundancia, pero se aplica a las respuestas predichas por los modelos penalizados, permitiendo explorar la estructura multirasgo inducida por los predictores genéticos y facilitar la interpretación conjunta de asociaciones complejas.

Dado que actualmente no existe una implementación directa del Modelo Logístico Multivariado Elastic Net en el software R, fue necesario abordar un desafío computacional específico para llevar a cabo la propuesta metodológica. En este escenario, la tesis desarrolla el marco teórico correspondiente y presenta una propuesta computacional original que posibilita el modelado simultáneo de múltiples respuestas binarias mediante penalización compartida y selección conjunta de marcadores. Con

el fin de evaluar la estabilidad de las asociaciones fenotipo–marcador y de disponer de un criterio inferencial para la selección de marcadores en el marco de modelos penalizados, se emplean esquemas de validación cruzada para la calibración de los parámetros de penalización, combinados con procedimientos de remuestreo bootstrap. En particular, el bootstrap se utiliza para estimar intervalos de confianza empíricos de los coeficientes del modelo, permitiendo identificar como significativos aquellos marcadores cuyos intervalos no incluyen el cero. Este enfoque suple la ausencia de pruebas de hipótesis analíticas en los métodos penalizados y permite caracterizar la variabilidad de los coeficientes estimados, aportando evidencia empírica sobre la robustez de las asociaciones detectadas. Estas estrategias resultan especialmente adecuadas en contextos de alta dimensionalidad, contribuyendo a mitigar el sobreajuste y fortaleciendo la validez de las variables seleccionadas. La utilidad del marco metodológico se evalúa mediante su aplicación a dos conjuntos de datos contrastantes: accesiones de quinoa del Noroeste Argentino y el panel *Rice Diversity Panel 1* de arroz. En ambos casos, los modelos Multivariados penalizados permitieron identificar asociaciones genéticas consistentes y biológicamente plausibles, mientras que los análisis *RDA-like* facilitaron la visualización e interpretación de patrones multirasgo, explicando una proporción sustancial de la variabilidad total. En arroz, además, se redetectaron marcadores previamente reportados en la literatura como asociados a genes clave involucrados en la determinación del tamaño del grano, lo que refuerza la coherencia biológica del enfoque.

En conjunto, esta tesis contribuye a la **Estadística aplicada** mediante el desarrollo de un enfoque metodológico que integra **modelización Multivariada penalizada, selección conjunta de predictores e interpretación estructural** en escenarios de alta dimensionalidad y respuestas mixtas. La propuesta amplía el alcance del **Elastic Net Multivariado** a contextos no gaussianos y articula modelos con técnicas clásicas de análisis Multivariado, ofreciendo una herramienta **unificada, reproducible y extensible** para el análisis de datos complejos, más allá del ámbito específico de la genómica.

Palabras clave: alta dimensionalidad; modelos multivariados penalizados; Elastic

Net; RDA-like; características de recursos genéticos; marcadores moleculares; Bancos de Germoplasma.

Abstract

Germplasm banks store accessions that have been characterized by phenotypic traits and molecular markers in different evaluation and conservation programs. This information is largely underutilized in terms of its potential for joint analysis. The availability of these data allows them to be reused through associative mapping strategies aimed at detecting associations between phenotypes and molecular markers. These associations constitute candidate signals that can subsequently be explored in greater depth and validated in specific studies. However, associative mapping presents methodological challenges in the presence of high correlation between variables, coexistence of quantitative and qualitative responses, and high dimensionality. In this context, traditional univariate and multivariate approaches are limited in their ability to capture patterns of joint association between molecular markers and traits, as well as to interpret possible pleiotropic signals.

This paper proposes a methodological analysis strategy that combines Multivariate Elastic Net regression models—in their Gaussian formulation for quantitative variables and their logistic extension for dichotomous variables—with an interpretative approach called *RDA-like*. This approach draws on the logic of Redundancy Analysis, but applies it to the responses predicted by the penalized models, allowing the exploration of the multitrait structure induced by genetic predictors and facilitating the joint interpretation of complex associations.

Given that there is currently no direct implementation of the Elastic Net Multivariate Logistic Model in R software, it was necessary to address a specific computational challenge in order to carry out the methodological proposal. In this scenario, the thesis develops the corresponding theoretical framework and presents an original computational proposal that enables the simultaneous modeling of multiple binary responses through shared penalty and joint marker selection. In order to evaluate the stability of phenotype-marker associations and to provide an inferential criterion

for marker selection in the context of penalized models, cross-validation schemes are used for the calibration of penalty parameters, combined with bootstrap resampling procedures. In particular, bootstrap is used to estimate empirical confidence intervals for the model coefficients, allowing markers whose intervals do not include zero to be identified as significant. This approach compensates for the absence of analytical hypothesis testing in penalized methods and allows the variability of the estimated coefficients to be characterized, providing empirical evidence on the robustness of the associations detected. These strategies are particularly suitable in high-dimensional contexts, helping to mitigate overfitting and strengthening the validity of the selected variables. The usefulness of the methodological framework is evaluated by applying it to two contrasting data sets: quinoa accessions from northwestern Argentina and the Rice Diversity Panel 1. In both cases, penalized multivariate models allowed us to identify consistent and biologically plausible genetic associations, while RDA-like analyses facilitated the visualization and interpretation of multitrait patterns, explaining a substantial proportion of the total variability. In rice, markers previously reported in the literature as associated with key genes involved in grain size determination were also redetected, reinforcing the biological coherence of the approach.

Together, this thesis contributes to applied statistics by developing a methodological approach that integrates penalized multivariate modeling and joint predictor selection. Overall, this thesis contributes to **applied statistics** by developing a methodological approach that integrates **penalized multivariate modeling, joint predictor selection**, and **structural interpretation** in high-dimensionality and mixed-response scenarios. The proposal extends the scope of **Multivariate Elastic Net** to non-Gaussian contexts and articulates models with classical multivariate analysis techniques, offering a **unified, reproducible, and extensible** tool for the analysis of complex data, beyond the specific field of genomics.

Keywords: high dimensionality; penalized multivariate models; Elastic Net; RDA-like; genetic resource characteristics; molecular markers; germplasm banks.

Publicaciones

- Lac Prugent, Verónica; Costa Tártara, Sabrina; Curti, Ramiro; Bramardi, Sergio. (2020). *Análisis factorial múltiple para la identificación de asociaciones entre caracteres fenotípicos cualitativos y marcadores moleculares*. En *XXV Reunión Científica del Grupo Argentino de Biometría*.
- Lac Prugent, Verónica; Costa Tártara, Sabrina; Curti, Ramiro; Bramardi, Sergio. (2022). *Estudio comparativo entre AFM y mapeo asociativo para identificar relaciones entre caracteres fenotípicos cuantitativos y marcadores moleculares*. En *XXVI Reunión Científica del Grupo Argentino de Biometría*. Córdoba, Argentina.
- Lac Prugent, Verónica; Giménez, Gustavo; Costa Tártara, Sabrina; Curti, Ramiro; Bramardi, Sergio. (2022). *Estudio comparativo de alternativas metodológicas para selección de marcadores moleculares en mapeo asociativo en quinoa*. En *XLIX Coloquio Argentino de Estadística y VII Jornada de Educación Estadística “Martha Aliaga”*.
- Lac Prugent, Verónica; Curti, Ramiro; Bramardi, Sergio. (2023). *Estrategias multivariadas para selección de marcadores moleculares en mapeo asociativo en quinoa*. En *XXVII Reunión Científica del Grupo Argentino de Bioestadística*.
- Lac Prugent, Verónica P.; Costa Tártara, Sabrina; Curti, Ramiro; Bramardi, Sergio. (2024). *Detección de asociaciones genéticas en accesiones de quinoa del noroeste argentino mediante mapeo asociativo con variable dependiente cualitativa*. En *XXVIII Reunión Científica del Grupo Argentino de Bioestadística*. Fortaleciendo la soberanía a través del conocimiento científico.
- Lac Prugent, Verónica; Costa Tártara, Sabrina; Curti, Ramiro; Ruminot, Sergio; Bramardi, Sergio. (2025). *Detección de asociaciones genéticas en acce-*

siones de quínoa del noroeste argentino mediante un modelo multivariado de regresión de mínimos cuadrados parciales (PLS). En XXIX Reunión Científica del Grupo Argentino de Bioestadística.

- Lac Prugent, Verónica; Bramardi, Sergio; Curti, Ramiro. (2026). *Comparación de mapeo asociativo naive, modelos Elastic Net y regresión PLS en accesiones de quínoa del noroeste argentino. Revista SaberEs. Manuscrito en evaluación.*

Dedicatorias

A mis padres Miguel y Graciela, por enseñarme que la libertad se construye desde el conocimiento.

A mi compañero de vida, Pablo, por acompañarme con su amor incondicional en todos mis proyectos.

A mis hijos Evana y Bastian, por su acompañamiento amoroso.

A mis hermanas y a mi hermano: Gabi, Marcos, Ani y Nahi, y a mi gran amigo Edu, por estar siempre.

A mis amigas Soledad Rey y Natalí Santoro, por su amistad incondicional, por acompañarme en los momentos más difíciles y por celebrar conmigo cada logro de este camino.

A mis tíos, los Drs. en Estadística Nora y Carlitos Lac Prugent, a quienes debo mi curiosidad por la Estadística, por haber sido una inspiración para seguir este camino científico y por sus valiosos consejos profesionales.

Agradecimientos

Quiero expresar mi más profundo agradecimiento a mi Director de Tesis, el Dr. Sergio Bramardi, por haber aceptado dirigir este trabajo y por su acompañamiento académico permanente a lo largo de todo el proceso de investigación. Destaco especialmente su rigurosa guía científica, su claridad conceptual y metodológica, y la generosidad con la que compartió sus conocimientos, los cuales fueron fundamentales para la definición, el desarrollo y la consolidación de esta tesis.

Asimismo, agradezco a mi Co Director de Tesis, el Dr. Ramiro Curti, por su valiosa disposición y aportes académicos.

Agradezco también al personal de los bancos de germoplasma del INTA, de donde provienen los datos de quinoa utilizados en esta tesis. En particular, deseo destacar a la Dra. Sabrina Costa Tártara, investigadora del CONICET, por su apoyo y colaboración en el área de la genética vegetal.

Del mismo modo, agradezco a mis compañeros y compañeras del Departamento de Estadística por el acompañamiento y el espacio académico compartido a lo largo de este proceso.

A nivel personal, quiero agradecer especialmente a mi tía, Nora Lac Prugent, por sus valiosos consejos, su apoyo constante y amoroso, y por la confianza que siempre depositó en mí.

A mi suegra, Lidia, por su interés, su lectura y por su permanente estímulo hacia el estudio.

A mis compañeras de doctorado Eugenia Bortolotto y Ana Paula del Médico ¡GRACIAS!, por el espacio compartido de vivencias, aprendizajes y motivación que hizo posible avanzar juntas.

Capítulo 1

Introducción

En un contexto mundial con gran crecimiento de población, y por ende nuevas demandas de alimentos, es de suma importancia cuidar, preservar y utilizar de manera sostenible los recursos genéticos. Entendiendo a recursos genéticos como el material hereditario que se encuentra en las especies, que tienen un valor económico, medioambiental, científico o social, real o potencial (Organización de las Naciones Unidas para la Agricultura y la Alimentación [FAO], [1989](#)). En este marco, la conservación de estos recursos resulta indispensable para ampliar la productividad agrícola y sustentar las demandas futuras de alimentos. Con esta finalidad es que existen los bancos de germoplasma. En nuestro país los bancos de germoplasma nacionales surgieron a mediados del siglo pasado, en el Instituto Nacional de Tecnología Agropecuaria (INTA). El INTA coordina la Red Nacional de Conservación de Recursos Fitogenéticos (REDGEN) cuyo objetivo es la conservación ex situ de germoplasma vegetal y la investigación en temas relacionados con la caracterización de este material. Esta red conserva aproximadamente 36.000 entradas de germoplasma vegetal de interés actual y potencial. La estructura de REDGEN incluye una colección base y colecciones activas distribuidas en 20 bancos activos y un jardín botánico, ubicados en diversas regiones del país (FAO, [2021b](#)). Además de la conservación y mantenimiento de los recursos fitogenéticos, la evaluación del material conservado es una etapa crítica para su posterior aprovechamiento. Los bancos de germoplasma no se limitan a la mera conservación a largo plazo, se complementan con un trabajo arduo y costoso de documentación, caracterización morfológica y agronómica, evaluación de la variabilidad genética y estudios filogenéticos. Las bases de datos de accesiones son producto de la recolección y preservación del material genético que lleva

a cabo el INTA. Una accesión es una muestra específica y claramente identificada de material vegetal, almacenada en un banco de germoplasma, que representa una población, variedad o línea genética particular, recolectada en un lugar y momento definidos (Engels & Visser, 2003). Estas bases de accesiones contienen información de variables cuantitativas como diámetro del tallo, altura de la planta, longitud de hoja, rendimiento, entre otras; y variables cualitativas como forma y color de fruto, resistencia a enfermedades, presencia de espinas, entre otras. Esta caracterización agronómica y morfológica de las accesiones se ha completado en los últimos años, al menos en parte del material conservado, con caracterización molecular. Establecer relaciones entre la caracterización agronómica-morfológica (fenotipo) de un conjunto de accesiones conservadas en un banco de germoplasma con la caracterización molecular (genotipo) es un paso preliminar, que, utilizando información preexistente, puede contribuir a los procesos de mejora genética de especies de interés alimentario. En la Mesa Redonda “Mejoramiento Genético” del 1º Congreso Argentino de Estadística (Buenos Aires, 2015) se expuso sobre “Desarrollos estadísticos actuales y futuros en el Mejoramiento Vegetal: integración de Información Fenotípica y Molecular”. En dicha presentación se ejemplificaron las metodologías utilizadas para la integración de variables agronómicas y moleculares, entre ellas el Análisis de Procrustes Generalizado (APG) y Análisis factorial Múltiple (AFM), concluyendo que existen nuevos desafíos en el área, tales como: distribuciones asimétricas muy dispersas, importante grado de correlación entre las variables, individuos no independientes, grandes bases de datos y escasa cantidad de repeticiones, entre otras (Bramardi et al., 2015). Numerosos estudios nacionales han abordado el tema entre ellos, el trabajo de Bramardi et al. (2017) presentado en el XLV Coloquio Argentino de Estadística de la Sociedad Argentina de Estadística (Rosario, 2017). En este trabajo se presentan métodos de análisis multivariado a tres vías para identificación de asociaciones entre caracteres fenotípicos de carácter mixtos y marcadores moleculares aplicados a una base de datos de quinoa, pero estas técnicas no permiten establecer el nivel de certidumbre de dichas relaciones, sino que constituyen un primer acercamiento de carácter descriptivo. Otro antecedente significativo es

la tesis doctoral de A. N. Peña Malavera (2015) en la que utiliza la metodología de modelos mixtos y simulaciones para evaluar distintos modelos de mapeo asociativos, con diferentes estructuras genéticas. Breseghello y Sorrells (2006) definen al Mapeo asociativo como una metodología estadística introducida en estudios de mejoramiento genético para identificar genes responsables de la variación de caracteres cuantitativos donde adquiere importancia establecer la significancia estadística de las asociaciones entre la información que proveen marcadores moleculares y características fenotípicas de interés. Una de las características de éstas fuentes de datos es su alta dimensionalidad, es decir suelen incluir cientos o incluso miles de marcadores moleculares frente a un número considerablemente menor de observaciones. Entre los diversos modelos de mapeo asociativo el primero en proponerse fue el SCE (sin corrección por estructura) o modelo Naive (ingenuo) que recurre a regresión lineal considerando a las variables independientes (marcadores moleculares) como efectos fijos, posteriormente se incorpora información acerca de la no independencia entre los individuos a través de un Análisis de Componentes Principales (Pritchard et al., 2000) al que se lo identifica como modelo P que da origen a los modelos EGP (Estructura Genética Poblacional), comenzándose a modelar la estructura genética de los individuos mediante una matriz entre los efectos aleatorios de los genotipos que se incorpora a la estructura de la variancias-covariancias de un Modelo Lineal Mixto (MLM) (Demidenko, 2004) y la estimación se hace por máxima verosimilitud restringida (REML) (Patterson & Thompson, 1971). Entre ellos podemos mencionar los modelos K o Kinship (parentesco) que utilizan una matriz de coancestría y los modelos Q que recurren a una matriz que presenta las probabilidades de pertenencia de los individuos a k grupos en los que se podría subdividir el material bajo estudio. También se encuentran modelos que resultan de la combinación de los modelos P, K y Q. Los modelos Kinship, quizás los más utilizados, fueron propuestos por Yu y Buckler (2006) y modificado posteriormente por Kang et al. (2008). Todos ellos planteados para una sola variable dependiente y de carácter cuantitativo. En el trabajo de J. Li y Ji (2005) se deja de manifiesto que en las pruebas múltiples correlacionadas de las investigaciones en genética surge el problema de multiplicidad

en los test de hipótesis de asociaciones entre cada marcador y el carácter fenotípico estudiado, en los cuales se tiene un error tipo I o nivel de significación por cada test, lo que aumenta significativamente al considerar todos los test de asociaciones. En relación con este problema, la bibliografía consultada cita a Benjamini y Hochberg (1997), quienes compararon distintas propuestas de procedimientos de incorporación de pesos en los test múltiples y analizaron sus tasas de error tipo I. Los autores concluyen que se requiere de trabajo adicional para evaluar el poder estadístico de los distintos procedimientos, así como la necesidad de compararlos con otros enfoques previamente publicados, tales como los de Rüger (1978), Hommel (1986), Shaffer (1986) y Bergmann y Hommel (1988).

Para abordar el problema de multiplicidad, varios autores han realizado correcciones o reglas, podemos citar a la regla de Bonferroni (1935), Benjamini y Hochberg (1995), Benjamini y Yekutieli (2001), que probaron su eficacia en caso de muestras independientes, sin embargo, tales procedimientos pueden no ser adecuados en situaciones de poblaciones estructuradas en diferentes niveles de parentesco (A. N. Peña Malavera, 2015). J. Li y Ji (2005) proponen una estimación más precisa que la propuesta por Cheverud (2001) para pruebas correlacionadas, sin embargo, A. N. Peña Malavera (2015) afirma que no se conoce hasta el momento como la presencia de la estructura genética poblacional podría impactar sobre el desempeño del método. A. N. Peña Malavera (2015, p. 6) afirma que “los procedimientos analíticos que se deben utilizar para mapeo asociativo tienen que considerar la presencia de tales correlaciones en los datos. Es así, como el cuerpo de análisis estadísticos utilizados en estudios de Mapeo Asociativo es muy variado y muchas veces no es claro para el investigador qué estrategia de modelación es más conveniente emplear para analizar estadísticamente sus datos moleculares en combinación con un carácter fenotípico de interés”.

Con el fin de disminuir los descubrimientos de falsas asociaciones que resultan significativas sólo por azar A. Peña Malavera et al. (2018) proponen un método de corrección por multiplicidad para los p-valores, basados en estimación del número efectivo de pruebas desde un modelo que ajusta por la estructura de correlación

subyacente (estructura genética). La propuesta está basada en el número de pruebas efectivas o independientes (similar a Li y Ji). Partiendo de tres Modelos de mapeo asociativo, uno que no contempla ninguna corrección por estructura genética poblacional (EGP) (modelo naive) y otros con distintos tipos de ajuste de la Estructura Genética de la población, la modificación que proponen para la corrección por multiplicidad, utiliza los ejes derivados de la descomposición espectral aplicada sobre la matriz de estadísticos de Mantel y Haenszel Mantel y Haenszel (1959) incorporando la información conocida de la estructura genética poblacional. El método propuesto fue comparado con otros métodos de corrección por multiplicidad usando datos simulados. Otros autores tales como, (A. Li & Barber, 2019), proponen el algoritmo de Benjamini-Hochberg adaptable a la estructura (SABHA por las siglas de Structure Adaptive Benjamini-Hochberg Algorithm) como una generalización de los métodos de control de la tasa de falsos positivos (FDR, por las siglas False Discovery Rate). A lo largo de la revisión de antecedentes, en todos los casos el mapeo asociativo se presenta para relacionar un conjunto de marcadores moleculares con una única variable fenotípica dependiente de carácter cuantitativo, pero es muy común encontrar situaciones en donde dicha variable es de carácter cualitativo, es más, es frecuente encontrar, sobre todo en datos relevados de bancos de germoplasma, varios caracteres fenotípicos que se quieren relacionar con un grupo de marcadores moleculares, con presencia simultánea de caracteres cualitativos y cuantitativos. No existe al momento un procedimiento estadístico de análisis para tales situaciones, por lo tanto, se concibe la necesidad de investigar y desarrollar metodología estadística que permita evaluar la relación entre variables morfológicas y agronómicas de carácter cualitativo y cuantitativo con variables moleculares (Mapeo Asociativo en contexto de variable dependiente multivariada y Mixta).

1.1 Presentación del Problema

El hombre, desde la prehistoria, ha seleccionado y mejorado especies animales y vegetales basándose en el fenotipo. La variabilidad genética, la heredabilidad del

carácter a mejorar, la intensidad y eficacia de la selección y el tiempo para realizar la selección, hicieron posible las mejoras genéticas. A pesar de esto, el número y efecto de los genes implicados en la expresión del carácter agronómico o morfológico bajo estudio, la localización de estos genes, y su función fisiológica resultaban ser aspectos desconocidos. Los criterios utilizados carecían muy a menudo de definición y objetividad y, en cualquier caso, esos fenotipos resultaban ser marcadores ambiguos debido a las influencias ambientales (Claros Díaz, 1998). La obtención de nuevos cultivares por la vía de la mejora clásica o convencional no garantiza la obtención de ese cultivar mejorado. La selección asistida por marcadores (SAM) (en inglés: Marker Assisted Selection, MAS), es una técnica moderna de mejora vegetal que ha experimentado en los últimos años una revolución silenciosa que contrasta con la atención pública acaparada por la ingeniería genética (Enciclopedia Colaborativa Cubana (ECURED), 2020) y constituye una estrategia poderosa para el mejoramiento que reduce los tiempos necesarios para desarrollar variedades. La SAM no suplanta a la mejora vegetal tradicional, la complementa y la hace más eficiente. Esta técnica realiza la selección “molecular” del carácter agronómico de forma que se independiza del fenotipo y del ambiente y puede hacerse en estadíos de plántula, con sólo buscar el marcador mediante una prueba de ADN, el que permite detectar de inmediato, por ejemplo, qué plántulas han adquirido resistencia y cuáles no (Lombardo et al., 2010). Para poder aplicar SAM previamente se debe encontrar esa relación carácter fenotípico-marcador molecular. Existen diversas técnicas experimentales para determinar dicha asociación la más difundida es el mapeo de ligamiento para identificación de QTLs. El concepto de ligamiento se refiere a la tendencia que tienen los alelos de loci cercanos en un mismo cromosoma a heredarse juntos durante la meiosis, debido a que no segregan de manera independiente. Como consecuencia, las combinaciones alélicas parentales aparecen con mayor frecuencia que lo esperado según la Ley de la Segregación Independiente de Mendel (Griffiths et al., 2021). Como afirma Guindon (2017, p. 29): “La posibilidad de que ocurra el entrecruzamiento está directamente relacionada con la distancia entre dos genes, cuanto menor sea la frecuencia de recombinación entre dos marcadores, más cerca

se sitúan en el cromosoma”. Esto permite determinar con precisión, en muchos casos, la porción del cromosoma que controla un determinado carácter fenotípico, pero tiene la desventaja de requerir complejos desarrollos de progenie. En contraposición el mapeo asociativo (MA) o GWAS (por sus siglas en inglés, Genome Wide Association Study) es una herramienta simple y poderosa para identificar marcadores ligados a variaciones fenotípicas en un carácter de interés a partir de un conjunto de genotipos que explota la diversidad existente en el germoplasma estudiado sin la necesidad de construir poblaciones de mapeo o poblaciones de cruzamientos controlados. La metodología del MA ha sido propuesto y desarrollado para asociar marcadores moleculares a una variable morfológica o agronómica de carácter cuantitativo (Jannink et al., [2001](#)), pero en las bases de datos disponibles, generalmente en los bancos de germoplasma, existe gran cúmulo de información referida a caracteres fenotípicos tanto cuantitativos como cualitativos que se encuentra subexplotadas en cuanto a su relación con marcadores moleculares. El estudio de estas asociaciones y sus modelos, presentan desafíos metodológicos tales como encontrar modelos de mapeos asociativos para variables dependientes cuantitativas y cualitativas en forma conjunta (multivariado mixto), reducir la cantidad de falsos positivos dado los múltiples test que implica esta metodología, así como definir adecuadamente el corte de los p-valores correspondientes a las variables independientes. Aparte esta tesis doctoral se inscribe en el marco de la capacitación y formación de recursos humanos en el ámbito de la investigación, específicamente dentro de los programas “Análisis Multivariado y Modelización de bases de datos con estructura a tres vías y de alta dimensionalidad” (2018-2022) y “Análisis y Modelización de bases de datos con estructura multivía y de alta dimensionalidad” (2023-2026), ambos desarrollados en la Universidad Nacional del Comahue, bajo la financiación de la Secretaría de Ciencia y Técnica de la universidad.

1.2 Objetivos e Hipótesis

Hipótesis: La información disponible en un banco de germoplasma puede ser utilizada para modelar, con un nivel de certidumbre aceptable, la relación entre un conjunto de variables morfológicas y agronómicas de carácter mixto (cuantitativas y cualitativas) con la correspondiente caracterización molecular.

Objetivo principal: Proponer una estrategia metodológica para el estudio de asociaciones entre variables fenotípicas y marcadores moleculares en el contexto de análisis multivariado de carácter mixto.

Objetivos Específicos:

- Desarrollar un modelo de mapeo asociativo para una variable dependiente de carácter cualitativo
- Implementar un modelo de mapeo asociativo para una variable dependiente multivariada
- Combinar en un solo procedimiento asociación de marcadores moleculares y variables cuantitativas y cualitativas
- Proponer y evaluar métodos para la corrección de valores-p por multiplicidad, para cuantificar la significancia de la asociación atributo fenotípico – marcador molecular
- Evaluar comparativamente sobre datos empíricos el modelo de mapeo asociativo multivariado mixto propuesto con los modelos empleados actualmente
- Programar las sentencias necesarias en R que permita utilizar las metodologías desarrolladas
- Caracterizar, mediante técnicas estadísticas multivariadas, fenotípica y molecularmente el material preexistente en bancos de germoplasma de quinoa y arroz

1.3 Descripción del trabajo

Esta investigación se organiza en diez capítulos que recorren un itinerario completo: desde los fundamentos teóricos de los modelos lineales univariados, multivariados y penalizados, hasta su implementación para el análisis de relaciones entre variables fenotípicas multidimensionales de carácter mixto con marcadores moleculares. El Capítulo 2 presenta el marco conceptual biológico-agronómico que sustenta el estudio, mientras que el Capítulo 3 desarrolla los aspectos estadísticos esenciales empleados a lo largo de la tesis. El Capítulo 4 describe los modelos de mapeo asociativo tradicionales que sirven como punto de partida metodológico.

Los resultados del estudio comparativo entre métodos univariados y multivariados aplicados a accesiones de quinoa se detallan en el Capítulo 6. Sobre esta base, el Capítulo 7 introduce tres propuestas metodológicas diseñadas para integrar de manera eficiente información proveniente de rasgos cuantitativos y cualitativos. Su aplicación y evaluación en dos escenarios contrastantes —quinoa y arroz— se desarrollan en el Capítulo 8.

El Capítulo 9 sintetiza y discute los hallazgos más relevantes. Finalmente, el Capítulo 10 presenta las conclusiones generales, destacando las implicancias del enfoque propuesto junto con las principales limitaciones del trabajo y delineando posibles líneas de investigación futura.

1.4 Alcance del Trabajo

Esta investigación se centra en el desarrollo, implementación y evaluación de un marco metodológico para el análisis de asociación genética con variables mixtas en contextos de alta dimensionalidad. El enfoque propuesto, basado en modelos Elastic Net multivariados para respuestas gaussianas y binarias, y técnicas de reducción de dimensionalidad (PCA y RDA), se aplica y evalúa en dos estudios de caso contrastantes en cuanto a especie, tamaño muestral y características genómicas, incluyendo diferencias en el tipo y número de marcadores moleculares empleados: un conjunto

de accesiones de quinoa y el **Rice Diversity Panel 1 (RDP1)** de arroz, ambos con rasgos cuantitativos y cualitativos. El alcance del trabajo se delimita específicamente a los métodos mencionados, excluyendo expresamente el análisis de modelos multinomiales.

Capítulo 2

Aspectos Biológicos y Agronómicos

2.1 Introducción

La diversidad genética, que conforma el ADN de las especies tiene influencia en la capacidad de los individuos futuros de desarrollar características de adaptabilidad. Los bancos de germoplasma cumplen un rol clave, siendo uno de sus objetivos principales conservar dicha variabilidad genética esencial para la biodiversidad, resguardando los recursos fitogenéticos. El concepto de germoplasma engloba la colección de material hereditario que constituye la base física y genética de una especie (Auditoría General de la Nación [AGN], 2019). Los recursos fitogenéticos son definidos como el material vegetal que posee valor real o potencial para la alimentación y la agricultura (FAO, 2010). Estos bancos tienen como fin primordial la preservación de la riqueza genética de las especies cultivables y sus parientes silvestres, abarcando diversas formas de conservación como bancos de semillas, almacenamiento en cámaras frías, mantenimiento de especies en condiciones naturales y la protección en jardines botánicos o arboretums. Las bases de datos de dichos bancos, guardan una extensa información sobre accesiones, que incluye datos sobre caracteres fenotípicos tanto cuantitativos (como el rendimiento) como cualitativos (como el color), y sobre marcadores moleculares o genotipos. El término genotipo se refiere al conjunto de genes o rasgos hereditarios de un individuo o variedad, mientras que el fenotipo representa la manifestación de estos genes en un entorno específico (Copelli, 2010). Es decir, los genotipos proporcionan la información genética contenida en el ADN, denominándose genoma en el contexto de especies, pool genético a nivel de poblaciones, y combinación de alelos para individuos específicos. Los fenotipos, que son los

rasgos visibles o medibles en el individuo o la población, pueden ser de naturaleza agronómica y variar entre cuantitativos (diámetro) y cualitativos (Color). Según la definición provista por la FAO ([2021a](#)), una accesión es una muestra única y claramente identificable de semillas que representan un cultivar, una línea de cría, o una población, almacenadas con el propósito de su conservación y utilización. Esta definición destaca la importancia de los bancos de germoplasma como reservorios de diversidad genética y como fuentes de información y material genético crítico para la investigación y el desarrollo agrícola. Los marcadores polimórficos se han utilizado desde hace décadas comenzando con los marcadores fenotípicos, continuando por los bioquímicos y más recientemente a nivel del ADN. El incremento del conocimiento biológico ha colaborado en su desarrollo e implementación (Villegas Castagnasso et al., [2021](#)). Actualmente, con las nuevas tecnologías se pueden detectar variaciones ínfimas en el material genético que resultan de utilidad en genética de microorganismos, vegetales y animales. Los conceptos y análisis de desequilibrio de ligamiento, locis de caracteres cuantitativos (QTL), Mapeo Asociativo y estudios de asociación genómica amplia (GWAS), constituyen la base teórica y metodológica para el estudio de la asociación entre fenotipo y caracteres moleculares. A continuación se desarrollan los conceptos mencionados, la cuantificación del desequilibrio de ligamiento mediante medidas estadísticas específicas, y se describirá la implementación de Modelos de mapeo asociativo como estrategia de identificación de marcadores moleculares asociados a caracteres de interés.

2.2 Bancos de Germoplasma

En los recientes años, la biodiversidad genética, que es la riqueza de genes presentes en los seres vivos, ha enfrentado amenazas crecientes debido a la contaminación ambiental, el cambio climático, monocultivo y la introducción de especies exóticas. Esta situación es particularmente preocupante en el ámbito de la flora, donde la variedad genética de las plantas se ve comprometida. La pérdida de esta diversidad genética tiene consecuencias directas en la seguridad alimentaria de futuras

generaciones, ya que reduce la capacidad de adaptación y resistencia de los cultivos a condiciones adversas y enfermedades. Por otro lado, el avance de monocultivos atenta contra la explotación de fuentes adicionales de diversidad, que promuevan una agricultura sostenible y la diversificación de la producción agrícola, a través del uso de nuevos cultivos y especies poco utilizadas (Food and Agriculture Organization of the United Nations, 2008). Para preservar la diversidad genética, se han establecido bancos de germoplasma. Banco de germoplasma es una instalación o unidad destinada a conservar material vegetal vivo (semillas, tejidos, órganos o células), manteniéndolo bajo condiciones controladas que garantizan su supervivencia, viabilidad genética y disponibilidad futura para programas de investigación, mejoramiento o reintroducción de especies (Engels & Visser, 2003). Estas instituciones tienen como función esencial la recolección y conservación a largo plazo de recursos genéticos, como las semillas, asegurando así la disponibilidad y el mantenimiento de la diversidad genética de especies de importancia agrícola y silvestre. Las tareas que llevan a cabo estos bancos, es extensa y abarca desde la documentación detallada y la caracterización morfológica y agronómica de los especímenes, hasta la evaluación de la variabilidad genética y la realización de estudios filogenéticos, representando un esfuerzo significativo tanto en términos de recursos como de costos. Este enfoque proactivo hacia la conservación y el uso sostenible de los recursos fitogenéticos es vital no sólo para contrarrestar las amenazas actuales a la biodiversidad genética, sino también para asegurar la alimentación y el bienestar de las generaciones venideras, permitiendo el desarrollo de cultivos más resilientes y adaptativos frente a los retos ambientales y climáticos del futuro. En nuestro país, los bancos de germoplasma surgieron a mediados del siglo pasado con el INTA. En el año 2016 la Argentina comienza a aplicar el Tratado Internacional sobre Recursos Fitogenéticos para la Alimentación y la Agricultura (TIRFAA), que da acceso gratuito a los investigadores, a los recursos genéticos incorporados al sistema multilateral. El INTA cuenta con la Red de Recursos Genéticos (RedGen) que tiene tres sub redes asociadas a los recursos fitogenéticos, zoogenéticos y microbiológicos. Esta red informó en el año 2017, un total de 58.699 entradas únicas documentadas, de las cuales 27.990

corresponden a recursos fitogenéticos conservados ex situ (AGN, [2019](#)).

2.3 Caracterización de recursos genéticos

Los bancos de germoplasma realizan tareas de caracterización de materiales genéticos que facilita su utilización. Según la FAO, la caracterización de los recursos genéticos comprende la descripción sistemática de accesiones utilizando atributos cuantitativos y/o cualitativos. Este proceso permite identificar duplicaciones, establecer relaciones genéticas y evaluar la variabilidad presente en las colecciones. (FAO, [2010](#), [2012](#)). En la bibliografía consultada esta caracterización se clasifica en:

1. Caracterización morfológica y agronómica
2. Caracterización citogenética
3. Caracterización bioquímica
4. Caracterización molecular

2.3.1 Caracterización morfológica y agronómica

Existen diferencias morfológicas y fisiológicas entre las plantas. La caracterización morfológica se funda en los atributos de un organismo a nivel celular o superior. Los procesos fisiológicos determinan los rasgos morfológicos y anatómicos. Los primeros estudios de las diferencias entre individuos, se basaban en la descripción de características morfológicas (Villegas Castagnasso et al., [2021](#)). En cuanto a la caracterización agronómica se refiere a atributos relacionados con el rendimiento, tolerancia a estrés, precocidad, resistencia a la sequía o a una determinada enfermedad. Los mejoradores, si bien tienen como objetivo seleccionar plantas con buen rendimiento, sanidad y calidad de cultivo, tradicionalmente basaban su elección en la observación de rasgos morfológicos. Los primeros mejoradores dependían de su intuición, y observación para realizar este trabajo. Utilizaban el método tradicional, para implementar una caracterización morfológica, que consistía en la identificación de especies vegetales, basándose en características fenotípicas visibles, tales como el

color de la hoja o del fruto. Como lo expresa Acquaaah (2012), los marcadores morfológicos son útiles cuando se lleva a cabo la selección visual, resultando ventajosa una aguda capacidad de observación para el proceso de selección. Este enfoque ha sido fundamental en el avance de la genética, contribuyendo al entendimiento del ligamiento genético y facilitando la creación de los primeros mapas genéticos en el ámbito vegetal. No obstante, este tipo de caracterización enfrenta diversas limitaciones, como la influencia ambiental y el tiempo requerido en la recolección de datos para su evaluación, ya que es indispensable disponer de la planta en una etapa de desarrollo avanzada, y en muchas ocasiones, en su fase adulta. Además la observación personal lleva a descripciones subjetivas que podían constituir una fuente de errores involuntarios por parte de los mejoradores. Por otro lado para muchos de éstos caracteres se desconoce su base genética, siendo algunos poligénicos. Si bien la caracterización morfológica es una herramienta valiosa en los estudios genéticos y de biodiversidad, sus limitaciones llevan a complementarla con otras técnicas más sofisticadas y precisas para la identificación y clasificación de las variaciones genéticas en el reino vegetal.

2.3.2 Caracterización citogenética

Como lo define Nikoloff y Ruiz de Arcaute (2021), la citogenética es la rama de la genética que se centra en el estudio de los cromosomas, incluyendo su estructura, número, morfología y comportamiento durante la división celular. En el contexto de los recursos fitogenéticos, ha impactado en la evaluación y utilización de la biodiversidad, en el mejoramiento genético y la ingeniería genética, permitiendo la localización de secuencias específicas, la identificación de cromosomas particulares, el estudio del origen y la estructura de los genomas de híbridos y poliploides, la comparación de regiones genómicas homólogas, el estudio del comportamiento cromosómico y el análisis de la expresión génica (Seijo et al., 2004).

2.3.3 Caracterización bioquímica

En años más recientes, el avance en el uso de marcadores bioquímicos, a través del análisis de compuestos químicos específicos, ha sido una alternativa a los métodos tradicionales de caracterización morfológica. Mediante esta técnica se conocen las características relacionadas con isoenzimas y polimorfismos de proteínas. A continuación, se detallan los marcadores bioquímicos basados en las isoenzimas y las proteínas de reserva. Se sigue el marco teórico de Martínez et al. (2004).

1. Marcadores basados en Isoenzimas: Las isoenzimas fueron originalmente definidas como formas moleculares múltiples de las enzimas que tienen funciones idénticas o similares y están presentes en el mismo individuo (Markert & Møller, 1959). La identificación de estas variantes es posible gracias a diferencias en su movilidad electroforética, lo cual permite detectar con precisión el impacto de modificaciones alélicas. Factores como el número de genes que codifican estas enzimas, los distintos alelos presentes, la estructura cuaternaria de los productos proteicos y su localización subcelular definen un patrón de bandas específico, conocido como zimograma. Una de las ventajas de este método es su capacidad para mostrar codominancia; es decir, la expresión simultánea de ambos alelos de un locus en un individuo. Esto facilita la comparación entre especies, poblaciones de una misma especie, y la detección de híbridos o la introgresión de genes (Paredes & Gepts, 1995). Su simplicidad, bajo costo y utilidad en estudios poblacionales vegetales para analizar la estructura y variabilidad genética son puntos fuertes y significativos. Sin embargo, presenta algunas desventajas, como un bajo nivel de polimorfismo. Además, cuando se analiza material de hojas y raíces (en lugar de semillas), las variaciones enzimáticas pueden verse influenciadas por las condiciones ambientales y la edad del tejido, lo que puede comprometer la reproducibilidad de los patrones de bandas observados.
2. Proteínas de reserva: Las proteínas de reserva son un conjunto heterogéneo

de proteínas presentes en las semillas de las plantas, desempeñando un papel crucial al suministrar energía al embrión durante sus etapas iniciales de crecimiento. Estas proteínas pueden ser extraídas de las semillas mediante métodos sencillos y luego separadas a través de electroforesis, un proceso que permite su visualización sin necesidad de detectar actividad enzimática específica. Similar a los marcadores basados en isoenzimas, la utilización de proteínas de reserva como marcadores genéticos es de fácil implementación en el laboratorio, además de ser un método de bajo coste y rápida ejecución. No obstante, una limitación importante de este enfoque es su baja cobertura del genoma. Esto significa que no se puede presumir independencia entre los loci, ya que muchos de los genes que codifican estas proteínas tienden a estar físicamente cercanos unos de otros en el genoma. Esta proximidad genética puede limitar la utilidad de las proteínas de reserva para algunas aplicaciones de mapeo genético y análisis de la diversidad, al reducir la capacidad de detectar variabilidad genética independiente entre diferentes loci.

2.3.4 Caracterización molecular. ADN

En años recientes, los bancos de germoplasma han ampliado su enfoque tradicional en caracterización agronómica y morfológica para incluir técnicas de caracterización molecular. Un marcador molecular se define como una secuencia de ADN localizada en un punto específico de un cromosoma, que puede formar parte de regiones codificantes o no codificantes del genoma. Un buen marcador se tiene que poder visualizar y medir. Un marcador molecular es un marcador genético. Según Rojas-Beltrán (2007) un marcador genético es una característica heredable, polimórfica, fácilmente visualizada o medida y utilizada en análisis genéticos, como por ejemplo: en estudios de diversidad genética, mapeos genéticos, revelar características de importancia económica o agronómica ligadas al marcador (QTLs). El polimorfismo se refiere al grado de variación de la característica (diferentes formas alternativas del carácter). Estos marcadores moleculares, biomoléculas asociadas a rasgos genéticos heredables, abarcan desde proteínas específicas como antígenos e isoenzimas hasta

diversas formas de ADN, ya sean genes de función conocida o segmentos de secuencia sin función determinada.

Con los avances en biología molecular, se han desarrollado métodos sofisticados para detectar polimorfismos genéticos directamente en el ADN, permitiendo una clasificación detallada de los tipos de marcadores moleculares según las técnicas empleadas en su detección. Estas técnicas modernas comienzan con la extracción del ADN genómico de la planta objeto de estudio, abriendo camino a una comprensión más profunda y precisa de la diversidad genética almacenada en los bancos de germoplasma (Martínez et al., 2004). Este enriquecimiento de la caracterización genética del germoplasma conservado fortalece la capacidad de investigación y desarrollo en el campo de la biología vegetal y la agronomía. Se sigue la propuesta de clasificación de marcadores moleculares de Martínez et al. (2004): marcadores moleculares y marcadores basados en SNP.

2.3.4.1 Marcadores moleculares

A continuación se hace un breve resumen de estas técnicas de extracción de ADN, según lo detallado en los marcadores moleculares por Martínez et al. (2004):

1. Marcadores basados en la hibridación del ADN:

- RFLP, siglas en inglés de Restriction Fragment Length Polymorphism (Polimorfismo en la Longitud de Fragmentos de Restricción), representa uno de los métodos más antiguos en la detección de variabilidad genética. Este método se basa en la comparación de perfiles de bandas obtenidos del ADN de distintos individuos tras su digestión con enzimas de restricción específicas. Las diferencias en los tamaños de los fragmentos resultantes reflejan variaciones genéticas entre los individuos, permitiendo así su identificación y clasificación basada en la presencia o ausencia de secuencias de reconocimiento para dichas enzimas.
- VNTR, sigla de Variable Number of Tandem Repeats (Número Variable de Repeticiones en Tándem), se conocen como minisatélites, consisten en

secuencias de ADN compuestas por repeticiones de 9 a 100 pares de bases. Estas secuencias, que raramente exceden las 1000 bases, son notables por su variabilidad en el número de repeticiones entre diferentes individuos, lo que las convierte en marcadores genéticos útiles. La detección de VNTR puede realizarse mediante dos métodos principales: hibridación, que implica la unión de secuencias de ADN complementarias, o técnicas de PCR (de sus siglas en inglés: Polymerase Chain Reaction, Reacción en Cadena de la Polimerasa), una metodología que permite la amplificación selectiva de segmentos de ADN, facilitando su identificación y análisis. Estas técnicas aprovechan la variabilidad de los VNTR para investigaciones en genética poblacional, forense y en estudios de paternidad

2. Marcadores basados en la amplificación arbitraria o semi arbitraria de ADN

- RAPD, acrónimo de Random Amplified Polymorphic DNA (ADN Polimórfico Amplificado Aleatoriamente), fueron desarrollados y popularizados en la investigación genética a finales de los años 80 y principios de los 90. Esta metodología fue descubierta por varios grupos de investigación, destacándose el trabajo realizado por Williams et al. ([1990](#)). Esta técnica, una variante de PCR, se caracteriza por ser un método de detección genética de bajo costo, sencillo en su ejecución, automatizable y no radioactivo. Los RAPD se valen de cebadores cortos de secuencia aleatoria para amplificar regiones específicas del ADN, lo que permite identificar variaciones genéticas sin necesidad de información previa sobre la secuencia del genoma. Ofrece ventajas significativas por su simplicidad y bajo costo. Su principal limitación radica en la baja reproducibilidad de los resultados entre diferentes laboratorios, un aspecto que puede afectar la comparabilidad y fiabilidad de los datos obtenidos mediante esta técnica. Este marcador es dominante debido a que no permite diferenciar individuos heterocigotas.
- AFLP, sigla de Amplified Fragment Length Polymorphism (Polimorfismo

en la Longitud de Fragmentos Amplificados), son marcadores que destacan por su capacidad excepcional para detectar variabilidad genética. La eficacia de los AFLP radica en su habilidad para identificar la presencia o ausencia de fragmentos de ADN de un tamaño específico. Aunque los AFLP comparten ciertas similitudes con otros marcadores, como los RAPD, en cuanto a que no permiten la detección directa de heterocigotos —actuando así como marcadores dominantes—, se diferencian en términos de robustez y fiabilidad. La implementación de AFLP requiere de una inversión moderada a alta y depende de la disponibilidad de infraestructura de laboratorio avanzada. Sin embargo, esta inversión se justifica por el alto nivel de resolución y la capacidad de estos marcadores para generar información detallada sobre la estructura genética y la diversidad dentro de las poblaciones estudiadas, superando en robustez y aplicabilidad a los marcadores RAPD (Martínez et al., 2004)

3. Marcadores moleculares basados en la amplificación sitio-específica de ADN: Como mencionan Martínez et al. (2004), éstos marcadores representan una avanzada categoría de herramientas genéticas que se distinguen de otros marcadores previamente mencionados, los cuales se enfocan en la amplificación general de ADN. Esta metodología exige el diseño cuidadoso de primers específicos que permiten la amplificación dirigida de un locus genético particular. Este enfoque selectivo posibilita una mayor precisión en la identificación y caracterización de variantes genéticas, al enfocarse exclusivamente en regiones de interés dentro del genoma. La capacidad de estos marcadores para aislar y amplificar secuencias específicas los hace excepcionalmente útiles en estudios de mapeo genético, diagnóstico de enfermedades y análisis de la diversidad genética, ofreciendo una mirada detallada y precisa sobre la estructura y función de los genes estudiados.

- Los microsatélites, también conocidos como SSR (Simple Sequence Repeats, Repeticiones Simples de Secuencia), SSRP (Simple Sequence Re-

peat Polymorphisms, Polimorfismos de Repeticiones Simples de Secuencia), SSLP (Simple Sequence Length Polymorphisms, Polimorfismos de Longitud de Secuencias Simples) o STMS (Sequence Tagged Microsatellite Sites, Sitios Microsatélites Etiquetados por Secuencia), son secuencias repetitivas cortas que se encuentran ampliamente distribuidas en el genoma de las especies eucariotas. La identificación de estos marcadores se realiza mediante PCR, técnica que permite distinguir entre diferentes microsatélites basándose en el número de repeticiones. Esta diferenciación se logra utilizando primers específicos, pequeños fragmentos de ADN de 20 a 30 pares de bases, que facilitan la amplificación de las secuencias de interés. Posteriormente, los fragmentos amplificados son separados por su longitud mediante electroforesis en gel de agarosa, con el tamaño de la banda reflejado por el número de repeticiones en tándem. Los microsatélites son marcadores codominantes, es decir, en cada locus un individuo puede presentar uno o más alelos, dependiendo de su dotación cromosómica. Esta característica los convierte en herramientas valiosas para el estudio de especies autóгамas, como el trigo, debido a su alto nivel de polimorfismo por locus. Ofrecen una mayor especificidad genómica y polimorfismo en comparación con métodos como los RFLP y RAPD, posicionándolos como marcadores ideales para una amplia gama de aplicaciones genéticas, desde análisis de parentesco hasta estudios poblacionales, gracias a su capacidad para generar perfiles genéticos únicos o huellas genéticas (Martínez et al., 2004)

Los microsatélites requieren una infraestructura y presupuesto comparables a los AFLP, siendo más exigentes que los RAPD. Una limitación a considerar es que la eficacia de amplificación de los loci puede disminuir a medida que aumenta la distancia genética entre las secuencias de interés, lo que podría afectar la resolución de algunos estudios genéticos.

- Otros marcadores basados en Microsatélites: Con el fin de diseñar pri-

mers específicos se debe conocer las secuencias que flanquean las repeticiones. Se han introducido innovaciones en el ámbito de los marcadores moleculares que permiten la exploración de las repeticiones de los microsatélites sin necesidad de secuenciar su ADN. Entre estos avances se encuentra el desarrollo del MP-PCR (Microsatellite Primer PCR), una técnica que emplea un primer específico que incluye repeticiones de un microsatélite determinado, facilitando así la amplificación del ADN objetivo. En la misma línea de innovación, se encuentran los marcadores ISSR (Inter-SSR amplification) o microsatélite anclado (AMP-PCR, Anchored Microsatellite-Primer PCR), que ofrecen una estrategia alternativa para la amplificación de microsatélites sin la necesidad previa de secuenciación. Otra técnica valiosa que amplía las posibilidades de identificación y caracterización genética es la SAMPLE (Selective Amplification of Microsatellite Polymorphic Loci). Estas metodologías avanzadas permiten un análisis genético más profundo y accesible, especialmente en contextos donde la información de secuencia es limitada.

- Los STS (Sequence Tagged Sites, Sitios Marcados por Secuencia), junto con sus variantes, los SCAR (Sequence Characterized Amplified Region, Región Amplificada Caracterizada por Secuencia) y los CAPS (Cleaved Amplified Polymorphic Sequence, Secuencia Polimórfica Amplificada y Clivada), son herramientas esenciales en la selección asistida por marcadores genéticos. Destacan por su metodología directa y su alta capacidad de reproducibilidad, lo cual facilita su implementación y transferencia entre distintos laboratorios de investigación. Su simplicidad de aplicación permite una detección eficaz y precisa de variantes genéticas específicas. Son marcadores indispensables en estudios de mapeo genético, selección de rasgos de interés en programas de mejora y en investigaciones de diversidad genética, proporcionando una base sólida y fiable para el avance de proyectos genéticos en una variedad de campos.

2.3.4.2 Marcadores SNP (Single Nucleotide Polymorphisms)

Los autores Martínez et al. (2004) observan que si bien estos marcadores son marcadores moleculares que se podrían incluir en el punto anterior, se los trata en otro apartado debido a que los SNP pueden derivar tanto de secuencias genómicas codificantes como no codificantes. Los marcadores moleculares denominados SNP, es decir polimorfismo de nucleótido, son empleados en diversos estudios de genética. Estos se basan en la detección de polimorfismos resultantes de la alteración de una única base en una secuencia de ADN. Aunque muchos SNP no causen efectos directos pueden estar cerca de genes relevantes y servir como indicadores independientes. Estos marcadores tomaron relevancia a partir la obtención de la secuenciación del genoma humano, en la que se descubrió que el 90 % de polimorfismo encontrado en el genoma eran SNP. Con la secuenciación a gran escala, en el área vegetal, se detectó la presencia de SNP en diferentes especies tales como el melón, la soja, el girasol, el maíz entre otras. Este tipo de marcadores son bialélicos y se pueden localizar en regiones codificantes, regulatorias y no codificantes. Su presencia en regiones codificantes muestran 100 % de asociación con el carácter de interés, siendo de utilidad en selección asistida y en el aislamiento de genes.

2.3.4.3 Secuenciación de ADN

La secuenciación de ADN es el proceso mediante el cual se determina el orden exacto de los nucleótidos en una molécula de ADN. Tal como lo nombra Acquaah (2012), la secuenciación de ADN no es un marcador genético, es una herramienta esencial de la investigación y el mejoramiento vegetal por su alta resolución y ausencia de sesgos. La secuenciación de ADN experimentó un avance decisivo en 1977 con el desarrollo de dos métodos pioneros: el método químico de Maxam y Gilbert (Maxam & Gilbert, 1977) y el método de terminación de cadena desarrollado por Sanger et al. (1977). Este último resultó más simple, seguro y práctico, convirtiéndose en la técnica predominante durante varias décadas. Este método dio origen a dos enfoques clásicos de secuenciación descritos por Acquaah (2012): el enfoque clon por clon, que fragmenta el genoma en partes conocidas y facilita el ensamblaje aunque

omite regiones repetitivas, y el enfoque shotgun, que secuencia el genoma completo con redundancia, pero presenta mayores dificultades de ensamblaje. Según Acquaah (2012), la secuenciación de genomas vegetales es costosa en comparación con otros sistemas de marcadores y enfrenta desafíos adicionales, tales como el gran tamaño de muchos genomas, la presencia de grandes cantidades de ADN repetitivo y la poliploidía (duplicaciones de su genoma, 3 o más juegos de cromosomas). La concepción de la secuenciación de Acquaah (2012), gracias al avance de las tecnologías ha evolucionado hacia lo que se denomina secuenciación de nueva generación o de sus siglas en inglés NGS (Next Generation Sequencing) que está orientado hacia el descubrimiento de marcadores moleculares. Tal como lo expresan Panahi et al. (2024) la NGS ha transformado la genómica vegetal al ofrecer tecnologías de secuenciación de alto rendimiento y bajo costo, que permiten identificar rápidamente variantes genéticas como SNP, indel, SSR y variantes estructurales. Estas tecnologías han superado las limitaciones de métodos tradicionales como la secuenciación Sanger, facilitando así el avance de programas de mejoramiento genético y de estudios de diversidad genética. Asimismo, Panahi et al. (2024) discuten estrategias como RNA-Seq, GBS (Genotyping by Sequencing) y RAD-seq (Restriction site Associated DNA sequencing), las cuales han permitido desarrollar marcadores a gran escala, especialmente en especies de interés agrícola como maíz, soja, arroz, trigo y otros cultivos. Los autores hacen énfasis en la fusión de bioinformática, inteligencia artificial (IA) y aprendizaje automático (ML, por sus siglas en inglés de Machine Learning) para el análisis de datos de NGS, lo que permite optimizar el descubrimiento de marcadores y potenciar la selección asistida por marcadores.

2.4 Desequilibrio de ligamiento

La posibilidad de identificar relaciones entre marcadores moleculares y rasgos fenotípicos depende en gran medida del *ligamiento genético*, entendido como la tendencia de loci cercanos en un cromosoma a heredarse juntos debido a la baja probabilidad de recombinación entre ellos. Este concepto, que amplió y reformuló la

visión mendeliana de segregación independiente, constituye la base de los estudios de *desequilibrio de ligamiento* (LD, por sus siglas en inglés de Linkage Disequilibrium) y en consecuencia del mapeo asociativo. Tal como señalan Kearsey y Pooni (1996), la ley de segregación independiente sólo se cumple para genes localizados en cromosomas distintos o suficientemente alejados en un mismo cromosoma; en caso contrario, se observa ligamiento. A nivel poblacional, este fenómeno se traduce en asociaciones no aleatorias entre alelos de distintos loci, lo que se conoce como LD.

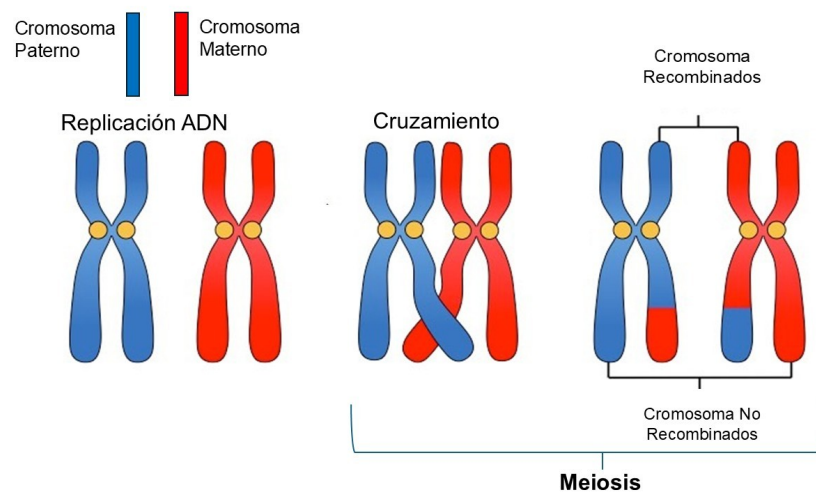


Figura 2.1: Esquema del entrecruzamiento durante la meiosis.

Durante la meiosis, los cromosomas transmitidos a la descendencia no son copias exactas de los parentales, ya que ocurren eventos de recombinación genética. En este proceso, las enzimas cortan y vuelven a unir el material genético, tal como se ilustra en la Figura 2.1, permitiendo que los genes de los cromosomas homólogos se mezclen y den lugar a gametos con combinaciones nuevas de alelos. Dichas combinaciones constituyen una fuente fundamental de variabilidad genética en las poblaciones. Sin embargo, cuando dos loci están físicamente próximos, la probabilidad de que sean separados por recombinación disminuye y tienden a heredarse como un bloque. En este sentido, Flint-Garcia et al. (2003) destacan su papel central en los estudios de *association mapping*, mientras que Zhu et al. (2008) subrayan cómo este enfoque, basado en los patrones de LD, se ha convertido en una estrategia prometedora para identificar loci asociados a caracteres cuantitativos y apoyar el mejoramiento de

cultivos.

2.4.1 Medidas de desequilibrio de ligamiento

La medida clásica de LD se representa con la letra D , definida como:

$$D = f(AB) - f(A)f(B) \quad (2.4.1)$$

donde $f(AB)$, $f(A)$ y $f(B)$ denotan las frecuencias poblacionales del haplotipo AB y de los alelos A y B , respectivamente. En la práctica, estas frecuencias se estiman a partir de frecuencias muestrales. Si $D = 0$, los alelos se consideran independientes; valores distintos de cero indican desequilibrio.

No obstante, el rango de D depende de las frecuencias alélicas, lo que dificulta su interpretación. Para resolverlo, Lewontin (1964) y Lewontin y Hubby (1966) propusieron la medida normalizada D' :

$$D' = \frac{D}{D_{\text{máx}}} \quad (2.4.2)$$

donde $D_{\text{máx}}$ es el valor máximo posible de D , en función de las frecuencias de los alelos. Así, D' se encuentra siempre entre -1 y 1 , independientemente de las frecuencias alélicas, y permite comparar distintos pares de loci. Un valor de D' igual a 1 o a -1 indica un desequilibrio de ligamiento completo.

Posteriormente, Hill y Robertson (1968) propusieron utilizar el cuadrado del coeficiente de correlación entre las frecuencias alélicas de dos loci (denotado como r^2) como otra medida de LD:

$$r^2 = \frac{D^2}{f(A)f(a)f(B)f(b)} \quad (2.4.3)$$

Este estadístico oscila entre 0 y 1 e indica la proporción de variabilidad genética explicada por el ligamiento entre dos loci. Mientras que D' refleja principalmente la historia recombinacional, r^2 incorpora tanto efectos de la recombinación como de la mutación y, por ello, suele preferirse en estudios de asociación (Gupta et al. (2005),

entre otros).

Además de estas medidas, el LD puede representarse gráficamente de forma resumida mediante gráficos de decaimiento o matrices de LD. En síntesis, las medidas D , D' y r^2 constituyen herramientas complementarias para describir el desequilibrio de ligamiento. Su correcta aplicación permite avanzar en la detección de asociaciones genéticas, facilitando la identificación de polimorfismos moleculares vinculados a caracteres agronómicos de interés.

2.4.2 Aplicaciones en mapeo asociativo

En estudios de asociación en plantas suele analizarse un gran número de polimorfismos dentro de un mismo gen o región genómica, dado que el polimorfismo causal rara vez es conocido de antemano. El mapeo asociativo se basa en la acumulación histórica de recombinación y en el patrón de desequilibrio de ligamiento presente en poblaciones naturales para identificar variantes genéticas relacionadas con caracteres complejos, sin necesidad de generar poblaciones experimentales específicas (Zhu et al., 2008).

Sin embargo, la extensión y distribución del LD en la población bajo estudio condicionan directamente la *resolución* (qué tan cerca del marcador está el polimorfismo causal) y la *potencia estadística* (capacidad de detectarlo). Por ello, caracterizar los patrones de LD constituye un paso previo indispensable en todo análisis de mapeo asociativo, ya que permite evitar asociaciones espurias y mejorar la interpretación biológica de los resultados (A. N. Peña Malavera, 2015).

En este contexto, Rogers y Huff (2009) demostraron que, en ausencia de haplotipos completos, el coeficiente de correlación de Pearson calculado sobre el número de alelos por locus en individuos diploides es una buena aproximación a la correlación gamética entre loci, siendo su r^2 equivalente al estadístico de LD normalizado. Más recientemente, Sharma et al. (2018) extendieron este enfoque a datos poliploides, mostrando que el uso de la correlación de Pearson sobre conteos de alelos resulta una herramienta práctica tanto para evaluar la estructura poblacional como para desarrollar modelos de mapeo asociativo más precisos.

2.5 QTL, Mapeo Asociativo y GWAS

Los caracteres importantes en la agricultura, tales como el rendimiento, la calidad de los productos y la intensidad a enfermedades, son controlados por múltiples genes y son conocidos como caracteres cuantitativos. Estos también se describen como poligénicos, multifactoriales o complejos. Las regiones genómicas que contienen genes asociados con estos caracteres se denominan loci de caracteres cuantitativos (QTL). Un QTL influye en la variación de rasgos cuantitativos en la población y se encuentra asociado con características específicas que varían de manera continua, como la altura, el peso o el rendimiento agrícola, y son afectados tanto por múltiples genes como por factores ambientales (Collard et al., 2005). Los caracteres cuantitativos se diferencian de los rasgos cualitativos, que son influenciados por uno o pocos genes con efectos grandes y bien definidos, resultando en variaciones discretas, como el color de los ojos (Kearsey & Pooni, 1996). La correlación directa entre varios marcadores moleculares y un rasgo genético específico facilita el establecimiento de un QTL. La selección asistida por marcadores (MAS), utiliza marcadores moleculares como sustitutos para asistir en la selección fenotípica, ofreciendo una manera más eficiente, efectiva, confiable y rentable en comparación con los métodos convencionales de mejoramiento (Collard et al., 2005; Moose & Mumm, 2008; Xu, 2010). Esta técnica permite realizar selecciones basadas en características agronómicas específicas, independientemente de las condiciones fenotípicas y ambientales. El mapeo de ligamiento es la técnica experimental predominante para identificar QTL, basada en la premisa de que la frecuencia de recombinación entre dos marcadores disminuye a medida que están más cercanos en el cromosoma, facilitando la localización precisa de regiones cromosómicas asociadas con rasgos específicos (Martínez et al., 2004). Alternativamente, el MA se basa en aprovechar la diversidad genética natural y la recombinación histórica para detectar asociaciones genotipo-fenotipo (Zhu et al., 2008). Esta metodología estadística permite la identificación de regiones genómicas asociadas con variaciones en rasgos cuantitativos, a través de su asociación con marcadores moleculares. Es fundamental aplicarla con rigurosidad para asegurar la

precisión y la relevancia de los hallazgos. Por otro lado los estudios de asociación de todo el genoma (GWAS, Genome-Wide Association Studies) son una forma especial de mapeo asociativo, aplicada a grandes conjuntos de marcadores de todo el genoma (Zhu et al., 2008). GWAS, al igual que el MA, busca correlaciones entre variaciones genéticas y rasgos fenotípicos. Estos estudios de asociación se han revelado como una herramienta poderosa para investigar rasgos complejos en cultivos.

GWAS resulta ser un enfoque integral para buscar sistemáticamente en el genoma variaciones genéticas causales. Se prueba un gran número de marcadores para su asociación con varios rasgos complejos, y no se requiere información previa sobre genes candidatos (Zhu et al., 2008). Esto contrasta con los estudios de *genes candidatos*, donde sí se seleccionan previamente loci de interés basados en conocimientos previos de rutas bioquímicas o análisis mutacionales.

En este contexto, es importante diferenciar los enfoques. Los QTL son las regiones genómicas donde se asocia variación genética con variación fenotípica, identificados clásicamente a partir de poblaciones biparentales mediante análisis de ligamiento. El MA, siguiendo a A. N. Peña Malavera (2015), se define como una estrategia basada en el desequilibrio de ligamiento que aprovecha la recombinación histórica en poblaciones naturales para detectar asociaciones entre variantes genéticas y caracteres, sin necesidad de poblaciones experimentales derivadas de cruzamientos específicos (Yu & Buckler, 2006). De este modo, puede entenderse la relación jerárquica entre estos conceptos: los QTL son las unidades de interés, el MA es la metodología que explota la variación natural y el desequilibrio de ligamiento para localizarlos, y los *GWAS* representan la aplicación a escala genómica completa de esta estrategia (A. N. Peña Malavera, 2015; Yu & Buckler, 2006; Zhu et al., 2008). La figura 2.2 representa la relación de estos conceptos y técnicas.

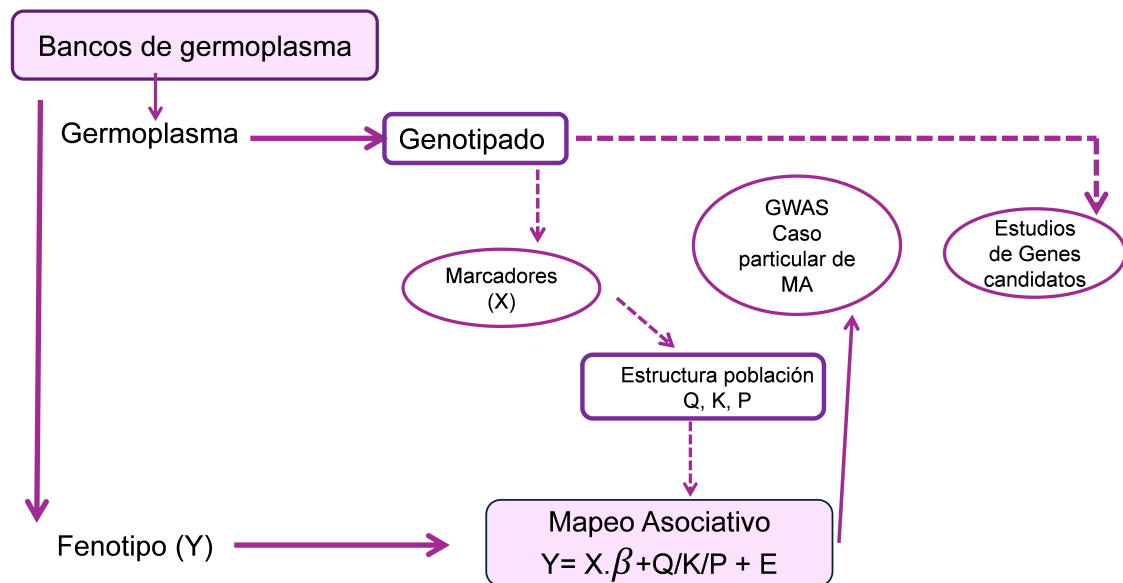


Figura 2.2: Esquema de enfoques de asociación: **MA** puede implementarse como **GWAS**, explorando sistemáticamente todo el genoma sin hipótesis previas, o bien mediante **estudios de genes candidatos**, que se centran en loci seleccionados a priori por conocimiento biológico. Adaptado de la Figura 3 de Zhu et al. (2008).

Estas técnicas representan herramientas fundamentales en la genética moderna para mejorar las características agrícolas, demostrando cómo la integración de enfoques moleculares y estadísticos puede acelerar y refinar el mejoramiento vegetal.

Capítulo 3

Aspectos Estadísticos

La Estadística como ciencia, tiene como objetivos principales describir, predecir y modelizar eventos del mundo real. Con estos fines se han desarrollado una diversidad de técnicas y métodos que permiten describir, visualizar y explicar conjuntos de datos cada vez más complejos. Entre éstas técnicas, se destaca el análisis Multivariado por su capacidad para abordar la interdependencia entre múltiples variables, mientras que los modelos lineales y no lineales resultan de interés para explicar de manera rigurosa relaciones entre variables. A continuación, se exponen algunas de estas metodologías, que fundamentan el desarrollo del presente trabajo doctoral.

3.1 Análisis Multivariado.

En los últimos años, el desarrollo y la accesibilidad a dispositivos como computadoras, tabletas, celulares, secuenciadores de ADN, y satélites de observación terrestre, ha generado de manera automática gran volumen de datos, tales como datos de consumidores, tendencias de compras, consultas en paginas web, datos moleculares, imagenes satelitales, que resultan útiles en la toma de decisiones en cada área de interés. Gran parte de estos datos, son almacenados en la nube y en servidores, poseen gran volumen y complejidad y requieren de métodos de análisis más sofisticados, como las técnicas multivariantes. Estas permiten transformar grandes volúmenes de información en conocimiento. Muchas de las técnicas desarrolladas en análisis multivariado son una extensión de técnicas univariadas. Tal como lo expresa Anderson (2003), entre los principales resúmenes estadísticos en el análisis multivariado se incluyen el vector de medias, la matriz de covarianzas y la matriz de correlaciones, que proporcionan información esencial sobre el comportamiento con-

junto de las variables. El análisis multivariado se focaliza en el estudio simultáneo de múltiples variables, que se miden a un conjunto de individuos u objetos. Tal conjunto puede ordenarse en una matriz, en la que cada fila representa un individuo u objeto y cada columna una variable. El vector fila de la matriz, corresponde a un punto en un espacio euclidiano, donde sus coordenadas representan un individuo u objeto mientras que las columnas son variables y se representan como vectores. La medición y análisis de la dependencia entre variables, entre conjuntos de variables, y entre variables y conjuntos de variables son fundamentales al análisis multivariado (Anderson, 2003). Cuadras (2019) define el análisis multivariado (AM) como una parte de la estadística y del análisis de datos que estudia, analiza, representa e interpreta los datos que resultan de observar más de una variable estadística sobre una muestra de individuos. En el contexto de fenómenos multidimensionales, el objetivo, por un lado, es reducir la dimensionalidad con la mínima pérdida de información y, por otro, agrupar y clasificar individuos u objetos con características afines, identificar relaciones entre las variables o llevar a cabo inferencias sobre parámetros multivariados. Entre los métodos clásicos de AM que trabajan con matrices de individuos por variables, esto es con datos de “dos vías o dos modos” se encuentran, entre otros, el Análisis de Componentes Principales (ACP), Análisis de Correspondencias (AC), Escalado Multidimensional, Análisis Discriminante, Análisis Factorial Múltiple (AFM), el conjunto de técnicas para la obtención de árboles, clúster o dendrogramas, y los modelos lineales múltiples multivariados. En tales situaciones la información se representa mediante una “matriz de dos vías” de orden $n \times p$, cuyas filas y columnas recogen los datos de n individuos y p variables respectivamente. Sin embargo, en muchas experiencias, las observaciones de las variables pueden ser realizadas bajo distintas situaciones o momentos en el tiempo, es decir, en q distintas condiciones. En tales casos se tiene para cada situación o momento una tabla de datos de dos vías; por consiguiente la información puede representarse mediante k matrices de orden $n \times p$, como también es posible concentrar la información proveniente de las diferentes tablas en una única matriz de tres vías o tres modos de orden $n \times p \times q$ donde cada elemento representa la obser-

vación originada por tres modos: individuos, variables y condiciones; constituyendo una matriz cúbica o tabla de tres entradas, cuyo elemento genérico es $\{x_{ijk}\}$ con $i = 1, \dots, n$; $j = 1, \dots, p$ y $k = 1, \dots, q$. La tabla $X = \{x_{ijk}\}$ puede considerarse como una sucesión de tablas de dos modos X_k , cada una de las cuales está referida a una condición, ocasión experimental o momento diferente. Por ello, se puede expresar como la sucesión $X_1, X_2, \dots, X_q$, donde x_{ij} es el elemento genérico de cada X_k . Kiers (1991) y posteriormente Kroonenberg (2008) diferencian los términos ‘modos’ y ‘vías’, asignándole al primero las dimensiones estructurales del conjunto de datos y al segundo el número de dimensiones que se están considerando simultáneamente en el análisis, es decir, indica la cantidad de subíndices necesarios para identificar un individuo en la base de datos. De esta forma clasifican a los datos de tres vías en «Datos de Tres Vías – Tres Modos» (*Three Way – Three Mode Data*) donde cada dato es originado por tres clases o entidades distintas (modos), como por ejemplo individuos, variables y condiciones, es decir, existe un solo conjunto de individuos, un solo conjunto de variables y un solo conjunto de condiciones; y en «Datos de Conjuntos Múltiples» o «Tres Vías – Dos Modos», cuando uno de los modos está compuesto por dos conjuntos, pudiendo presentarse dos situaciones diferentes:

1. Dos conjuntos de variables y un mismo número de individuos y condiciones. Es decir, en cada condición se miden diferentes variables a los mismos individuos.
2. Dos conjuntos de individuos, un solo conjunto de variables y un solo conjunto de condiciones. En este caso, en cada condición se miden las mismas variables a diferentes grupos de individuos.

Los Datos de Conjuntos Múltiples se diferencian de los Datos de Tres Vías porque los tres modos no están totalmente cruzados. A la extensión de una base de datos a más de tres vías recibe la denominación de «Datos Multivía» o «Multiway Set of Data».

Dentro de los métodos multivariados a tres vías se encuentra el AFM, STATIS, Análisis Multivariado de Medidas Repetidas y Análisis de Redundancia, entre otros. En la Figura 3.1 se muestra como podrían clasificarse las técnicas multivariadas



Figura 3.1: Clasificación del Análisis Multivariado. *Fuente: Adaptada de los apuntes del curso "Métodos Estadísticos para el Análisis de datos en estructuras de tres vías", dictado por el Dr. Bramardi S. y Lavalle A. en el Doctorado en Estadística, Universidad Nacional de Rosario.*

según distintos criterios y en la Figura 3.2 como se clasifican usualmente en función de la estructura de la base de datos a analizar.

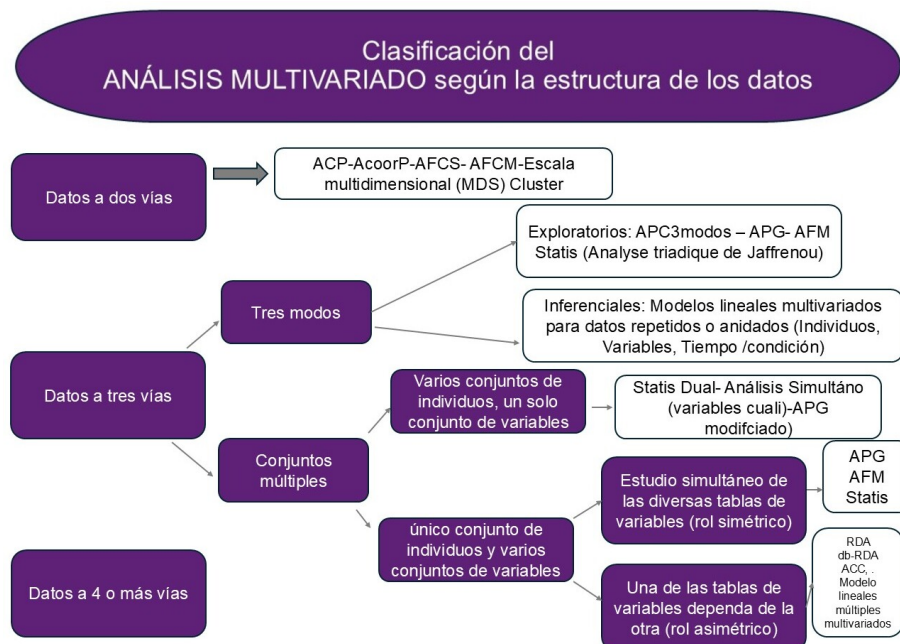


Figura 3.2: Clasificación del análisis multivariado según la estructura de los datos. *Fuente: Adaptada de los apuntes del curso "Métodos Estadísticos para el análisis de datos en estructuras de tres vías" dictado por Dr. Bramardi S. y Lavalle A. Doctorado en Estadística. Universidad Nacional de Rosario.*

Como explica Cuadras (2019), el análisis multivariado se aplica a variables que son homogéneas y correlacionadas, ya que su fundamento es explotar las relaciones de dependencia entre ellas. En ausencia de correlación (cuando la matriz de correlaciones coincide con la identidad), estos métodos pierden sentido, pues no existe estructura que interpretar. Por lo tanto, antes de aplicar un análisis multivariado

resulta fundamental verificar si las variables están suficientemente correlacionadas entre sí como para justificar una reducción de dimensión. Para ello, se emplea el **test de esfericidad de Bartlett**.

3.1.1 Test de esfericidad de Bartlett

El test es una prueba estadística que evalúa si la matriz de correlación entre las variables difiere significativamente de la matriz identidad. Bajo la hipótesis nula, se asume que las variables no están correlacionadas, es decir, que son ortogonales. Rechazar esta hipótesis sugiere que existe suficiente correlación entre las variables como para aplicar técnicas de reducción de dimensión como el ACP. Formalmente, el test evalúa:

$$H_0 : \mathbf{R} = \mathbf{I}_p$$

donde $\mathbf{R}$ es la matriz de correlación entre p variables, y $\mathbf{I}_p$ es la matriz identidad de orden p . Si se rechaza H_0 , se concluye que existe estructura correlacional suficiente.

La estadística de prueba fue propuesta originalmente por Bartlett (1950, 1951) y se basa en el determinante de la matriz de correlación. Está dada por:

$$\chi^2 = - \left(n - 1 - \frac{2p + 5}{6} \right) \ln |\mathbf{R}|$$

donde n es el número de observaciones y p el número de variables. Esta estadística se distribuye aproximadamente como una χ^2 con $p(p-1)/2$ grados de libertad bajo la hipótesis nula. Un p -valor bajo (usualmente $< 0,05$) indica que la matriz de correlación difiere significativamente de la identidad, lo que valida el uso de métodos multivariados basados en correlaciones.

3.2 Análisis Multivariado a dos vías.

En los métodos a dos vías (o dos modos), los datos se organizan en una matriz bidimensional donde las filas representan a n individuos u objetos y las columnas a p variables medidas sobre ellos. Uno de los más conocidos es el ACP.

3.2.1 Componentes Principales

El método de ACP, desarrollado por Hotelling en 1933, ha tomado relevancia a partir de la década de los 80 y 90, con el desarrollo de software especializado. Se parte de una matriz de datos multivariantes $X_{n \times p}$, siendo n la cantidad de individuos y p la cantidad de variables:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{n \times p}$$

Las componentes principales son variables no correlacionadas tales que unas pocas explican la mayor variabilidad de X (Cuadras, 2019). A partir de la formulación en 1936 de la descomposición en valores singulares (DVS) (Eckart & Young, 1936), la forma más común de aplicar ACP es a través de los resultados de DVS. Este consiste en una descomposición matricial mediante el producto de tres matrices que contienen los autovectores de $XX^\top$, los valores singulares y los autovectores de $X^\top X$ de la matriz de datos.

Descomposición en valores singulares: Sea una matriz $X_{p \times q}$ compuesta por elementos reales. Es posible identificar matrices ortogonales U y V , además de una matriz diagonal Λ . En Λ , los valores singulares λ_i cumplen con la relación $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_s$, donde $s = \min(p, q)$. De esta forma, la descomposición en valores singulares se expresa como

$$X = U \Lambda V', \quad (3.2.1)$$

cuando se consideran todas las componentes, o bien

$$X \approx U_k \Lambda_k V_k', \quad (3.2.2)$$

si se retienen únicamente las $k < s$ primeras, según el resultado de Eckart y Young (1936).

donde:

- U es la matriz de autovectores asociada a XX' .
- V es la matriz de autovectores asociada a $X'X$.
- Λ es la matriz diagonal de valores singulares, tales que λ_i^2 son los autovalores de XX' y $X'X$.

3.3 Análisis Multivariado a tres vías.

Los datos a tres vías (Kroonenberg, 2008) hacen referencia a un conjunto n de observaciones por p variables y q condiciones, momentos en el tiempo o diferentes grupos, es decir en arreglos tridimensionales. Esto permite analizar simultáneamente las relaciones entre individuos, variables y condiciones. A continuación se resumen los métodos multivariados a tres vías más usados.

3.3.1 Análisis Factorial Multiple (AFM)

El AFM (Abdessemed & Escofier, 1996; Escofier & Pagès, 2008) es un método multivariado a tres vías que analiza observaciones para varios bloques o conjuntos de variables. Todos los grupos miden a los mismos individuos, pero pueden tener distinto número de variables y naturaleza (cuantitativa o cualitativa), aunque las variables dentro de cada grupo deben ser de la misma naturaleza (Abdi & Williams, 2010). Para los individuos, se utilizará el subíndice i (con $i = 1, 2, \dots, n$); para las variables, el subíndice j (con $j = 1, 2, \dots, p_K$); y para las condiciones, el subíndice K (donde $K = 1, 2, \dots, q$).

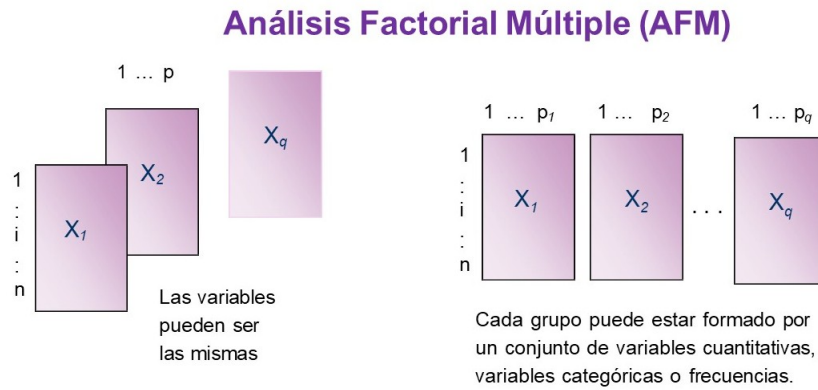


Figura 3.3: AFM. Fuente: en base a apuntes del curso «Métodos Estadísticos para el Análisis de datos en estructuras de tres vías», dictado por S. Bramardi y A. Lavalle. Doctorado en Estadística, Universidad Nacional de Rosario.

Los objetivos del AFM son (Bramardi & Lavalle, 2021):

- Estudiar las similitudes de los individuos respecto a todas las variables.
- Analizar las relaciones entre variables.
- Encontrar una estructura común respecto a los grupos que ponga en evidencia las especificidades de cada grupo.
- Comparar las tipologías de individuos obtenidas con cada grupo de variables.

El AFM utiliza la técnica de descomposición en valores singulares de dos vías, procesa las matrices de datos iniciales (se ponderan) organizándolas en una estructura bimodal, mediante un promedio ponderado o mediante la yuxtaposición de estas matrices. Esta metodología permite tratar las matrices compuestas como entidades de dos modos en el análisis subsecuente.

Considere $X_{n \times p}$, una matriz de datos reales (centrados o estandarizados y ponderados) correspondiente a n individuos descritos por p variables. Entonces, la descomposición de la matriz en valores singulares se expresa como en la ecuación (3.2.1). Esta matriz X da lugar a dos nubes de puntos, una de puntos filas o individuos y otra de puntos variables o puntos columnas.

A continuación se realiza el ajuste de la nube de individuos en el espacio de las

variables ($\mathbb{R}^p$) Dada la descomposición $X = U \Lambda V'$, con X de dimensión $n \times p$:

$$X'X = (U \Lambda V')'(U \Lambda V') = V \Lambda U'U \Lambda V' = V \Lambda^2 V'.$$

Los elementos diagonales de la matriz Λ^2 , es decir, λ_i^2 , constituyen los autovalores de la matriz $X'X$.

Estos autovectores nos brindan las coordenadas de los individuos en el i -ésimo eje y se definen como F_i :

$$F_i = X \mathbf{v}_i,$$

donde $\mathbf{v}_i$ es la i -ésima columna de la matriz V .

Siguiendo el mismo criterio se obtiene el ajuste de la nube de variables en el espacio de los individuos ($\mathbb{R}^n$) que esta dado por:

$$XX' = U \Lambda V'(U \Lambda V')' = U \Lambda V'V \Lambda U' = U \Lambda^2 U'.$$

Con el mismo criterio utilizado se obtienen las coordenadas de las variables en el i -ésimo eje en el nuevo espacio:

$$G_i = X' u_i,$$

donde u_i es la i -ésima columna de la matriz U .

Estos dos ajustes están intrínsecamente relacionados, ya que se derivan de la misma matriz X . En efecto, existe una correspondencia que permite transformar las representaciones del espacio de individuos al espacio de variables. Dichas correspondencias, denominadas *relaciones de transición*, vinculan los autovectores de un espacio con los autovectores del otro. Las relaciones de transición entre $\mathbf{u}_i$ y $\mathbf{v}_i$ se pueden escribir como:

$$\mathbf{v}_i = \frac{1}{\lambda_i} X^\top \mathbf{u}_i, \quad \mathbf{u}_i = \frac{1}{\lambda_i} X \mathbf{v}_i,$$

Definiciones de $\mathbf{F}_i$ y $\mathbf{G}_i$:

$$\begin{aligned}\mathbf{F}_i &= \lambda_i \mathbf{u}_i, & \mathbf{G}_i &= \frac{1}{\lambda_i} X^\top \mathbf{F}_i, \\ \mathbf{G}_i &= \lambda_i \mathbf{v}_i, & \mathbf{F}_i &= \frac{1}{\lambda_i} X \mathbf{G}_i.\end{aligned}$$

Estas relaciones de transición permiten trabajar en un único espacio, de manera que es posible derivar las coordenadas de los individuos a partir de las de las variables y viceversa. Por ejemplo, si nos ubicamos en el espacio generado por XX' , se pueden obtener las coordenadas de las variables; asimismo, los vectores U proporcionan las coordenadas de los individuos, estableciendo así una correspondencia recíproca entre ambos espacios. En resumen hay tres formas de obtener las coordenadas de los individuos y/o las variables. Estas distintas expresiones se utilizan en la interpretación del AFM. Estas dos relaciones de transición permiten interpretar de manera conjunta tanto el plano de individuos como el círculo de correlación donde se representan las variables. En consecuencia, la descomposición dada en la ecuación (3.2.1) facilita la construcción del biplot, en el cual se representan simultáneamente individuos y variables. Como todos los grupos de variables están definidos sobre el mismo conjunto de individuos, las tablas pueden ser yuxtapuestas y formar así una única tabla que cruza individuos y variables (Escofier & Pagès, 1992), tal como se muestra en la siguiente figura:

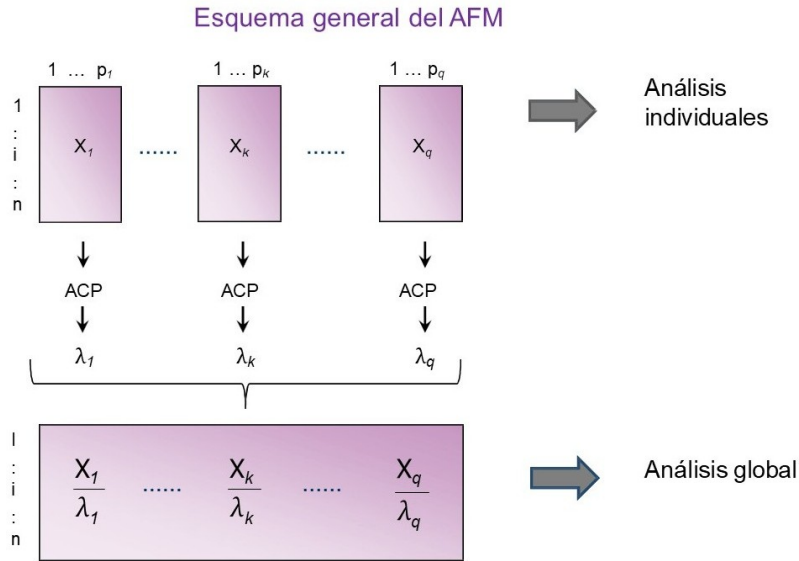


Figura 3.4: AFM esquema general cuando las variables de la tercera vía son todas cuantitativas. *Fuente: Apuntes del curso "Métodos Estadísticos para el Análisis de datos en estructuras de tres vías" dictado por Dr. Bramardi S. y Lavallo A.*

El objetivo del AFM, consiste en comparar nubes de individuos, cada uno asociado con uno de los grupos de variables. Mediante este análisis, se puede observar en un sólo gráfico las observaciones y la relación entre los grupos. El AFM, se concibe como una extensión ponderada del ACP. A cada tabla de individuos con sus variables (todas cuantitativas) se le aplica un ACP, esta etapa se denomina análisis individuales. Del ACP de cada grupo de variables $X_1, X_2, \dots, X_q$, se toma λ_i el primer autovalor (el más grande) de la matriz de covarianzas (o correlaciones) correspondiente al grupo i . A cada una de esas matrices la vamos a ponderar por el inverso del autovalor y luego las vamos a concatenar en una única matriz. Sobre la matriz única concatenada se realiza un ACP, que se denomina análisis global. La concatenación ponderada de las tablas se escribe como:

$$[X_1/\lambda_1, X_2/\lambda_2, \dots, X_q/\lambda_q].$$

Esta ponderación tiene tres propiedades básicas

Propiedades de esta ponderación

- Se preserva la estructura de cada grupo.
- Ningún grupo, por sí mismo, puede generar la primera dimensión global.
- Para cada grupo, la varianza de la primera dimensión principal es igual a 1.

Si algún grupo de variables está constituida por variables cualitativas, es posible, transformándolas convenientemente, obtener una equivalencia formal entre los métodos de análisis. En particular, el Análisis de Correspondencias (AC) puede interpretarse como un ACP aplicado a los perfiles fila y columna de una tabla de contingencia con métrica χ^2 , mientras que el Análisis de Correspondencias Múltiple (ACM) se corresponde con un ACP de la matriz de Burt. De este modo, tanto ACP como AC y ACM comparten una misma formulación algebraica basada en la descomposición en valores singulares, lo que refuerza la idea de que se trata de técnicas análogas para datos cuantitativos y cualitativos, respectivamente (Cuadras, 2019; Rencher, 2002).

Métodos ICI: Interestructura – Compromiso – Intraestructura

ra El AFM, junto con STATIS, pertenece a la familia de los métodos ICI, ya que permite estudiar de manera integrada la *interestructura* entre tablas, la estructura de *compromiso* (o consenso) y la *intraestructura* de individuos y variables (Abdi & Williams, 2010; Bramardi & Lavalle, 2021). A partir del AFM se obtienen representaciones conjuntas de los grupos de variables, los individuos y las variables, así como medidas sintéticas que describen la contribución relativa de cada grupo al consenso global (Escoufier & Pagès, 2008).

1. **Interestructura.** El análisis de la interestructura evalúa el grado de similitud entre las distintas tablas de datos, comparando las estructuras de inercia que aportan los grupos de variables (Escoufier, 1973; Robert & Escoufier, 1976). Para evaluar la estructura común entre los grupos se utiliza el coeficiente de ligazón entre grupos (Liaison groupes),

$$\text{Lg}(X_1, X_2) = \text{traza}((X_1 X_1')(X_2 X_2')),$$

que mide la cantidad de estructura común entre dos grupos de variables. Si $Lg = 0$, las variables de un grupo no están correlacionadas con las del otro grupo y si $Lg > 0$, indica que los grupos comparten dimensiones de inercia relevantes.

La dimensionalidad propia de cada grupo se mide mediante:

$$Ng(j) = \sqrt{Lg(X_j, X_j)},$$

y la intensidad relativa de la relación entre la estructura de dos grupos, se expresa a través del coeficiente de correlación vectorial (Escoufier, 1973; Robert & Escoufier, 1976) denominado *Coefficiente RV de Escoufier*

$$RV(X_1, X_2) = \frac{Lg(X_1, X_2)}{Ng(1) Ng(2)} = \frac{\text{tr}(X_1 X_1^\top X_2 X_2^\top)}{\sqrt{\text{tr}((X_1 X_1^\top)^2) \text{tr}((X_2 X_2^\top)^2)}}, \quad (3.3.1)$$

cuyos valores están acotados entre 0 (ausencia de relación estructural) y 1 (estructuras homotéticas).

2. **Compromiso.** Esta etapa construye una estructura consenso que sintetiza la información de todos los grupos de variables (Abdi & Williams, 2010; Escoufier & Pagès, 2008). Cada tabla X_k se centra y estandariza, y luego se pondera por el inverso de su primer valor singular σ_1^k , asegurando una contribución equilibrada. Las tablas ponderadas $Z_k = X_k / \sigma_1^k$ se concatenan en una matriz global $Z = [Z_1, Z_2, \dots, Z_q]$. Sobre Z se aplica una descomposición en valores singulares (equivalente a un ACP), obteniendo los ejes factoriales y las coordenadas consenso de los individuos.
3. **Intraestructura.** Esta etapa permite representar y analizar, dentro del espacio de compromiso, la configuración específica de cada grupo original. A partir de la descomposición en valores singulares de la matriz consenso Z , se obtienen las coordenadas de los individuos consenso F . Los individuos parciales de cada tabla Z_k se proyectan sobre este mismo espacio factorial, facilitando la

comparación directa entre condiciones y la identificación de concordancias o discrepancias respecto de la estructura común. Para cuantificar este ajuste, se analizan medidas de inercia, como el cociente de inercia, que determina el grado de alineación de cada individuo o grupo con el consenso. Paralelamente, las coordenadas de las variables se calculan como la correlación entre cada variable original y los ejes factoriales, $G_s(j) = r(j, F_s)$, lo que permite interpretar su contribución a la estructura global.

3.4 Modelos Lineales.

Los modelos lineales son modelos estadísticos que constituyen la base de una amplia gama de metodología estadística (Faraway, 2016). Estos tienen el objetivo principal de describir la relación entre una o más variables dependientes (o respuesta) y una o más variables independientes (también llamadas predictores o explicativas) a través de una combinación lineal de los parámetros desconocidos. La característica esencial es que el modelo es lineal en los parámetros. El modelo lineal en notación matricial se expresa como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (3.4.1)$$

donde:

- y es el vector o matriz de observaciones de la variable respuesta.
- X es la matriz de diseño.
- β es el vector de parámetros.
- ε es el vector de términos de error.

La dimensión de los vectores y matrices empleados en el modelo depende fundamentalmente del enfoque estadístico que se adopte. A partir de este marco general, pueden derivarse distintos tipos de modelos, como los de regresión (simple, múltiple o multivariante), donde la $\mathbf{Y}$ y la $\mathbf{X}$ son cuantitativas, modelos de regresión penalizados, modelos de regresión de Mínimos Cuadrados Parciales, modelos Lineales

Generalizados, entre otros. Cada uno de estos enfoques presenta supuestos y estructuras de dependencia específicas, lo que los hace apropiados para distintos tipos de datos y objetivos de investigación. Cuando las variables explicativas son categóricas (factores) en vez de continuas, estamos ante un modelo lineal denominado Análisis de Varianza (ANOVA, por sus siglas en inglés de Analysis of Variance), que compara las medias entre grupos categóricos. Si $p=1$ se denomina ANOVA unifactorial y si $p>1$ se denomina ANOVA multifactorial. Si en lugar de tener una sola variable respuesta continua se tiene dos o más, el modelo es ANOVA Multivariado (MANOVA) de uno o varios factores. Cuando en el análisis se tienen variables explicativas categóricas y continuas, estamos ante un modelo lineal denominado Análisis de Covariancia, que combina ANOVA y regresión lineal.

3.4.1 Modelos de Regresión

Uno de los enfoques más conocidos dentro del ámbito de los modelos lineales son los denominados modelos de regresión. Estos se utilizan para describir la relación entre una o varias variables continuas de respuesta (Y) y una o más variables independientes (X). La ecuación del modelo lineal (3.4.1), se puede expresar tal como la plantean Hastie et al. (2009):

$$y_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j + \epsilon_i \quad (3.4.2)$$

En forma matricial:

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \dots, n. \quad (3.4.3)$$

3.4.1.1 Regresión lineal simple

Cuando se trabaja con una sola variable independiente x ($p = 1$) y una variable dependiente y ,

$$y_1 = \beta_0 + \beta_1 x_1 + \epsilon_1, \quad (3.4.4)$$

$$y_2 = \beta_0 + \beta_1 x_2 + \epsilon_2, \quad (3.4.5)$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_n + \epsilon_n. \quad (3.4.6)$$

Usando la notación matricial:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Por lo tanto, el modelo se puede expresar como:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (3.4.7)$$

las dimensiones de los vectores y parámetros involucrados en el modelo son:

- $\mathbf{y}$ es un vector de respuestas de dimensión $n \times 1$.
- $\mathbf{x}$ es un vector de predictores de dimensión $n \times 2$ si se incluye la columna de 1's para el intercepto (o $n \times 1$ si se maneja el intercepto por separado), con una variable predictora
- $\boldsymbol{\beta}$ es el vector de coeficientes de dimensión 2×1 (incluyendo intercepto β_0 y pendiente β_1).
- $\boldsymbol{\epsilon}$ es el vector de errores de dimensión $n \times 1$.

3.4.1.2 Regresión lineal múltiple

En el caso de la regresión lineal múltiple, con una única variable dependiente y varias variables predictoras, cada observación y_i , $i = 1, \dots, n$, se modela como una función lineal de las variables predictoras más un término de error aleatorio:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_q x_{1p} + \epsilon_1, \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_q x_{2p} + \epsilon_2, \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_q x_{np} + \epsilon_n. \end{aligned}$$

Usando la notación matricial:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

el modelo puede escribirse como:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

donde normalmente $\mathbf{X}$ incluye una columna de unos para el intercepto. Las dimensiones de los vectores y matrices son las siguientes:

- $\mathbf{y}$ es un vector de dimensión $n \times 1$, con n observaciones de la variable respuesta.
- $\mathbf{X}$ es una matriz de dimensión $n \times (p + 1)$ si se incluye la columna de 1's para el intercepto (o $n \times p$ si se maneja el intercepto por separado), con p variables predictoras.
- $\boldsymbol{\beta}$ es un vector de dimensión $(p + 1) \times 1$ (o $p \times 1$ sin intercepto dentro de $\mathbf{X}$).

- ϵ es un vector de errores de dimensión $n \times 1$.

De acuerdo con el método de Mínimos Cuadrados Ordinarios (MCO), los estimadores de los coeficientes se obtienen como:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (3.4.8)$$

A partir de estos, los valores ajustados de la variable respuesta pueden escribirse como:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}. \quad (3.4.9)$$

3.4.1.3 Regresión lineal múltiple multivariada

En la regresión lineal múltiple multivariada se modelan varias variables de respuesta simultáneamente. Las observaciones multivariadas pueden disponerse en una matriz $\mathbf{Y}$ de dimensión $n \times q$, cuyas filas corresponden a los vectores de respuestas de cada individuo:

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nq} \end{pmatrix} = \begin{pmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_n \end{pmatrix}.$$

Los n valores de $X_1, X_2, \dots, X_p$ se pueden colocar en una matriz:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

La expresión del modelo múltiple multivariado (Renchert, 2002) es:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \quad (3.4.10)$$

donde:

- $\mathbf{Y}$ es una matriz de respuestas de dimensión $n \times q$, con n observaciones y q variables dependientes.
- $\mathbf{X}$ es la matriz de predictores de dimensión $n \times (p+1)$ si se incluye el intercepto (o $n \times p$ si se maneja el intercepto por separado).
- $\mathbf{B}$ es la matriz de coeficientes de dimensión $(p+1) \times q$ (o $p \times q$ sin intercepto en $\mathbf{X}$).
- $\mathbf{E}$ es la matriz de errores aleatorios de dimensión $n \times q$.

En cada uno de estos modelos, los parámetros β (o $\mathbf{B}$) y los errores ϵ (o $\mathbf{E}$) son desconocidos y deben estimarse a partir de los datos, mientras que las variables $\mathbf{y}$ (o $\mathbf{Y}$) y $\mathbf{x}$ (o $\mathbf{X}$) son conocidas y forman parte de las observaciones disponibles. Existen varios métodos para estimar los parámetros de un modelo de regresión, siendo los más empleados los MCO y el Método de Máxima Verosimilitud (MV).

El método de MCO, introducido por Carl Friedrich Gauss, consiste en minimizar la suma de los cuadrados de los residuos sin necesidad de suponer distribuciones específicas para los errores. Bajo los supuestos clásicos de Gauss–Markov (linealidad, independencia y homocedasticidad de los errores), este procedimiento garantiza que los estimadores obtenidos son los mejores estimadores lineales insesgados (MELI, del inglés *Best Linear Unbiased Estimators*, BLUE).

En contraste, el método de MV requiere suponer explícitamente una distribución para los errores, usualmente Normal (Normal multivariada para regresión multivariada). En ese caso, la maximización de la función de verosimilitud conduce a los mismos estimadores que MCO, pero además permite obtener sus distribuciones exactas y, por ende, realizar inferencia estadística mediante pruebas de hipótesis e intervalos de confianza.

En notación matricial, las ecuaciones normales del modelo multivariado son:

$$\mathbf{X}'\mathbf{X}\hat{\mathbf{B}} = \mathbf{X}'\mathbf{Y}, \quad (3.4.11)$$

y, cuando $\mathbf{X}$ tiene rango columna completo ($\text{rank}(\mathbf{X}) = p+1$), es decir, cuando $\mathbf{X}'\mathbf{X}$ es invertible, las ecuaciones normales admiten una única solución cerrada:

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \quad (3.4.12)$$

En consecuencia, las predicciones y residuos son

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}, \quad \hat{\mathbf{E}} = \mathbf{Y} - \hat{\mathbf{Y}}. \quad (3.4.13)$$

En el caso univariado ($q = 1$), si además de los supuestos de Gauss–Markov se asume que los errores son normales e independientes, el modelo se denomina modelo lineal normal univariado:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n),$$

y se cumple que

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right). \quad (3.4.14)$$

En el contexto multivariado ($q > 1$), los supuestos de normalidad implican que cada vector fila de errores $\mathbf{e}_i$ sigue una distribución normal multivariada,

$$\mathbf{e}_i \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma}), \quad (3.4.15)$$

con $\boldsymbol{\Sigma}$ una matriz de covarianza $q \times q$ que permite la existencia de correlaciones entre las variables de respuesta. Lo cual conduce a que el modelo completo pueda representarse como una normal matricial:

$$\hat{\mathbf{B}} \sim \mathcal{MN}_{p+1,q}\left(\mathbf{B}, (\mathbf{X}'\mathbf{X})^{-1}, \boldsymbol{\Sigma}\right). \quad (3.4.16)$$

Esta distribución permite derivar inferencia multivariada conjunta sobre todos los coeficientes, en lugar de analizar ecuaciones separadas.

Para evaluar la calidad del ajuste en el caso univariado ($q = 1$), se emplea la suma de cuadrados del error residual:

$$SCE = \sum_{i=1}^n (y_i - \hat{y}_i)^2. \quad (3.4.17)$$

El estimador insesgado de σ^2 es

$$\hat{\sigma}^2 = \frac{SCE}{n - p}. \quad (3.4.18)$$

Asimismo, el coeficiente de determinación R^2 indica la proporción de la variación en la variable dependiente explicada por las variables explicativas:

$$R^2 = 1 - \frac{SCE}{SCT}, \quad (3.4.19)$$

donde SCT es la suma total de cuadrados de Y .

Esta medida tiende a aumentar con el número de predictores, por lo que se emplea también el coeficiente de determinación ajustado (R_{aj}^2), más adecuado para comparar modelos de distinta complejidad (Gujarati, 2009; Rencher, 2002).

En el contexto multivariado ($q > 1$), la evaluación del ajuste no se resume en un único R^2 , sino que se consideran medidas multivariadas globales de significación. Entre las más utilizadas se encuentran:

- **Lambda de Wilks** (Λ): mide la proporción de varianza no explicada por el modelo.
- **Rastro de Pillai** (Pillai's Trace): basado en la traza de una matriz de efectos explicados, robusto en muestras pequeñas.
- **Rastro de Hotelling–Lawley**.
- **Mayor raíz de Roy**.

Estos contrastes se derivan de la descomposición en autovalores de las matrices

de variación explicada y no explicada, y permiten evaluar la significación conjunta de todos los predictores sobre todas las variables respuesta.

3.4.2 Selección de variables

Cuando el número de observaciones $n \gg p$ las estimaciones por mínimos cuadrados, tienden a presentar una varianza baja, sin embargo cuando el número de variables predictoras, p , es mayor que el número de observaciones, n , la estimación de los coeficientes deja de ser única, lo que imposibilita el uso del método de mínimos cuadrados ordinarios (James et al., 2013). Este último caso, se denomina de alta dimensionalidad, y es un escenario frecuente que tiene importantes desafíos tanto en la estimación como en la interpretación de resultados. En estas situaciones, las técnicas de regresión tradicionales resultan insuficientes, y una opción clásica ha sido recurrir a la selección de variables. James et al. (2013) consideran, entre los métodos de selección de subconjuntos de predictores, los siguientes:

- **Selección de subconjuntos completos (Best Subset Selection):** En este método se evalúa todos los posibles subconjuntos de variables predictoras p , es decir se ajusta el modelo lineal de mínimos cuadrados a los p modelos con un predictor, luego se ajusta con dos predictores y así sucesivamente. Luego se elige el modelo que optimice algún criterio de bondad del ajuste, como el R^2 ajustado, el AIC (Criterio de información de Akaike) o el BIC (Criterio de información bayesiano). Este método cuando el número de predictores es grande, demanda gran recurso de la computadora.
- **Selección hacia adelante (Forward Selection):** Se inicia con un modelo que no contenga predictores (modelo nulo) y se van añadiendo, de forma secuencial, hasta que todos los predictores estén en el modelo. En cada paso se incluye la variable que proporcione la mayor mejora estadísticamente significativa.
- **Selección hacia atrás (Backward Elimination):** se parte de un modelo que incluye todas las variables, se remueven iterativamente aquellas cuyo

aporte al modelo no es significativo, hasta alcanzar un modelo óptimo. En particular, cuando $p \gg n$ no se puede aplicar el método de selección hacia atrás, dado que no es posible ajustar inicialmente un modelo completo con más variables que observaciones.

- **Selección por pasos (Stepwise Selection):** Combina la selección hacia adelante y hacia atrás. En cada iteración se evalúa la posibilidad de agregar o eliminar variables.

El método *Stepwise*, en el caso en que el número de variables predictoras p es muy elevado, genera procesos en los que se necesita el uso de equipos con gran capacidad de memoria y velocidad. Este método puede generar una alta varianza en las estimaciones de los coeficientes y conducir a un sobreajuste. Por otra parte, ni la selección hacia adelante (forward selection) ni la selección hacia atrás (backward selection) aseguran que el modelo obtenido incluya una cantidad de p predictores de forma óptima. Los métodos de selección hacia adelante y el procedimiento *Stepwise* resultan ser mejores opciones con escenarios de alta dimensionalidad. Estos métodos se basan en el método de mínimos cuadrados para ajustar modelos lineales que contiene un subconjunto de predictores.

3.4.3 Modelos de regresión penalizados

El trabajo de Tikhonov (1943), orientado a resolver ecuaciones integrales con características particulares, introdujo el concepto de regularización, noción que posteriormente se extendió a diversos ámbitos, en especial a la estadística, la optimización y el aprendizaje automático. En estas áreas, la regularización se ha consolidado como una estrategia fundamental para controlar el sobreajuste y mejorar la estabilidad de los métodos de estimación. Desde el punto de vista estadístico, la motivación surge a partir de las limitaciones del estimador de mínimos cuadrados definido en (3.4.12). Si bien dicho estimador posee la varianza más baja dentro de la clase de estimadores lineales insesgados, ello no garantiza que dicha varianza sea suficientemente pequeña. En situaciones de alta multicolinealidad o cuando el número de

predictores p es comparable o incluso mayor que el número de observaciones n , el estimador de mínimos cuadrados resulta inestable o carece de solución única, ya que la matriz de diseño X pierde rango completo (Hastie et al., 2009, p. 46). En tales casos, puede ser preferible recurrir a estimadores sesgados que, a cambio, logren un error cuadrático medio menor que el de $\hat{\beta}$. Este compromiso entre sesgo y varianza constituye el fundamento de las técnicas de regularización, las cuales imponen restricciones o penalizaciones sobre los coeficientes de regresión. Al añadir un término de penalización, se modifica la matriz de información a $X'X + \lambda I$, lo que garantiza la existencia y unicidad de la solución y mejora la estabilidad del ajuste.

En este contexto surgen distintos métodos de regularización penalizada, que se distinguen por la forma en que aplican dichas restricciones. Entre los más utilizados se encuentran la *Regresión Ridge*, que penaliza la magnitud de los coeficientes; la *Regresión LASSO*, que además favorece la selección de variables; y el *Elastic Net*, que combina ambas estrategias. A continuación se presentan sus fundamentos teóricos y principales propiedades.

3.4.3.1 Modelo de Regresión Ridge

Los modelos *Ridge* fueron introducidos por Hoerl y Kennard (1970). Este método se desarrolló como una técnica de estimación sesgada que podría abordar problemas de multicolinealidad en modelos de regresión Múltiple, donde las variables independientes están altamente correlacionadas. En MCO se busca los $\beta_0, \beta_1, \dots, \beta_p$ que minimizan las ecuaciones normales (3.4.11), en *Ridge* se adiciona a la suma de cuadrados del error a minimizar un término proporcional al cuadrado de los coeficientes de regresión (penalización ℓ_2) que penalizará los coeficientes grandes (el intercepto β_0 , permanece excepto de penalización), evitando sobreajuste y mejorando la estabilidad del modelo. En el caso de una única variable respuesta y p predictoras, la función objetivo de la regresión Ridge se define como:

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \frac{\lambda}{2} \sum_{j=1}^p \beta_j^2, \quad (3.4.20)$$

donde el primer término corresponde a la suma de cuadrados de los residuos y el segundo constituye la penalización ℓ_2 , que restringe la magnitud de los coeficientes.

A partir de esta formulación, la extensión al caso multivariado (con q variables respuesta) se expresa como:

$$\sum_{i=1}^n \sum_{k=1}^q \left(y_{ik} - \beta_{0k} - \sum_{j=1}^p \beta_{jk} x_{ij} \right)^2 + \frac{\lambda}{2} \sum_{j=1}^p \sum_{k=1}^q \beta_{jk}^2. \quad (3.4.21)$$

Donde las observaciones varían de $i = 1, \dots, n$; los predictores de, $j = 1, \dots, p$; las variables respuestas de, $k = 1, \dots, q$; β_{0k} es el intercepto para la k -ésima respuesta; β_{jk} es el coeficiente del j -ésimo predictor sobre la k -ésima respuesta; $\lambda > 0$ es el parámetro global de penalización. En particular:

- Cuando $\lambda = 0$, la regresión *Ridge* coincide con la estimación por mínimos cuadrados.
- A medida que $\lambda \rightarrow \infty$, el estimador $\hat{\beta}_{\lambda}^R$ tiende al vector $\mathbf{0}$.
- Para valores intermedios de λ , se alcanza un compromiso entre el ajuste lineal y la reducción de los coeficientes.

La ecuación (3.4.21) puede reformularse en términos de la formulación de Lagrange como sigue:

$$\min_{\mathbf{B} \in \mathbb{R}^{p \times q}} \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{XB}\|_F^2 + \lambda \|\mathbf{B}\|_F^2 \right\}, \quad (3.4.22)$$

donde $\|\cdot\|_F$ denota la norma de Frobenius, es decir, $\|\mathbf{M}\|_F^2 = \sum_{i,j} m_{ij}^2$ para una matriz $\mathbf{M} = (m_{ij})$. Este planteo corresponde a un caso particular de regularización de Tikhonov (Tikhonov, 1943), en donde se penaliza la norma euclídea de los coeficientes. En el caso univariado ($q = 1$), esto se reduce a la penalización ℓ_2 sobre el vector de coeficientes β . Al resolver la ecuación (3.4.22), se obtiene una solución cerrada a partir del sistema de ecuaciones que resulta de anular el gradiente de la función objetivo (condiciones de primer orden):

$$\frac{\partial}{\partial \mathbf{B}} \{ \|\mathbf{Y} - \mathbf{XB}\|_F^2 + \lambda \|\mathbf{B}\|_F^2 \} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{XB} + 2\lambda\mathbf{B} = 0. \quad (3.4.23)$$

En el caso general multivariado, el estimador *Ridge* resulta:

$$\hat{\mathbf{B}}^{Ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1} \mathbf{X}'\mathbf{Y}, \quad (3.4.24)$$

donde $\mathbf{I}_p$ es la matriz identidad de dimensión $p \times p$.

En el caso univariado ($q = 1$), esta expresión se reduce al estimador *Ridge* clásico:

$$\hat{\beta}^{Ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1} \mathbf{X}'\mathbf{y}. \quad (3.4.25)$$

El término adicional $\lambda\mathbf{I}_p$ asegura que la matriz $\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p$ sea definida positiva, pues todos sus autovalores son $\lambda_i + \lambda > 0$, garantizando así la invertibilidad y mejorando la estabilidad numérica de la solución incluso cuando $\mathbf{X}'\mathbf{X}$ es singular. Por otro lado, la regresión *Ridge* encoge todos los coeficientes hacia cero y reduce su variabilidad, por ello es más útil para casos de predicción que para interpretar importancia relativa de las variables independientes, que es nuestro objetivo. El parámetro de penalización (o regularización) λ , controla el equilibrio entre la minimización de la suma de cuadrados del error y la penalización impuesta a los coeficientes. Este estimador $\hat{\beta}^{Ridge}$ existe y está bien definido para todo $\lambda > 0$, incluso en los casos en que X no es de rango completo y el estimador MCO no es único. Esto se debe a que $X'X + \lambda I$ es definida positiva, y por tanto invertible para todo $\lambda > 0$. Como se discute en Hastie et al. (2009, Chapter 3), ridge puede verse como una proyección del problema original en una base ortonormal determinada por los autovectores de $X'X$, con un reescalamiento de las componentes según la magnitud de los autovalores y el parámetro λ .

Por lo tanto, $\hat{\beta}_\lambda^R$ es un estimador sesgado de β , a menos que $\lambda = 0$. En ese caso, $\hat{\beta}_\lambda^R = \hat{\beta}$. Se parte de la descomposición espectral de la matriz de información $X'X$, siendo Λ la matriz diagonal de autovalores de $X'X$, y P la transformación ortogonal

tal que:

$$X'X = P' \Lambda P,$$

y definamos $\alpha = P \beta$.

El Error Cuadrático Medio (ECM) de $\hat{\beta}_\lambda^R$ puede descomponerse como:

$$\text{ECM}(\hat{\beta}_\lambda^R) = \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + \lambda)^2} + \lambda^2 \sum_{i=1}^p \frac{\alpha_i^2}{(\lambda_i + \lambda)^2}, \quad (3.4.26)$$

donde λ_i son los autovalores de $X'X$ y α_i son los componentes de α .

Para mayor claridad, podemos separar esta expresión en dos términos:

$$\text{ECM}_1 = \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + \lambda)^2}, \quad \text{ECM}_2 = \lambda^2 \sum_{i=1}^p \frac{\alpha_i^2}{(\lambda_i + \lambda)^2}.$$

El término ECM_1 representa la varianza total del estimador $\hat{\beta}_\lambda^R$, dada la penalización impuesta en la regresión *Ridge*. Esta varianza es menor que la correspondiente al estimador $\hat{\beta}$ sin penalización. La componente ECM_2 se puede interpretar como el cuadrado del sesgo introducido al utilizar $\hat{\beta}_\lambda^R$ en lugar de $\hat{\beta}$ para estimar β . La regresión *Ridge* retiene los p predictores en el modelo, esto resulta ser una desventaja en comparación con la técnica de selección de subconjuntos completos, o de los métodos de selección hacia adelante (*forward*) y hacia atrás (*backward*), que tienden a incluir solo un subconjunto de las variables. Es decir *Ridge* reduce los coeficientes haciendo que muchos se acerquen a cero, pero ninguno de ellos se anula por completo. Esta característica complica la interpretación del modelo y dificulta la identificación de variables relevantes, en especial cuando el número de predictores p es muy elevado. Esta deducción del error cuadrático medio que se ha presentado en el caso univariado, de forma análoga puede extenderse al caso multivariado ($q > 1$), manteniendo la misma interpretación en términos de la descomposición entre varianza y sesgo al aplicar la penalización ℓ_2 .

3.4.3.2 Modelo de Regresión LASSO

La regresión LASSO propuesta por Tibshirani (1996), es una alternativa a *Ridge*, que tiene la ventaja de realizar selección automática de variables, ya que la penalización de LASSO (ℓ_1) fuerza a algunos coeficientes a ser exactamente cero (Tibshirani, 1996). LASSO mejora la precisión del modelo *Ridge*, al mismo tiempo que controla el problema de multicolinealidad. En el caso de una única variable respuesta y p predictores, los coeficientes LASSO, $\hat{\beta}_\lambda^L$, se obtienen minimizando la siguiente expresión:

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| = S(\beta) + \lambda \sum_{j=1}^p |\beta_j|. \quad (3.4.27)$$

Al comparar (3.4.20) con (3.4.27) se observan expresiones similares. La única diferencia es que el término β_j^2 en *Ridge* (3.4.20) es reemplazado por $|\beta_j|$ (3.4.27), denominado término de penalización ℓ_1 . La función objetivo de LASSO, en su forma de Lagrange, se puede expresar como:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2N} \|y - X\beta\|^2 + \lambda \|\beta\|_1 \right\} \quad (3.4.28)$$

La función de (3.4.28), es convexa en términos de β . A diferencia de la regresión *Ridge*, la regresión LASSO no admite una solución cerrada mediante ecuaciones normales, ya que la función valor absoluto no es diferenciable en cero. En consecuencia, no es posible obtener las condiciones de primer orden de la forma tradicional, y la estimación de los coeficientes requiere recurrir a algoritmos iterativos de optimización, tales como descenso de coordenadas (*coordinate descent*) que utiliza como paso clave la aplicación de la función de umbralización suave (*soft-thresholding*), el algoritmo LARS (*Least Angle Regression*) o técnicas de optimización convexa como el método del gradiente con *soft-thresholding* (gradiente proximal). La función de umbralización suave utilizada en el paso de actualización del algoritmo de descenso por coordenadas (se desarrolla en la subsubsección 3.4.3.4.) para el LASSO se define como en Friedman et al. (2010). En este contexto, el estimador penalizado se obtiene

aplicando dicho operador al estimador de mínimos cuadrados parcial.

En particular, sea $\hat{\beta}_{jk}^{\text{MCO}}$ el estimador de mínimos cuadrados ordinarios parcial para el coeficiente β_{jk} . En el contexto del algoritmo de descenso por coordenadas para el LASSO, en cada paso se calcula primero este estimador $\hat{\beta}_{jk}^{\text{MCO}}$ manteniendo fijos todos los demás coeficientes. Luego, sobre este valor se aplica la función de umbralización suave (*soft-thresholding*) para obtener el estimador penalizado:

$$\hat{\beta}_{jk}^{\text{LASSO}} = S_{\lambda}(\hat{\beta}_{jk}^{\text{MCO}}) = \begin{cases} \hat{\beta}_{jk}^{\text{MCO}} - \lambda, & \text{si } \hat{\beta}_{jk}^{\text{MCO}} > \lambda, \\ 0, & \text{si } |\hat{\beta}_{jk}^{\text{MCO}}| \leq \lambda, \\ \hat{\beta}_{jk}^{\text{MCO}} + \lambda, & \text{si } \hat{\beta}_{jk}^{\text{MCO}} < -\lambda, \end{cases} \quad (3.4.29)$$

o, de forma compacta,

$$S_{\lambda}(\hat{\beta}_{jk}^{\text{MCO}}) = \text{sign}(\hat{\beta}_{jk}^{\text{MCO}}) \left(|\hat{\beta}_{jk}^{\text{MCO}}| - \lambda \right)_+, \quad (3.4.30)$$

donde $(\cdot)_+ = \max(\cdot, 0)$.

Aparte la ecuación (3.4.27) se puede generalizar al caso multivariado con q variables respuestas como sigue:

$$\sum_{i=1}^n \sum_{k=1}^q \left(y_{ik} - \beta_{0k} - \sum_{j=1}^p \beta_{jk} x_{ij} \right)^2 + \lambda \sum_{j=1}^p \sum_{k=1}^q |\beta_{jk}|. \quad (3.4.31)$$

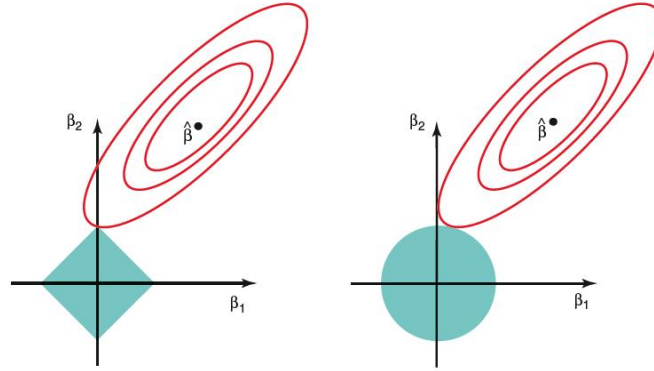


Figura 3.5: **Geometría de las restricciones Ridge y LASSO.** Las áreas en azul representan las regiones de restricción: $|\beta_1| + |\beta_2| \leq s$ (LASSO, izquierda) y $\beta_1^2 + \beta_2^2 \leq s$ (Ridge, derecha). Las elipses rojas corresponden a los contornos de la suma de cuadrados residuales (SCE).

Fuente: James et al. (2013, Figura 6.7).

La Figura 3.5 representa el caso con dos coeficientes (β_1, β_2) en el espacio de parámetros, utilizado con fines didácticos para ilustrar la diferencia geométrica entre LASSO y *Ridge*; la misma interpretación se extiende al caso general con p predictores. Geométricamente (véase Figura 3.5), la penalización ℓ_1 define una región de factibilidad con forma de diamante, lo que induce soluciones en los vértices de dicha región, donde algunos coeficientes se anulan exactamente. De este modo, LASSO no sólo controla la magnitud de los coeficientes, sino que también realiza selección automática de variables.

3.4.3.3 Modelos de Regresión Red Elástica (Elastic Net)

Las metodologías de regularización *Ridge* y LASSO constituyen enfoques complementarios: *Ridge* aporta estabilidad en presencia de multicolinealidad, mientras que LASSO introduce un mecanismo de selección automática de variables mediante la penalización ℓ_1 . No obstante, ambos presentan limitaciones cuando se aplican en forma aislada. En particular, el LASSO sólo puede seleccionar como máximo n predictores cuando $p > n$, y en contextos con grupos de variables fuertemente correlacionadas tiende a conservar una sola, descartando las restantes. Además, en escenarios con $n > p$ y alta correlación entre predictores, su capacidad predictiva suele ser superada por *Ridge* (Efron et al., 2004; Tibshirani, 1996). Para superar

estas restricciones, Zou y Hastie (2005) propusieron la *Red Elástica* (*Elastic Net*), que combina en una única penalización, las propiedades de *shrinkage* de *Ridge* con la capacidad de selección de LASSO. Este método permite incorporar grupos completos de predictores correlacionados y mejora la precisión en situaciones de alta dimensionalidad y fuerte multicolinealidad. Los parámetros se estiman resolviendo la siguiente función objetivo en el caso univariado:

$$\min_{(\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^p} \left\{ \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - x'_i \beta)^2 + \lambda \left[\frac{1}{2} (1 - \delta) \|\beta\|_2^2 + \delta \|\beta\|_1 \right] \right\} \quad (3.4.32)$$

Donde $\delta \in [0, 1]$ el parámetro de mezcla entre *Ridge* y LASSO. Cuando se considera el aporte de un coeficiente individual β_j en la penalización, se obtiene:

$$\lambda \left[\frac{1}{2} (1 - \delta) \beta_j^2 + \delta |\beta_j| \right].$$

Se observa que:

- Cuando $\delta = 1$, la red elástica equivale a la regresión LASSO.
- Cuando $\delta = 0$, la red elástica equivale a la regresión *Ridge*.
- Para $\delta < 1$ y $\lambda > 0$, el problema de optimización de (3.4.32) es estrictamente convexo, lo cual garantiza la existencia de una única solución. Esta propiedad resulta una ventaja sobre LASSO, ya que evita indeterminaciones en presencia de predictoras duplicadas o altamente correlacionadas.

La ecuación (3.4.32) plantea un problema de optimización que ha sido abordado en la literatura mediante la propuesta de distintos algoritmos iterativos, cuya revisión será detallada en la Subsubsección 3.4.3.4 de la presente tesis.

En el contexto multivariado, los coeficientes se obtienen resolviendo el problema

de optimización que se define como:

$$\begin{aligned} \{\hat{\beta}_{0k}, \hat{\beta}_{jk}\} = \arg \min_{\beta_{0k}, \beta_{jk}} & \left\{ \sum_{i=1}^n \sum_{k=1}^q (y_{ik} - \beta_{0k} - \sum_{j=1}^p \beta_{jk} x_{ij})^2 \right. \\ & \left. + \lambda \left[\delta \sum_{j=1}^p \sum_{k=1}^q |\beta_{jk}| + \frac{1-\delta}{2} \sum_{j=1}^p \sum_{k=1}^q \beta_{jk}^2 \right] \right\}. \end{aligned} \quad (3.4.33)$$

En esta expresión las observaciones varían de $i = 1, \dots, n$, los predictores de, $j = 1, \dots, p$ las variables respuestas de, $k = 1, \dots, q$; el parámetro β_{0k} corresponde al intercepto para la k -ésima respuesta; β_{jk} es el coeficiente del j -ésimo predictor sobre la k -ésima respuesta; $\lambda > 0$ es el parámetro que controla la intensidad global de la penalización; y $\delta \in [0, 1]$ determina la proporción entre los términos ℓ_1 y ℓ_2 . Esta formulación multivariada sigue a Friedman et al. (2010). Cuando $\delta = 0$ el modelo se reduce a *Ridge* puro, con coeficiente de penalización $\lambda_{\text{Ridge}} = \frac{\lambda}{2}$, mientras que cuando $\delta = 1$ se obtiene el modelo LASSO. En forma matricial, la ecuación (3.4.33) puede escribirse como:

$$\{\hat{\beta}_0, \hat{\mathbf{B}}\} = \arg \min_{\beta_0 \in \mathbb{R}^q, \mathbf{B} \in \mathbb{R}^{p \times q}} \left\{ \|\mathbf{Y} - \mathbf{1}\beta_0' - \mathbf{X}\mathbf{B}\|_F^2 + \lambda \left[(1-\delta)\|\mathbf{B}\|_F^2 + \delta\|\mathbf{B}\|_1 \right] \right\}, \quad (3.4.34)$$

donde $\beta_0 = (\beta_{01}, \dots, \beta_{0q})'$ es el vector de interceptos y $\hat{\mathbf{B}}^{EN} = (\hat{\beta}_{jk}^{EN})_{p \times q} \in \mathbb{R}^{p \times q}$ la matriz de coeficientes estimados del modelo de *Elastic Net*, obtenida al resolver (3.4.34) mediante un método iterativo.

En *Elastic Net*, la selección de los parámetros de penalización (λ, δ) se realiza habitualmente mediante validación cruzada (CV, por sus siglas en inglés *cross-validation*) (3.5.0.1), ya que no existen criterios analíticos que determinen de manera óptima dichos valores. Al igual que otros métodos de regularización, *Elastic Net* no proporciona directamente medidas de significación estadística para los coeficientes estimados. No obstante, es posible evaluar la estabilidad e importancia de las variables seleccionadas a través de procedimientos de remuestreo, como el *bootstrap* (3.5.0.2) o el *stability selection* (Meinshausen & Bühlmann, 2010), los cuales permiten obtener estimaciones de incertidumbre y mayor robustez en la interpretación de

los resultados.

3.4.3.4 Métodos de Optimización

Los modelos de regresión penalizados y multivariados no suelen admitir soluciones en forma cerrada, lo que hace necesario recurrir a algoritmos numéricos eficientes para su ajuste. En este contexto, se han desarrollado diversas estrategias de optimización convexa, entre las que destacan el descenso por coordenadas, los métodos de gradiente proximal (como el algoritmo iterativo de reducción y umbralización, ISTA, por sus siglas en inglés *Iterative Shrinkage-Thresholding Algorithm*, y su versión acelerada Algoritmo Iterativo Rápido de Reducción y Umbralización FISTA, por sus siglas en inglés, *Fast Iterative Shrinkage-Thresholding Algorithm*), la regresión de ángulo mínimo (LARS) y regresión regularizada por partición (SplitReg). Estos algoritmos han demostrado ser altamente efectivos para problemas de alta dimensionalidad y penalizaciones no diferenciables, tales como la norma ℓ_1 , consolidándose como métodos de referencia en implementaciones ampliamente empleadas, como el paquete `glmnet` del software R (Beck & Teboulle, 2009; Friedman et al., 2010; Hastie et al., 2009).

a) Descenso por coordenadas.

Esta idea fue propuesta por Fu (1998) y más tarde generalizada por Friedman et al. (2007) y Wu y Lange (2008). Este método consiste en optimizar iterativamente cada coeficiente de manera cíclica, fija la penalización λ en su forma de lagrange (3.4.28) y optimiza sucesivamente sobre cada parámetro, manteniendo los demás parámetros fijos (Hastie et al., 2009). En el caso de penalizaciones no diferenciables, como la norma ℓ_1 , cada actualización se resuelve mediante un paso proximal, que corresponde al operador de *soft-thresholding* como se define en la ecuación (3.4.29). Esto se repite hasta que la solución converge. El mismo tipo de algoritmo puede aplicarse en otros modelos en los que la penalización es una suma de funciones separables de los parámetros individuales, como *Elastic Net* (Friedman et al., 2010). En particular, los

autores muestran que este enfoque resulta altamente eficiente cuando se penaliza modelos de regresión lineal, regresión logística binaria y la regresión logística multinomial. Siguiendo a Hastie et al. (2009), el problema de LASSO univariado puede escribirse, para un coeficiente β_j , como la minimización del residuo parcial:

$$R(\tilde{\beta}(\lambda), \beta_j) = \frac{1}{2} \sum_{i=1}^n \left(y_i - \sum_{r \neq j} x_{ir} \tilde{\beta}_r(\lambda) - x_{ij} \beta_j \right)^2 + \lambda \sum_{r \neq j} |\tilde{\beta}_r(\lambda)| + \lambda |\beta_j|, \quad (3.4.35)$$

donde $\tilde{\beta}_r(\lambda)$ denota la estimación actual en el paso iterativo.

Definiendo el residuo parcial

$$\tilde{y}_i^{(j)} = y_i - \sum_{r \neq j} x_{ir} \tilde{\beta}_r(\lambda),$$

la actualización de β_j admite una solución cerrada en términos del operador de *soft-thresholding*:

$$\tilde{\beta}_j(\lambda) \leftarrow S \left(\sum_{i=1}^n x_{ij} (y_i - \tilde{y}_i^{(j)}), \lambda \right), \quad S(t, \lambda) = \text{sign}(t) (|t| - \lambda)_+. \quad (3.4.36)$$

Cuando la respuesta es multivariada en el caso gaussiano, $Y \in \mathbb{R}^{n \times q}$, cada predictor X_j está asociado a un vector de coeficientes $\beta_j \in \mathbb{R}^q$. El problema de optimización con penalización ℓ_1 elemento a elemento (selección tipo LASSO) se formula como:

$$\min_{\{\beta_j\}_{j=1}^p} \frac{1}{2} \|Y - XB\|_F^2 + \lambda \sum_{j=1}^p \sum_{k=1}^q |\beta_{jk}|, \quad (3.4.37)$$

donde $B = (\beta_1, \dots, \beta_p)^\top \in \mathbb{R}^{p \times q}$ y $\|\cdot\|_F$ denota la norma de Frobenius.

Definiendo el residuo parcial multivariado

$$\tilde{Y}^{(j)} = Y - \sum_{r \neq j} X_r \beta_r^\top,$$

la actualización por coordenadas de cada coeficiente β_{jk} se obtiene aplicando el *soft-thresholding* escalar:

$$\beta_{jk} \leftarrow \text{sign}(z_{jk}) (|z_{jk}| - \lambda)_+, \quad z_{jk} = X_j^\top \tilde{y}_k^{(j)}, \quad (3.4.38)$$

para $k = 1, \dots, q$.

Esta ecuación (3.4.38) constituye la extensión natural de la actualización escalar (3.4.36) al caso multivariado lineal con penalización ℓ_1 . Es la implementación utilizada en el paquete `glmnet` cuando se emplea la opción `family = "mgauasian"` (Friedman et al., 2010).

El procedimiento descrito puede resumirse en el siguiente algoritmo:

Algoritmo (descenso por coordenadas proximal multivariado con ℓ_1 , usando *warm start*)

1. (Estandarización) Para cada respuesta Y_k centrar $y_{\cdot k}$; centrar cada columna de X y escalar a norma unitaria.
2. (Grilla de penalización) Construir una grilla decreciente de valores $\lambda_1 = \lambda_{\text{máx}} > \dots > \lambda_M = \varepsilon \lambda_{\text{máx}}$ (con $\varepsilon \in [10^{-4}, 10^{-2}]$). Aquí $\lambda_{\text{máx}} = \frac{1}{n} \max_{j,k} |X_j^\top y_{\cdot k}|$.
3. (Inicialización global) $\hat{B}(\lambda_1) \leftarrow \mathbf{0}_{p \times q}$.
4. **Para** $m = 1, \dots, M$ **hacer**
 - a) (*Warm start*) Si $m > 1$, fijar $\hat{B}^{(0)} \leftarrow \hat{B}(\lambda_{m-1})$ como inicialización para λ_m .
 - b) **Repetir** (bucle cíclico por predictores $j = 1, \dots, p$ y respuestas $k = 1, \dots, q$)
 - 1) Residuo parcial: $\tilde{r}^{(j,k)} \leftarrow y_{\cdot k} - \sum_{r \neq j} X_r \hat{\beta}_{rk}$.
 - 2) Estadístico: $z_{jk} \leftarrow \frac{1}{n} X_j^\top \tilde{r}^{(j,k)}$.

3) Actualización proximal (soft-thresholding escalar):

$$\hat{\beta}_{jk} \leftarrow S(z_{jk}, \lambda_m), \quad S(t, \lambda) = \text{sign}(t) (|t| - \lambda)_+.$$

c) **hasta** convergencia (p.ej., $\max_{j,k} |\hat{\beta}_{jk}^{\text{nuevo}} - \hat{\beta}_{jk}^{\text{viejo}}| < \delta$ o condiciones Karush-Kuhn-Tucker(KKT)).

d) Guardar $\hat{B}(\lambda_m)$.

5. Devolver la trayectoria $\{\hat{B}(\lambda_m)\}_{m=1}^M$.

Notas: (i) Este procedimiento puede interpretarse como la extensión directa del LASSO de Tibshirani (1996) al caso multivariado gaussiano, posteriormente formalizado en la literatura y actualmente implementado en `glmnet` bajo la opción `family="mgaussian"` (Friedman et al., 2010). aplicando penalización ℓ_1 elemento a elemento sobre los coeficientes β_{jk} . (ii) El *warm start* acelera la convergencia a lo largo de la grilla de penalizaciones (Friedman et al., 2010) (véase también (Friedman et al., 2010; Hastie et al., 2009)).

Condiciones KKT para Elastic Net multivariado:

Dado el problema penalizado para cada respuesta $k = 1, \dots, q$:

$$\min_{\beta_k} \left\{ \frac{1}{2n} \|\mathbf{y}_k - X\beta_k\|_2^2 + \lambda \left(\alpha \|\beta_k\|_1 + \frac{1-\alpha}{2} \|\beta_k\|_2^2 \right) \right\},$$

las condiciones de Karush-Kuhn-Tucker (KKT) que debe satisfacer la solución óptima $\hat{\beta}_k = (\hat{\beta}_{1k}, \dots, \hat{\beta}_{pk})'$ son, para cada predictor $j = 1, \dots, p$:

$$\frac{1}{n} X_j^\top (\mathbf{y}_k - X\hat{\beta}_k) = \begin{cases} \lambda\alpha \cdot \text{sign}(\hat{\beta}_{jk}) + \lambda(1-\alpha)\hat{\beta}_{jk}, & \text{si } \hat{\beta}_{jk} \neq 0, \\ z_{jk} \in [-\lambda\alpha, \lambda\alpha], & \text{si } \hat{\beta}_{jk} = 0, \end{cases}$$

donde $z_{jk} := \frac{1}{n} X_j^\top (\mathbf{y}_k - X\hat{\beta}_k)$ representa el estadístico de correlación parcial entre el predictor X_j y el residuo multivariado ajustado.

Nota. Estas condiciones garantizan que la solución obtenida es óptima respecto de la función objetivo penalizada. En la práctica, pueden utilizarse como criterio de parada

alternativo al umbral de convergencia numérica en los algoritmos de coordenadas cíclicas, verificando que se satisfacen (dentro de una tolerancia) en todos los coeficientes.

1. Métodos de Gradiente Proximal

El punto de partida es el descenso por gradiente, cuya idea básica consiste en actualizar todos los coeficientes simultáneamente en la dirección opuesta al gradiente de la función objetivo. En el caso multivariado con penalización ℓ_1 , la función objetivo puede escribirse como:

$$F(B) = f(B) + g(B), \quad (3.4.39)$$

donde $f(B)$ es la función de pérdida diferenciable (por ejemplo, la suma de pérdidas logísticas sobre las q respuestas simultáneas) y $g(B) = \lambda \sum_{j=1}^p \sum_{k=1}^q |\beta_{jk}|$ corresponde a la penalización LASSO extendida al caso multivariado.

Los métodos de gradiente proximal combinan un paso de gradiente sobre la parte diferenciable f con la aplicación del operador proximal correspondiente a g . En este contexto, el operador proximal actúa elemento a elemento sobre los coeficientes de la matriz B :

$$\text{prox}_{\lambda \|\cdot\|_1}(z_{jk}) = \text{sign}(z_{jk}) (|z_{jk}| - \lambda)_+, \quad (3.4.40)$$

lo que equivale a aplicar la regla de *soft-thresholding* a cada coeficiente β_{jk} .

Ejemplos representativos son el algoritmo ISTA y su versión acelerada *FISTA* (Beck & Teboulle, 2009). En el caso multivariado, ISTA actualiza iterativamente la matriz de coeficientes B según

$$B^{(k+1)} = \text{prox}_{tg} \left(B^{(k)} - t \nabla f(B^{(k)}) \right), \quad (3.4.41)$$

donde $\nabla f(B^{(k)})$ denota el gradiente de la pérdida multivariada respecto de B , $t > 0$ es el tamaño de paso (inverso de la constante de Lipschitz del gradiente de

f), y el operador proximal se aplica elemento a elemento sobre los coeficientes de B .

La versión acelerada FISTA introduce un término de extrapolación tipo Nesterov, que mejora la tasa de convergencia de $O(1/k)$ en ISTA a $O(1/k^2)$:

$$B^{(k+1)} = \text{prox}_{tg} \left(Y^{(k)} - t \nabla f(Y^{(k)}) \right), \quad (3.4.42)$$

$$\theta_{k+1} = \frac{1 + \sqrt{1 + 4\theta_k^2}}{2}, \quad (3.4.43)$$

$$Y^{(k+1)} = B^{(k+1)} + \frac{\theta_k - 1}{\theta_{k+1}} (B^{(k+1)} - B^{(k)}), \quad (3.4.44)$$

donde $Y^{(k)}$ es la matriz extrapolada obtenida a partir de las dos iteraciones previas.

Esta formulación permite aplicar los métodos de gradiente proximal a problemas de regresión multivariada penalizada, en los que la función de pérdida corresponde a una suma de términos diferenciables (como las pérdidas cuadrática o logística) y la penalización ℓ_1 se impone sobre todos los coeficientes de la matriz de regresión. De este modo, ISTA y FISTA constituyen alternativas generales al descenso por coordenadas para resolver problemas convexos con penalizaciones separables, logrando tasas de convergencia de $O(1/k)$ y $O(1/k^2)$ respectivamente (Beck & Teboulle, 2009).

2. Regresión de Ángulo Mínimo (LARS):

El método LARS constituye una alternativa eficiente para ajustar trayectorias de modelos lineales penalizados. Su funcionamiento es similar a un procedimiento secuencial: comienza incorporando la variable más correlacionada con la respuesta y, a medida que nuevas variables alcanzan una correlación equivalente con los residuos, se van añadiendo al conjunto activo. Los coeficientes de las variables activas se actualizan en bloque hasta completar el ajuste de mínimos cuadrados.

3. **Regresión Regularizada por Partición (SplitReg):** También existen otros métodos de optimización como la Regresión Regularizada por Partición. Esta técnica también denominada *Split-Regularized Regression (SplitReg)* consiste en dividir el conjunto de predictores en varios grupos y ajustar un modelo regularizado en cada uno, incorporando además una penalización que fomenta la diversidad entre los modelos. Las estimaciones resultantes se combinan en un promedio, funcionando como un ensamble (*ensemble*) lineal que mejora la precisión respecto al estimador regularizado de base (Christidis et al., 2020).

3.4.4 Modelos de Regresión de mínimos cuadrados parciales (PLS)

Fue propuesto por H. Wold (1982), en base al trabajo de su padre H. O. A. Wold (1966) donde introduce un enfoque de mínimos cuadrados iterativos (Iterative Least Squares) que permite la obtención de componentes principales en análisis de datos de alta dimensionalidad sin necesidad de calcular la matriz de variancia-covariancia completa (generalmente singular, cuando p es cercano a n), lo que se conoce como algoritmo NIPALS (*Non-linear Iterative Partial Least Squares*). Geladi y Kowalski (1986) ofrecen un tutorial del método de PLS. Siguiendo el marco teórico de los autores mencionados, se quiere predecir q variables de respuesta continuas, denotadas por $Y_1, \dots, Y_q$, utilizando p variables predictoras continuas $X_1, \dots, X_p$. Se tiene una muestra de datos de n observaciones. Tanto las variables X como las Y se estandarizan previamente. Las variables $x_i = (x_{i1}, \dots, x_{ip})'$ se organizan en una matriz X de dimensión $n \times p$. De manera similar, la matriz Y de dimensión $n \times q$ contiene a las variables $y_i = (y_{i1}, \dots, y_{iq})'$. PLS proyecta las variables X e Y en un espacio de menor dimensión mediante la construcción iterativa de componentes latentes (*scores*). A diferencia del ACP, cuyas combinaciones lineales se seleccionan para capturar la mayor variabilidad de los predictores, en PLS los componentes se eligen para maximizar la covarianza entre X e Y , garantizando que la variación de X retenida esté directamente asociada con la respuesta (Kuhn & Johnson, 2013,

p. 114). El modelo PLS se construye mediante las siguientes descomposiciones:

$$X = T P^\top + E, \quad Y = U Q^\top + F \quad (3.4.45)$$

donde: T y U son las matrices de *scores* (componentes latentes), P y Q son las matrices de *loadings* (cargas), E y F son las matrices residuales. En cada iteración, se construyen las variables latentes, estimadas mediante el algoritmo NIPALS, tal como se describe en Abdi (2010) y Geladi y Kowalski (1986). El primer paso consiste en inicializar las matrices residuales como $E = X$ y $F = Y$, las cuales se centran por columnas y se escalan a puntuaciones (*Z-scores*) antes de iniciar las iteraciones del algoritmo NIPALS. A continuación, para extraer cada par de componentes latentes (t, u) , NIPALS itera los pasos siguientes (“ $\propto$ ” indica que se normaliza cada vector para que su norma sea 1, por ejemplo $\|w\| = 1$):

1. $w \propto E^\top u$ (estima los pesos de X).
2. $t \propto E w$ (estima los *scores* de X).
3. $c \propto F^\top t$ (estima los pesos de Y).
4. $u \propto F c$ (estima los *scores* de Y).

Una vez que t ha convergido, se calcula el escalar $b = t^\top u$ y el *loading* $p = E^\top t$. A continuación, las matrices residuales se actualizan (deflacionan) con $E = E - t p^\top$ y $F = F - b t c^\top$, almacenándose los vectores t, u, w, c, p y el escalar b en sus respectivas matrices (T, U, W, C y P). Si E no es la matriz nula, el algoritmo retorna al primer paso para extraer la siguiente componente; de lo contrario, finaliza el procedimiento. La predicción de Y a partir de X se realiza mediante $\hat{Y} = T \beta C^\top = X \beta^{\text{PLS}}$, donde β^{PLS} es la matriz de coeficientes estimada por PLS. Los coeficientes de regresión β^{PLS} se obtienen como:

$$\beta^{\text{PLS}} = W (P^\top W)^{-1} C^\top \quad (3.4.46)$$

donde $W = [w_1, \dots, w_L]$ son los vectores de pesos de X y $P = [p_1, \dots, p_L]$ sus *loadings*. En la práctica, el número de componentes latentes se selecciona mediante CV, ya que aunque la teoría establece $L \leq \min(n, p)$, el modelo suele truncarse antes para evitar sobreajuste. Cabe señalar, además, que el método no proporciona directamente medidas de significación estadística para los coeficientes estimados. Sin embargo, dicha significación puede evaluarse aplicando procedimientos de remuestreo, tales como el *bootstrap* o pruebas de permutación, que permiten obtener intervalos de confianza y contrastes sobre los parámetros estimados (Acharjee et al., 2013; Boulesteix & Strimmer, 2007).

Rosipal y Trejo (2001) introducen el Kernel PLS (KPLS) un algoritmo que arroja coeficientes y predicciones $\hat{Y}$ algebraica y numéricamente equivalentes a los obtenidos con el algoritmo NIPALS clásico (Hovde Liland et al., 2024). El algoritmo de KPLS, se formula a partir de la versión NIPALS mediante la aplicación del método de *kernels* (núcleo en español), lo que permite obtener componentes en espacios de muy alta (incluso infinita) dimensión sin necesidad de cálculos explícitos en dicho espacio. Este método aplica un *kernel* $\phi: \mathbb{R}^p \rightarrow \mathcal{H}$ (lineal o gaussiano) para proyectar las observaciones $\mathbf{X}$ a un espacio de mayor dimensión, de modo que en lugar de calcular directamente $\mathbf{X}^\top \mathbf{X}$ se emplea la matriz de Gram

$$K = \langle \phi(X), \phi(X) \rangle \in \mathbb{R}^{n \times n},$$

donde $K: \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ es un *kernel* positivo semidefinido. Existe un espacio de Hilbert asociado $\mathcal{H}$ que satisface

$$\langle \phi(x), \phi(x') \rangle_{\mathcal{H}} = K(x, x'),$$

En este contexto, $\mathcal{H}$ es un espacio funcional donde los métodos lineales clásicos (p.ej., NIPALS-PLS o PCR) operan mediante funciones núcleo, sin necesidad de conocer las coordenadas $\phi(x)$. Así, las componentes latentes se obtienen resolviendo el problema PLS sobre K en lugar de sobre $\mathbf{X}$. A modo de síntesis, el funcionamiento

del algoritmo PLS puede observarse en la Figura [3.6](#)

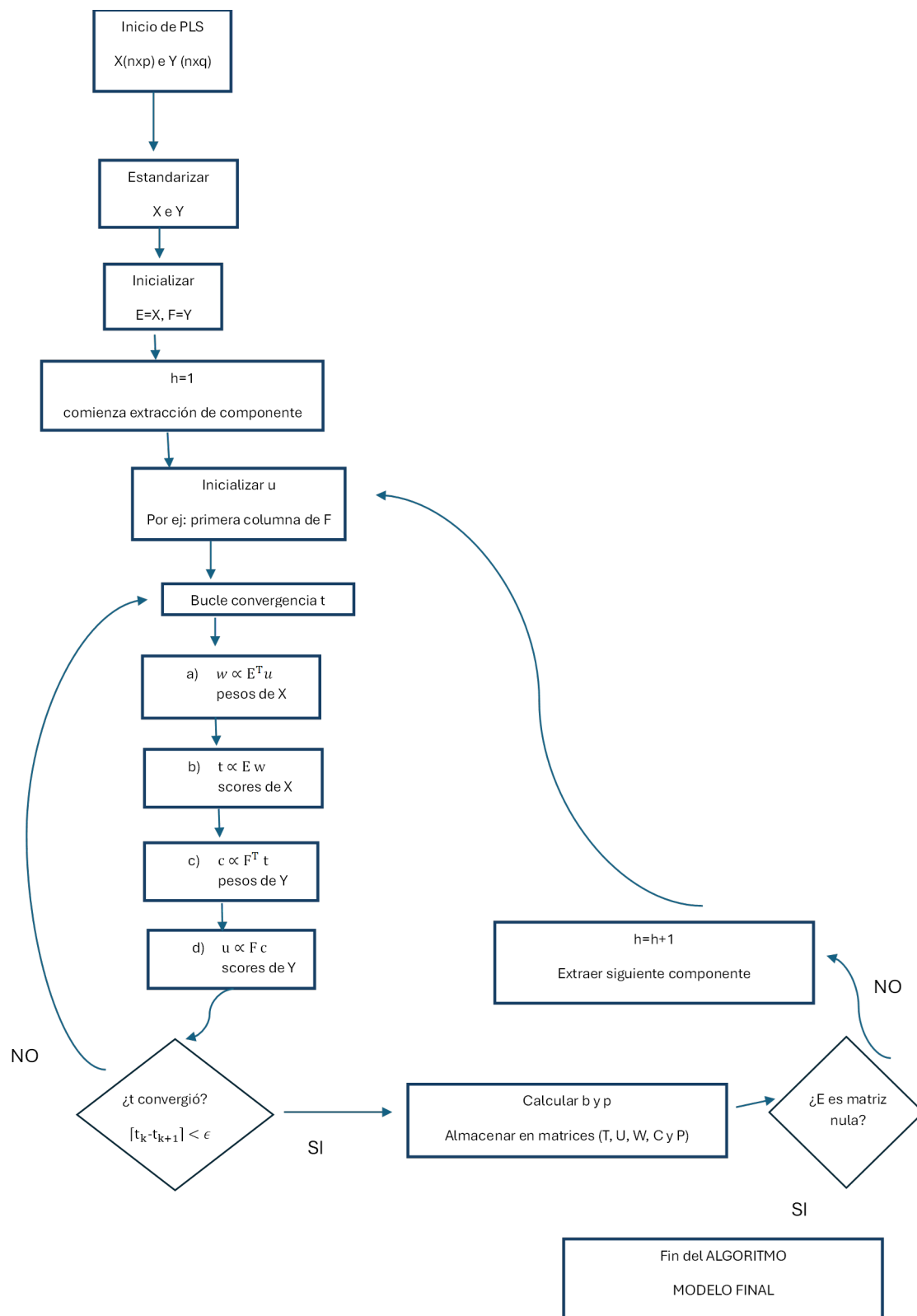


Figura 3.6: Diagrama del algoritmo PLS (NIPALS): extracción iterativa de componentes y deflación.

Nota. Elaboración propia.

3.4.5 Modelos Lineales Generalizados (MLG)

Los Modelos Lineales Generalizados (MLG) introducidos por Nelder y Wedderburn (1972), constituyen una extensión natural del modelo lineal clásico que permite analizar variables respuesta cuya distribución no necesariamente es normal y cuya varianza puede depender de la media. Mientras que en el modelo lineal clásico se asume que la variable respuesta es continua, normalmente distribuida y con varianza constante, los MLG amplían este marco al admitir distribuciones pertenecientes a la familia exponencial, tales como binomial, Poisson o gamma, entre otras.

Desde el punto de vista estructural, los MLG conservan el concepto central del predictor lineal, en el cual las predictoras se combinan linealmente mediante un conjunto de parámetros desconocidos. La diferencia fundamental radica en que la media de la variable respuesta no se vincula directamente con dicho predictor, sino a través de una función de enlace que permite modelar adecuadamente la relación entre ambos. De este modo, los MLG generalizan el modelo lineal clásico manteniendo su interpretabilidad y flexibilidad, al tiempo que amplían su campo de aplicación a una amplia variedad de tipos de datos.

En lo que sigue, se presenta la formulación general de los MLG a partir de sus tres componentes fundamentales. Posteriormente, se describen las distribuciones pertenecientes a la familia exponencial consideradas en este trabajo (Cuadro 3.2), los modelos específicos asociados a cada una de ellas y, finalmente, los procedimientos de estimación de parámetros e inferencia estadística.

Siguiendo el enfoque de los modelos lineales clásicos, se considera una variable aleatoria Y compuesta por n observaciones independientes, normalmente distribuidas y con varianza constante σ^2 . En este contexto, la esperanza de la variable respuesta se modela como

$$E(\mathbf{Y}) = \boldsymbol{\mu}, \quad \text{donde} \quad \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}. \quad (3.4.47)$$

Los MLG amplían este esquema manteniendo la estructura del predictor lineal,

pero permitiendo que la variable respuesta siga una distribución distinta de la normal y que la varianza dependa de la media. Esta generalización puede describirse a partir de tres componentes fundamentales (Faraway, 2016; Nelder & Wedderburn, 1972):

1. **Componente aleatoria:** la variable respuesta Y sigue una distribución perteneciente a la familia exponencial, con media $E(Y) = \mu$.
2. **Componente sistemática:** las predictoras $\mathbf{x}_1, \dots, \mathbf{x}_p$ se combinan linealmente en el predictor lineal

$$\eta = \sum_{j=1}^p \mathbf{x}_j \beta_j.$$

3. **Función de enlace:** una función monótona y diferenciable $g(\cdot)$ que vincula la media de la respuesta con el predictor lineal,

$$g(\mu_i) = \eta_i.$$

De este modo, un MLG relaciona la media de la variable respuesta con el predictor lineal mediante una función de enlace, extendiendo el modelo lineal clásico, que constituye el caso particular en el cual se asume normalidad, varianza constante y función de enlace identidad, es decir, $\mu = \eta$.

3.4.5.1 Familia exponencial y modelos asociados

Una característica central de los MLG es que la distribución de la variable respuesta pertenece a la familia exponencial. En su forma general, la función de densidad (o de probabilidad) puede escribirse como

$$f(y \mid \theta, \phi) = \exp \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right], \quad (3.4.48)$$

donde θ es el parámetro canónico, ϕ es el parámetro de dispersión, y $a(\cdot)$, $b(\cdot)$ y $c(\cdot)$ son funciones conocidas que determinan la distribución específica.

Esta formulación permite expresar, dentro de un mismo marco teórico, distribuciones comúnmente utilizadas en el modelado estadístico, tales como la normal,

Poisson, binomial (incluye Bernoulli), gamma y normal inversa, entre otras. En el contexto de los MLG, la media de la variable respuesta $\mu = E(Y)$ se vincula con el parámetro canónico θ y con el predictor lineal a través de la función de enlace.

Componente aleatoria	Enlace	Predictor	Modelo
Normal	Identidad	Continua	Regresión lineal
Normal	Identidad	Categórica	ANOVA
Normal	Identidad	Mixta	ANCOVA
Poisson	Log	Mixta	Regresión log-lineal
Binomial (incluye Bernoulli)	Logit	Mixta	Regresión logística
Multinomial	Logit generalizado	Mixta	Logit multinomial

Tabla 3.1: Principales modelos lineales generalizados

Fuente: adaptado de Agresti (2002, p. 118).

Tabla 3.2: Familias exponenciales univariadas en el marco de los MLG. Funciones de enlace canónicas y funciones de varianza.

Familia	Enlace canónico	Función de varianza
Normal	$\eta = \mu$	1
Poisson	$\eta = \log(\mu)$	μ
Binomial (incluye Bernoulli)	$\eta = \log\left(\frac{\mu}{1-\mu}\right)$	$\mu(1-\mu)$
Gamma	$\eta = \mu^{-1}$	μ^2
Normal Inversa	$\eta = \mu^{-2}$	μ^3

Nota: En este cuadro, el término univariado se refiere a modelos con un único predictor lineal y un parámetro natural escalar. En consecuencia, no se incluye la distribución multinomial, cuya formulación implica un parámetro natural vectorial y una estructura de varianza-covarianza matricial.

Fuente: adaptado de (Faraway, 2016, p. 154).

El caso Bernoulli corresponde a una distribución Binomial con un único ensayo, por lo que comparte el enlace canónico y la función de varianza dentro del marco de los MLG. De acuerdo con el tratamiento estándar de los MLG, este caso se incluye dentro de la familia binomial y no se presenta como una familia separada (Agresti, 2002; Dobson, 2002; McCullagh & Nelder, 1989). A continuación se describen brevemente las principales distribuciones pertenecientes a la familia exponencial consideradas en el marco de los MLG del cuadro 3.2, destacando su media, estructura de varianza y la función de enlace asociada.

Distribución Normal. La distribución normal se utiliza para variables respuesta continuas, simétricas y definidas en todo $\mathbb{R}$. En este caso, se asume que

$$Y \sim \mathcal{N}(\mu, \sigma^2),$$

donde la media μ no depende de la varianza, la cual es constante. La función de varianza es $V(\mu) = 1$ y la función de enlace canónica es la identidad, $g(\mu) = \mu$. Bajo estas condiciones, el MLG normal coincide con el modelo lineal clásico, incluyendo los modelos de regresión, análisis de la varianza y análisis de covarianza.

Distribución Poisson. La distribución Poisson es adecuada para modelar datos de conteo definidos sobre los enteros no negativos. Si $Y \sim \text{Poisson}(\mu)$, entonces

$$E(Y) = \mu, \quad \text{Var}(Y) = \mu,$$

lo que implica que la varianza coincide con la media. La función de enlace canónica es el logaritmo,

$$g(\mu) = \log(\mu),$$

tal como se resume en el Cuadro 3.2. Este modelo da lugar a los denominados modelos de regresión log-lineal, ampliamente utilizados en el análisis de tasas y frecuencias.

Distribución Binomial (incluida Bernoulli) y modelos logísticos. En el contexto de los MLG, cuando la respuesta Y_i representa el número de éxitos en m_i ensayos independientes, se asume formalmente que

$$Y_i \sim \text{Binomial}(m_i, \pi_i),$$

donde π_i denota la probabilidad de éxito en cada ensayo. La media y la varianza de la respuesta están dadas por:

$$\mathbb{E}(Y) = m\pi, \quad \text{Var}(Y) = m\pi(1 - \pi),$$

La función de enlace canónica se expresa en términos de π como

$$g(\pi) = \log\left(\frac{\pi}{1 - \pi}\right),$$

en concordancia con el enlace logit indicado en el Cuadro 3.2. El caso particular $m_i = 1$ corresponde a respuestas dicotómicas individuales y se asocia a la distribución Bernoulli. En este contexto, la media coincide con la probabilidad de éxito,

$$\mathbb{E}(Y_i) = \pi_i,$$

y la varianza se reduce a

$$\text{Var}(Y_i) = \pi_i(1 - \pi_i).$$

Bajo este marco, se introducen los modelos de regresión logística para el análisis de respuestas binarias y proporciones. En particular, el modelo de regresión logística para una variable explicativa X se expresa como:

$$\pi(x) = P(Y = 1 \mid X = x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)},$$

o, de manera equivalente, en términos del predictor lineal,

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \alpha + \beta x. \quad (3.4.49)$$

Notar que el enlace logit es una función lineal del predictor. La ecuación (3.4.49) combina la función de enlace con el predictor lineal del modelo. El signo del coeficiente β determina si la probabilidad de éxito $\pi(x)$ aumenta o disminuye a medida que X crece. En particular, si $\beta = 0$, la variable respuesta Y es independiente de la covariable X . Cuando X es una variable cuantitativa y $\beta > 0$, la función $\pi(x)$ adopta la forma de la función de distribución acumulada de la distribución logística.

En el caso de regresión logística múltiple, es decir, cuando se consideran $X_1, X_2, \dots, X_p$ variables explicativas, el predictor lineal se expresa como

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip},$$

y el modelo queda dado por

$$\text{logit}(\pi(x_i)) = \log\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}.$$

Además de las pruebas de hipótesis basadas en la devianza, es habitual emplear medidas que permitan evaluar el ajuste relativo y la capacidad predictiva de los modelos logísticos. Una de las más difundidas es el pseudo- R^2 de McFadden (McFadden, 1974), definido como

$$R_{\text{McFadden}}^2 = 1 - \frac{\ell(\hat{\pi} \mid y)}{\ell(\pi_0 \mid y)}, \quad (3.4.50)$$

donde $\ell(\hat{\pi} \mid y)$ representa la log-verosimilitud del modelo ajustado y $\ell(\pi_0 \mid y)$ la correspondiente al modelo nulo con solo intercepto. Aunque este coeficiente no debe interpretarse en los mismos términos que el R^2 de la regresión lineal, valores comprendidos entre 0,2 y 0,4 suelen indicar un ajuste razonable.

Otra herramienta ampliamente utilizada es la curva ROC (*Receiver Operating Characteristic*), que representa la sensibilidad frente a $1 - \text{especificidad}$ para distintos puntos de corte de probabilidad. El área bajo la curva (AUC, *Area Under the Curve*) resume esta información en un único valor numérico, con rango entre 0,5 (clasificación aleatoria) y 1 (clasificación perfecta), y se define como

$$\text{AUC} = \int_0^1 \text{TPR}(t) d\text{FPR}(t), \quad (3.4.51)$$

donde

$$\text{TPR}(t) = \frac{VP(t)}{VP(t) + FN(t)} \quad \text{y} \quad \text{FPR}(t) = \frac{FP(t)}{FP(t) + VN(t)}$$

corresponden a la tasa de verdaderos positivos (TPR, del inglés *True Positive Rate*,

también denominada sensibilidad) y a la tasa de falsos positivos (FPR, del inglés *False Positive Rate*, equivalente a $1 - \text{especificidad}$), respectivamente.. En conjunto, el pseudo- R^2 de McFadden y el AUC proporcionan criterios complementarios para contrastar el ajuste y la capacidad predictiva de los modelos logísticos.

En situaciones en las que la respuesta no se limita a un único indicador binario, sino que presenta más de dos categorías mutuamente excluyentes, el marco de la regresión logística puede extenderse naturalmente a los modelos logit multinomiales. estos modelos se presentan a continuación.

Modelos Logit Multinomial y Ordinal. Además de los casos binarios y de conteo, los MLG contemplan situaciones en las que la variable respuesta es categórica con más de dos categorías. Si las categorías son nominales, es decir, sin un orden natural, corresponde utilizar el modelo logit multinomial, que extiende la regresión logística binaria al comparar cada categoría con una categoría de referencia (Faraway, 2016, p. 129). En cambio, cuando las categorías presentan un orden inherente, se emplea el modelo logit ordinal, donde se modelan probabilidades acumuladas mediante una función de enlace apropiada (Faraway, 2016, p. 139).

Modelo logit multinomial (respuesta nominal). Sea $Y_i \in \{1, \dots, J\}$ y sea J la categoría de referencia. Para $j = 1, \dots, J - 1$ se define

$$\Pr(Y_i = j \mid \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \beta_j)}{1 + \sum_{k=1}^{J-1} \exp(\mathbf{x}_i^\top \beta_k)}, \quad (3.4.52)$$

y para la categoría de referencia:

$$\Pr(Y_i = J \mid \mathbf{x}_i) = \frac{1}{1 + \sum_{k=1}^{J-1} \exp(\mathbf{x}_i^\top \beta_k)}. \quad (3.4.53)$$

Modelo logit ordinal (respuesta ordinal). Sea $Y_i \in \{1, \dots, J\}$ con orden natural. El modelo logit ordinal se plantea sobre probabilidades acumuladas:

$$\Pr(Y_i \leq j \mid \mathbf{x}_i) = \gamma_{ij} = \frac{\exp(\alpha_j + \mathbf{x}_i^\top \beta)}{1 + \exp(\alpha_j + \mathbf{x}_i^\top \beta)}, \quad j = 1, \dots, J - 1, \quad (3.4.54)$$

donde α_j son puntos de corte que definen las probabilidades acumuladas y β representa el efecto común de los predictores sobre la respuesta ordinal.

Distribución Gamma. La distribución gamma es apropiada para modelar variables continuas positivas y asimétricas, como tiempos o montos. En el marco de los MLG se asume

$$E(Y) = \mu > 0, \quad \text{Var}(Y) = \phi \mu^2,$$

donde ϕ es un parámetro de dispersión. La función de enlace canónica es la inversa,

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

tal como se presenta en el Cuadro 3.2, aunque en aplicaciones prácticas también se emplea con frecuencia el enlace logarítmico por su facilidad de interpretación.

Distribución normal inversa. La distribución normal inversa se utiliza para variables continuas positivas cuya variabilidad aumenta más rápidamente que en el caso gamma. En este modelo,

$$E(Y) = \mu > 0, \quad \text{Var}(Y) = \phi \mu^3,$$

y la función de enlace canónica viene dada por

$$g(\mu) = \mu^{-2},$$

de acuerdo con la función de varianza indicada en el Cuadro 3.2. Este modelo resulta útil en contextos donde la dispersión crece de manera pronunciada con el nivel medio de la respuesta.

3.4.5.2 Modelo Bernoulli multivariado condicionado (Modelo logístico multivariado conjunto para respuestas binarias)

Cuando la respuesta Y está compuesta por varios indicadores binarios observados en forma simultánea, se plantea un modelo logístico multivariado conjunto en el que se estima un vector de funciones de enlace de manera simultánea. En este marco, se modela directamente la probabilidad de ocurrencia conjunta de múltiples variables binarias, admitiendo dependencia entre ellas (Dai et al., 2013). Este modelo puede interpretarse como una extensión natural de la regresión logística binaria al caso multivariado, en la cual la distribución de la respuesta corresponde a una Bernoulli multivariada condicionada a los predictores.

Desde el punto de vista de la familia exponencial, dicha formulación es equivalente a un modelo log-lineal binario multivariado, en el que se incorporan términos de interacción entre las respuestas para capturar asociaciones condicionales residuales (Dai et al., 2013). En presencia de covariables, esta formulación resulta matemáticamente equivalente a un modelo de tipo Ising condicional con predictores, una representación ampliamente estudiada en estadística gráfica y aprendizaje automático para describir estructuras de dependencia entre variables binarias condicionadas a información externa (Besag, 1974; Wainwright & Jordan, 2008), incluyendo desarrollos en alta dimensión basados en regularización (Ravikumar et al., 2010).

Siguiendo a Dai et al. (2013), consideremos primero la distribución base sin variables predictoras. Sea

$$Y = (Y_1, \dots, Y_q)^\top \in \{0, 1\}^q.$$

Para cada subconjunto no vacío $S \subseteq \{1, \dots, q\}$ definimos la estadística suficiente

$$Y_S = \prod_{j \in S} Y_j.$$

La distribución Bernoulli multivariada (sin predictoras) admite la forma expo-

nencial

$$\log p(y) = \sum_{\emptyset \neq S \subseteq \{1, \dots, q\}} \theta_S y_S - A(\theta), \quad (3.4.55)$$

donde los θ_S son los parámetros canónicos (efectos principales e interacciones) y $A(\theta)$ es la función generadora de cumulantes, también conocida como función de partición logarítmica, definida como

$$A(\theta) = \log \sum_{y \in \{0,1\}^q} \exp \left(\sum_{\emptyset \neq S \subseteq \{1, \dots, q\}} \theta_S y_S \right). \quad (3.4.56)$$

Los momentos de la distribución se obtienen como derivadas de la función $A(\theta)$. En particular, la media marginal de Y_j y la covarianza entre Y_j y Y_k están dadas por

$$\mu_j = \mathbb{E}(Y_j) = \frac{\partial A(\theta)}{\partial \theta_{\{j\}}}, \quad \text{Cov}(Y_j, Y_k) = \frac{\partial^2 A(\theta)}{\partial \theta_{\{j\}} \partial \theta_{\{k\}}}, \quad (3.4.57)$$

respectivamente.

El cálculo de la función $A(\theta)$ en la ecuación (3.4.56) requiere sumar sobre los 2^q posibles vectores binarios, lo que se vuelve computacionalmente prohibitivo para valores moderados o grandes de q . En la práctica, este problema se aborda mediante aproximaciones numéricas o restringiendo el modelo a interacciones de bajo orden.

Para incorporar el efecto de variables predictoras $x \in \mathbb{R}^p$, los parámetros canónicos se modelan como funciones lineales de x :

$$\theta_S(x) = \alpha_S + x^\top \beta_S, \quad \emptyset \neq S \subseteq \{1, \dots, q\}, \quad (3.4.58)$$

donde $\alpha_S \in \mathbb{R}$ es un intercepto y $\beta_S \in \mathbb{R}^p$ es un vector de coeficientes que cuantifica el efecto de las predictoras sobre el parámetro canónico θ_S . Nótese que para cada subconjunto S (efecto principal si $|S| = 1$ o interacción si $|S| \geq 2$) se estima un β_S distinto. Una parametrización muy usada es la truncada a interacciones de segundo

orden (análoga a un modelo de Ising con predictoras):

$$\log p(y | x) = \sum_{j=1}^q y_j \eta_j(x) + \sum_{1 \leq j < k \leq q} y_j y_k \phi_{jk} - A(\eta(x), \phi), \quad \eta_j(x) = \alpha_j + x^\top \beta_j, \quad (3.4.59)$$

donde $\eta_j(x)$ son efectos principales dependientes de x y ϕ_{jk} captura asociaciones residuales entre Y_j y Y_k (si $\phi_{jk} = 0$ para todo $j \neq k$, las respuestas son condicionalmente independientes dado x).

Dadas observaciones independientes $\{(y_i, x_i)\}_{i=1}^n$, la log-verosimilitud es

$$\ell(\Theta) = \sum_{i=1}^n \left\{ \sum_{\emptyset \neq S} y_{i,S} \theta_S(x_i) - A(\theta(x_i)) \right\}, \quad (3.4.60)$$

con $y_{i,S} = \prod_{j \in S} y_{ij}$ y $\Theta = \{\alpha_S, \beta_S\}$.

La estimación por máxima verosimilitud se realiza numéricamente. El caso de independencia condicional (sin interacciones, $\theta_S \equiv 0$ para $|S| \geq 2$) reduce la ecuación (3.4.55) a q logits univariados separados; el esquema con interacciones en la ecuación (3.4.59) permite modelar asociaciones condicionales entre respuestas binarias simultáneas.

3.4.5.3 Inferencia en MLG

Modelo nulo y modelo saturado. En el marco de los modelos lineales generalizados, el modelo nulo corresponde al modelo más simple, que incluye únicamente un intercepto y asume que los predictores no explican la variabilidad de la respuesta. En el extremo opuesto, el modelo completo o saturado es aquel que reproduce exactamente los datos observados, alcanzando la máxima log-verosimilitud posible. Estos dos modelos delimitan el rango de ajuste dentro del cual se evalúa el desempeño de un modelo propuesto (Faraway, 2016).

Devianza. La calidad del ajuste de un modelo lineal generalizado se evalúa mediante una estadística basada en la razón de verosimilitudes que compara el modelo

ajustado con el modelo saturado. Esta estadística se define como

$$2 \{l(y, \phi \mid y) - l(\hat{\mu}, \phi \mid y)\},$$

donde $l(y, \phi \mid y)$ denota la log-verosimilitud del modelo saturado y $l(\hat{\mu}, \phi \mid y)$ la correspondiente al modelo ajustado, siendo $\hat{\mu}$ la media estimada a partir de los parámetros del predictor lineal. El factor 2 se introduce por conveniencia asintótica.

Para observaciones independientes pertenecientes a la familia exponencial y cuando $a_i(\phi) = \phi/w_i$, esta cantidad puede escribirse como

$$D(y, \hat{\mu})/\phi,$$

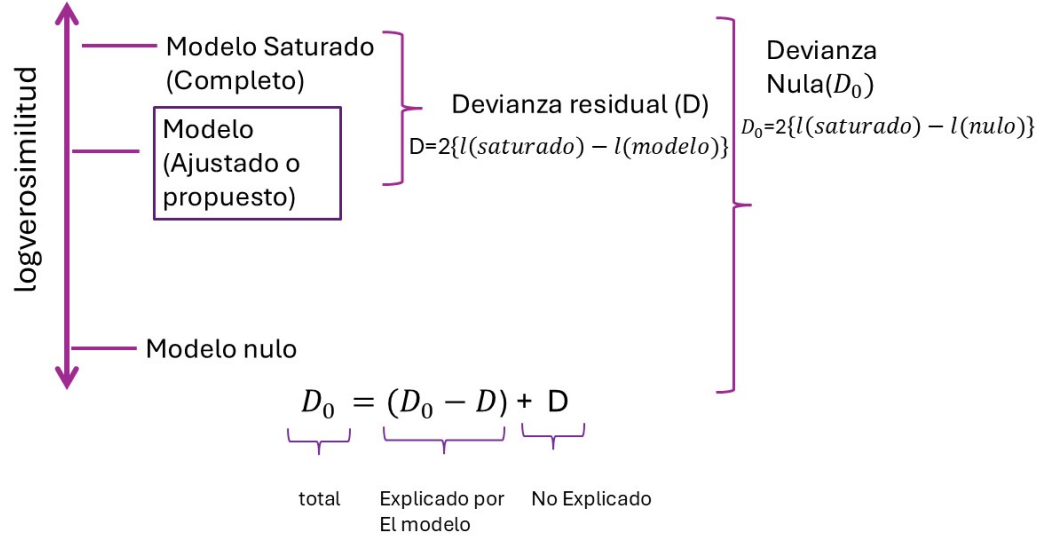
donde $D(y, \hat{\mu})$ se denomina devianza y $D(y, \hat{\mu})/\phi$ corresponde a la devianza escalada. En este contexto, la devianza cuantifica la discrepancia entre el modelo ajustado y el modelo saturado.

En particular, al considerar el modelo nulo y el modelo propuesto, se definen la devianza nula y la devianza residual, dadas respectivamente por

$$D_0 = 2 \{l(y, \phi \mid y) - l(\hat{\mu}_0, \phi \mid y)\} \quad \text{y} \quad D = 2 \{l(y, \phi \mid y) - l(\hat{\mu}, \phi \mid y)\},$$

donde $\hat{\mu}_0$ denota la media estimada bajo el modelo nulo. A partir de estas cantidades pueden definirse distintas medidas de devianza, como se ilustra en la Figura 3.7. Las expresiones explícitas de la devianza para algunos MLG comunes se presentan en el Cuadro 3.3. Las desvianzas correspondientes a los MLG comunes se presentan en el Cuadro 3.3.

Pruebas de hipótesis y comparación de modelos. En el marco de los MLG, se consideran principalmente dos tipos de pruebas de hipótesis. La primera es la prueba de bondad de ajuste, que evalúa si el modelo propuesto describe adecuadamente los datos observados. Bajo ciertas condiciones de regularidad, si el modelo es correcto, tanto la devianza escalada como la estadística χ^2 de Pearson son asintóticamente

Figura 3.7: Tipos de devianzas. *Elaboración propia.*

MLG	Devianza
Gaussiano	$\sum_i (y_i - \hat{\mu}_i)^2$
Poisson	$2 \sum_i \left[y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right]$
Binomial	$2 \sum_i \left[y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) + (m - y_i) \log \left(\frac{m - y_i}{m - \hat{\mu}_i} \right) \right]$
Gamma	$2 \sum_i \left[-\log \left(\frac{y_i}{\hat{\mu}_i} \right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right]$
Gaussiana inversa	$\sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2 \nu_i}$

Nota: La distribución multinomial no se incluye en el cuadro debido a que la respuesta y el parámetro medio son de naturaleza vectorial, lo que conduce a una expresión de la devianza que no admite una forma escalar simple.

Tabla 3.3: Devianza de los MLG comunes. Fuente: Faraway (2016, p. 157).

distribuciones χ^2 , con grados de libertad iguales al número de observaciones menos el número de parámetros estimados.

La segunda prueba corresponde a la comparación entre modelos anidados. Sea Ω un modelo más general y ω un modelo anidado más simple; entonces, la diferencia entre sus desvianzas escaladas,

$$D_\omega - D_\Omega,$$

se distribuye asintóticamente como una χ^2 , con grados de libertad iguales a la diferencia en el número de parámetros identificables entre ambos modelos (Faraway, 2016). Este contraste permite evaluar la contribución estadística de subconjuntos de

predictores dentro del modelo.

La segunda prueba corresponde a la comparación entre modelos anidados. Sea Ω un modelo más general y ω un modelo anidado más simple; entonces, la diferencia entre sus desvianzas escaladas,

$$D_\omega - D_\Omega,$$

se distribuye asintóticamente como una χ^2 , con grados de libertad iguales a la diferencia en el número de parámetros identificables entre ambos modelos (Faraway, 2016). Este contraste permite evaluar la contribución estadística de subconjuntos de predictores dentro del modelo, y constituye un procedimiento estándar para realizar inferencia sobre los parámetros β en el marco de los modelos lineales generalizados.

3.5 Métodos de Remuestreo

Tal como lo expresan James et al. (2013, p. 175):

Los métodos de remuestreo consisten en extraer repetidamente muestras de un conjunto de entrenamiento y ajustar un modelo de interés en cada muestra con el fin de obtener información adicional sobre el modelo ajustado. Por ejemplo, para estimar la variabilidad de un ajuste de regresión lineal, se puede extraer repetidamente diferentes muestras de los datos de entrenamiento, ajustar una regresión lineal en cada nueva muestra y luego examinar hasta qué punto difieren los ajustes resultantes. Este enfoque puede permitirnos obtener información que no estaría disponible si ajustáramos el modelo solo una vez utilizando la muestra de entrenamiento original.

Los métodos de remuestreo se originan a mitad del siglo XX con la técnica de *Jackknife* introducida por Quenouille (1949, 1956) y luego Tukey (1958) la popularizó. El término *Jackknife* —acuñado por Tukey como una “navaja de usos múltiples”— se basa en repetir el cálculo de un estimador de una muestra, eliminando una observación distinta cada vez para estudiar el sesgo o la varianza del estimador

original. Luego apareció la validación cruzada en diversos trabajos, entre ellos Allen (1974), Stone (1974) y Geisser (1975). Por su parte, el *bootstrap* fue propuesto por Efron (1979) y más tarde difundido ampliamente en aplicaciones estadísticas por Efron y Tibshirani (1993).

A continuación, se presentarán con mayor detalle dos de las técnicas de remuestreo más utilizadas: la CV y el bootstrap.

3.5.0.1 Validación cruzada

La validación cruzada *K-fold* consiste en dividir el conjunto de datos en K particiones aproximadamente del mismo tamaño y, de forma iterativa, emplear una de ellas como conjunto de validación mientras que las restantes se utilizan para entrenar el modelo (Hastie et al., 2009). Si bien sus primeras formulaciones datan de la década de 1970, el método se difundió y consolidó en décadas posteriores como una herramienta estándar en la evaluación de modelos.

Siguiendo la notación propuesta en Hastie et al. (2009), se define

$$\kappa : \{1, \dots, n\} \rightarrow \{1, \dots, K\}$$

como una función de indexación que indica la partición a la cual la observación i es asignada por la aleatorización.

Sea $f(x, \theta)$ un modelo parametrizado por θ . Denotemos por

$$\hat{f}^{-k}(x, \theta)$$

el modelo ajustado con parámetro θ tras excluir la k -ésima parte de los datos. Entonces, la estimación del error de predicción mediante CV es

$$\text{CV}(\hat{f}, \theta) = \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{f}^{-\kappa(i)}(x_i, \theta)), \quad (3.5.1)$$

donde:

- n es el número total de observaciones.

- $L(\cdot, \cdot)$ es la función de pérdida (p. ej., error cuadrático medio, devianza, etc.).
- $\kappa(i)$ indica la partición (*fold*) a la que pertenece la observación i .
- $\hat{f}^{-\kappa(i)}(x_i, \theta)$ es el modelo ajustado sin incluir los datos de la partición $\kappa(i)$.

Usualmente, se elige K igual a 5 o 10, ya que estos valores ofrecen un equilibrio adecuado entre sesgo y varianza. El caso extremo $K = n$ se conoce como validación cruzada *leave-one-out* (LOOCV), en el cual $\kappa(i) = i$, es decir, para la i -ésima observación, el ajuste se realiza utilizando todos los datos excepto la propia observación i .

La función $CV(\hat{f}, \theta)$ proporciona una estimación de la curva de error de prueba. Luego, se busca el parámetro de ajuste $\hat{\theta}$ que la minimice. El modelo final elegido es $f(x, \hat{\theta})$, ajustado a la totalidad de los datos.

Si $K=5$ el escenario es:

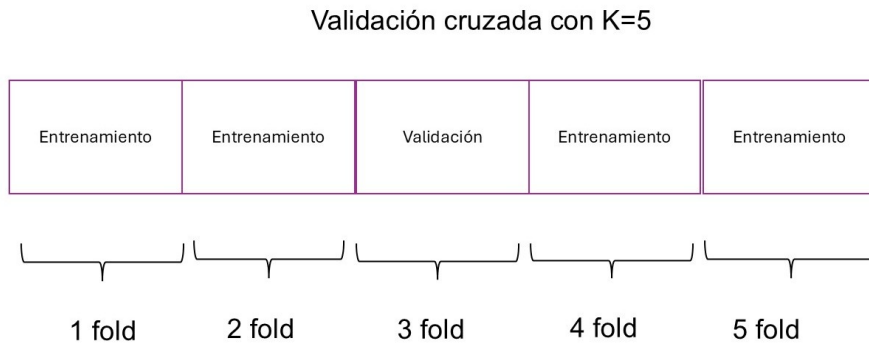


Figura 3.8: Validación Cruzada

En el contexto de un modelo lineal con variable respuesta cuantitativa, resulta fundamental calcular el error cuadrático medio (ECM), ya que esta medida cuantifica la diferencia promedio entre los valores observados y los predichos. En cada iteración del procedimiento de CV, se obtiene un valor del error a partir del conjunto de validación correspondiente, denotado como $ECM_{\text{sub}1}$. Tras repetir el proceso K veces, se generan K estimaciones: $ECM_1, ECM_2, \dots, ECM_K$. La estimación del CV de k -

folds, se estima promediando esos valores (James et al., 2013):

$$CV(k) = \frac{1}{k} \sum_{i=1}^k ECM_i. \quad (3.5.2)$$

Este proceso se repite iterativamente y se calculan las curvas de CV. Se observa en dichas curvas el punto mínimo. Es decir el que arroja el menor ECM.

3.5.0.2 Bootstrap

Al introducir el *bootstrap*, Efron (1979) enfrentó uno de los desafíos centrales de la estadística: obtener un estimador para un parámetro de interés y, al mismo tiempo, evaluar su precisión. Para ello propuso un procedimiento de remuestreo que permite estimar el error estándar y construir intervalos de confianza de manera no paramétrica. Con el tiempo, el *bootstrap* se consolidó como una herramienta estadística fundamental, cuya utilidad trasciende la evaluación de la precisión de un estimador. Sus aplicaciones abarcan la estimación de la tasa de error en análisis discriminante, selección de subconjuntos en problemas de regresión, regresión logística y clasificación, análisis de conglomerados, *kriging* (modelado espacial), regresión no lineal, análisis de series de tiempo, procesamiento de encuestas complejas, ajuste del valor p en pruebas múltiples, así como en análisis de supervivencia y confiabilidad (Chernick, 2008, p. 8).

El *bootstrap* se destaca en su aplicabilidad a contextos en los que las medidas de variabilidad o de precisión de los estimadores son complejas de obtener o no se generan de forma automática. Un ejemplo representativo de esta situación lo constituyen los modelos de regresión de mínimos cuadrados parciales (PLS), en los cuales resulta inviable calcular los p-valores de los coeficientes estimados.

Siguiendo el marco teórico presentado en Hastie et al. (2009), considérese una muestra de n vectores aleatorios independientes $X_1, X_2, \dots, X_n$, y un estimador $\hat{\theta}$ de un parámetro de interés. La función de distribución empírica F_n asigna probabilidad $1/n$ a cada observación, y al muestrear con reemplazo de F_n se generan muestras *bootstrap*. La distribución de $\hat{\theta}^*$ obtenida a partir de estas remuestras permite apro-

ximar la distribución muestral del estimador, proporcionando así una estimación del error estándar y posibilitando la construcción de intervalos de confianza de manera no paramétrica.

Desde un punto de vista computacional, el procedimiento consiste en:

1. Generar una muestra *bootstrap* de tamaño n mediante muestreo con reemplazo a partir de los datos originales.
2. Calcular el estimador $\hat{\theta}^*$ en la muestra *bootstrap*.
3. Repetir los pasos 1 y 2 un total de B veces (por ejemplo, $B = 1000$).

De este modo se obtienen B estimaciones $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$. A partir de ellas se puede estudiar la variabilidad de cualquier estadístico $S(Z)$ calculado sobre los datos. En particular, la varianza se estima como:

$$\widehat{\text{Var}}[S(Z)] = \frac{1}{B-1} \sum_{b=1}^B \left(S(Z^{*b}) - \bar{S}^* \right)^2, \quad (3.5.3)$$

donde:

$$\bar{S}^* = \frac{1}{B} \sum_{b=1}^B S(Z^{*b}).$$

Existen distintas variantes para construir intervalos de confianza a partir del *bootstrap*. Una de las más simples es el método percentil (Chernick, 2008), que utiliza directamente los percentiles de la distribución empírica de las estimaciones $\hat{\theta}_b^*$. Por ejemplo, un intervalo de confianza al 90 % para θ puede definirse a partir de los percentiles 5 y 95 de la distribución $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$.

3.6 Análisis de Redundancia

El Análisis de Redundancia, conocido por sus siglas en inglés **RDA** (*Redundancy Analysis*), es una técnica estadística que combina regresión lineal múltiple con métodos multivariados de tipo canónico. Su objetivo es explicar la variabilidad de un conjunto de variables respuesta $\mathbf{Y}$ a partir de un conjunto de predictores $\mathbf{X}$, proyec-

tando dicha relación en un espacio de dimensiones reducidas, de manera análoga al ACP. En este sentido, el RDA integra el poder explicativo de los modelos lineales con la capacidad de síntesis de las técnicas de reducción de dimensión, constituyendo una herramienta valiosa para analizar relaciones complejas dentro de una estructura de predictores.

Rao (1973) en su libro, denomina a lo que hoy conocemos como análisis canónico de redundancia componentes principales de variables instrumentales, y lo propone como tema de futuras investigaciones. Posteriormente, van den Wollenberg (1977) retoma el tema y lo denomina análisis de redundancia, en referencia al índice de redundancia que es la proporción de varianza de la matriz de la variable respuesta explicada por la matriz de $\mathbf{X}$ (índice de redundancia de Stewart y Love (1968)).

Es en el campo de la biología y ciencias ambientales, donde por lejos, más se ha empleado el análisis de redundancia para relacionar la composición de comunidades biológicas, medidas por número de especies o abundancia en diferentes sitios, con un variado número de variables ambientales como clima, suelo, nutrientes, contaminación, etc.

3.6.1 Fundamentos metodológicos

Formalmente, sea $\mathbf{Y}_{n \times q}$ la matriz que contiene las q variables dependientes observadas en n individuos, y $\mathbf{X}_{n \times p}$ la matriz de p variables explicativas correspondientes.

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nq} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Ambas matrices son previamente estandarizadas y el proceso de análisis de redundancia puede explicarse a partir de la Figura 3.9.

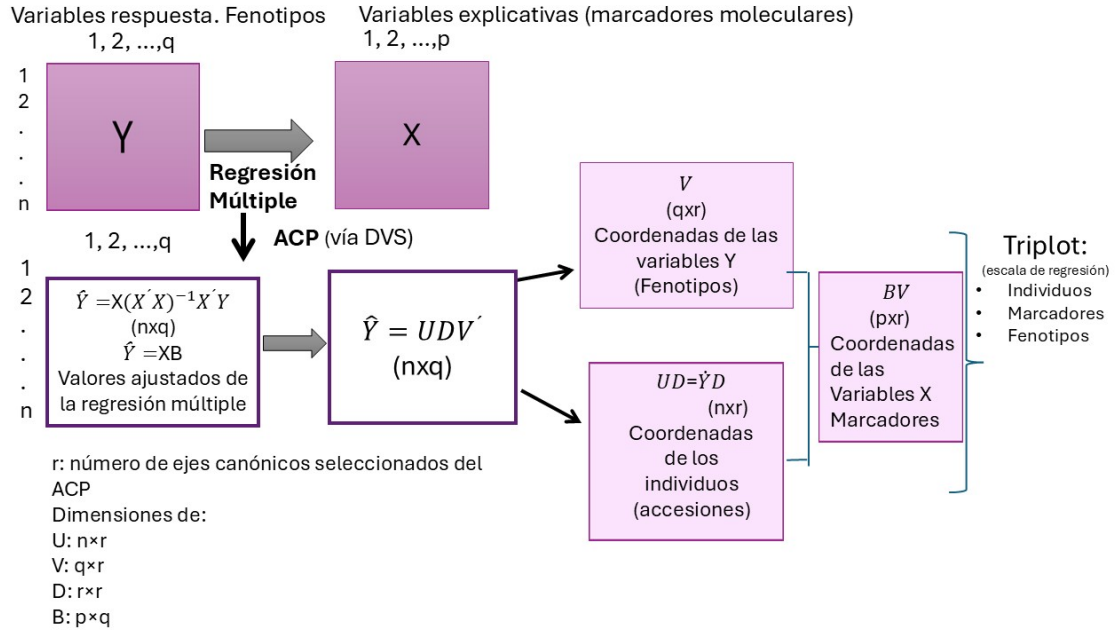


Figura 3.9: Proceso del análisis de redundancia (RDA). Adaptado de Legendre y Legendre (2012).

Nota sobre la escala de representación. Los softwares de análisis multivariado suelen representar los triplots de RDA mediante proyecciones basadas en correlaciones (o cosenos direccionales) entre las variables y los ejes canónicos. En el esquema presentado se adopta una escala de regresión, en la cual las coordenadas de los individuos corresponden a las proyecciones directas UD , las de los fenotipos a los vectores V , y las de los marcadores a los coeficientes canónicos proyectados BV .

El primer paso, siguiendo el enfoque metodológico de Legendre y Legendre (2012), consiste en ajustar q regresiones lineales múltiples independientes, una para cada variable respuesta Y_j en función de $\mathbf{X}$. La matriz que reúne todos los coeficientes de regresión puede escribirse como en la ecuación (3.4.8), donde $\mathbf{B}$ es la matriz de coeficientes de dimensión $p \times q$. En consecuencia, los valores ajustados de cada regresión, $\hat{Y}_j$, se disponen como columnas para conformar la matriz de valores ajustados $\hat{\mathbf{Y}}$, la cual se expresa como en la ecuación (3.4.9).

El carácter multivariado del RDA emerge al analizar conjuntamente los valores ajustados $\hat{\mathbf{Y}}$ mediante un ACP. En este trabajo, el ACP se implementa computacionalmente mediante la descomposición en valores singulares de la matriz $\hat{\mathbf{Y}}$, lo cual

es algebraicamente equivalente a la descomposición espectral de $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$. En particular, se obtiene la factorización

$$\hat{\mathbf{Y}} = UDV',$$

donde la matriz $V \in \mathbb{R}^{q \times r}$ contiene los autovectores asociados a la matriz $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$, que definen las direcciones canónicas en el espacio de las variables respuesta, y $D \in \mathbb{R}^{r \times r}$ es la matriz diagonal asociada a los autovalores correspondientes. La matriz $U \in \mathbb{R}^{n \times r}$ contiene las coordenadas de los individuos en el espacio canónico restringido.

Bajo esta descomposición, las coordenadas de los individuos (accesiones) están dadas por la matriz UD , mientras que las coordenadas de las variables respuesta (fenotipos) corresponden a las columnas de V . Las coordenadas de las variables explicativas (marcadores moleculares) se obtienen proyectando los coeficientes de regresión sobre los ejes canónicos, resultando en la matriz BV . Esta representación corresponde a una escala de regresión, tal como se ilustra en la Figura 3.9.

Índice de redundancia. La medida básica de ajuste es la redundancia estadística, análoga al coeficiente de determinación en regresión múltiple:

$$R_{Y|X}^2 = \frac{SC(\hat{\mathbf{Y}})}{SC(\mathbf{Y})}$$

Donde $SC(\hat{\mathbf{Y}})$ es la suma de cuadrados de $\hat{Y}$ y $SC(\mathbf{Y})$ es la suma de cuadrados de las observaciones Y . En la práctica suele reportarse también su versión ajustada R_a^2 , que corrige el sesgo de sobreestimación al incluir muchos predictores (Legendre & Legendre, 2012).

3.6.2 Visualización mediante triplot

La representación gráfica más utilizada en el RDA es el **triplot**, que combina en un mismo gráfico tres conjuntos de elementos: individuos, variables respuesta y variables explicativas. Continuando con la descomposición matricial ilustrada en la

Figura 3.9.

En esta representación, el triplot se construye a partir de:

1. **Coordenadas de los individuos** (sitios o accesiones), que corresponden a las puntuaciones en el espacio canónico restringido y se obtienen como las proyecciones directas de los valores ajustados sobre los ejes canónicos.
2. **Vectores de las variables respuesta $\mathbf{Y}$** , que representan las direcciones canónicas asociadas a la variación explicada, y se corresponden con los auto-vectores de la matriz $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$.
3. **Vectores de las variables explicativas $\mathbf{X}$** , que reflejan la contribución de cada predictor a los ejes canónicos y se obtienen proyectando los coeficientes de regresión sobre dichas direcciones.

En términos matriciales, las coordenadas de los individuos en el espacio canónico restringido están dadas por el producto UD . Por su parte, las coordenadas de las variables respuesta corresponden a las columnas de la matriz V . Finalmente, las coordenadas de las variables explicativas se obtienen proyectando la matriz de coeficientes de regresión sobre los ejes canónicos, lo que conduce al producto BV . Esto da lugar a una representación en escala de regresión, en la cual las longitudes y direcciones de los vectores de $\mathbf{X}$ se interpretan en términos de la contribución de los predictores a los ejes canónicos, y no como correlaciones directas.

De este modo, el triplot constituye una síntesis gráfica de la estructura del RDA: los puntos representan la posición relativa de los individuos en el espacio canónico restringido, los vectores de las variables respuesta describen las direcciones principales de la variación explicada, y los vectores de las variables explicativas indican cómo cada predictor contribuye a dicha variación multivariada. En la práctica, la implementación del triplot de RDA en distintos paquetes de software suele basarse en proyecciones construidas a partir de correlaciones o cosenos direccionales con los ejes canónicos.

Capítulo 4

Mapeo Asociativo

El Mapeo Asociativo (MA) es definido por Breseghello y Sorrells (2006) como una metodología estadística introducida en estudios de mejoramiento genético para identificar genes responsables de la variación de caracteres cuantitativos donde adquiere importancia establecer la significancia estadística de las asociaciones entre la información que proveen marcadores moleculares y características fenotípicas de interés.

4.1 Modelo sin corrección por estructura (*Naive*)

El primer modelo propuesto para el MA fue el denominado modelo ingenuo o *Naive*, conocido como SCE (sin corrección por estructura), el cual utiliza regresiones lineales simples en las que los marcadores moleculares se consideran como variables independientes con efectos fijos y cada fenotipo como variable dependiente. Es decir se ajusta un modelo separado para combinación de fenotipo y marcador molecular generando tantos modelos como la multiplicación de la cantidad de fenotipos por la cantidad de marcadores moleculares. El modelo *Naive* puede representarse mediante la ecuación (3.4.7). Alternativamente, es posible expresar este modelo de forma vectorial como sigue:

$$\vec{y}_{n \times 1} = X_{n \times 2} \vec{\beta}_{2 \times 1} + \vec{\epsilon}_{n \times 1}, \quad (4.1.1)$$

donde n representa la cantidad de observaciones, $\vec{y}$ representa el valor observado del carácter fenotípico (variable dependiente o respuesta), que se relaciona con cada marcador molecular x (variable independiente o regresora) mediante el coeficiente de regresión $\vec{\beta}$, el cual debe estimarse a partir de los datos disponibles para cada

marcador individualmente, y $\vec{\epsilon}$ corresponde al término de error aleatorio, bajo el supuesto de independencia y distribución normal. Este modelo no incluye ninguna corrección por estructura genética poblacional.

4.2 Modelo por estructura genética poblacional (EGP)

P y Q

Posteriormente al modelo *Naive*, se desarrollaron los modelos conocidos como modelos de Estructura Genética Poblacional (EGP). Dentro de este grupo están los modelos P y Q. En los modelos P, se incorpora información sobre la dependencia genética existente entre individuos mediante técnicas basadas en ACP (Pritchard et al., 2000). La matriz P se construye a partir de las componentes principales obtenidas del ACP sobre los datos moleculares, seleccionándose aquellas componentes estadísticamente significativas según el criterio de *Tracy-Widom*. En los modelos Q, se emplea información basada en las probabilidades de pertenencia de los individuos a grupos genéticos inferidos previamente, es decir, la matriz Q se construye a partir de la probabilidad de cada individuo de integrar un grupo genéticamente predeterminado. Los modelos P y Q, pueden expresarse:

$$(\text{Modelo P}) \quad \vec{y}_{n \times 1} = X_{n \times 2} \vec{\beta}_{2 \times 1} + P_{n \times k} \vec{u}_{k \times 1} + \vec{\epsilon}_{n \times 1}. \quad (4.2.1)$$

$$(\text{Modelo Q}) \quad \vec{y}_{n \times 1} = X_{n \times 2} \vec{\beta}_{2 \times 1} + Q_{n \times k} \vec{u}_{k \times 1} + \vec{\epsilon}_{n \times 1}. \quad (4.2.2)$$

donde n es la cantidad de observaciones, k es la cantidad de componentes que se incluyen en el modelo P y la cantidad de grupos genéticos de pertenencia en el modelo Q, $\vec{y}$ es el vector fenotípico, X es la matriz de diseño que contiene el marcador que se incluye en el modelo, $\vec{\beta}$ es el vector de parámetros (o efectos fijos), P o Q son las matrices que indican la estructura genética, $\vec{u}$ es el vector de efectos de la población y $\vec{\epsilon}$ es el vector de errores. Además se han propuesto enfoques que

combinan simultáneamente las metodologías de los modelos P y Q.

4.3 Modelo kinship (K)

Los modelos *Kinship* (parentesco) fueron propuestos por Yu y Buckler (2006) y posteriormente modificados por Kang et al. (2008). Estos utilizan una matriz de coancestría que presenta las probabilidades de pertenencia de los individuos a k grupos en los que se podría subdividir el material bajo estudio. Esta matriz de coancestría indica las relaciones de parentesco entre los efectos aleatorios. El modelo se expresa como:

$$\vec{y}_{n \times 1} = X_{n \times 2} \vec{\beta}_{2 \times 1} + Z_{n \times m} \vec{u}_{m \times 1} + \vec{\epsilon}_{n \times 1}, \quad (4.3.1)$$

donde n es la cantidad de observaciones, $\vec{y}$ es el vector fenotípico, X es la matriz de diseño que contiene el marcador molecular, $\vec{\beta}$ es el vector de parámetros (o efectos fijos), $\vec{u}$ es el vector de efectos poligénicos Z es una matriz de incidencia asociada a $\vec{v}$ de efectos poligénicos, $\vec{\epsilon}$ vector de errores Los supuestos del modelo son:

- El vector $\mathbf{u}$ es independiente del vector de errores $\vec{\epsilon}$, su matriz de varianzas y covarianzas está dada por $\text{Var}(\mathbf{u}) = \sigma_g^2 \mathbf{K}$.
- La varianza del vector de errores es

$$\text{Var}(\epsilon) = \sigma_\epsilon^2 \mathbf{I}.$$

- La matriz de varianzas y covarianzas de los fenotipos será:

$$V = \sigma_g^2 Z K Z' + \sigma_\epsilon^2 I.$$

σ_g^2 representa la varianza genética asociada a los efectos aleatorios de origen genético y σ_ϵ^2 es la varianza de los errores

Existen propuestas de modelos mixtos que combinan los modelos K con P y Q.

4.4 Métodos de corrección por multiplicidad

En las investigaciones genéticas es usual realizar pruebas de hipótesis de asociación entre cada marcador y el fenotipo estudiado. Debido a que la cantidad de marcadores genéticos es muy grande, se enfrenta el problema de la multiplicidad al evaluar hipótesis de asociación, en la que se asigna un error tipo I a cada test y, en consecuencia, aumenta la probabilidad de errores cuando se consideran todos los test (J. Li & Ji, 2005). Para corregir el error tipo I y minimizar la incidencia de falsos positivos, se emplean los denominados métodos de corrección por multiplicidad, entre los cuales por orden cronológico encontramos los siguientes: Bonferroni (1935), Holm (1979), Bergmann y Hommel (1988), Hochberg (1988), Benjamini y Hochberg (1995), Benjamini y Yekutieli (2001), Cheverud (2001), J. Li y Ji (2005), Foster y Stine (2008), A. N. Peña Malavera (2015) y Ramdas et al. (2019). En este trabajo se adoptó el método propuesto por J. Li y Ji (2005), el cual constituye una extensión del ajuste de Bonferroni que incorpora explícitamente la dependencia entre marcadores genéticos. En estudios de asociación, los marcadores moleculares suelen presentar correlaciones sustanciales debido a fenómenos como el desequilibrio de ligamiento (LD), violando así el supuesto de independencia que subyace al ajuste de Bonferroni clásico (Cheverud, 2001). El método de Li y Ji aborda este problema estimando un número efectivo de pruebas independientes (m_{eff}) mediante un ACP aplicado a la matriz de correlación entre marcadores. Para ello se realiza un ACP sobre la matriz de correlación entre marcadores moleculares y se calcula el número efectivo de pruebas independientes m_{eff} (por ejemplo, contabilizando los autovalores mayores o iguales que uno, o bien reteniendo las componentes principales necesarias para explicar un porcentaje elevado de la variabilidad total, como el 99 %). Este ajuste produce un umbral de significancia menos conservador que el de Bonferroni, pero sin perder el control riguroso del error tipo I. Los autores demostraron, mediante simulaciones y aplicaciones a datos reales, que este enfoque incrementa sustancialmente la potencia para detectar asociaciones verdaderas en comparación con la corrección tradicional (J. Li & Ji, 2005). Gracias a este equilibrio entre control

de falsos positivos y sensibilidad, el método de Li y Ji se ha consolidado como una opción ampliamente recomendada en genética molecular y estudios de asociación con alto número de marcadores correlacionados (Foster & Stine, 2008; A. N. Peña Malavera, 2015).

El valor de m_{eff} se utiliza luego en las fórmulas de Bonferroni para determinar el nivel de significancia corregido α_p :

$$\alpha_p = 1 - (1 - \alpha_{\text{global}})^{1/m_{\text{eff}}} \quad (4.4.1)$$

$$\alpha_p \approx \frac{\alpha_{\text{global}}}{m_{\text{eff}}} \quad (4.4.2)$$

donde α_p es el error tipo I particular de cada prueba.

Dado que estos valores de error tipo I para cada prueba suelen ser muy pequeños, es común expresarlos en términos logarítmicos y construir un gráfico de Manhattan (*Manhattan plot*) en el que se representen los valores de $-\log_{10}(p)$ correspondientes a cada marcador molecular, considerando como significativos aquellos que superen el umbral $-\log_{10}(\alpha_p)$.

Capítulo 5

Materiales

Para llevar a cabo esta investigación se utilizaron dos conjuntos de datos fundamentales. El primero consiste en una colección de accesiones de quinoa, caracterizadas a nivel morfológico, agronómico y molecular, proveniente de los bancos de germoplasma del INTA. El segundo es el **Rice Diversity Panel 1 (RDP1)**, un panel de diversidad de arroz que cuenta con una caracterización integral análoga. Estos recursos permiten analizar la relación genotipo-fenotipo en dos cultivos de gran relevancia.

5.1 Base de datos de Quinoa

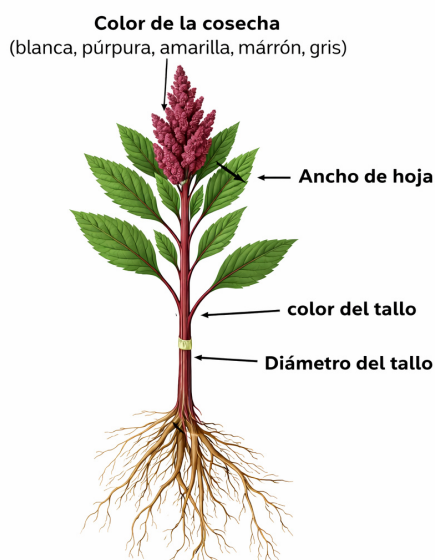
Se trata de una base de datos constituida por 25 accesiones de quinoa provenientes del Noroeste Argentino. La caracterización del germoplasma se llevó a cabo en la localidad de Caleta, Departamento de Humahuaca (23° 12' S, 65° 20' O; 2939 msnm. Jujuy-Argentina); durante el ciclo de cultivo 2008-2009. Se condujo un experimento donde las accesiones fueron sembradas en un diseño en bloques completo aleatorio con dos repeticiones. Cada parcela consistió en tres surcos de 5 m de longitud, separados por 0,5 m. La densidad de siembra fue de 14 semillas m⁻¹, equivalente a 280.000 semillas ha⁻¹. El experimento se mantuvo libre de malezas y plagas, y se realizaron las prácticas culturales clásicas para el cultivo. La cosecha se realizó según la precocidad de las accesiones (Curti et al., [2012](#)). Se utilizó la lista de descriptores de especies (IBPRG 1981, actualmente Bioversity International) para medir (cuantitativamente) y calificar (cualitativamente) los descriptores fenotípicos en 10 plantas (5 en cada parcela replicada) elegidas aleatoriamente de la fila central.

Se relevaron 17 atributos cuantitativos y 11 atributos cualitativos nominales.

Las variables cuantitativas consideradas fueron: diámetro del tallo, número de ramas, relación longitud/ancho de la hoja superior, número de dientes, longitud del pecíolo, longitud de la hoja, ancho de la hoja, longitud de la panoja, longitud del glomérulo, longitud del cotiledón, longitud del hipocótilo, días desde la siembra hasta la emergencia, días desde la siembra hasta botón floral, días desde la siembra hasta la floración, días desde la siembra hasta madurez fisiológica, altura de planta y diámetro del grano. Las variables cualitativas nominales registradas fueron: presencia de axilas, presencia de estrías, color de las estrías, color del tallo (verde, gris o púrpura), presencia de ramas, posición de las ramas, color de la panoja al final de la antesis, color de la panoja en cosecha, tipo de panoja, forma de panoja y presencia de pigmento en los cotiledones. Para los descriptores fenotípicos se registró el promedio del valor obtenido en las 10 planta, mientras que para los estadios de desarrollo se registró la fecha cuando al menos el 50 % de las plantas en las parcelas habían alcanzado dicho estadio

Los datos moleculares se obtuvieron mediante 22 loci microsatélites (SSR), los cuales originaron un total de 313 alternativas alélicas, con un rango de 2 a 20 alelos por locus, en cada uno de los alelos identificados se registró la frecuencia alélica relativa de una muestra de 5 a 6 hojas elegidas en forma aleatoria para cada accesión (Costa Tártara et al., [2012](#)). A modo ilustrativo, en la siguiente imagen se presentan algunos de los caracteres morfológicos evaluados, tales como color de panoja, ancho de hoja, color del tallo y diámetro del tallo.

Figura 5.1: Algunos caracteres morfológicos evaluados



En el cuadro 5.1, se muestran las variables, el código utilizado, y las unidades de medida o modalidades.

Tabla 5.1: Variables morfológicas y agronómicas utilizadas en el estudio.

Variable	Codigo	Tipo	Modalidad/unidades
diametro de tallo	DiamTallo	Cuantitativa	cm
numero de ramas	NroRamas	Cuantitativa	unidades
forma hoja superior	FormaHojaSu	Cuantitativa	relación longitud-ancho
numero de dientes	NroDientes	Cuantitativa	unidades
longitud peciolo	LogPeciolo	Cuantitativa	cm
longitud hoja	LogHoja	Cuantitativa	cm
ancho hoja	AnchoHoja	Cuantitativa	cm
longitud panoja	LogPanoja	Cuantitativa	mm
longitud glomerulo	LogGlomer	Cuantitativa	mm
longitud cotiledon	LogCotil	Cuantitativa	mm
longitud hipocotilo	LogHipocotil	Cuantitativa	mm
dias siembra-emergencia	DiasSE	Cuantitativa	días
dias siembra-boton floral	DiasSBF	Cuantitativa	días
dias siembra-floracion	DiasSF	Cuantitativa	días
dias siembra-madurez fisiologica	DiasSMF	Cuantitativa	días
altura de planta	AltPlanta	Cuantitativa	cm
diametro de grano	DiamGrano	Cuantitativa	mm
presencia de axilas	Axilas	cualitativa nominal	Si y No
presencia de estrias	Estrias	cualitativa nominal	Si y No
color de estrias	Cestrias	cualitativa nominal	0 sin estrias, 1 amarillo, 2 verde, 3 purpura
color tallo	CTallo	cualitativa nominal	2 verde, 3 gris, 5 púrpura
presencia de rama	Rama	cualitativa nominal	Si y No
posicion de ramas	PosRama	cualitativa nominal	0 sin rama, 1 oblicuas, 2 curvadas
color fin antesis	CFA	cualitativa nominal	1 blanco, 3 púrpura, 7 gris
color cosecha	Ccos	cualitativa nominal	1 blanco, 3 púrpura, 4 amarillo, 6 marrón, 7 gris
tipo de panoja	Tpanoja	cualitativa nominal	1 diferenciada, 2 no diferenciado
forma de panoja	Fpanoja	cualitativa nominal	1 glomerulada, 2, amarantiforme
pigmento cotiledones	PigCot	cualitativa nominal	Si y No

En el cuadro 5.2 se observa cada *primers* con la cantidad de alelos.

Tabla 5.2: Variables moleculares utilizadas en el estudio.

PRIMERS	Q074	Q088	Q001	Q097	Q076	Q087	Q112	Q050	Q027	Q062	Q106	KGA03	Q022	KGA20	QGA02	Q024	Q078	Q084	QGA03	QCA67	Q026	Q071	Total
Alternativas alelicas	11	20	18	17	18	18	10	13	19	17	12	8	24	13	9	17	11	8	15	6	12	17	313

5.2 Base de datos de Arroz

Se partió de los datos genotípicos crudos provenientes de la matriz **High-Density Rice Array (HDRA)** desarrollada por McCouch et al., 2016, que contiene aproximadamente 700.000 polimorfismos de nucleótido simple (SNPs) caracterizados en 1.568 accesiones de arroz, disponibles en el portal público *SNP-Seek* del *International Rice Informatics Consortium (IRIC)*.

De este conjunto, se trabajó específicamente con las 421 accesiones que conforman el **Rice Diversity Panel 1 (RDP1)**, una colección purificada y homocigota de *Oryza sativa* L. que captura la amplia variabilidad genética de la especie Eizenga et al., 2014. Este panel incluye variedades tradicionales (*landraces*) y cultivares élite, agrupados en cinco subpoblaciones: **indica** (95), **aus** (60), **tropical japónica** (106), **templada japónica** (111) y **aromática** (16); 33 accesiones adicionales se clasificaron como accesiones de ancestría mixta.

Para garantizar la calidad del conjunto de datos de asociación, se aplicaron filtros estándar sobre estas 421 accesiones: Frecuencia Alélica Menor ($MAF \geq 0,05$), un *call rate* por SNP $\geq 0,95$ —es decir, una proporción mínima del 95 % de genotipos no faltantes para cada marcador— y la exclusión de individuos con más del 10 % de datos genotípicos faltantes. El *call rate* por marcador controla la calidad de cada SNP, mientras que la exclusión por porcentaje de datos faltantes asegura la confiabilidad global de cada individuo; ambos filtros actúan de manera complementaria para depurar el conjunto genotípico. Tras este proceso, se obtuvo el conjunto final de 186.833 SNPs para 342 accesiones retenidas.

Para este estudio, la base de datos utilizada en los análisis se constituyó entonces en 342 accesiones, caracterizadas por 18 variables cuantitativas, 3 variables cualitativas dicotómicas y 186.833 SNPs. Las variables cuantitativas evaluadas fue-

ron: contenido de amilosa del grano (%), valor de dispersión alcalina, contenido de proteína en grano sin cáscara (%), días a espigazón (días), altura de planta (cm), longitud y ancho de la hoja bandera (cm), número de panículas por planta, longitud de la panícula (cm), número de ramas primarias por panícula, número total de flores por panícula, número de granos vacíos y llenos por panícula, longitud y ancho del grano con cáscara (mm), relación largo/ancho del grano (adimensional), volumen del grano con cáscara (mm^3) y peso de 100 granos (g). Las variables cualitativas dicotómicas (codificadas como 1 = Sí, 0 = No) fueron: desgrane espontáneo, presencia de aristas y pubescencia foliar.

A modo ilustrativo, en la siguiente imagen 5.2 se presentan algunos de los caracteres evaluados, tales como color de panícula, longitud de la panícula y número de ramas primarias por panícula.

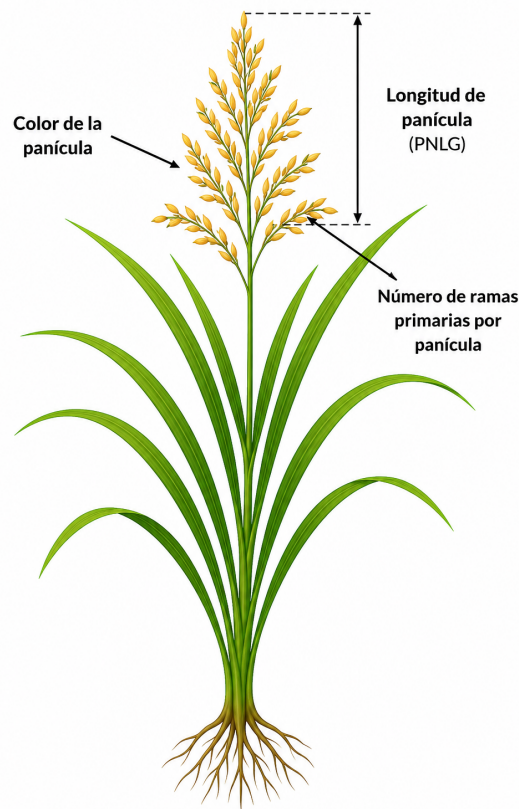


Figura 5.2: Algunos caracteres morfológicos evaluados

En el cuadro 5.3, se muestran las variables, el código utilizado, y las unidades de

medida o modalidades.

Tabla 5.3: Variables morfológicas y agronómicas utilizadas en el estudio.

Variable	Codigo	Tipo	Modalidad/unidades
Porcentaje de Amilosa Grano	%Amilosa	Cuantitativa	Porcentaje%
Dispersión alcalina	DisperALC	Cuantitativa	De 1 a 7
Porcentaje de Proteína	%Prot	Cuantitativa	Porcentaje%
Días a floración	DiasFI	Cuantitativa	días
Altura	Alt	Cuantitativa	cm
Logitud Bandera	LogBan	Cuantitativa	cm
Ancho de hoja	AnchHoj	Cuantitativa	cm
Número de panículas	Npan	Cuantitativa	número
Longitud de la panícula	LogPan	Cuantitativa	cm
Número de ramas	Nrama	Cuantitativa	número
Total de flores	TotFI	Cuantitativa	número
Número de granos vacíos	NgranVac	Cuantitativa	número
Número de granos llenos	Ngranlle	Cuantitativa	número
Longitud del grano con cáscara	LoggranCC	Cuantitativa	mm
Ancho del grano con cáscara	AnchGranCC	Cuantitativa	mm
Relacion largo/ancho grano	Rellaranch	Cuantitativa	adimensional
Volumen del grano con cáscara	VolGranCC	Cuantitativa	mm ³
Peso de 100 granos	Pes100gran	Cuantitativa	g
Desgrane espontáneo	DesgrE	cualitativa nominal	Si y No
Presencia de aristas	PresAris	cualitativa nominal	Si y No
Pubescencia foliar	PubAris	cualitativa nominal	Si y No

5.3 Entorno de trabajo y Herramientas Computacionales

Los análisis de esta investigación se realizaron en el entorno de programación R (v. 4.4.1; R Core Team, 2024) y en el entorno de desarrollo integrado RStudio (v. 2025.09.2+418). Algunas etapas exploratorias se desarrollaron bajo versiones anteriores, pero todos los análisis finales que sustentan los resultados reportados se ejecutaron con las versiones aquí indicadas. Esta plataforma ofrece un entorno que facilita desde el manejo y la depuración de datos, hasta la ejecución de análisis multivariados, el ajuste de modelos estadísticos, la generación de gráficos de alta calidad y la programación de nuevas alternativas metodológicas de análisis.

Se aplicaron diferentes paquetes especializados según la naturaleza de cada análisis. Para la exploración y análisis descriptivo de los datos se utilizaron `Hmisc`, `psych`,

`dplyr`, `tibble` y `tidyr`. La construcción y visualización de matrices de correlación al cuadrado se realizó con `Hmisc` (`rcorr()`), y su adecuación se evaluó mediante el test de esfericidad de Bartlett (`psych`, `cortest.bartlett()`). La representación gráfica de las matrices de correlación (y su versión cuadrada) se efectuó con `corrplot` (`corrplot()` y `corrplot.mixed()`).

Para los análisis multivariados, incluyendo ACP y AFM, se emplearon los paquetes `FactoMineR`, `factoextra`, `vegan` y `cluster`. El ajuste de modelos lineales y generalizados se realizó mediante `MASS`, `nnet`, `glmnet` y `pls`. La evaluación de supuestos y robustez de resultados se complementó con herramientas adicionales como `boot` (procedimientos de remuestreo), `MVN` (verificación de normalidad multivariada) y `grid` (personalización de gráficos).

La exportación de resultados y tablas se efectuó en formato Excel mediante `openxlsx` y `writexl`. Asimismo, para asegurar la reproducibilidad y trazabilidad de los flujos de análisis, se incorporaron librerías de apoyo tales como `ggplot2`, `stringr` y `ggrepel`, utilizadas en la elaboración de gráficos de dispersión, triplots y visualizaciones multivariadas.

Además de las librerías de uso general, se desarrollaron funciones y rutinas propias en R adaptadas a los requerimientos específicos de esta investigación. Entre ellas destacan:

- **Ajuste multivariante conjunto** (`multilogit_joint`), implementado mediante un algoritmo de gradiente proximal acelerado (FISTA con aceleración de Nesterov) y una penalización de tipo *Elastic Net*, encapsulado en la función `fista_multilogit_joint()`.
- **Predicción marginal y conjunta de probabilidades** a través de la función `predict_multilogit_joint()`, que estima probabilidades condicionales y conjuntas a partir de los coeficientes estimados y de la estructura de dependencias entre respuestas.
- **Validación cruzada multivariante**, implementada en

`cv_multilogit_joint()`, para la selección de hiperparámetros de regularización mediante la minimización de la pérdida promedio y la aplicación de la regla de un error estándar.

- **Procedimientos de remuestreo por bloques o “chunks”**, destinados a la evaluación de estabilidad de los coeficientes mediante esquemas de *bootstrap*, con ejecución paralela y control explícito de la reproducibilidad.
- **Rutinas y scripts de control** para el monitoreo de ejecuciones, la detección de iteraciones incompletas y la generación de salidas intermedias y finales, lo que permite retomar procesos interrumpidos sin pérdida de información.

El conjunto de estos desarrollos conformó un flujo computacional reproducible y documentado, estructurado en directorios jerárquicos y controlado mediante semillas de generación aleatoria (`set.seed()` y `RNGkind("L'Ecuyer-CMRG")`), lo que asegura la replicabilidad y trazabilidad de todos los resultados presentados en esta tesis.

El desarrollo de las funciones asociadas al ajuste multilogit conjunto (`fista_multilogit_joint()`, `cv_multilogit_joint()`, `predict_multilogit_joint()` y rutinas de control por “chunks”) se realizó de manera iterativa, combinando programación manual y asistencia automatizada. En particular, se emplearon herramientas de inteligencia artificial generativa, como ChatGPT (OpenAI), con el objetivo de optimizar la estructura de los algoritmos, depurar errores sintácticos y documentar los procedimientos de validación y remuestreo. Dicho apoyo se limitó estrictamente a la etapa de implementación computacional. Las decisiones metodológicas y la interpretación de los resultados quedan exclusivamente bajo la responsabilidad de la autora.

Capítulo 6

Estudio comparativo de alternativas metodológicas disponibles para evaluar asociación de caracteres fenotípicos con marcadores moleculares

En este estudio se comparan distintas metodologías estadísticas de mapeo asociativo aplicadas a accesiones de quínoa del Noroeste Argentino. Como etapa inicial, se llevó a cabo un análisis preliminar que incluyó el estudio del desequilibrio de ligamiento entre los marcadores moleculares y un análisis factorial múltiple para caracterizar de manera conjunta la información genotípica y fenotípica. Posteriormente se aplicó la metodología clásica y más difundida de mapeo asociativo para caracteres cuantitativos bajo dos enfoques: modelo *Naive* y modelo de mapeo P, ambos con corrección por multiplicidad de J. Li y Ji (2005). A continuación se compararon estos resultados con las alternativas metodológicas: Regresión *Elastic Net* multivariada, regresión por mínimos cuadrados parciales multivariada. En el caso de caracteres cualitativos, se ajustaron modelos de mapeo asociativo *Naive* logísticos para las variables dicotómicas y modelos de mapeo asociativo *Naive* logit multinomiales para las variables politómicas, también con la corrección por multiplicidad propuesta por J. Li y Ji (2005).

El conjunto de datos analizados en este capítulo, incluyó información fenotípica de cuatro variables cuantitativas (Altura de planta, Ancho de hoja, Días hasta la floración y Días hasta el botón floral visible), seis cualitativas nominales dicotómicas

(Presencia de axilas, Presencia de estrías, Presencia de ramas, Forma de la panoja, Tipo de panoja y Pigmentación de los cotiledones) y cinco cualitativas nominales politómicas (Color de las estrías, Color del tallo, Posición de las ramas, Color fin anthesis y Color de la cosecha), junto con información genotípica proveniente de los 313 marcadores moleculares tipo SSR.

6.1 Análisis preliminar

6.1.1 Desequilibrio de ligamiento

Se calculó la matriz de correlaciones al cuadrado (r^2) entre los marcadores moleculares utilizando la librería **stats** de R (R Core Team, 2025). Se observó que el 84 % de las correlaciones presentan valores de $r^2 \leq 0,05$, indicando una asociación muy débil entre la mayoría de los pares de marcadores. Por su parte, el 9,4 % de las correlaciones presentan valores comprendidos entre $0,05 < r^2 \leq 0,2$, el 5,38 % exhiben valores moderados ($0,2 < r^2 \leq 0,8$) y el 1,22 % (598 correlaciones sobre un total de 48.828 posibles correlaciones) presentan valores de $r^2 > 0,8$, indicando una asociación alta entre dichos pares de marcadores.

Por otro lado, la aplicación del test de esfericidad de Bartlett —implementado con la función `cortest.bartlett` del paquete **psych** (Revelle, 2025)— indicó el rechazo de la hipótesis nula (H_0): la matriz de correlaciones poblacional entre marcadores moleculares es la matriz identidad, lo cual evidencia que hay asociaciones no aleatorias entre alelos de distintos loci y se habilita el mapeo asociativo.

6.1.2 Análisis Multivariado: Factorial Múltiple (AFM)

Se integró la información de Quinoa mediante un análisis factorial múltiple realizado con la función `MFA` de la librería **FactoMineR** de R (Lê et al., 2008), que extrae un espacio común de baja dimensión a partir de las tablas de las variables cuantitativas, cualitativas y moleculares medidas sobre las mismas accesiones, equilibrando la contribución de cada bloque mediante la métrica de compromiso. Operacional-

mente, se aplicó PCA a las variables cuantitativas, se construyó la matriz disyuntiva completa y se realizó un ACM para las cualitativas, y AC para las variables moleculares. La información ponderada de los bloques se concatenó y representó en un *triplot* que proyecta simultáneamente individuos, vectores de variables cuantitativas y centroides de categorías cualitativas (Fig. 6.1).

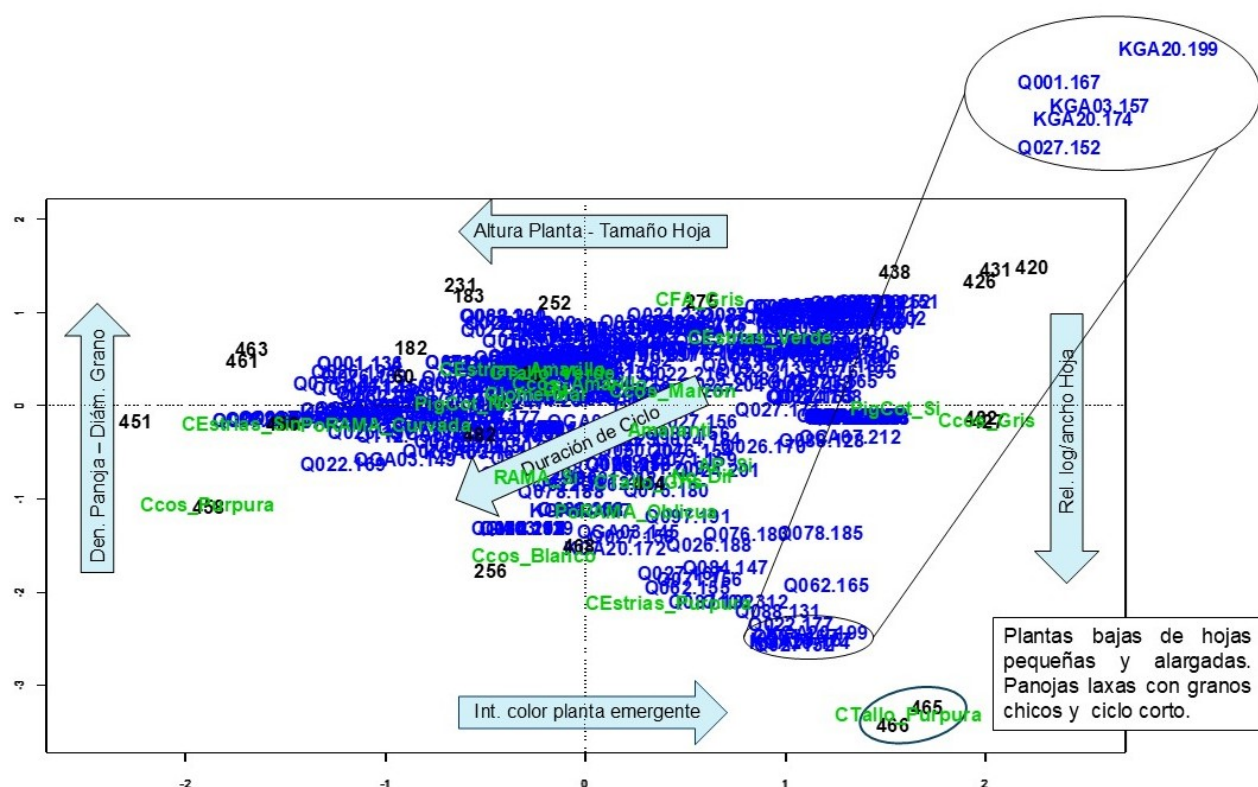


Figura 6.1: Mapa AFM. Los números en negro corresponden a las accesiones, mientras que en azul se representan las alternativas alélicas de los marcadores moleculares. Las flechas celestes indican los cinco principales gradientes, es decir las direcciones de mayor valor para las variables cuantitativas; las etiquetas verdes corresponden a las categorías de las variables cualitativas. La posición relativa de las accesiones respecto a las flechas permite interpretar gradientes y la cercanía a las variables cualitativas permite describir las características de esas accesiones.

El análisis de la figura 6.1 muestra como la posición de las alternativas alélicas (azul) se asocia con los gradientes morfológicos de las variables cuantitativas (flechas celestes). Esta asociación se estructura en cinco ejes de variación conjunta la altura de planta y tamaño de hoja; la duración del ciclo; la relación largo/ancho de la hoja; diámetro del grano y número de dientes de la panoja. y por último la intensidad del

color. La proximidad de grupos de accesiones a determinadas alternativas alélicas y a categorías cualitativas (verdes) permitió definir perfiles fenotípicos integrados. Por ejemplo, las accesiones 465 y 466 que se visualizan abajo a la derecha son plantas bajas de hojas pequeñas y más alargadas panojas laxas con granos chicos, ciclo corto y color del tallo púrpura. Estos fenotipos están cercanos a un grupo de marcadores: KGA20.199, Q001.167, KGA20.175 y Q027.152 Esta organización visual muestra una clara correspondencia entre la señal molecular (alternativas alélicas) y los gradientes fenotípicos complejos.

Este análisis preliminar permitió visualizar proximidad entre marcadores moleculares y categorías cualitativas, así como su relación con los vectores gradientes cuantitativos. La síntesis exploratoria puso de manifiesto conexiones consistentes entre los rasgos y destacó grupos de accesiones con perfiles moleculares y morfológicos concordantes. De este modo, proporcionó una base sólida y sugirió relaciones específicas que se formularon como hipótesis concretas para los posteriores modelos de mapeo asociativo.

No obstante, es pertinente señalar que, si bien en esta situación particular fue posible visualizar asociaciones entre variables fenotípicas y moleculares, en bases de datos de mayor envergadura —como es habitual en este tipo de estudios— esta alternativa de análisis se vuelve impracticable debido a la elevada densidad de puntos que presentaría el gráfico. Además, se trata de un enfoque estrictamente descriptivo, y no inferencial, a diferencia de las técnicas que se abordan a continuación.

6.2 Mapeo Variables Cuantitativas

6.2.1 Coeficiente de Escoufier

El coeficiente de escoufier R_v calculado entre la matriz de respuestas multivariadas Y y la matriz de marcadores moleculares X con la función `RV.rtest` de la librería `ade4` de R (Dray et al., 2007), fué de 0,3804 (p-valor = 0,001). Este valor indica una similitud moderada pero estadísticamente significativa entre las dos estructuras de

datos. Es decir, existe una asociación moderada entre la variabilidad conjunta de las variables fenotípicas cuantitativas y la de los marcadores moleculares.

6.2.2 Modelos univariados *Naive*

Se ajustaron modelos individuales de mapeo asociativo *Naive* con la función `lm` de la librería `stats` de R, previa estandarización de las variables respuestas, considerando un modelo por cada una de las variables cuantitativas combinadas con cada variable molecular, totalizando 1252 modelos. Se extrajo el p-valor de cada modelo. Estos p-valores se utilizaron para construir el gráfico de Manhattan. El número efectivo de pruebas M_{eff} fue de 24, reflejando la dimensionalidad real de la información disponible. Para $\alpha_{\text{global}} = 0,05$ resultó un α_p de 0,00208. En la figura 6.2 se muestra los marcadores significativos al aplicar el modelo *Naive* y los marcadores significativos luego de la corrección por multiplicidad de Li y Ji. El modelo *Naive* detectó en total 48 asociaciones entre marcadores y fenotipos estudiados, luego de corregir por J. Li y Ji (2005) quedaron 26. En el cuadro 6.1, asociado a la figura 6.2, se presentan las asociaciones significativas para cada variable luego de aplicar la corrección por multiplicidad, junto con la raíz del error cuadrático medio y el coeficiente de determinación de cada modelo.

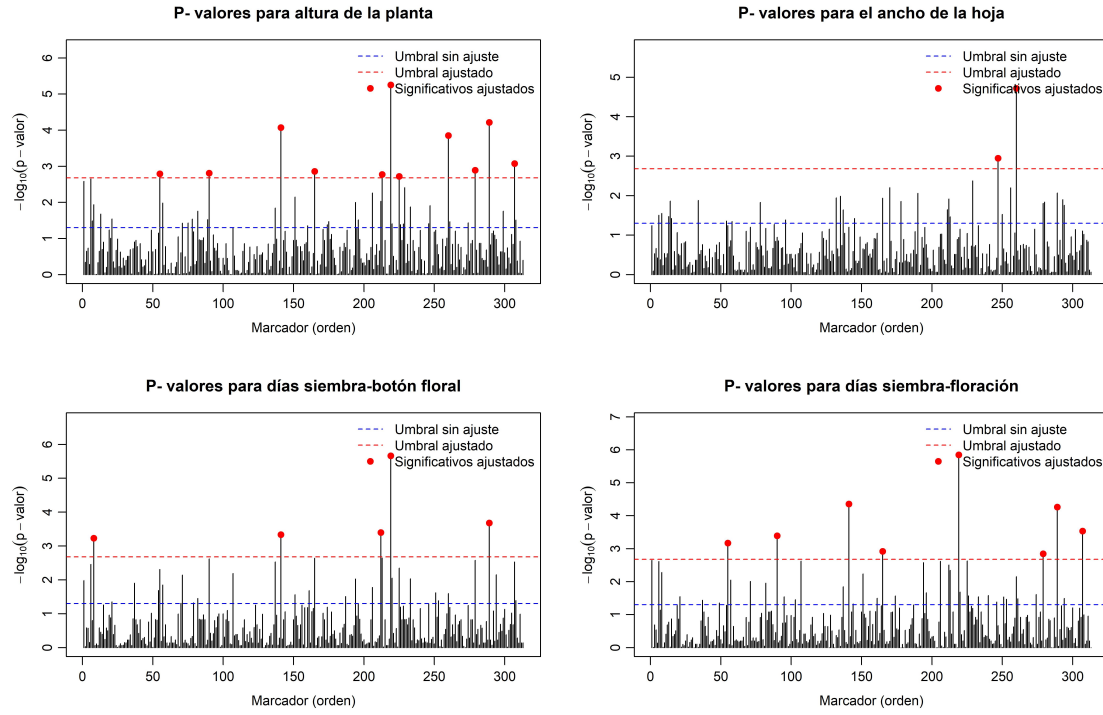


Figura 6.2: Modelos de mapeo asociativo *Naive*. Gráficos de Manhattan correspondientes a las pruebas de asociación entre cada marcador molecular y la variable fenotípica analizada. El eje horizontal corresponde al índice de los 313 marcadores (del 1 al 313), y el eje vertical al valor de $-\log_{10}(p\text{-valor})$. La línea azul indica el umbral de significación $\alpha_{\text{global}} = 0,05$, equivalente a $-\log_{10}(0,05) = 1,30$ en la escala representada en el eje vertical, mientras que la línea punteada superior corresponde al umbral ajustado mediante la corrección de Li y Ji, equivalente a $-\log_{10}(\alpha_p) = 2,68$, con $\alpha_p = 0,00208$. Los puntos rojos resaltan aquellos marcadores cuyos valores de $-\log_{10}(p\text{-valor})$ superan dicho umbral y, por tanto, resultan estadísticamente significativos tras la corrección por multiplicidad.

Variable	Marcador	$\sqrt{\text{CME}}$	R^2	$\sqrt{\text{ECM}}$
Altura de la planta	Q097_182	0,82	0,36	0,79
	Q087_190	0,82	0,36	0,78
	Q027_175	0,72	0,50	0,70
	Q106_300	0,81	0,37	0,78
	KGA20_180	0,82	0,35	0,79
	QGA02_186	0,65	0,60	0,62
	QGA02_204	0,82	0,35	0,79
	Q084_165	0,74	0,47	0,71
	QCA67_197	0,81	0,37	0,78
	Q026_176	0,72	0,51	0,69
	Q071_171	0,80	0,39	0,77
Ancho de hoja	Q078_185	0,81	0,38	0,77
	Q084_165	0,68	0,56	0,65
Días siembra-boton floral	Q074_196	0,79	0,41	0,75
	Q027_175	0,78	0,42	0,75
	KGA20_178	0,77	0,43	0,74
	QGA02_186	0,62	0,63	0,60
Días siembra-floracion	Q026_176	0,75	0,46	0,72
	Q097_182	0,79	0,40	0,76
	Q087_190	0,77	0,43	0,74
	Q027_175	0,71	0,52	0,68
	Q106_300	0,81	0,37	0,78
	QGA02_186	0,61	0,64	0,59
	QCA67_197	0,82	0,36	0,78
	Q026_176	0,71	0,51	0,68
	Q071_171	0,76	0,44	0,73

Tabla 6.1: Modelos *Naive*. Marcadores significativos con corrección de J. Li y Ji (2005), raíz del cuadrado medio del error residual ($\sqrt{\text{CME}}$), coeficiente de determinación (R^2) y raíz del error cuadrático medio ($\sqrt{\text{ECM}}$). Los marcadores que aparecen en más de un modelo fueron resaltados con un mismo color, facilitando así su identificación y comparación a lo largo de las diferentes variables fenotípicas.

Se observan 11 marcadores moleculares significativos para la altura de la planta, 2 marcadores significativos para el ancho de hoja (LEAFWI), 5 marcadores significativos para días desde la siembra hasta botón floral (SBF) y 8 marcadores significativos para días desde la siembra hasta la floración (SFL). Dado que varios marcadores moleculares se repiten en más de una variable cuantitativa (modelo), las 26 asociaciones significativas identificadas corresponden en realidad a 14 marcadores de los 313 evaluados. En los 26 modelos observados en el cuadro 6.1, la raíz de los CME varía entre 0,61 y 0,82, mientras que la raíz de los ECM osciló entre 0,59 y 0,79. Por su parte los coeficientes de determinación tomaron valores entre el 35 % y el 65 %.

6.2.3 Modelos univariados P

Se realizó un ACP a la matriz de marcadores moleculares con la función PCA de la librería FactoMineR de R, y se seleccionaron dos componentes estadísticamen-

te significativas según el criterio de *Tracy-Widom*, evaluado con la función `ptw` de la librería `RMTstat`. Se ajustaron modelos individuales de mapeo asociativo P, previa estandarización de las variables respuestas, considerando un modelo por cada una de las variables cuantitativas combinadas con cada variable molecular más dos componentes significativas, totalizando 1252 modelos. Se extrajo el p-valor de cada modelo. De forma análoga al modelo *Naive*, se construyó el gráfico de Manhattan. En la figura 6.3 se muestran los modelos en que los marcadores resultaron significativos al aplicar el modelo P, luego de aplicar la corrección por multiplicidad de Li y Ji. El modelo P detectó en total 124 asociaciones entre marcadores y fenotipos estudiados, luego de corregir por J. Li y Ji (2005) quedaron 24. De esos 24 modelos, en 23 resultaron significativas ambas componentes principales. En el cuadro 6.2, asociado a la figura 6.3, se presentan las asociaciones que resultaron significativas para cada variable luego de la corrección por multiplicidad, junto con la raíz del error cuadrático medio y el coeficiente de determinación ajustado de cada modelo.

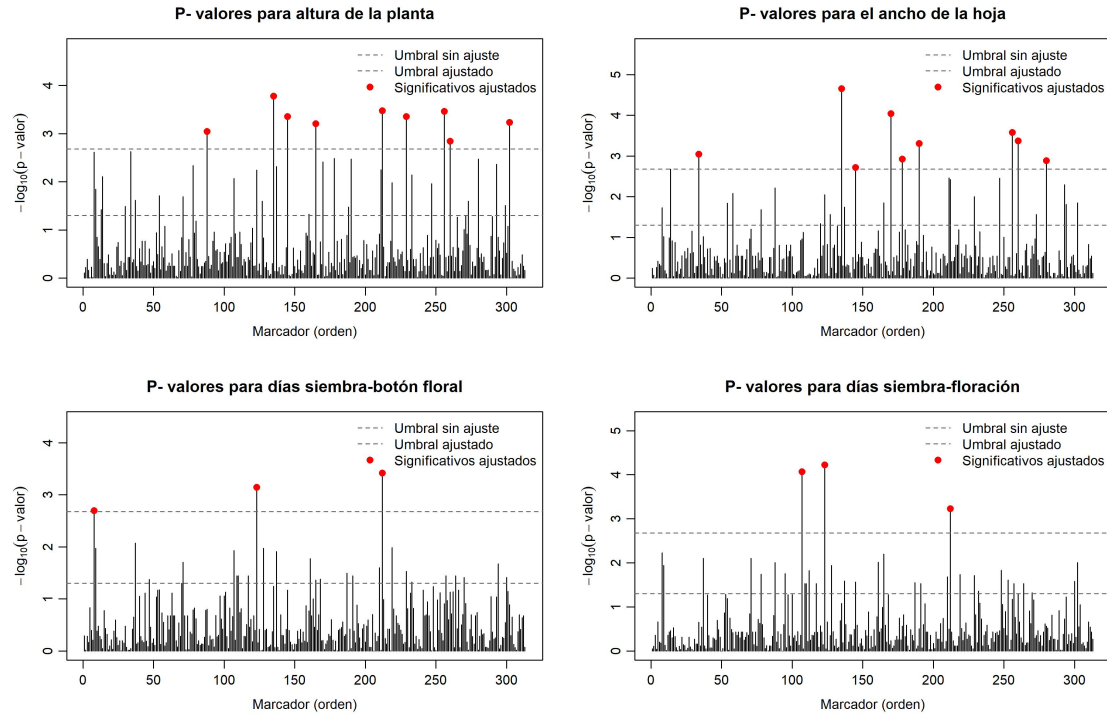


Figura 6.3: Modelos de MA P. Gráficos de Manhattan correspondientes a las pruebas de asociación entre cada marcador molecular y la variable fenotípica analizada. El eje horizontal corresponde al índice de los 313 marcadores (del 1 al 313), y el eje vertical al valor de $-\log_{10}(\text{p-valor})$. La línea azul indica el umbral de significación $\alpha_{\text{global}} = 0,05$, equivalente a $-\log_{10}(0,05) = 1,30$ en la escala representada en el eje vertical, mientras que la línea punteada superior corresponde al umbral ajustado mediante la corrección de Li y Ji, equivalente a $-\log_{10}(\alpha_p) = 2,68$, con $\alpha_p = 0,00208$. Los puntos rojos resaltan aquellos marcadores cuyos valores de $-\log_{10}(\text{p-valor})$ superan dicho umbral y, por tanto, resultan estadísticamente significativos tras la corrección por multiplicidad.

Variable	Marcador	$\sqrt{\text{CME}}$	R^2_{adj}	$\sqrt{\text{ECM}}$
Altura de la planta	Q087_182	0,52	0,73	0,47
	Q027_167	0,48	0,77	0,44
	Q062_155	0,50	0,75	0,46
	Q106_300	0,51	0,74	0,47
	KGA20_178	0,49	0,76	0,45
	Q024_201	0,50	0,75	0,46
	Q084_147	0,50	0,75	0,45
	Q084_165	0,53	0,72	0,48
Ancho de hoja	Q071_156	0,51	0,74	0,46
	Q001_167	0,71	0,50	0,65
	Q027_167	0,60	0,64	0,55
	Q062_155	0,73	0,47	0,67
	Q106_312	0,64	0,60	0,58
	KGA03_157	0,72	0,49	0,66
	Q022_177	0,69	0,53	0,63
	Q084_147	0,67	0,55	0,61
Días siembra-boton floral	Q074_196	0,60	0,64	0,55
	Q050_210	0,57	0,67	0,53
	KGA20_178	0,56	0,69	0,51
Días siembra-floracion	Q112_197	0,47	0,78	0,43
	Q050_210	0,47	0,78	0,43
	KGA20_178	0,52	0,73	0,47

Tabla 6.2: Modelos P. Marcadores significativos con corrección de J. Li y Ji (2005), raíz del cuadrado medio del error residual ($\sqrt{\text{CME}}$), coeficiente de determinación ajustado R^2_{adj} y raíz del error cuadrático medio ($\sqrt{\text{ECM}}$). Los marcadores que aparecen en más de un modelo fueron resaltados con un mismo color, facilitando así su identificación y comparación a lo largo de las diferentes variables fenotípicas.

Se observan 9 marcadores moleculares significativos para la altura de la planta, 8 marcadores significativos para el ancho de hoja (LEAFWI), 3 marcadores significativos para días desde la siembra hasta botón floral (SBF) y 3 marcadores significativos para días desde la siembra hasta la floración (SFL). Dado que varios marcadores moleculares se repiten en más de una variable cuantitativa (modelo), las 23 asociaciones significativas identificadas corresponden en realidad a 17 marcadores de los 313 evaluados. En los 23 modelos observados en el cuadro 6.2, la raíz de los CME varía entre 0,47 y 0,73, mientras que la raíz de ECM osciló entre 0,43 y 0,67. Por su parte los coeficientes de determinación ajustados tomaron valores entre el 47 % y el 78 %.

6.2.4 Modelo de Regresión Elastic Net Múltiple Multivariado

De la prueba de Mardia para las variables dependientes, realizada con la función `mvn` de la librería `MVN` de R, se concluyó que se cumple la normalidad multivariada. Luego se ajustó un modelo de regresión múltiple multivariado *Elastic Net*, con la función `cv.glmnet` de la librería `glmnet`, con k-fold ($k = 5$) en la validación cruzada.

Como se puede observar en la figura 6.4 el λ_{opt} fue de 0,2171 para un valor de δ de 0,9. En la figura 6.5 se muestran predictores significativos para cada variable de respuesta.

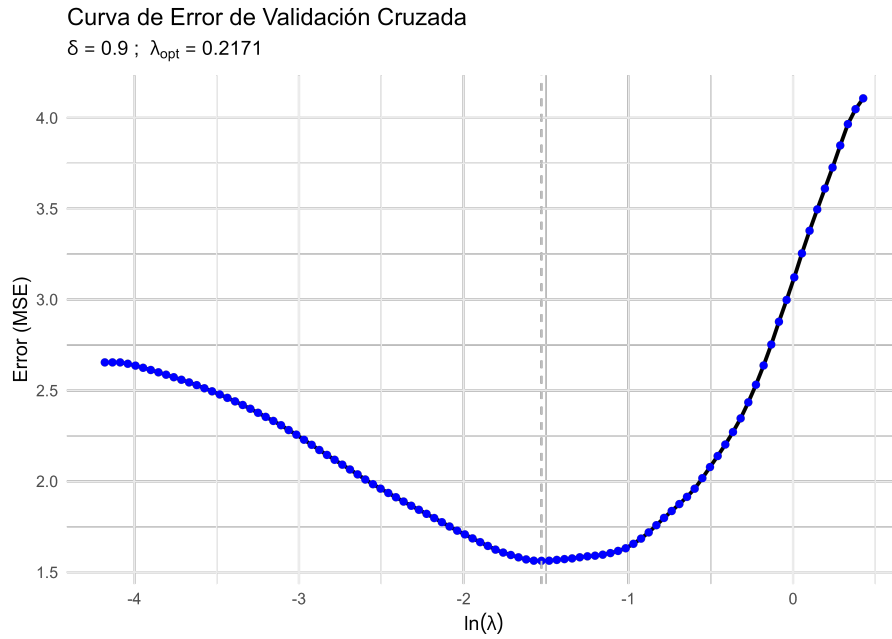


Figura 6.4: Curva de validación cruzada del modelo Elastic Net Multivariado correspondiente a $\delta = 0,9$. El eje de las ordenadas representa el error promedio de predicción, calculado como el promedio de los errores cuadráticos medios individuales de cada variable respuesta. El eje de las abscisas representa $\ln(\lambda)$, donde $\ln$ denota el logaritmo neperiano.

De un total de 1.252 coeficientes posibles del modelo de *Elastic Net* multivariado, el 90,73 % (1.136) fueron reducidos a cero por el efecto de la penalización. De los 116 restantes (distintos de cero), el análisis de remuestreo *bootstrap* (1.000 repeticiones) permitió identificar 18 coeficientes estadísticamente significativos.

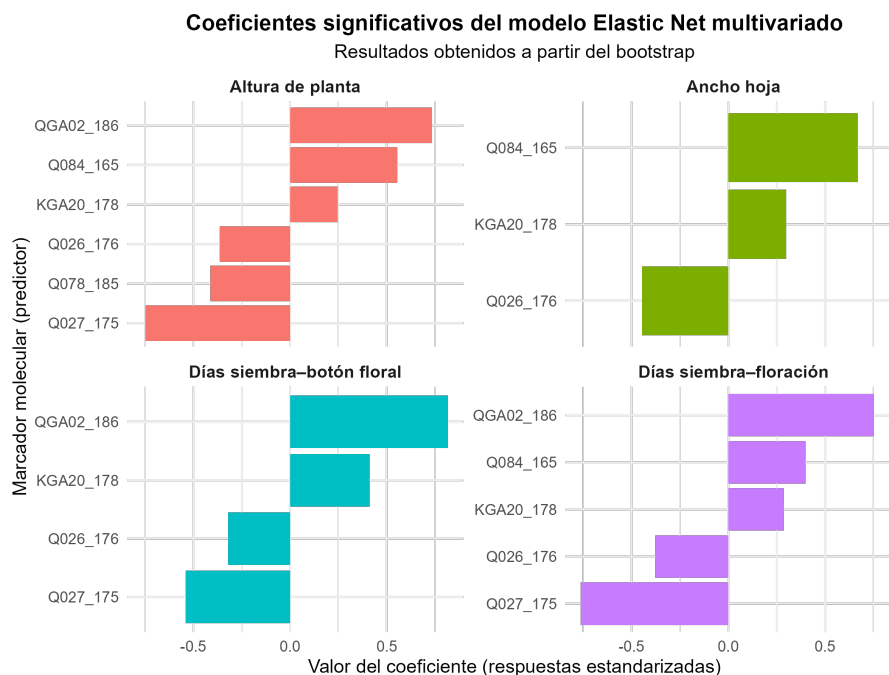


Figura 6.5: Coeficientes del modelo Elastic Net Múltiple Multivariado.

La figura 6.5 muestra estas 18 asociaciones, 6 para la altura de la planta, 3 para el ancho de la hoja, 4 para días siembra-botón floral y 5 para días de siembra de floración; sin embargo, dado que algunos marcadores son significativos en más de una variable, estas 18 asociaciones corresponden en realidad a 6 marcadores de los 313 evaluados. Los marcadores Q026-176 y KGA20-178 resultaron significativos en las cuatro variables analizadas. En este modelo se observan 12 asociaciones coincidentes con los modelos Naive y 4 coincidencias con el Modelo P. En el cuadro 6.3 se visualizan la raíz del ECM hallados en este modelo y los coeficientes de determinación de CV, de cada una de las variables.

Variable	$\sqrt{\text{ECM}}$	R^2
Altura de la planta	0,50	0,74
Ancho de hoja	0,71	0,48
Días siembra-boton floral	0,66	0,55
Días siembra-floracion	0,61	0,61

Tabla 6.3: Raíz del Error cuadrático medio ($\sqrt{\text{ECM}}$) y coeficiente de determinación (R^2) del modelo Elastic Net multivariado.

6.2.5 Modelo Multivariado de Regresión de Mínimos Cuadrados Parciales (PLS)

Con el método de *widekernelpls* de la librería `pls` de R se determinó que el número óptimo de componentes eran 6. El modelo explicó el 36,47 % de la varianza en los predictores moleculares y explicó el 99,15 % de la variabilidad en la altura de la planta, el 97,14 % en el ancho de hoja, el 98,88 % en los días de siembra de floración, y el 97,67 % en los días siembra-botón floral. Para facilitar la interpretación visual de los resultados se emplearon gráficos tipo *forest plot* contruidos con la función `ggplot` de la librería `ggplot2`. Estos gráficos muestran aquellos coeficientes que alcanzaron significancia estadística. Se observa en la figura 6.6 las estimaciones de los coeficientes junto con sus intervalos de confianza.

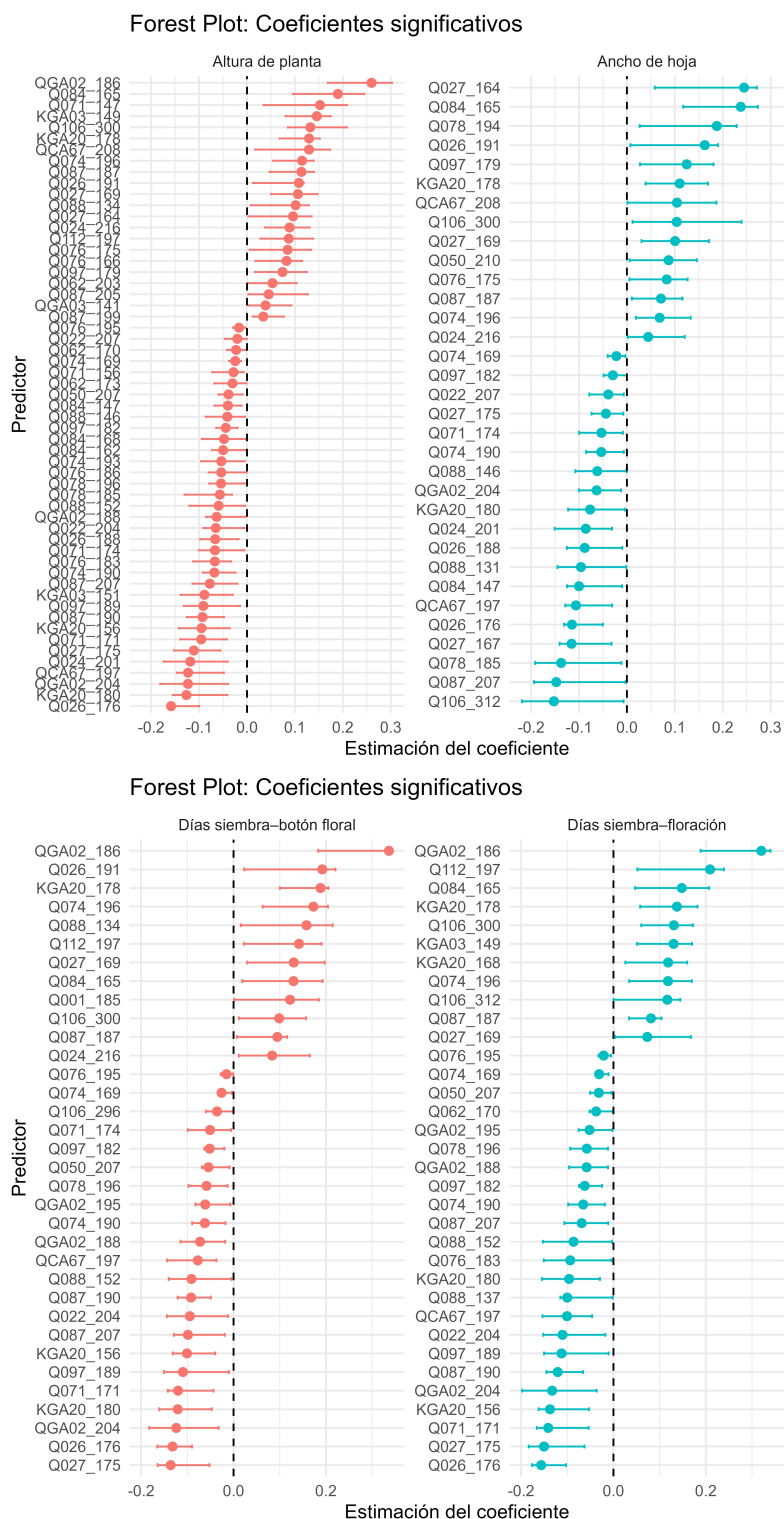


Figura 6.6: Intervalos de Confianza de los coeficientes del modelo PLS

En total se detectaron 158 asociaciones significativas entre marcadores moleculares y variables cuantitativas, 57 para altura de planta, 33 para ancho de hoja, 34

para días siembra–botón floral y 34 para días siembra–floración; sin embargo, dado que algunos marcadores son significativos en más de una variable, estas 158 asociaciones corresponden en realidad a 67 marcadores de los 313 evaluados. Se observa que todos los marcadores significativos en los modelos *Naive*, resultaron significativos en el modelo PLS multivariado de 6 componentes y 13 con el modelo P. Por otro lado en este modelo PLS se incluyen las 18 asociaciones del modelo de *Elastic Net* múltiple multivariado. En el cuadro 6.4 se muestra la raíz del ECM y los coeficientes de determinación de CV de cada variable.

Variable	$\sqrt{\text{ECM}}$	R^2
Altura de la planta	0,60	0,62
Ancho de hoja	0,73	0,45
Días siembra-boton floral	0,72	0,46
Días siembra-floracion	0,70	0,49

Tabla 6.4: Raíz de los errores cuadráticos medios, coeficientes de determinación del modelo PLS multivariado

6.2.6 Comparación de los modelos

La cantidad total de asociaciones detectadas por cada modelo, así como el número de coincidencias entre ellos, se resumen en la figura 6.7, un diagrama de Venn generado con la función `ggVennDiagram` de la librería `ggVennDiagram` de R (exportado con `ggplot2::ggsave`); adicionalmente, se incluyó una variante basada en áreas de Euler construida con `eulerr::euler` y `plot`. Allí se observa que el modelo PLS multivariado detectó la mayor cantidad de asociaciones (158) mientras que el *Naive* aportó 26, el Modelo P identificó 23 y el *Elastic Net* multivariado fue la alternativa más conservadora que detectó la menor cantidad (18), aunque todas sus asociaciones fueron confirmadas por al menos una de las otras propuestas. En cuanto a coincidencias, solo 2 asociaciones fueron comunes a las cuatro metodologías. Sin embargo, al considerar únicamente PLS, *Naive* y *Elastic Net*, se observan 12 coincidencias (2 coinciden con modelo P), lo que indica un mayor solapamiento entre estas tres metodologías. Si bien se observan coincidencias con el Modelo P, estas asociaciones

resultaron exclusivas o compartidas de manera puntual con cada una de las otras metodologías. Por otro lado, los cuadros 6.1, 6.2, 6.3, 6.4, muestran que el modelo *Elastic Net* múltiple multivariado tiene las raíces de los ECM más bajos y los coeficientes de determinación más altos entre los métodos comparados, lo que indica una buena capacidad predictiva para las variables fenotípicas analizadas.

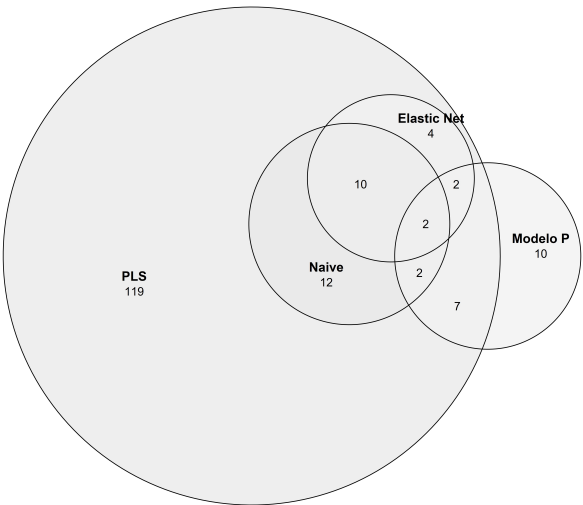


Figura 6.7: Cantidad de asociaciones significativas de cada metodología propuesta y número de coincidencias entre ellas

Para el análisis de solapamiento se consideraron únicamente los tres enfoques consistentes —PLS multivariado, modelos *Naive* con corrección de multiplicidad y *Elastic Net* multivariado—, excluyendo el modelo P debido a su baja replicabilidad y escasa coincidencia con los demás métodos. El grado de coincidencia entre los marcadores detectados por estas tres técnicas se sintetiza en el diagrama de Venn presentado en el cuadro 6.5.

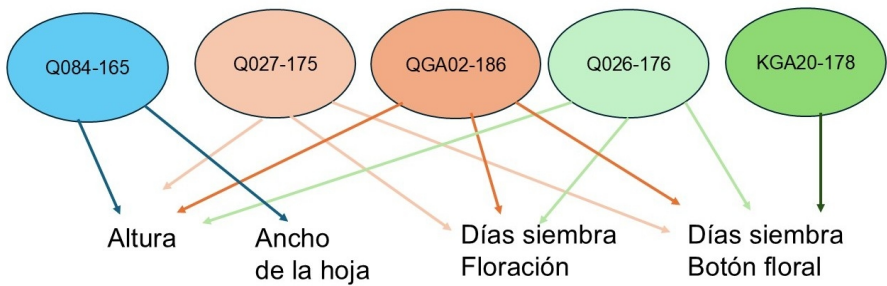


Tabla 6.5: Marcadores moleculares coincidentes en los tres modelos

6.3 Mapeo Variables Cualitativas

6.3.1 Modelos *Naive* Logísticos

Se ajustaron modelos logísticos individuales de mapeo asociativo *Naive* con la función `glm` de la librería `stats` de R, considerando un modelo por cada una de las variables dicotómicas combinadas con cada variable molecular, totalizando 1.878 modelos. Se extrajo el p-valor de cada modelo. Estos p-valores se utilizaron para construir el gráfico de Manhattan de los valores $-\log_{10}(p\text{-valores})$. Se utilizó el mismo nivel de significancia ajustado α_p , calculado a partir del número efectivo de pruebas (M_{eff}), de manera consistente con el criterio adoptado en el modelo *Naive* y en el modelo P de variables cuantitativas. La figura 6.8 muestra los marcadores significativos en el modelo logístico *Naive* después de aplicar la corrección por multiplicidad de Li y Ji. Se excluye el gráfico correspondiente a la variable tipo de panoja, ya que ningún marcador resultó significativo tras la corrección.

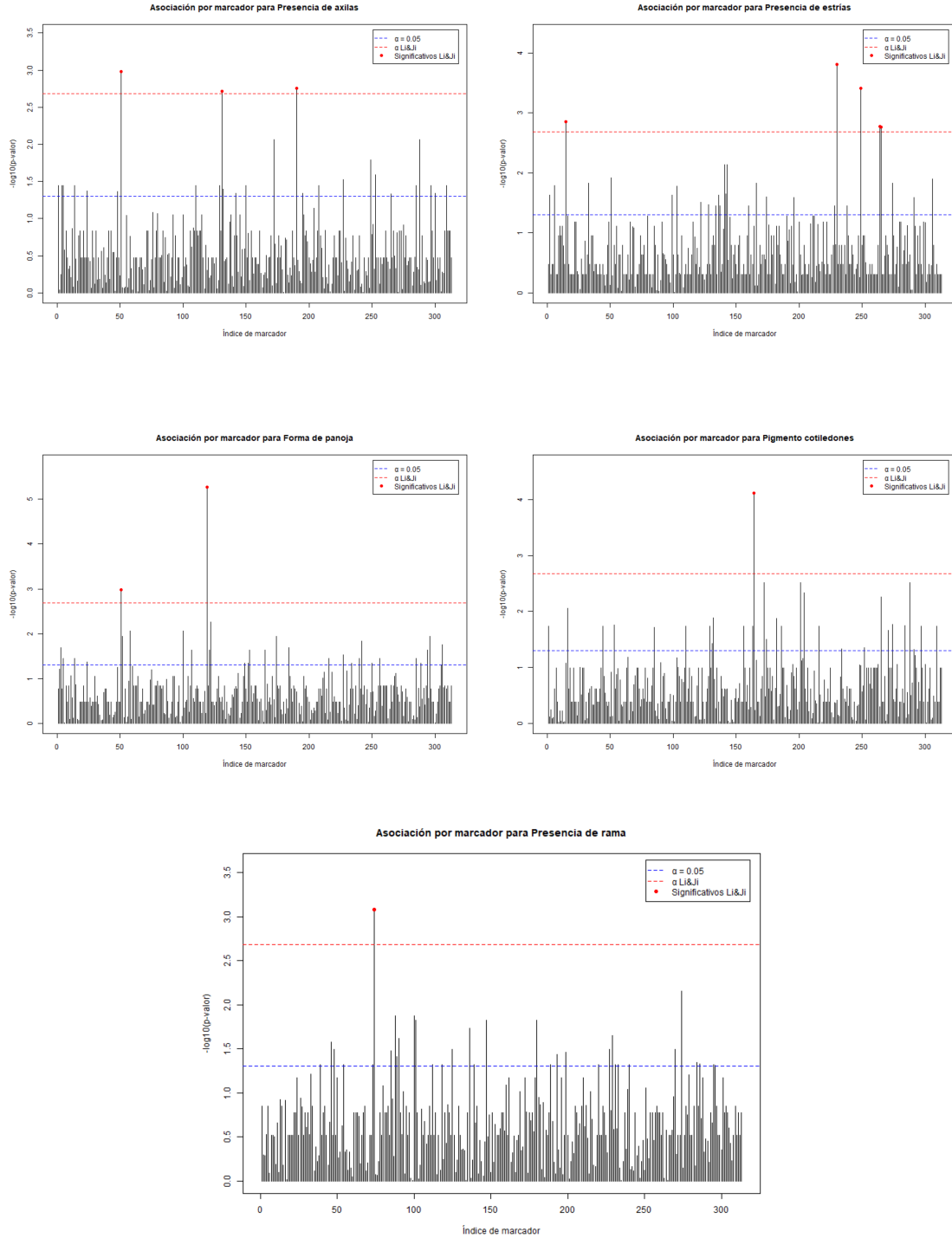


Figura 6.8: Modelos logísticos de mapeo asociativo *Naive*. Gráficos de Manhattan correspondientes a las pruebas de asociación entre cada marcador molecular y la variable fenotípica analizada. El eje horizontal corresponde al índice de los 313 marcadores (del 1 al 313), y el eje vertical al valor de $-\log_{10}(p)$. La línea azul indica el umbral de significación $\alpha_{\text{global}} = 0,05$, equivalente a $-\log_{10}(0,05) = 1,30$ en la escala representada en el eje vertical, mientras que la línea punteada superior corresponde al umbral ajustado mediante la corrección de Li y Ji, equivalente a $-\log_{10}(\alpha_p) = 2,68$, con $\alpha_p = 0,00208$. Los puntos rojos resaltan aquellos marcadores cuyos valores de $-\log_{10}(p)$ superan dicho umbral y, por tanto, resultan estadísticamente significativos tras la corrección por multiplicidad.

Variable	Marcador	R2_McFadden	AUC_ROC
Presencia de axilas	Q062_170	0,29	0,72
	Q022_201	0,33	0,78
	Q078_185	0,30	0,75
Presencia de estrías	Q088_140	0,57	0,88
	Q097_179	0,50	0,96
	Q106_297	0,39	0,85
	Q106_300	0,39	0,94
	KGA20_17	0,41	0,87
Presencia de ramas	Q024_232	0,33	0,75
Forma de la panoja	Q050_204	0,63	0,93
	Q022_201	0,33	0,78
Pigmento de los cotiledones	Q074_190	0,53	0,85

Tabla 6.6: Modelos Logísticos Naive. Marcadores significativos con corrección de J. Li y Ji (2005), coeficiente de determinación de McFadden (R^2_{McFadden}) y el área bajo la curva ROC (AUC-ROC). Los marcadores que aparecen en más de un modelo fueron resaltados con un mismo color, facilitando así su identificación y comparación a lo largo de las diferentes variables fenotípicas.

El modelo logístico *Naive* detectó en total 172 asociaciones entre marcadores y fenotipos estudiados, luego de corregir por J. Li y Ji (2005) quedaron 12. En el cuadro 6.6, relacionado con la figura 6.8, se presentan las asociaciones significativas con cada variable fenotípica luego de aplicar la corrección de Li y Ji. Además, se informa el coeficiente de determinación de McFadden (R^2_{McFadden}) y el área bajo la curva (AUC) correspondiente a cada modelo, lo que permite evaluar tanto la bondad de ajuste como la capacidad discriminante de los marcadores significativos. Se observaron 3 marcadores significativos asociados con la presencia de axilas, 5 asociados con la presencia de estrías, 1 asociado con la presencia de ramas, 2 asociados con la forma de la panoja y 1 asociado con el pigmento de los cotiledones. Dado que el marcador Q022_201 se encuentra asociado tanto a la presencia de axilas como a la forma de la panoja, resultaron significativos 11 marcadores de los 313 analizados. Se observaron valores de R^2_{McFadden} entre 0,29 y 0,63, lo cual indica un poder explicativo moderado a alto de los modelos ajustados (Hosmer et al., 2013). Asimismo, los valores del área bajo la curva ROC (AUC-ROC) oscilaron entre 0,72 y 0,96, lo que refleja una capacidad de discriminación adecuada a excelente de los marcadores evaluados (Agresti, 2013).

6.3.2 Modelos Naive Logit Multinomiales

Para cada rasgo categórico multiclase: color de estrías, color del tallo, posición de ramas, color al fin de antesis y color de cosecha, se ajustaron modelos logit multinomiales univariados de MA *Naive*, con la función `multinom` del paquete `nnet` de R (Ripley & Venables, 2023; Venables & Ripley, 2002) con la categoría modal de cada rasgo como referencia, considerando un modelo por cada una de las variables categóricas combinadas con cada variable molecular, totalizando 1.565 modelos. Se extrajo el p-valor de cada modelo y se utilizó el mismo nivel de significancia ajustado α_p , calculado a partir del número efectivo de pruebas (M_{eff}), de manera análoga a la empleada en los modelos anteriores. La figura 6.9 muestra los marcadores significativos en el Modelo logit multinomial *Naive* y después de aplicar la corrección por multiplicidad de Li y Ji. Se excluyen los gráficos correspondientes a las variables color al fin de antesis y color de cosecha, ya que ningún marcador resultó significativo tras la corrección.

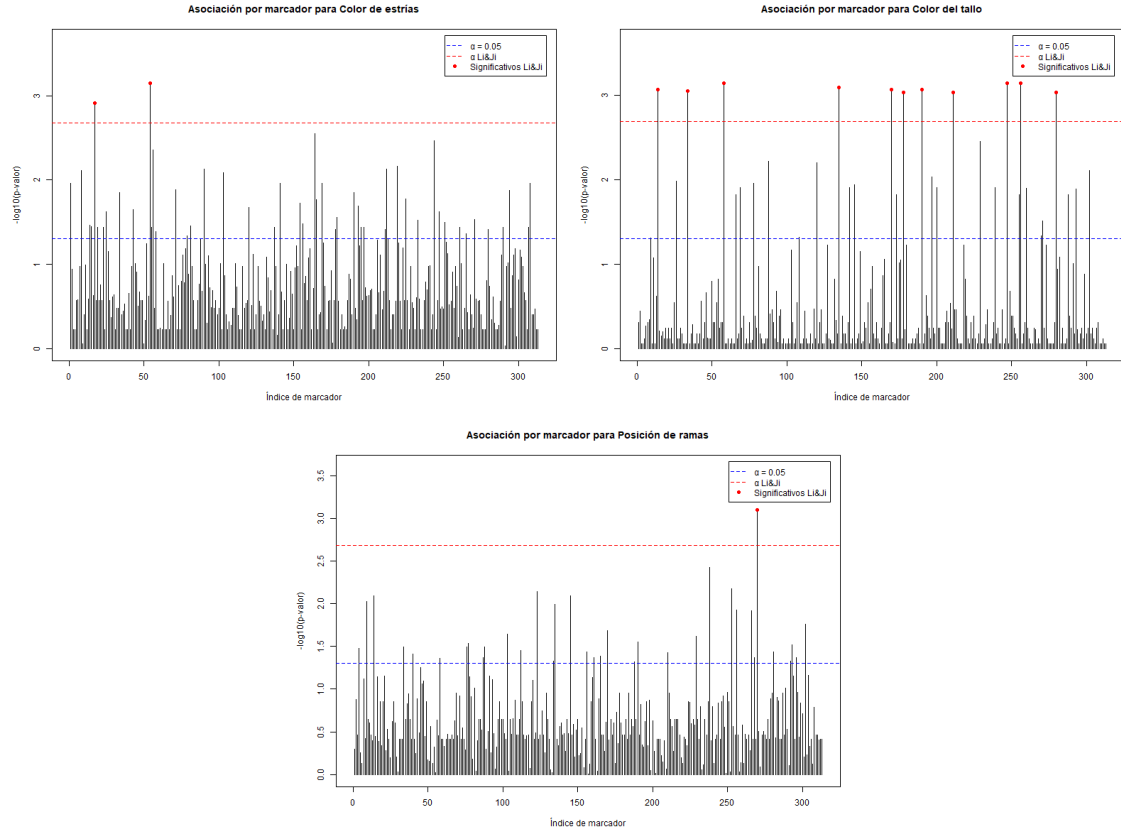


Figura 6.9: Modelos logit multinomiales de MA *Naive*. Gráficos de Manhattan correspondientes a las pruebas de asociación entre cada marcador molecular y la variable fenotípica analizada. El eje horizontal corresponde al índice de los 313 marcadores (del 1 al 313), y el eje vertical al valor de $-\log_{10}(p)$. La línea azul indica el umbral de significación $\alpha_{\text{global}} = 0,05$, equivalente a $-\log_{10}(0,05) = 1,30$ en la escala representada en el eje vertical, mientras que la línea punteada superior corresponde al umbral ajustado mediante la corrección de Li y Ji, equivalente a $-\log_{10}(\alpha_p) = 2,68$, con $\alpha_p = 0,00208$. Los puntos rojos resaltan aquellos marcadores cuyos valores de $-\log_{10}(p)$ superan dicho umbral y, por tanto, resultan estadísticamente significativos tras la corrección por multiplicidad.

Variable	Marcador	R ² _McFadden
Color de Estrías	Q097_179	0,26
	Q088_140	0,24
Color del Tallo	Q097_191	0,65
	Q084_147	0,65
	Q078_185	0,65
	Q027_167	0,64
	Q088_131	0,64
	Q022_177	0,64
	Q106_312	0,64
	Q001_167	0,63
	QCA67_202	0,63
	KGA03_157	0,63
	KGA20_174	0,63
Posición de Ramas	QGA03_153	0,26

Tabla 6.7: Modelos logit multinomiales *Naive*. Marcadores significativos con corrección de J. Li y Ji (2005), coeficiente de determinación de McFadden (R^2_{McFadden}).

El modelo logit multinomial *Naive* detectó en total 164 asociaciones entre marcadores y fenotipos estudiados, luego de corregir por J. Li y Ji (2005) quedaron 14. En el cuadro 6.7, relacionado con la figura 6.9, se presentan las asociaciones significativas con cada variable fenotípica luego de aplicar la corrección de Li y Ji. Además, se informa el coeficiente de determinación de McFadden (R^2_{McFadden}) correspondiente a cada modelo, lo que permite evaluar la bondad de ajuste. Se observan 2 marcadores significativos asociados con el color de las estrías, 11 marcadores asociados con el color del tallo y 1 marcador asociado con la posición de las ramas. Dado que ningún marcador se repite entre las variables, resultaron significativos 14 marcadores de los 313 analizados. Se observaron valores de R^2_{McFadden} entre 0,24 y 0,65, lo cual indica un poder explicativo moderado a alto de los modelos ajustados (Hosmer et al., 2013).

6.4 Consideraciones finales del estudio comparativo

En el análisis preliminar se exploró la coherencia entre la estructura multivariada global del AFM y las asociaciones marcador-fenotipo identificadas mediante modelos de mapeo asociativo. Para ello, se examinó la configuración conjunta de los 313 marcadores moleculares, las modalidades de las variables fenotípicas cualitativas (representadas con etiquetas verdes) y los gradientes correspondientes a las varia-

bles cuantitativas en el plano definido por las dos primeras dimensiones factoriales (Figura 6.1).

La comparación de este subconjunto con los marcadores seleccionados por los modelos *Naive* logístico y *Naive* logit multinomial mostró una coherencia espacial total: todos los marcadores reportados como significativos por estos modelos se ubicaron entre los 30 más próximos a su fenotipo correspondiente en el AFM. Un caso destacable es el del marcador Q024-232, que — siendo el único asociado significativamente a la presencia de ramas en los modelos logísticos *Naive* — resultó ser también el más cercano a la categoría «no presencia de rama» en el espacio factorial, reforzando y visualizando geométricamente dicha asociación estadística. En este análisis no se aplicaron modelos P para las variables cualitativas, dado que este tipo de enfoques - basados en la incorporación de componentes principales derivados de los marcadores moleculares como predictoras para corregir estructura poblacional— fueron originalmente desarrollados en el contexto de rasgos cuantitativos (Pritchard et al., 2000), y no cuentan con una formulación equivalente para respuestas categóricas múltiples.

En este marco, este alto grado de concordancia entre el modelo multivariado (AFM) y los modelos de asociación univariados valida la solidez de los marcadores identificados para rasgos cualitativos. No obstante, una pregunta central de este estudio era comprender cómo se relacionan entre sí los hallazgos provenientes de las distintas familias de modelos aplicadas a todo el espectro de variables respuesta. El análisis comparativo de los resultados para variables cuantitativas, binarias y multinomiales revela que la coincidencia de marcadores significativos entre tipos de respuesta es limitada, aunque no inexistente. Mientras que en los modelos cuantitativos se observa un conjunto relativamente consistente y recurrente de marcadores entre enfoques (como QGA02-186, Q084-165, KGA20-178, Q026-176 y Q027-175), en los modelos cualitativos —tanto binarios como multinomiales— la mayoría de los marcadores identificados son específicos de cada rasgo.

Sin embargo, se identificaron solapamientos puntuales pero relevantes, en particular los marcadores Q078-185 y Q106-300, significativos tanto para rasgos cuan-

titativos como cualitativos. Esta coincidencia sugiere la posible existencia de efectos pleiotrópicos o de regiones genómicas en desequilibrio de ligamiento que influyen en más de un tipo de carácter. En conjunto, estos hallazgos indican que cada clase de modelo captura dimensiones parcialmente distintas de la relación marcador–fenotipo. La evidencia convergente entre enfoques, aunque poco frecuente, resulta especialmente valiosa para priorizar marcadores candidatos en etapas posteriores de validación.

En síntesis, este estudio comparativo permitió evaluar un conjunto de enfoques para detectar asociaciones entre marcadores moleculares y caracteres fenotípicos, poniendo de manifiesto diferencias conceptuales y operativas según la naturaleza de la variable respuesta. Para rasgos cuantitativos, los enfoques clásicos de mapeo asociativo han mostrado un desempeño ampliamente documentado en la identificación de variantes genéticas asociadas a fenotipos complejos (Yang & Wang, 2012), y la literatura ha desarrollado extensiones multivariadas que explotan las correlaciones entre múltiples rasgos para incrementar el poder estadístico y mitigar el impacto de la corrección por pruebas múltiples (Turley et al., 2018; Yang & Wang, 2012). No obstante, en contextos de alta dimensionalidad $p \gg n$, métodos tradicionales como la regresión multivariada múltiple o el RDA se ven limitados por la singularidad de las matrices de covarianza, lo que ha motivado el uso de versiones penalizadas (por ejemplo, *Elastic Net*) y de enfoques basados en componentes latentes como alternativas más estables frente a la colinealidad estructural de los predictores.

En contraste, el panorama es marcadamente menos desarrollado para caracteres cualitativos. Aunque se han propuesto modelos específicos para fenotipos binarios u ordinales —incluyendo formulaciones logísticas mixtas orientadas al control del error tipo I en grandes bases de datos (Bi et al., 2021)— aún no existe un marco multivariado ampliamente aceptado que permita analizar, de manera conjunta, múltiples rasgos cualitativos en estudios de asociación genética. Si bien revisiones como la de Yang y Wang (2012) reconocen la posibilidad de abordar respuestas mixtas (continuas y categóricas), también enfatizan la complejidad y escasa estandarización de estas propuestas cuando los componentes de respuesta no son todos continuos. Esta

carencia metodológica conduce, en la práctica, a abordajes fragmentados y univariados, que dificultan la evaluación de efectos pleiotrópicos y restringen la capacidad de explotar la estructura fenotípica compartida entre rasgos.

En consecuencia, los resultados de este capítulo no sólo evidencian las limitaciones de los enfoques disponibles para ambos tipos de variables, sino que revelan una doble necesidad metodológica: por un lado, reforzar el tratamiento de rasgos cuantitativos en escenarios de alta dimensionalidad mediante estrategias penalizadas y estables; y por otro, avanzar hacia el desarrollo de un marco multivariado para variables cualitativas múltiples, actualmente ausente en la práctica del mapeo asociativo. Esta brecha constituye la motivación central del capítulo siguiente, donde se propone un marco unificado basado en modelos penalizados multivariados, orientado a ampliar el alcance analítico del mapeo asociativo y a ofrecer herramientas más integradoras y consistentes con la complejidad biológica de los sistemas estudiados.

Capítulo 7

Propuesta Metodológica

7.1 Introducción

El análisis de asociación genética se enfrenta al desafío de desarrollar metodologías capaces de capturar la complejidad inherente a variables fenotípicas y agronómicas de interés, caracterizadas por la acción conjunta de numerosos loci de pequeño efecto sobre ellas. Si bien los enfoques univariados han sido ampliamente difundidos y utilizados por su simplicidad e interpretabilidad, presentan una limitación importante: analizan cada rasgo de manera aislada. Esta descomposición puede dificultar la detección de efectos pleiotrópicos, reducir la sensibilidad a señales débiles distribuidas en múltiples fenotipos y, en consecuencia, afectar la capacidad de descubrimiento. Para superar estas restricciones, los modelos multivariados emergen como una alternativa sólida, ya que integran la información de múltiples rasgos de forma simultánea. Dentro de este conjunto de técnicas, los métodos penalizados, como *Elastic Net*, se distinguen por su capacidad para manejar escenarios de alta dimensionalidad, llevando a cabo de manera integrada la selección de variables y la estimación de coeficientes.

Aunque el enfoque PLS multivariado también permite integrar múltiples rasgos, su mecanismo se basa en una reducción de dimensionalidad mediante componentes latentes que, si bien es útil, carece de un criterio de penalización explícito. Esta característica limita el control directo sobre la complejidad del modelo y la selección de marcadores individuales. Como consecuencia, PLS tiende a generar un espectro más amplio de asociaciones potenciales, que lo sitúa como una herramienta más exploratoria que confirmatoria, y es propenso al sobreajuste en contextos de alta

colinealidad entre predictores.

En contraste, *Elastic Net* al combinar las penalizaciones ℓ_1 (LASSO) y ℓ_2 (ridge), permite seleccionar un subconjunto reducido y estable de marcadores al tiempo que controla la multicolinealidad. Esta dualidad lo convierte en un candidato ideal para construir un marco metodológico a la vez robusto y reproducible.

En este contexto, el presente capítulo tiene como objetivo principal proponer un marco metodológico unificado para el mapeo asociativo basado en modelos multivariados *Elastic Net*. Dada la prevalencia de caracteres cuantitativos y binarios en los estudios genéticos iniciales, el análisis se centrará de manera exhaustiva en estos tipos de respuesta. La extensión a escenarios multinomiales multivariados se posterga para desarrollos futuros. Esta delimitación no solo permite concentrar los esfuerzos en perfeccionar y validar las propuestas para los casos de uso más frecuentes, sino que también establece una base sólida sobre la cual expandir el marco metodológico en el futuro.

7.2 Limitaciones y alcances de las metodologías aplicadas en el estudio comparativo

En el capítulo anterior se realizó un estudio comparativo de diversas alternativas metodológicas disponibles para evaluar la asociación entre caracteres fenotípicos y marcadores moleculares. Dichos enfoques comprenden:

- Análisis factorial múltiple
- Modelos de mapeo asociativo clásico *Naive* y modelo P con ajuste de correcciones para comparaciones múltiples.
- Modelos lineales generalizados (MLG). Regresión logística y regresión logit multinomial, con ajuste de correcciones para comparaciones múltiples.
- Modelos penalizados multivariados para caracteres fenotípicos cuantitativos.

7.2.1 Limitaciones en el tratamiento de variables cuantitativas

Si bien cada estrategia aporta información valiosa, se identificaron limitaciones tales como: pérdida de potencia en los métodos univariados, dificultades computacionales y de interpretación; e incapacidad de los métodos tradicionales para captar la naturaleza multivariada y mixta de los caracteres fenotípicos en contextos de alta dimensionalidad ($p \gg n$). En particular, dos técnicas que naturalmente fueron desarrolladas para estudiar relaciones entre grupos de variables como los modelos de Regresión Múltiple multivariada y el Análisis de Redundancia (RDA) al requerir la inversión de la matriz de covarianzas de los predictores, hace que solo sea posible cuando el número de observaciones n supera al número de variables p . En contextos de alta dimensionalidad, como en casi la totalidad de caracterizaciones moleculares, esta condición no se cumple: la matriz de predictores $\mathbf{X}'\mathbf{X}$ resulta singular, lo que conduce a estimadores inestables e impide la aplicación directa de dichas técnicas. No obstante, esta restricción fue abordada mediante el empleo de modelos penalizados multivariados (*Elastic Net*, PLS), que permiten obtener estimadores estables aun bajo la condición $p \gg n$. Por lo tanto, si bien las metodologías tradicionales presentan serias restricciones en este contexto, las variantes penalizadas constituyen una alternativa viable y adecuada para este tipo de estudios.

7.2.2 Limitaciones en el tratamiento de variables cualitativas binarias y politómicas

En el caso de caracteres fenotípicos de naturaleza cualitativa, ya sean dicotómicos o politómicos (multiclase), existen restricciones metodológicas que impiden replicar de manera directa las estrategias desarrolladas para variables continuas. El método PLS clásico admite una matriz de respuestas Y siempre que éstas sean continuas; sin embargo, no dispone de una extensión estándar que permita modelar de forma multivariada múltiples respuestas categóricas bajo un enlace logit.

De modo similar, el enfoque penalizado mediante *Elastic Net*, implementado en `glmnet` con `family = "mgaussian"`, está limitado a respuestas continuas. En

consecuencia, el tratamiento conjunto de variables binarias o politómicas carece de un marco multivariado consolidado.

En un plano más general, dentro de los modelos logísticos multivariados se han propuesto alternativas teóricas, como los modelos de Ising con covariables, que permiten abordar respuestas binarias correlacionadas (Cheng et al., 2014). No obstante, estas propuestas aún no cuentan con implementaciones en el marco de PLS ni de *Elastic Net* multivariado, y su aplicación práctica sigue siendo incipiente.

En síntesis, mientras que las estrategias multivariadas aplicadas a variables continuas (PLS multivariado y *Elastic Net* multivariado) han alcanzado un grado de madurez metodológica y computacional, las respuestas cualitativas binarias y politómicas continúan sin un equivalente directo. Actualmente, las soluciones disponibles se restringen a modelaciones univariadas o multitarea, lo que limita la explotación conjunta de la correlación entre caracteres. El desarrollo de enfoques robustos para este tipo de respuestas constituye aún un área abierta de investigación metodológica.

7.3 Fundamento de la propuesta

La propuesta metodológica se orienta a superar las limitaciones señaladas en la sección anterior mediante un abordaje multivariado y penalizado, capaz de ofrecer una alternativa unificada para variables tanto cuantitativas como cualitativas. En este sentido, los ejes conceptuales que sustentan la propuesta son los siguientes:

- Modelo de regresión multivariado múltiple. Constituye el punto de partida para capturar la estructura de covariación entre múltiples respuestas cuantitativas y un conjunto de predictores moleculares. Ante la condición $p \gg n$, la estimación ordinaria se torna inestable o directamente inviable, lo que motiva el uso de regularización.
- Modelos logísticos multivariados (binario). Se plantea su inclusión conceptual para el tratamiento de caracteres cualitativos, dicotómicos. El logit binario permite abordar rasgos dicotómicos. Si bien las implementaciones multivaria-

das penalizadas aún son incipientes en este ámbito, estos modelos ofrecen el andamiaje teórico necesario para integrar adecuadamente los fenotipos cualitativos en un esquema multivariado.

- Modelos multivariados (*Elastic Net*). Se adopta esta estrategia como mecanismo eficiente para la selección de predictores relevantes y la estabilización de los coeficientes, al combinar la penalización ℓ_1 (LASSO) con la ℓ_2 (ridge). De esta manera, el *Elastic Net* permite afrontar la alta dimensionalidad, la multicolinealidad entre predictores y la necesidad de identificar subconjuntos estables de marcadores asociados a los caracteres de interés.
- Análisis de redundancia (RDA). Se utiliza como marco de referencia para proyectar el espacio de las respuestas sobre el subespacio generado por los predictores, sintetizando la relación entre fenotipos y marcadores. En el caso clásico, su aplicación se ve restringida por la alta dimensionalidad y la colinealidad entre predictores; aquí se lo retoma únicamente como base conceptual para definir un análogo penalizado.

A continuación se desarrollan las propuestas metodológicas específicas para el análisis de variables cuantitativas y para variables cualitativas binarias.

7.4 Propuesta para variables cuantitativas: Análisis de Redundancia Análogo (*RDA-like*) basado en *Elastic Net* multivariado

En el capítulo anterior se aplicó un modelo de Regresión multivariada penalizado mediante *Elastic Net*, el cual permitió identificar asociaciones entre caracteres cuantitativos y predictores moleculares en un marco de alta dimensionalidad. Pese a su utilidad para el análisis comparativo de técnicas, este modelo dejó abierta la posibilidad de explorar, a partir de las respuestas estimadas, una estructura multivariada subyacente. En este capítulo se aborda dicha estructura formulando una extensión

del RDA clásico, denominada *RDA-Like* multivariado para variables cuantitativas. Este enfoque consiste en someter las predicciones fuera de muestra (*out-of-fold*, OOF) $\hat{Y}_i^{\text{OOF}}$, obtenidas mediante validación cruzada del modelo *Elastic Net*, a un ACP. A partir de este ACP se construye un triplot que integra simultáneamente observaciones, caracteres fenotípicos y marcadores moleculares significativos, ofreciendo así una representación análoga al RDA clásico, pero adaptada al contexto de modelos penalizados multivariados. El esquema de la propuesta se visualiza en la figura 7.1.

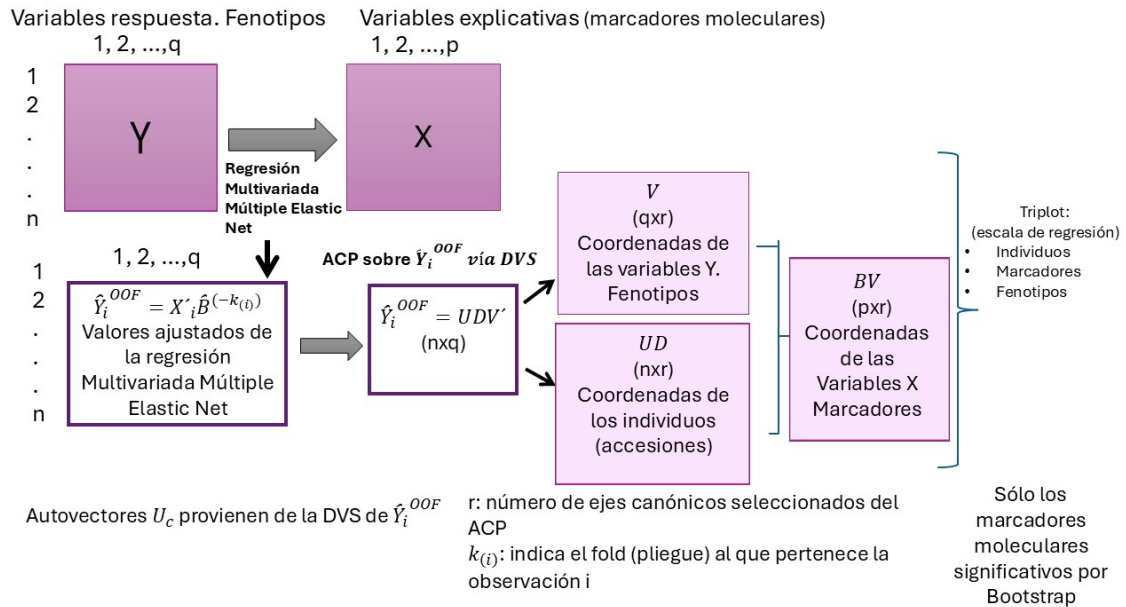


Figura 7.1: Proceso del análisis de redundancia (*RDA-like*) basado en *Elastic Net* multivariado

A continuación se describen los procedimientos para la aplicación de la propuesta metodológica, organizada en los siguientes pasos, cuyos resultados se presentan en el capítulo siguiente.

1. Preparación de datos. Se considera un conjunto de n observaciones caracterizadas por p predictores moleculares y múltiples variables fenotípicas cuantitativas de interés. Las matrices de predictores (X) y de respuestas (Y) se estandarizan (media cero y varianza unitaria) a fin de eliminar efectos de es-

cala y asegurar la comparabilidad entre variables.

2. Selección de parámetros de regularización. Se explora una grilla de valores para el parámetro de mezcla $\delta \in \{0, 1, 0.2, \dots, 0.9\}$ mediante validación cruzada de 5 pliegues. Para cada valor de δ se identifica el λ (intensidad de penalización) que minimiza el error de validación. Se determina que la combinación óptima (δ^*, λ^*) corresponde al valor que produce el menor error promedio de validación cruzada. Este procedimiento se implementa en R utilizando la función `cv.glmnet()` del paquete `glmnet`, con el argumento `family = "mgaussian"`. *Nota:* en la función `glmnet`, el parámetro δ corresponde al argumento `alpha`.
3. Ajuste del modelo de regresión multivariada *Elastic Net*. Con los parámetros óptimos (δ, λ) se ajusta el modelo definitivo según la formulación presentada en la Ecuación (3.4.10). Los coeficientes $\hat{B}$ se estiman mediante el algoritmo de descenso por coordenadas cíclico implementado en la librería `glmnet`. Se extraen y almacenan los coeficientes de regresión asociados a cada variable fenotípica, los cuales se exportan en tablas para su análisis posterior.
4. Evaluación de la significación de los coeficientes mediante *bootstrap*. Se implementa un procedimiento de remuestreo *bootstrap* no paramétrico con $R = 1000$ réplicas (`set.seed(2025)`), re-ajustando en cada réplica el modelo *Elastic Net* multivariado con los parámetros fijos (δ^*, λ^*) . Se extraen los coeficientes sin intercepto para cada respuesta y se obtienen intervalos de confianza al 95 % por percentiles:

$$[\hat{\beta}_{jk}^{*(0,025)}, \hat{\beta}_{jk}^{*(0,975)}].$$

Se considera significativo un coeficiente cuando el intervalo no incluye al cero. La matriz $\mathbf{X}$ se reduce al subconjunto de marcadores para los cuales al menos un coeficiente estimado $\hat{\beta}_{jk}$ resulta significativo.

5. Análisis multivariado de las predicciones. Con la combinación óptima (δ^*, λ^*) se obtienen las predicciones $\hat{Y}_i^{\text{OOF}}$ del modelo para todas las observaciones,

correspondientes a las predicciones *out-of-fold* generadas durante la validación cruzada. Dichas predicciones se asocian a un modelo *Elastic Net* multivariado ajustado sobre el conjunto completo de marcadores moleculares, bajo penalización.

La matriz *out-of-fold* $\hat{\mathbf{Y}}^{\text{OOF}}$ se somete a un ACP, con el fin de explorar patrones multivariados en las respuestas estimadas y describir la estructura de la variación explicada por los marcadores moleculares.

A partir de este ACP se extraen los autovalores, las coordenadas de individuos y de variables: las primeras se utilizan para representar cada accesión en el espacio definido por los dos primeros componentes principales, mientras que las segundas se grafican como vectores que indican la dirección y magnitud de la asociación de cada rasgo con dichos ejes.

Este procedimiento constituye un análogo al RDA clásico, con la diferencia de que se emplea $\hat{\mathbf{Y}}_i^{\text{OOF}}$ —proveniente de la regresión multivariada penalizada *Elastic Net*— en lugar de la matriz $\hat{\mathbf{Y}}$ obtenida mediante regresión múltiple ordinaria.

6. Construcción del gráfico de triplot estilo RDA y visualización de resultados.

A partir del ACP de $\hat{\mathbf{Y}}_i^{\text{OOF}}$ se elabora un triplot integrando observaciones, variables de respuesta y predictores moleculares significativos. Las coordenadas de los individuos se utilizan para representar las observaciones en el plano definido por las dos primeras componentes principales.

Las variables de respuesta se representa mediante sus coordenadas en dicho espacio. Los predictores moleculares se proyectan sobre el espacio canónico a partir de sus correlaciones con los ejes principales, permitiendo visualizar su asociación con la estructura multivariada de las respuestas. Luego se filtra para graficar únicamente aquellos marcadores identificados como significativos en el *bootstrap*.

El triplot final se construye con `ggplot2`, integrando capas de individuos, va-

riables de respuesta y predictores significativos.

7. Integración de resultados: identificación de predictores. A partir del gráfico triplot, se evalúan las observaciones y los predictores moleculares significativos cercanos a los fenotipos analizados.

En el Apéndice A.2 se muestra la función desarrollada en R para la propuesta *RDA-Like* aplicada a variables cuantitativas.

7.4.1 Supuestos y condiciones de aplicación

La propuesta metodológica se apoya en los siguientes supuestos:

- Existencia de correlación entre los caracteres fenotípicos, lo que justifica un abordaje multivariado.
- Relación aproximadamente lineal entre predictores y respuestas, entendida en el marco de modelos penalizados.
- Presencia de contextos de alta dimensionalidad ($p \gg n$ o $p \approx n$), que requieren regularización para garantizar estimación estable.

El enfoque resulta aplicable en:

- Estudios genómicos de alta dimensionalidad, en particular mapeo asociativo y predicción fenotípica a partir de marcadores moleculares.
- Contextos analíticos en los que se aborda de forma simultánea el modelado de múltiples respuestas correlacionadas, en función de un elevado número de variables independientes superior a la cantidad de observaciones.

7.5 Propuesta *RDA-Like* para variables dicotómicas

En la sección anterior se presentó una estrategia *RDA-Like* para el tratamiento de variables cuantitativas, basada en la combinación de un modelo multivariado

penalizado *Elastic Net* y ACP aplicado a las respuestas predichas. Sobre esta base, en la presente sección se extiende el planteo hacia el caso de variables cualitativas dicotómicas. Para ello se ajustará un modelo logístico multivariado penalizado mediante *Elastic Net*, del cual se obtendrán las predicciones $\hat{Y}_i^{\text{OOF}}$ generadas en el marco de la validación cruzada. Estas predicciones servirán como insumo para un análisis de componentes principales, a partir del cual se construirá un triplot integrando observaciones, variables de respuesta y predictores moleculares significativos. A esta extensión metodológica la denominamos *RDA-Like* multivariado para variables dicotómicas, en referencia a su carácter análogo al RDA clásico. A diferencia del modelo multivariado *Elastic Net* para variables cuantitativas, actualmente no existe una librería en R que implemente de manera directa el modelo logístico propuesto. Por ello, en lo que sigue se presenta el desarrollo teórico del modelo y se plantea una propuesta metodológica para su implementación mediante funciones en R-Studio. Posteriormente, se aplicará el modelo logístico multivariado con penalización *Elastic Net*, a los datos analizados en el capítulo anterior, considerando de manera simultánea las variables cualitativas dicotómicas en las que los modelos logísticos univariados evidenciaron asociaciones significativas.

7.5.1 Modelo logístico multivariado conjunto para respuestas binarias con penalización *Elastic Net*

El primer paso de la estrategia consiste en el ajuste de un modelo logístico multivariado conjunto para respuestas binarias, penalizado mediante *Elastic Net*. Desde el punto de vista probabilístico, el modelo se formula como un Bernoulli multivariado condicionado a los predictores moleculares (Dai et al., 2013), lo que permite modelar simultáneamente la dependencia de las respuestas dicotómicas respecto de los predictores y las asociaciones condicionales entre las propias respuestas. Sobre esta formulación probabilística se introduce un esquema de regularización penalizada que combina penalizaciones de tipo ridge y LASSO sobre los coeficientes de regresión, permitiendo abordar escenarios de alta dimensionalidad y fuerte colineali-

dad entre predictores. La formulación de la función objetivo penalizada y el esquema de optimización asociado, basado en métodos de gradiente proximal acelerado, se adoptan en línea con el marco general de regularización y optimización utilizados para modelos multivariados penalizados. Esta formulación se expresa mediante una verosimilitud conjunta de tipo log-lineal, que incorpora términos de interacción entre respuestas para capturar dependencias residuales no explicadas por los predictores. La implementación computacional del modelo se realiza mediante funciones desarrolladas ad hoc en R, integrando la formulación probabilística condicional con el esquema de regularización penalizada. En particular, la función de verosimilitud adopta la siguiente expresión:

$$\log p(y | x) = y^\top \eta(x) + \frac{1}{2} y^\top \Theta y - A(\eta(x), \Theta), \quad \eta(x) = B^\top x + b_0, \quad (7.5.1)$$

donde $Y = (Y_1, \dots, Y_K)^\top \in \{0, 1\}^K$ representa el vector de respuestas dicotómicas, $X \in \mathbb{R}^p$ el vector de predictores, $B \in \mathbb{R}^{p \times K}$ y $b_0 \in \mathbb{R}^K$ los parámetros de regresión, y $\Theta \in \mathbb{R}^{K \times K}$ una matriz simétrica con diagonal nula que describe las asociaciones residuales entre caracteres. El término $A(\eta(x), \Theta)$ denota la función de normalización (log-partición), que garantiza que $p(y | x)$ defina una distribución de probabilidad válida sobre $\{0, 1\}^K$.

La estimación se plantea por máxima verosimilitud penalizada, minimizando la pérdida logarítmica conjunta promedio más términos de regularización:

$$\begin{aligned} \min_{B, b_0, \Theta} & -\frac{1}{n} \sum_{i=1}^n \left(y_i^\top \eta_i + \frac{1}{2} y_i^\top \Theta y_i - A(\eta_i, \Theta) \right) \\ & + \frac{1-\delta}{2} \lambda \|B\|_F^2 + \delta \lambda \mathcal{P}_B(B) \\ & + \frac{1-\delta_\Theta}{2} \lambda_\Theta \|\Theta\|_F^2 + \delta_\Theta \lambda_\Theta \|\Theta\|_{1, \text{off}}, \end{aligned} \quad (7.5.2)$$

donde los parámetros δ y δ_Θ controlan, respectivamente, el balance entre las penalizaciones de tipo ridge y LASSO aplicadas a B y a Θ , permitiendo regular de manera independiente la intensidad de suavizado y de esparsidad en cada conjunto

de parámetros. Aquí $\|\Theta\|_{1,\text{off}} = \sum_{k < \ell} |\Theta_{k\ell}|$ promueve una red de asociaciones residuales escasa. Para B se utiliza la penalización *Elastic Net* (Zou & Hastie, 2005), que combina ridge (Hoerl & Kennard, 1970) y LASSO (Tibshirani, 1996) mediante el parámetro de mezcla $\delta \in [0, 1]$. Se consideran dos opciones para $\mathcal{P}_B(B)$:

- ℓ_1 **independiente**: $\mathcal{P}_B(B) = \|B\|_1$, que posibilita selección por rasgo.
- ℓ_{21} **grupal (multi-tarea)**: $\mathcal{P}_B(B) = \sum_{j=1}^p \|B_{j\cdot}\|_2$, que induce selección compartida de predictores entre respuestas.

La elección de hiperparámetros $(\lambda, \lambda_\Theta)$ se plantea mediante validación cruzada *K-fold*, minimizando la pérdida conjunta en validación y reportando tanto la solución $\lambda_{\min}$ como la correspondiente a la regla de un error estándar (1-SE), que selecciona el modelo más parsimonioso entre aquellos cuyo error de validación se encuentra dentro de un error estándar del mínimo observado.

El ajuste permite obtener probabilidades marginales condicionales $E[Y_k \mid X]$, identificar el patrón conjunto más probable en $\{0, 1\}^K$, y, en escenarios con un número reducido de caracteres dicotómicos, recuperar la distribución conjunta completa sobre $\{0, 1\}^K$.

En términos de implementación computacional, hasta donde alcanza nuestro conocimiento, no se dispone actualmente de una librería en R que permita ajustar de manera directa un modelo Bernoulli multivariado condicionado con interacciones entre respuestas, estimado mediante maximización de la verosimilitud conjunta completa e incorporando regularización simultánea tipo *Elastic Net* sobre los coeficientes de regresión y penalización ℓ_1 fuera de la diagonal en la matriz de asociaciones residuales. Por este motivo, el procedimiento de estimación se implementa de manera ad hoc, utilizando un esquema de optimización de gradiente proximal acelerado.

Obtención de respuestas predichas

A partir del ajuste del modelo se obtienen las probabilidades predichas $\hat{Y}_{ij} = P(Y_{ij} = 1 \mid X_i)$ para cada respuesta j y cada accesión i . Estas matrices de valores ajustados constituyen el análogo de los $\hat{Y}$ de los modelos lineales multivariados.

Implementación algorítmica y flujo de trabajo (desarrollo propio con apoyo de IA)

El ajuste del modelo logístico multivariado penalizado de (7.5.2) se implementa ad hoc en R mediante un esquema de optimización tipo FISTA (Por sus siglas en inglés: *Fast Iterative Shrinkage-Thresholding Algorithm*) con *backtracking* y aceleración de Nesterov. La parte suave de la función objetivo incluye la verosimilitud conjunta promedio y los términos ridge en B y Θ ; las no-suaves se resuelven con operadores proximales: (i) ℓ_1 por coeficiente ($\|B\|_1$) o $\ell_{2,1}$ por fila ($\sum_j \|B_{j\cdot}\|_2$) para selección compartida entre respuestas, y (ii) ℓ_1 en el triángulo superior de Θ para inducir una red de dependencia condicional rala entre las respuestas. En cada iteración, los gradientes se calculan vía momentos exactos del Bernoulli multivariado ($E[Y | x]$, $E[YY^\top | x]$), obtenidos por enumeración explícita de los 2^K estados posibles.

Secuencia de λ y validación cruzada

Para estabilizar la trayectoria de soluciones, $\lambda_{\text{máx}}$ se determina a partir de los datos mediante el gradiente evaluado en $B=0$ (y $\Theta=0$): norma ℓ_∞ (penalización ℓ_1) o norma-2 por fila (penalización $\ell_{2,1}$). Se usa una grilla log-espaciada $\lambda \in [\lambda_{\text{mín}}, \lambda_{\text{máx}}]$ con n_λ puntos. La selección de hiperparámetros se lleva a cabo mediante validación cruzada *K-fold* minimizando la pérdida conjunta en validación (verosimilitud negativa sin penalizaciones) y reportando tanto $\lambda_{\text{mín}}$ como la solución más parsimoniosa bajo la regla 1-SE.

Predicciones

El ajuste entrega: (i) probabilidades marginales $\hat{P}(Y_k=1 | X)$; (ii) el modo conjunto (patrón más probable en $\{0,1\}^K$); y (iii) la distribución conjunta completa (factible para K pequeño). Las matrices de predicciones marginales $\hat{Y}$ constituyen la base para un análisis *RDA-like*, desarrollado en la subsección siguiente.

Bootstrap e inferencia

Se implementa un *bootstrap* no paramétrico con $B = 500$ réplicas (paralelizado) fijando λ^* en λ_{ISE} (o $\lambda_{\text{mín}}$). Para cada parámetro se reportan intervalos percentiles al 95 %, utilizándose la exclusión del valor nulo como criterio para evaluar evidencia de asociación. Se exportan tablas largas de B (marcador $\times$ rasgo) y de pares (k, ℓ) de Θ .

Diagnósticos de consistencia

Se verifica la identidad diagonal $\text{diag}\{E[YY^\top \mid x]\} = E[Y \mid x]$ (binarias) y la simetría numérica de $E[YY^\top \mid x]$, junto con la coherencia entre los residuos medios de co-ocurrencia $(E[YY^\top] - YY^\top)$ y los pesos estimados en Θ . Se comparan además las soluciones correspondientes a $\lambda_{\text{mín}}$ y λ_{ISE} en términos de del grado de dispersión de los coeficientes de B y número de aristas en Θ .

Reproducibilidad y computación

La reproducibilidad se garantiza mediante el uso de `set.seed` y `RNGkind("L'Ecuyer-CMRG")`, y la ejecución se lleva a cabo en forma paralela utilizando `future.apply` en *Windows*. Se guardan automáticamente las curvas de validación cruzada (PNG/PDF/TIFF), los coeficientes, la matriz Θ y las métricas de CV en archivos `.xlsx`. La complejidad computacional de los momentos exactos escala como $O(n 2^K)$; por este motivo, el análisis se restringe a valores moderados de K , mientras que la extensión a aproximaciones variacionales para valores grandes de K se plantea como una línea de trabajo futuro.

7.5.1.1 Aplicación del modelo logístico multivariado con penalización *Elastic Net*

Tal como se hizo en el punto 6.2.4 con las variables cuantitativas a continuación se hace una aplicación de esta propuesta metodológica con el fin de evaluar empíricamente su performance. Para ello se ajustó un modelo logístico multivariado conjunto penalizado mediante *Elastic Net*, según la descripción del punto anterior, al conjunto de datos de quinoa, considerando simultáneamente seis variables cua-

litativas dicotómicas- presencia de axilas, presencia de estrías, presencia de ramas, forma de la panoja, tipo de la panoja y pigmento de los cotiledones- en función de 313 predictores moleculares. Dado que se consideran $K = 6$ respuestas binarias, el espacio de configuraciones conjuntas $\{0, 1\}^K$ contiene $2^K = 64$ estados posibles, correspondientes a todos los patrones binarios que pueden asumir simultáneamente las variables respuesta.

Resultados del ajuste multivariado

El procedimiento de estimación se llevó a cabo mediante el algoritmo FISTA con aceleración de Nesterov, incorporando penalización compartida por predictor ℓ_{21} ($\alpha = 0,5$) en la matriz de coeficientes B —para favorecer la selección conjunta de marcadores relevantes— y penalización ℓ_1 en la matriz Θ , induciendo una red residual dispersa de asociaciones entre caracteres. El parámetro de mezcla δ se calibró mediante una búsqueda en grilla en el intervalo $[0, 1, 0, 9]$, donde el valor $\delta = 0,40$ mostró el mejor desempeño al minimizar el error de validación cruzada con $K = 5$ *folds*, por lo que fue seleccionado para el modelo final. La regularización de los coeficientes B se evaluó sobre una secuencia de 60 valores de λ (`nlambda = 60`), generados entre $\lambda_{\text{máx}}$ y $\lambda_{\text{mín}} = 0,01 \lambda_{\text{máx}}$. El proceso de ajuste, realizado con semilla fija (`seed = 123`) y predictores estandarizados, identificó como valor óptimo $\lambda = 0,447$ bajo el criterio $\lambda_{\text{mín}}$, logrando un equilibrio adecuado entre capacidad predictiva y parsimonia. Paralelamente, la matriz de precisión residual Θ incorporó una penalización LASSO con parámetros $\lambda_{\Theta} = 0,01$ y $\delta_{\Theta} = 1,0$, con el objetivo de capturar dependencias residuales significativas entre las respuestas. Con esta configuración, el modelo produjo una matriz B con 612 coeficientes no nulos (32,6 % del total de 1.878 coeficientes), correspondientes a 102 marcadores únicos asociados al menos a un rasgo. La significancia de estas asociaciones se evaluó mediante un procedimiento de *bootstrap* con 1.000 réplicas, obteniendo intervalos de confianza percentilares al 95 %. Este análisis identificó únicamente cinco coeficientes estadísticamente significativos, que evidencian asociaciones relevantes entre el marcador Q050–204 con los rasgos presencia de estría, forma de la panoja y tipo de la pa-

noja, y entre el marcador Q112–185 con presencia de estría y presencia de rama. Para el resto de los coeficientes, los intervalos de confianza incluyeron al cero, por lo que no se cuenta con evidencia suficiente para afirmar su efecto bajo este criterio conservador (véase Figura 7.2).

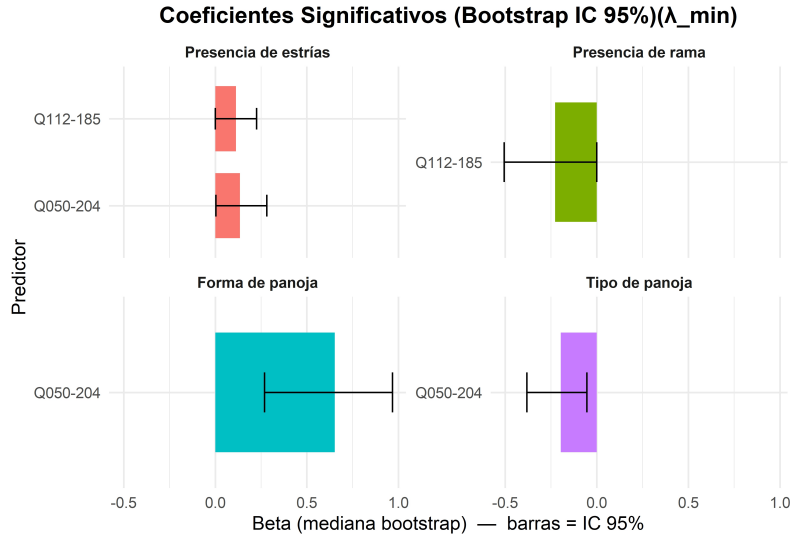


Figura 7.2: Intervalos de Confianza de los coeficientes significativos del modelo logístico multivariado penalizado mediante *Elastic Net*, *bootstrap* con $B = 1,000$ fijando $\lambda = \lambda_{1\min}$

El análisis *bootstrap* de los coeficientes de la matriz de precisión residual Θ permitió evaluar las asociaciones condicionales entre las variables cualitativas, una vez controlados los efectos de los predictores genéticos. Se identificaron asociaciones positivas entre presencia de estrías y forma de panoja, y entre presencia de ramas y tipo de panoja, mientras que la relación entre forma de panoja y tipo de panoja fue negativa. Estas tres conexiones son las únicas significativas. Los detalles de la implementación computacional pueden consultarse en el apéndice A.1, donde se presenta el código desarrollado en R. En la Figura 7.3 se identifican asociaciones positivas entre presencia de estrías y forma de panoja, así como entre presencia de rama y tipo de panoja. Asimismo, se observa una asociación negativa entre forma de panoja y tipo de panoja. Estas tres conexiones corresponden a las líneas más gruesas en la red, dado que son las únicas significativas; las restantes aristas se muestran en el gráfico para completar la estructura de Θ , pero sus intervalos contienen el cero y, por tanto, no representan asociaciones estadísticamente significativas.

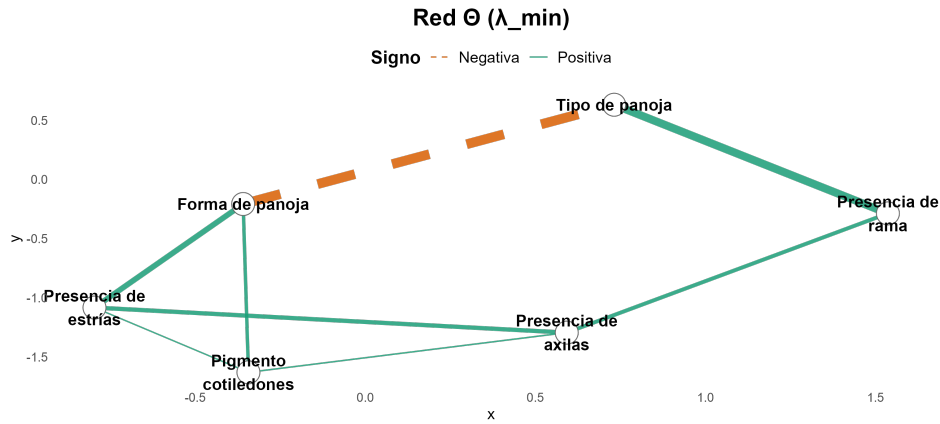


Figura 7.3: Red de asociaciones residuales entre variables cualitativas estimada a partir de los coeficientes de la matriz de precisión residual Θ del modelo multilogístico conjunto ajustado con $\lambda_{\min}$. Los nodos representan los rasgos cualitativos evaluados y las aristas indican dependencias condicionales significativas luego de controlar por los efectos de los marcadores moleculares. El grosor de las aristas es proporcional a la magnitud absoluta de los coeficientes $|\Theta_{ij}|$, reflejando la intensidad de la asociación residual entre pares de rasgos. Las líneas verdes continuas corresponden a asociaciones positivas y las líneas naranjas discontinuas a asociaciones negativas.

Comparación con los modelos univariados logísticos

La comparación entre los modelos univariados con corrección por multiplicidad de Li y Ji y el modelo logístico multivariado *Elastic Net* (EN) con selección por *bootstrap* pone de manifiesto una diferencia metodológica central. Mientras los enfoques univariados evalúan cada rasgo de manera independiente —identificando 12 asociaciones en este caso—, el modelo multivariado aplica una penalización conjunta sobre el conjunto de respuestas, privilegiando marcadores con efectos consistentes entre varios rasgos. Este esquema, combinado con la validación por *bootstrap*, introduce un umbral de estabilidad más exigente para las señales, lo que reduce el número total de asociaciones retenidas (5 en este estudio), pero produce un conjunto más parsimonioso y estructuralmente coherente con la naturaleza multivariada del enfoque.

En línea con este enfoque, el modelo multivariado recuperó marcadores con evidencia compartida: Q050–204 se asoció con presencia de estrías, forma de la panoja y tipo de la panoja; Q112–185 lo hizo con presencia de estrías y presencia de ramas. Solo Q050–204–forma de la panoja coincidió exactamente con los modelos

univariados, mientras que Q112-185 constituye un hallazgo exclusivo del enfoque multivariado. En contraste, señales presentes únicamente en los modelos univariados —como Q022-201 para forma de la panoja y varios marcadores en presencia de estrías (Q088-140, Q106-297, Q106-300, KGA20-178)— no fueron respaldadas bajo el criterio más restrictivo del modelo conjunto. En síntesis, la aplicación del enfoque multivariado con penalización ℓ_{21} —que favorece la selección grupal— resultó en un conjunto reducido de marcadores con efectos transversales, priorizando la estabilidad entre rasgos sobre el número de asociaciones.

7.5.2 Construcción del análisis *RDA-Like*

A partir de la matriz de respuestas predichas $\hat{Y}$ obtenida mediante el ajuste del modelo logístico penalizado *Elastic Net*, se propone construir un esquema de análisis de redundancia análogo al aplicado en el caso de caracteres cuantitativos (Figura 7.1). Este procedimiento permitirá explorar la estructura multivariada de las probabilidades predichas y su relación con los caracteres fenotípicos originales. En particular, se considerarán los siguientes pasos:

1. Centrar y escalar la matriz $\hat{Y}$ para homogeneizar la varianza entre respuestas.
2. Aplicar un ACP sobre $\hat{Y}$.
3. Extraer los autovalores y la proporción de varianza explicada.
4. Obtener las coordenadas de los individuos (accesiones) en el espacio de los componentes principales.
5. Construcción del gráfico de triplot estilo RDA y visualización de resultados.

A partir del ACP de la matriz de probabilidades predichas $\hat{Y}$ se elaborará un triplot integrando observaciones, variables de respuesta dicotómicas y predictores moleculares significativos. Para ello, se utilizarán las coordenadas de los individuos como puntos en el espacio de los dos primeros componentes principales.

Las variables de respuesta dicotómicas se representarán a partir de sus coordenadas. Los predictores moleculares se proyectarán en el espacio canónico calculando sus correlaciones con los ejes principales (`cor(X, scores_ind[, c("Dim.1", "Dim.2")])`). Luego se filtrarán para graficar únicamente aquellos marcadores identificados como significativos en el *bootstrap*.

El triplot final se construirá con `ggplot2`, integrando capas de individuos, variables de respuesta y predictores significativos.

6. Comparar la configuración obtenida con la de los caracteres cuantitativos, con el fin de evaluar patrones comunes o diferenciales en la estructura de covariación.

Este enfoque constituye un análisis *RDA-like* basado en predicciones de un modelo logístico multivariado penalizado, que integra la capacidad predictiva del modelo con herramientas exploratorias multivariadas.

Apoyo de IA (ChatGPT) y trazabilidad

El desarrollo algorítmico y la documentación del flujo de trabajo se llevó a cabo con apoyo de herramientas de inteligencia artificial, específicamente *ChatGPT* (OpenAI, 2025).

7.5.3 Supuestos y condiciones de aplicación

La propuesta metodológica se apoya en los siguientes supuestos de modelado:

- Los caracteres fenotípicos presentan correlación entre sí, lo que justifica un abordaje multivariado en lugar de un tratamiento univariado independiente.
- La relación entre predictores moleculares y respuestas cualitativas puede aproximarse mediante un modelo logístico lineal en los parámetros, con regularización que controla la complejidad.
- El escenario de trabajo corresponde a contextos de alta dimensionalidad ($p \gg$

n o $p \approx n$), lo que hace necesaria la incorporación de penalizaciones para garantizar estabilidad y evitar sobreajuste.

El enfoque propuesto es extensible a:

- Otros cultivos u organismos donde se disponga de datos de naturaleza fenotípica y genómica de estructura comparable.
- Estudios ómicos de alta dimensionalidad en los que las respuestas cualitativas (binarias) estén correlacionadas.
- Situaciones en las que se requiera modelar simultáneamente múltiples respuestas dicotómicas en presencia de un gran número de predictores.

7.6 Propuesta *RDA-Like* multivariado mixto: integración de respuestas cuantitativas y cualitativas.

7.6.1 Objetivo e idea central

El objetivo es sintetizar, en un único espacio canónico, la variación explicada por los predictores genéticos en ambos conjuntos de rasgos: cuantitativos (tratados mediante un modelo Gaussiano multivariado penalizado *Elastic Net*) y cualitativos dicotómicos (tratados mediante un modelo logístico multivariado penalizado *Elastic Net*).

En el primer caso, las respuestas estimadas $\hat{Y}_{\text{cuanti}}$ corresponden directamente a valores esperados de las variables cuantitativas. En el segundo, el modelo logístico multivariado proporciona estimaciones de probabilidades de ocurrencia para cada rasgo dicotómico, las cuales constituyen variables reales continuas en el intervalo $(0, 1)$.

En consecuencia, ambas componentes de $\hat{Y}$ se expresan en una escala cuantita-

tiva, lo que permite su integración en una única matriz de respuestas predichas,

$$\hat{Y}_{\text{mix}} = [\hat{Y}_{\text{cuanti}}, \hat{Y}_{\text{cuali}}],$$

que resume de manera conjunta la información explicada por los predictores genéticos sobre rasgos de naturaleza mixta. Sobre esta matriz $\hat{Y}_{\text{mix}}$, previamente centrada y escalada por columnas, se aplica un ACP. La representación gráfica obtenida puede interpretarse como una extensión del triplot clásico, en la cual se integran simultáneamente las observaciones, las variables respuesta —compuestas por caracteres fenotípicos de naturaleza mixta (cuantitativa y cualitativa)— y las variables explicativas moleculares. En este sentido, si bien se incorporan múltiples tipos de información, la estructura conceptual del gráfico se mantiene dentro del marco de los triplots, adaptado a un contexto de datos mixtos. El resultado se presenta como un triplot con respuesta mixta, que integra: (i) accesiones (puntos), (ii) vectores de rasgos (flechas verdes para cuantitativos y rojas para cualitativos) y (iii) vectores de marcadores moleculares significativos (flechas diferenciadas según el origen del hallazgo).

7.6.2 Construcción del espacio mixto

1. **Predicciones por bloque.** A partir de los modelos penalizados se obtienen:

$\hat{Y}_{\text{cuanti}} \in \mathbb{R}^{n \times K_q}$ que contiene los valores ajustados para los K_q rasgos cuantitativos, y $\hat{Y}_{\text{cuali}} \in \mathbb{R}^{n \times K_c}$ que contiene las probabilidades predichas para las K_c respuestas cualitativas.

2. **Estandarización y balanceo.** Cada bloque se centra y escala por columna. Para evitar que el bloque de mayor variabilidad domine el ACP, se aplica un balanceo por bloque que iguala, en promedio, la contribución de cada grupo de variables (traza de la covarianza comparable).
3. **ACP sobre la matriz conjunta.** Se forma $\hat{Y}_{\text{mix}} = [\hat{Y}_{\text{cuanti}}^*, \hat{Y}_{\text{cuali}}^*]$ y se realiza un ACP sin reescalado adicional (pues ya se estandarizó). Se reportan

autovalores y proporción de varianza explicada; en particular, se informa la proporción acumulada explicada por las dos primeras componentes.

7.6.3 Proyección de marcadores moleculares y armado del triplot

1. **Selección de marcadores moleculares significativos.** Se toman los conjuntos identificados por cada modelo penalizado (cuantitativo y cualitativo) bajo el umbral inferencial elegido (p.ej., *bootstrap*).
2. **Proyección como correlaciones con los ejes.** Dado el puntaje de individuos $S \in \mathbb{R}^{n \times 2}$ (PC1–PC2), cada marcador molecular x_j se proyecta vía correlaciones $r_j = \text{cor}(x_j, S)$, produciendo un vector en el círculo de correlación. Las flechas de marcadores moleculares se escalan para compartir el mismo marco que las flechas de variables respuesta.
3. **Visualización.** El triplot combina: puntos (accesiones, con etiquetas opcionales), flechas de respuestas (cuantitativa/cualitativa) y flechas de marcadores moleculares (coloreadas por origen: hallados en el modelo cuantitativo o cualitativo). Se dibuja además el círculo de correlación para facilitar la interpretación angular y de magnitud.

7.6.4 Interpretación

- **Direcciones de rasgos.** La orientación de cada flecha de respuesta indica su correlación con las componentes principales; su longitud refleja la calidad de representación en el plano (cuanto más próxima al círculo, mayor representación).
- **Accesiones.** La posición de cada accesión resume su perfil combinado respecto de $[\hat{Y}_{\text{cuanti}}, \hat{Y}_{\text{cuali}}]$. Accesiones próximas comparten patrones de respuestas predichas similares.

- **Marcadores moleculares.** Las direcciones de los marcadores moleculares muestran asociación con los ejes canónicos derivados de las respuestas. La alineación de un marcador molecular con determinadas flechas de rasgos sugiere contribución genética coherente con esos caracteres.
- **Coherencia entre bloques.** La proximidad angular entre flechas de rasgos cuantitativos y cualitativos revela patrones mixtos de co-variación: por ejemplo, alineación entre tipo/forma de panoja y variables de crecimiento.

7.6.5 Implementación y salidas

La función `RDA_mixto()` (ver Apéndice A.3) ejecuta los pasos anteriores y produce: (i) el triplot en formatos PNG/TIFF de alta resolución; (ii) un archivo `.xlsx` con autovalores (`eig`), coordenadas de individuos y variables (`ind_coord`, `var_coord`), y el cuadro final de flechas (`arrows`); (iii) tablas con correlaciones brutas marcador–componente por origen (cuantitativo/cualitativo) para evaluación de la proyección.

7.6.6 Ventajas y alcances

- Integra, en un único plano, la información explicada por modelos penalizados específicos por tipo de respuesta (gaussiano para cuantitativas y logístico para cualitativas), evitando las restricciones del RDA clásico cuando $p \gg n$.
- Permite inspeccionar coherencias entre bloques de respuestas multivariadas y vincularlas con predictores moleculares seleccionados.
- El balanceo por bloque y la estandarización controlan diferencias de escala entre conjuntos de variables, favoreciendo interpretaciones comparables.

7.6.7 Limitaciones

- Dado que el espacio canónico se construye a partir de las respuestas predichas por los modelos, los rasgos con bajo desempeño predictivo tienden a estar dé-

bilmente representados, lo que reduce en la práctica el conjunto de marcadores que contribuyen de manera apreciable a la configuración del plano.

- La proyección de marcadores moleculares basada en correlaciones con las componentes principales caracteriza asociaciones lineales con los ejes canónicos, pero tiene un alcance descriptivo y no sustituye los procedimientos inferenciales empleados para la identificación formal de asociaciones significativas.
- La interpretación suele centrarse en las primeras componentes, lo que puede omitir estructura relevante capturada en ejes de orden superior; por ello, la proporción de varianza acumulada debe ser evaluada para determinar la necesidad de considerar componentes adicionales.

Capítulo 8

Aplicación de la propuesta metodológica de mapeo asociativo para variables mixtas

8.1 Datos de quinoa

Se trabajó con la base de datos constituida por 25 accesiones de quinoa provenientes del Noroeste Argentino, descripta en la sección de Materiales 5 de esta tesis. Se consideraron las 17 variables cuantitativas: diámetro de tallo, número de ramas, forma de la hoja superior, número de dientes, longitudes de pecíolo, hoja, panoja, glomérulo, cotiledón e hipocótilo, junto con los días de siembra de emergencia, botón floral, floración y madurez fisiológica, además de ancho de hoja, altura de planta y diámetro de grano y las 6 variables cualitativas dicotómicas: presencia de axilas, presencia de estrías, presencia de rama, forma de panoja, tipo de panoja y pigmento de cotiledones. El conjunto de datos no presentó valores faltantes en las variables analizadas.

8.1.1 *RDA-Like* Multivariado para variables cuantitativas

El modelo multivariado *Elastic Net* se aplicó de manera conjunta a las 17 variables de respuesta cuantitativas, utilizando un total de 313 predictores moleculares. El ajuste se realizó mediante la librería `glmnet`, implementando una validación cruzada con cinco pliegues ($k = 5$) para comparar nueve valores del parámetro de mezcla δ (desde 0,1 hasta 0,9). Como resultado de este proceso, se identificó $\delta = 0,8$ en

combinación con un valor de penalización $\lambda = 1,01709$ como la configuración óptima, dado que presentó el menor error de validación cruzada (15.218,24), tal como se muestra en la figura 8.1.

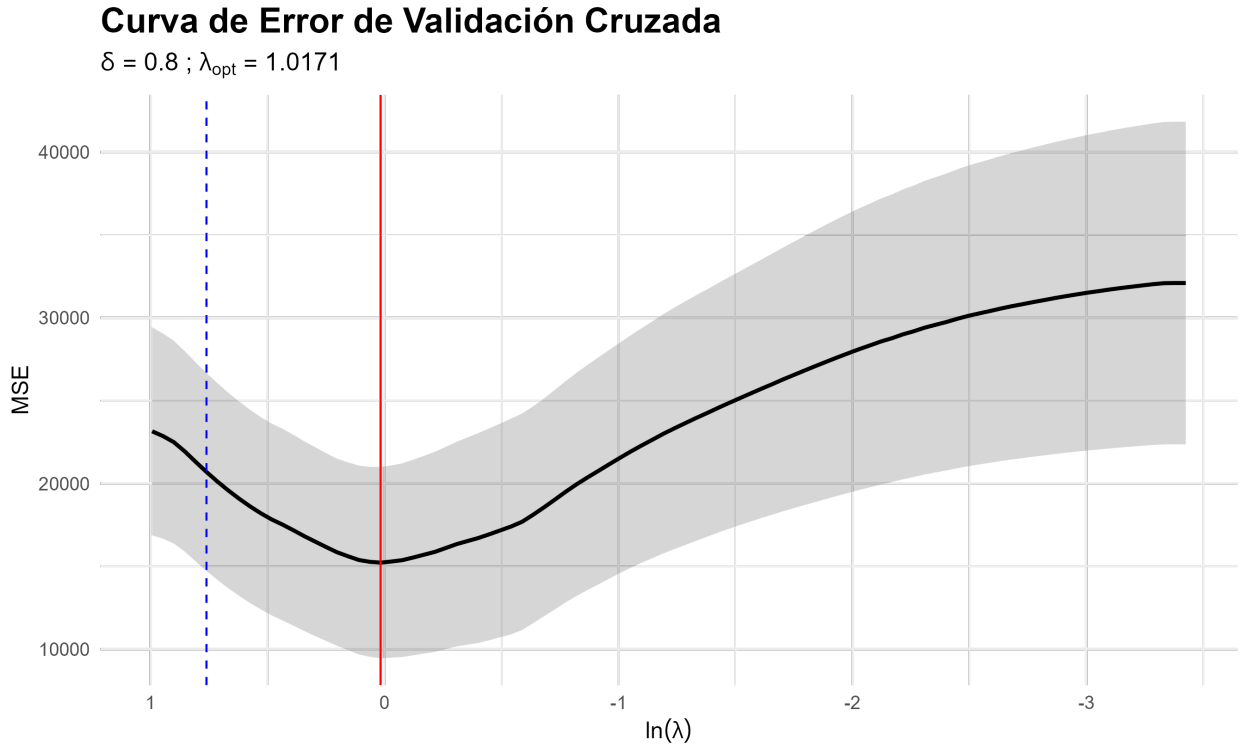


Figura 8.1: Curva de validación cruzada del modelo *Elastic Net* multivariado correspondiente a $\delta = 0,8$. El eje de las ordenadas representa el error promedio de predicción, calculado como el promedio de los errores cuadráticos medios individuales de cada variable respuesta. El eje de las abscisas representa $\ln(\lambda)$, donde $\ln$ denota el logaritmo neperiano.

El modelo final retuvo 374 coeficientes distintos de cero sobre un total de 5.321 posibles, es decir el 92,97% fueron reducidos a cero por efecto de la penalización. De los 374 restantes (distintos de cero), el análisis de remuestreo *bootstrap* (1.000 repeticiones) permitió identificar 27 coeficientes estadísticamente significativos. En las Figuras 8.2, 8.3 y 8.4 se muestran las 27 asociaciones detectadas, correspondientes a los tres marcadores seleccionados, para cada una de las variables respuesta y clasificadas según su grupo biológico. Dado que algunos marcadores resultan significativos en más de una variable respuesta, estas 27 asociaciones corresponden en realidad a tres marcadores distintos con efectos compartidos sobre múltiples rasgos. A partir de las matrices de predicciones obtenidas, de dimensiones 25×17 , se reali-

zó un ACP. Las dos primeras componentes explicaron el 73,5% y el 15,7% de la variabilidad total, respectivamente. En el triplot 8.5 se visualizan simultáneamente los individuos (accesiones), las variables de respuesta y los predictores significativos identificados (3 de los 313 evaluados).

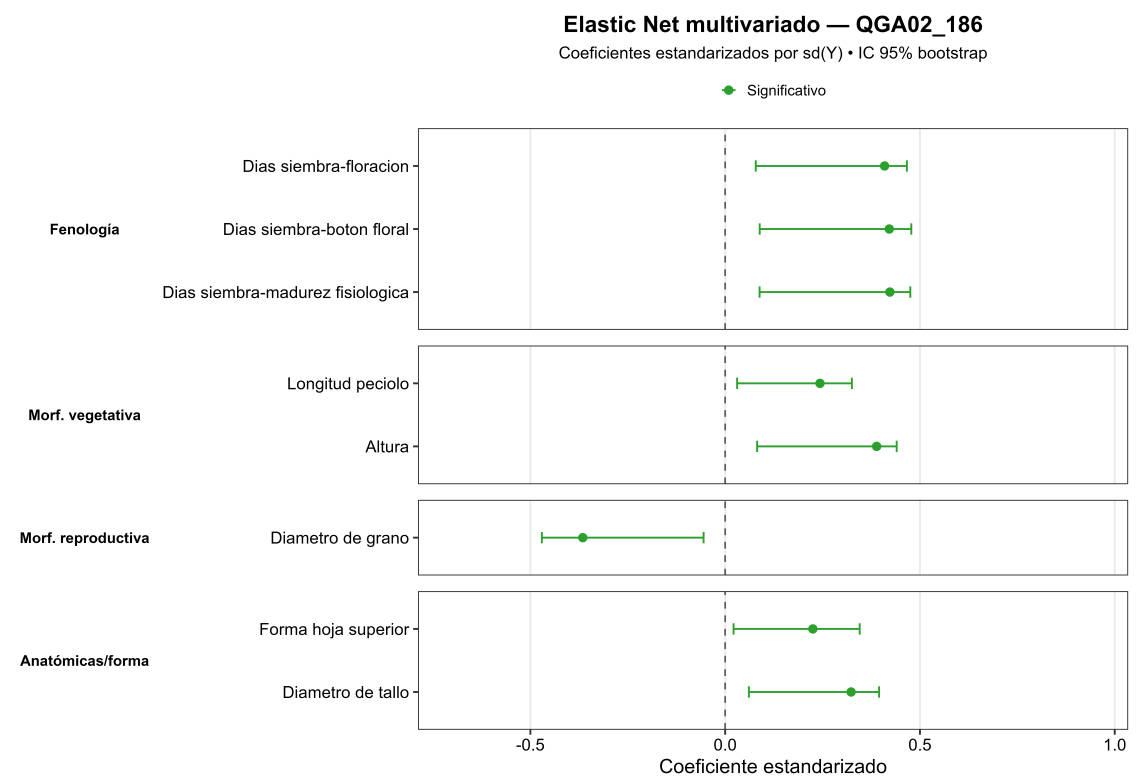


Figura 8.2: Marcador QGA02_186.

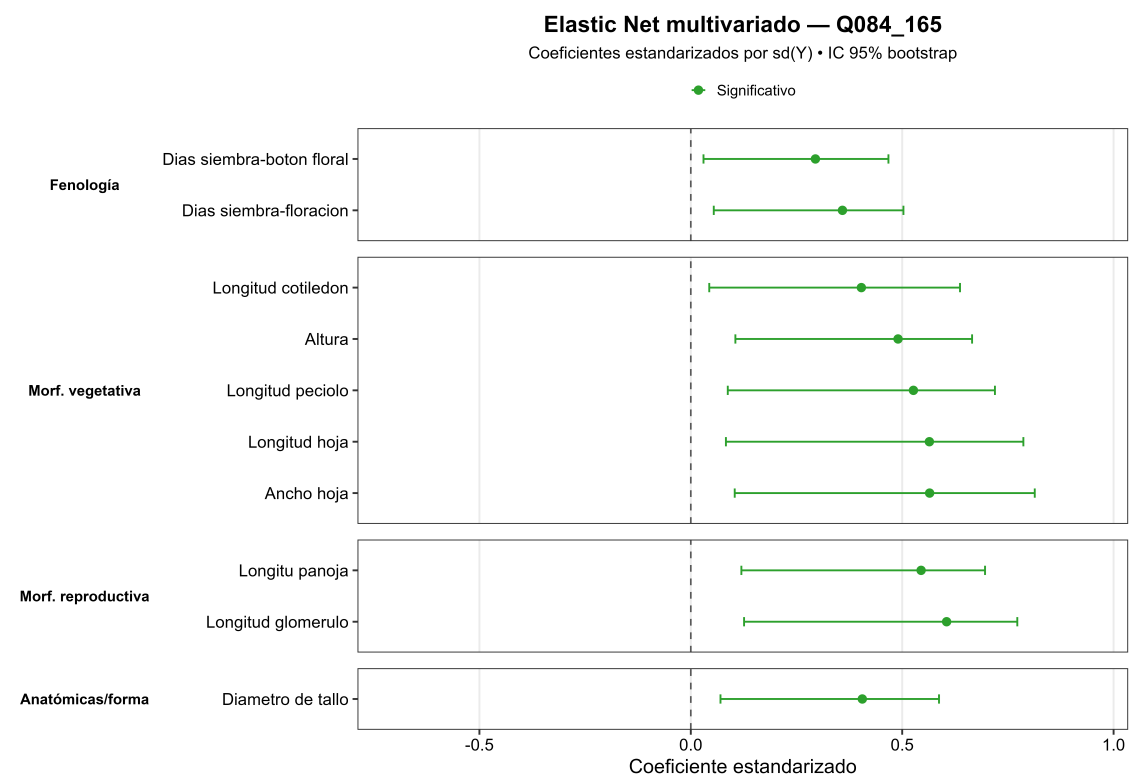


Figura 8.3: Marcador Q084_165.

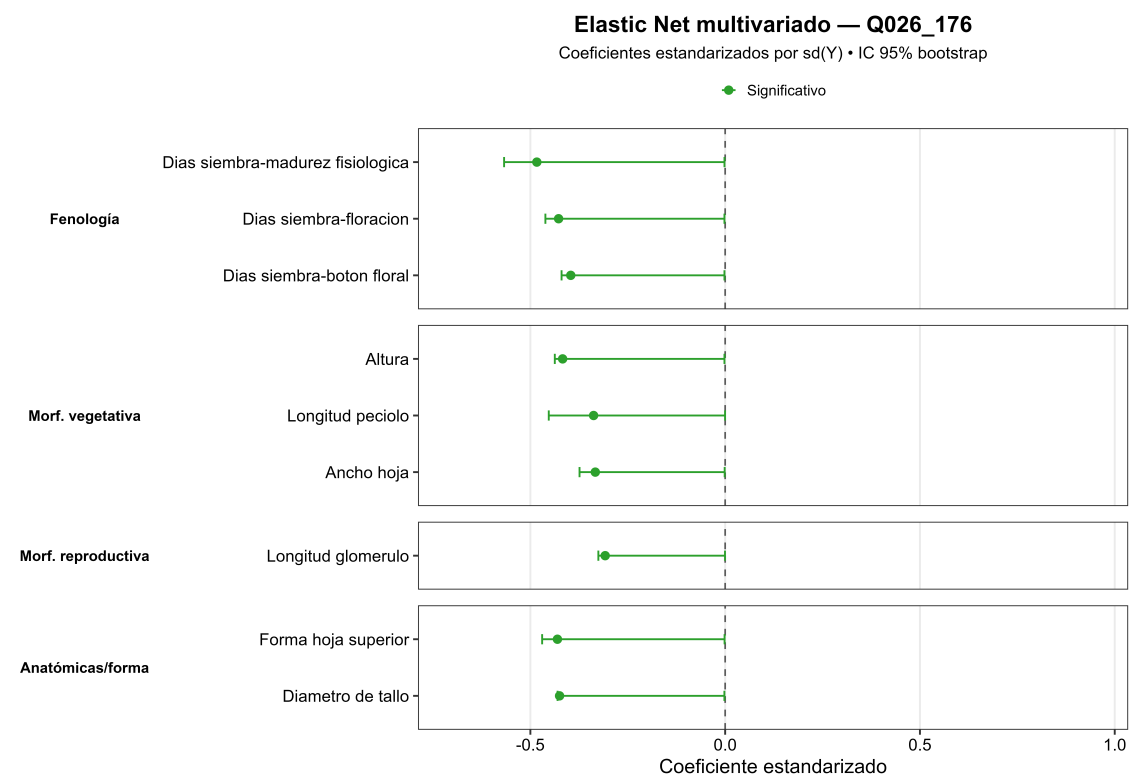


Figura 8.4: Marcador Q026_176.



Nota. El triplot se construyó a partir de un análisis de componentes principales (PCA) aplicado a la matriz de valores predichos $\hat{Y}$ obtenida a partir de un modelo *Elastic Net* multivariado ajustado con `glmnet` (`family = "mgaussian"`). Los puntos representan las coordenadas (*scores*) de los individuos en el plano definido por PC1–PC2, que establece la escala principal del gráfico. Las flechas azules corresponden a las variables fenotípicas y las flechas rojas a los marcadores moleculares, proyectados mediante la correlación de Pearson entre cada predictor estandarizado y los scores de los individuos. Para facilitar la visualización conjunta, las flechas fueron re-escaladas a la escala de los individuos usando $r_{\text{circ}} = 0,95 r_{\text{ind}}$, donde $r_{\text{ind}} = \max_i \sqrt{\text{PC1}_i^2 + \text{PC2}_i^2}$; esta re-escala es exclusivamente gráfica.

Los marcadores moleculares que resultaron significativos fueron Q026-176, Q084-165 y QGA02-186. El eje 1 concentra la mayor parte de la estructura multivariada y está fuertemente asociado a variables de tamaño y desarrollo como altura de la planta, longitud del glomérulo, días siembra floración, diámetro del tallo y días siembra botón floral, que se orientan en una misma dirección y definen el gradiente morfológico principal.

El eje 2 explica una proporción menor de la variación (15,7 %) y se asocia con la longitud de los hipocotilo.

Los marcadores seleccionados por *Elastic Net* se proyectan coherentemente junto a los rasgos con los que se asociaron, lo que confirma la consistencia multivariada de las asociaciones detectadas.

En el Cuadro 8.1 se presentan la raíz del error cuadrático medio (ECM) obtenida para este modelo y los coeficientes de determinación de la validación cruzada correspondientes a cada una de las variables de respuesta que mostraron asociación con al menos uno de estos marcadores.

Grupo biológico	Variable	$\sqrt{\text{ECM}}$	R^2
Fenología	Días siembra botón floral	0,74	0,43
	Días siembra floración	0,71	0,48
	Días siembra madurez biológica	0,74	0,43
Morfología Vegetativa	Altura de la planta	0,59	0,64
	Ancho de la hoja	0,77	0,39
	Longitud del cotiledon	0,99	0,00
	Longitud de la hoja	0,82	0,30
	Longitud del Pecíolo	0,81	0,32
Morfología Reproductiva	Longitud de la Panoja	0,85	0,25
	Longitud del glomérulo	0,77	0,38
	Diámetro del grano	0,88	0,19
Anatómica / forma	Diámetro del tallo	0,65	0,56
	Forma de hoja superior	0,89	0,17

Tabla 8.1: Raíz del error cuadrático medio ($\sqrt{\text{ECM}}$) y coeficiente de determinación (R^2) del modelo multiple multivariado *Elastic Net*.

8.1.2 *RDA-Like* multivariado para variables cualitativas dicotómicas

A partir de los resultados obtenidos en la subsubsección 7.5.1.1, donde se ajustó un modelo de regresión logístico multivariado conjunto para respuestas binarias *Elastic Net* para los seis rasgos binarios en función de todos los predictores, se procedió a aislar la estructura multivariada atribuible a X mediante un análisis *RDA-like* en dos pasos. En primer lugar, se calcularon probabilidades marginales predichas sin red ($\hat{Y}_{X\text{-only}} = \text{logit}^{-1}(XB + b_0)$), de modo de excluir las dependencias residuales

entre respuestas capturadas por Θ . Esto evita confundir correlación residual con estructura explicada por los predictores genéticos. Luego, se aplicó un ACP sobre $\hat{Y}_{X\text{-only}}$ (estandarizado por rasgo). El triplot resultante integra: (i) individuos (accesiones) en el plano de las dos primeras componentes, (ii) flechas rojas para los rasgos cualitativos (Y), y (iii) flechas azules para los marcadores significativos por *bootstrap*. La representación *RDA-like* para las variables cualitativas (Figura 8.6) muestra que los dos primeros componentes explican el 75,3 % de la variación total (PC1 = 43,1 %, PC2 = 32,2 %). El primer eje está principalmente asociado a los rasgos presencia de estrías, pigmento de los cotidellones y presencia de axilas, que se orientan hacia el lado positivo del eje y definen el gradiente dominante entre las características cualitativas evaluadas.

El eje 2, que explica el 32,2 % de la variación, está asociado a rasgos como presencia de rama y tipo de panoja, que se proyectan en la dirección positiva de este componente, diferenciando accesiones según la arquitectura de la planta.

Los marcadores moleculares significativos se proyectan de forma coherente con las respuestas con las que fueron asociados: en particular, Q112_185 con presencia de estrías y Q050_204 se alinea con presencia de estrías y forma de la panoja, lo que confirma la consistencia multivariada de las asociaciones detectadas por el modelo.

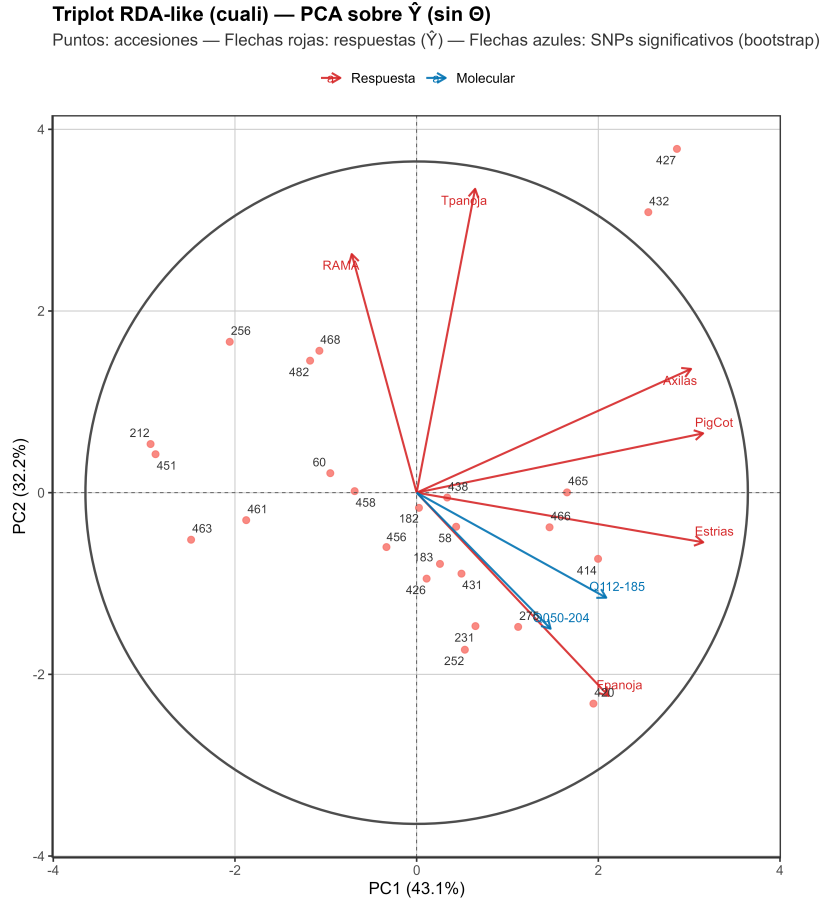


Figura 8.6: Triplot. Variables cualitativas de accesiones de quinoa.

En conjunto, los resultados apoyan un equilibrio adecuado entre parsimonia y capacidad explicativa: un subconjunto reducido pero consistente de marcadores con soporte de *bootstrap*, una red residual Θ interpretable y estable, y componentes principales (*RDA-like*) que capturan los ejes dominantes de variación fenotípica explicados por X .

8.1.3 *RDA-Like* multivariado mixto

Con el objetivo de sintetizar la estructura multivariada explicada por los predictores genéticos en ambos conjuntos de rasgos, se realizó un análisis *RDA-like mixto* aplicando un ACP sobre la matriz conjunta de respuestas predichas $[\hat{Y}_{\text{cuanti}}, \hat{Y}_{\text{cuali}}]$, obtenida a partir de los modelos *Elastic Net* multivariados: uno de tipo gaussiano para las diecisiete variables cuantitativas y otro logístico para las seis variables cua-

litativas. La Figura 8.7 muestra el triplot con respuesta mixta, correspondiente a las dos primeras componentes principales, que explicaron conjuntamente el 61,3 % de la variabilidad total. En el plano se observan las direcciones asociadas a los rasgos cuantitativos (flechas verdes), como altura de planta, ancho de hoja y días siembra-floración, y a las variables cualitativas (flechas rojas), entre las que destacan forma de panoja, tipo de panoja, presencia de ramas y pigmento de cotiledones. Los marcadores moleculares significativos identificados por ambos modelos se representaron con flechas azules turquesas (cualitativas) y violetas (cuantitativas): **Q026-176**, **Q084-165** y **QGA02-186** (identificados en el modelo cuantitativo) se orientaron en la dirección de los caracteres de crecimiento y desarrollo, mientras que **Q050-204** y **Q112-185** (asociados a las respuestas cualitativas) mostraron alineación con las variables estructurales de la panoja y el pigmento de cotiledones. La proyección conjunta de las accesiones en este espacio revela patrones de coexpresión fenotípica y asociaciones consistentes entre los ejes genéticos y los rasgos explicados por ambos conjuntos de modelos. El eje 1 se asocia principalmente de manera positiva con la presencia de axilas y el pigmento de cotiledones, mientras que muestra una asociación negativa con el ancho de hoja, la altura de la planta y la longitud del pecíolo. Por su parte, el eje 2 está caracterizado fundamentalmente por la presencia de ramas y el tipo de panoja, que presentan las mayores cargas sobre este componente.

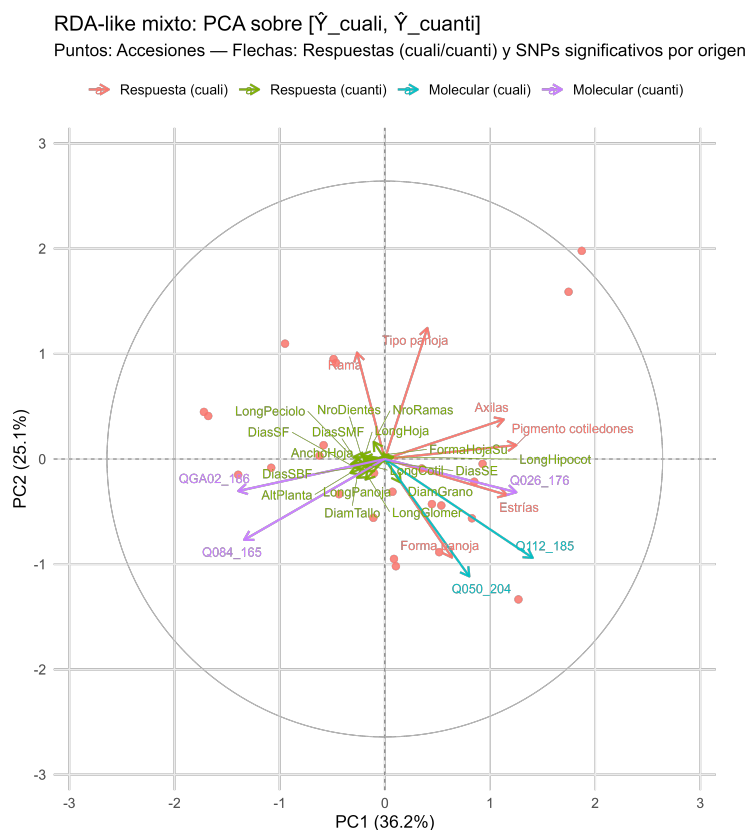


Figura 8.7: Triplot con respuesta mixta. Variables cuantitativas y cualitativas de accesiones de quinoa.

8.2 Datos de arroz

Se trabajó con la base de datos de arroz correspondiente al *Rice Diversity Panel 1* (RDP1), integrada por 342 accesiones descritas previamente en la sección de Materiales (capítulo 5). El conjunto fenotípico incluyó 18 variables cuantitativas que abarcan rasgos morfológicos y productivos: contenido de amilosa en grano (%), valor de dispersión alcalina, contenido de proteína en grano sin cáscara (%), días a espigazón (días), altura de planta (cm), longitud y ancho de la hoja bandera (cm), número de panículas por planta, longitud de la panícula (cm), número de ramas primarias por panícula (PBRNB), número total de flores por panícula (FLNBPPN), número de granos vacíos y llenos por panícula, longitud y ancho del grano con cáscara (mm), relación largo/ancho del grano (adimensional), volumen del grano (mm³) y peso de 100 granos (g).

Adicionalmente, se analizaron tres variables cualitativas dicotómicas (1 = Sí, 0 = No): desgrane espontáneo, presencia de aristas y pubescencia foliar. Como variables explicativas se utilizaron 186.833 SNPs.

8.2.1 *RDA-Like* multivariado para variables cuantitativas

Se ajustó un modelo de regresión múltiple multivariado *Elastic Net* mediante la función `cv.glmnet` de la librería `glmnet`, utilizando una validación cruzada *k-fold* con $k = 5$. El parámetro de mezcla δ se seleccionó empíricamente a partir de una búsqueda en una grilla de valores $\delta \in [0, 1, 0, 9]$. Para cada valor de δ , se calculó el error cuadrático medio (ECM) promedio obtenido mediante validación cruzada. El menor ECM promedio se obtuvo para $\delta = 0, 9$, valor que fue seleccionado para el ajuste final del modelo.

En la figura 8.8 se presenta la curva de validación cruzada correspondiente al parámetro de mezcla, $\delta = 0, 9$. Para este valor de δ , la función `cv.glmnet` calcula, para cada λ de la grilla, el ECM obtenido a partir de la validación cruzada *k-fold* ($k = 5$). El valor de λ seleccionado fue aquel que minimizó el ECM promedio de validación cruzada. En este caso, se obtuvo $\lambda = 0, 1817$ (línea azul discontinua). La línea roja indica λ_{1SE} , que corresponde a la mayor penalización cuyo error promedio no difiere significativamente del mínimo, siguiendo el criterio estándar propuesto por Friedman et al. (2010).

El modelo consideró un total de 3.362.994 coeficientes posibles (186.833 marcadores moleculares y 18 variables cuantitativas dependientes). Como resultado del proceso de regularización, el 99,65, % de los coeficientes fueron reducidos a cero, mientras que solo 11.628 coeficientes conservaron valores distintos de cero, representando potenciales asociaciones entre los marcadores moleculares y las variables fenotípicas. Los coeficientes no nulos del modelo multivariado *Elastic Net*, obtenidos bajo los hiperparámetros óptimos de regularización ($\delta = \delta_{\text{final}}$ y λ_{min}), fueron sometidos a un proceso de *bootstrap* percentil no paramétrico para evaluar su estabilidad y significancia. Mediante este procedimiento, se generaron $R = 500$ réplicas *bootstrap*, calculándose intervalos de confianza al 95 % para cada coeficiente β_{jk} . Un coeficiente

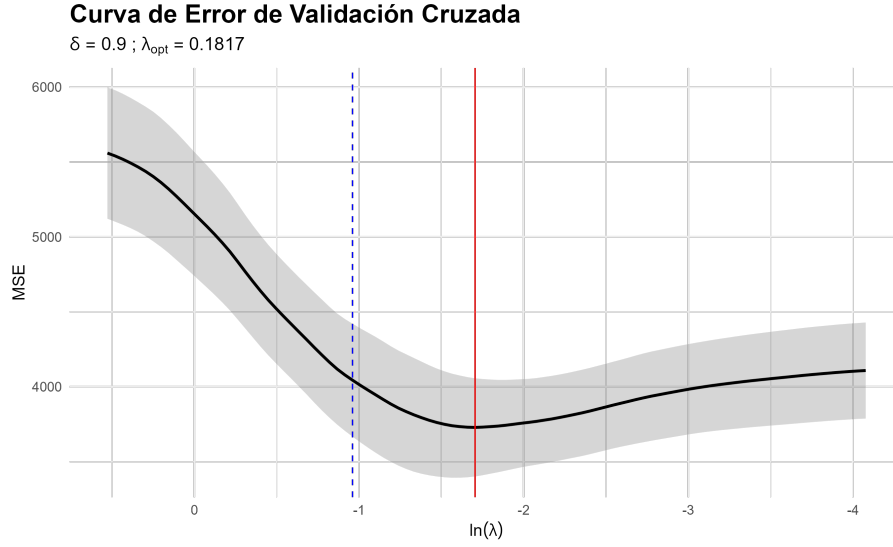


Figura 8.8: Curva de validación cruzada del modelo *Elastic Net* multivariado aplicado a los datos de arroz, con $\delta = 0.9$ y validación cruzada k -fold ($k = 5$). El eje de las ordenadas representa el error cuadrático medio (ECM) promedio de validación cruzada y el eje de las abscisas el logaritmo natural de λ . La línea azul indica el valor de λ que minimiza el ECM promedio y la línea roja corresponde a λ_{1SE} .

se consideró estadísticamente significativo cuando su intervalo de confianza *bootstrap* percentil $[IC_{2.5\%}, IC_{97.5\%}]$ no contenía al cero, indicando una asociación robusta entre el marcador molecular y la variable respuesta correspondiente.

Para abordar la alta dimensionalidad característica de los datos genómicos de Arroz (escenario $p \gg n$, con cientos de miles de marcadores SNP), el protocolo de *bootstrap* requirió adaptaciones computacionales significativas respecto del enfoque utilizado en los datos de quinoa. Se implementó un sistema de procesamiento por bloques mediante la función `run_bootstrap_chunk()`, que divide las réplicas en segmentos manejables de 30 iteraciones, con guardado incremental a través de `safe_write_rds()` para evitar el colapso de la memoria RAM.

A diferencia del enfoque clásico de *bootstrap* descrito por Davison y Hinkley (1997), implementado en la función `boot()` del paquete `boot`, en este trabajo se empleó un esquema adaptado que incorpora un control explícito de la convergencia del algoritmo de optimización, siguiendo las recomendaciones de Friedman et al. (2010). En particular, se ajustaron los parámetros del optimizador mediante `glmnet::glmnet.control(mxiter = 50000, fdev = 1e-4)` con el fin de estabilizar la

convergencia en un escenario ultra-dimensional, en el que la cantidad de predictores excede ampliamente el número de observaciones. Esta configuración resulta esencial para garantizar que cada réplica *bootstrap* alcance una solución numéricamente estable, evitando oscilaciones del algoritmo de coordenadas cíclicas cuando el número de coeficientes posibles se encuentra en el orden de millones.

Estas estrategias resultaron indispensables: sin la división en bloques y sin una gestión del uso de memoria cuidadosamente controlada, el análisis se tornaba computacionalmente inviable en equipos de escritorio convencionales. Bajo el esquema propuesto, cada bloque requiere aproximadamente 1 GB de RAM, mientras que la implementación estándar supera holgadamente los 10 GB, imposibilitando la ejecución completa del procedimiento *bootstrap*.

Una vez obtenidas las matrices de predicciones $\hat{Y}$, de dimensiones 342×18 , se realizó un análisis de componentes principales. Las dos primeras componentes explicaron el 35,2 % y el 23,3 % de la variabilidad total, respectivamente. En el triplot presentado en la Figura 8.9 se visualizan simultáneamente los individuos (accesiones), las variables respuesta y los predictores significativos identificados por el modelo (17 de los 186.833 SNP originales), permitiendo caracterizar de manera integrada los patrones de covariación entre genotipo y fenotipo. El eje 1 se asocia de manera positiva a variables como longitud de la bandera, días de floración y longitud de la panícula.

El eje 2, que explica el 23,3 % de la variación, se vincula de manera negativa con número de panículas y de forma positiva con el número de ramas.

Los marcadores moleculares significativos fueron: SNP-1.3458690, SNP-2.20400120, SNP-2.7090398, SNP-3.16707153, SNP-3.16732086, SNP-4.30882902, SNP-5.5371749, SNP-5.5372932, SNP-6.1735459, SNP-6.1764761, SNP-6.6709536, SNP-7.22062161, SNP-7.22189832, SNP-7.25324722, SNP-7.28315724, SNP-8.28098245, SNP-9.21502292. En el cuadro 8.2 se visualizan la raíz del ECM hallados en este modelo y los coeficientes de determinación de CV de cada una de las variables que tuvieron significación con algún marcador molecular, según su grupo biológico.

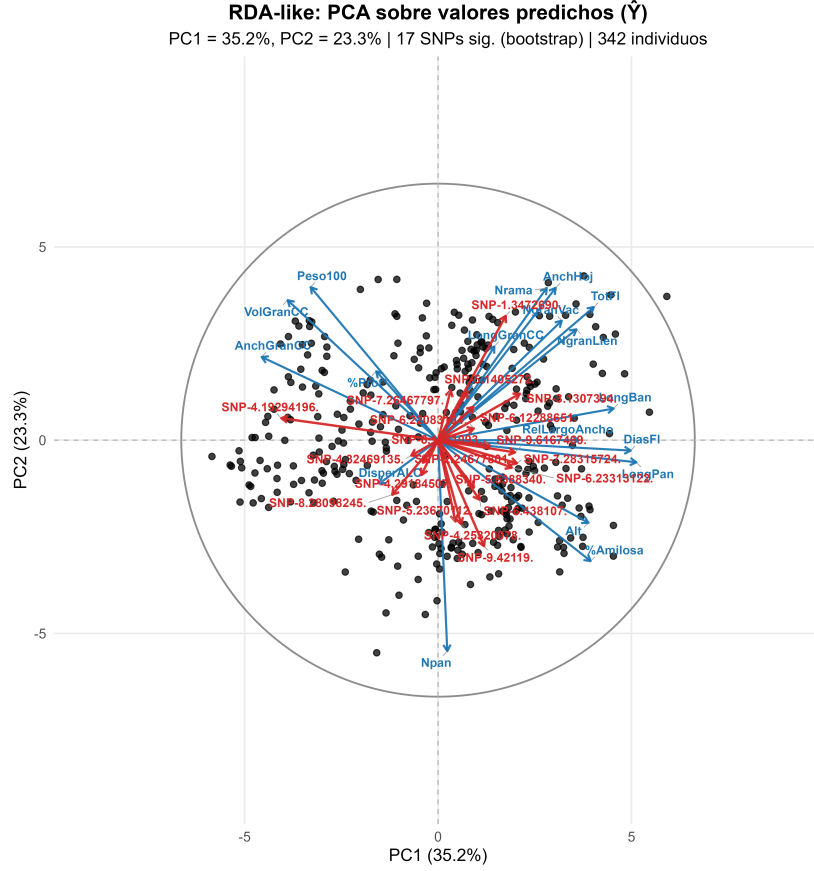


Figura 8.9: Triplot. Variables cuantitativas de accesiones de arroz.

Nota. El triplot se construyó a partir de un ACP aplicado a la matriz de valores predichos $\hat{Y}$ obtenida a partir de un modelo *Elastic Net* multivariado ajustado con `glmnet` (`family = "mgaussian"`). Los puntos representan las coordenadas (*scores*) de los individuos en el plano definido por PC1–PC2, que establece la escala principal del gráfico. Las flechas azules corresponden a las variables fenotípicas y las flechas rojas a los marcadores moleculares, proyectados mediante la correlación de Pearson entre cada predictor estandarizado y los *scores* de los individuos. Para facilitar la visualización conjunta, las flechas fueron re-escaladas a la escala de los individuos usando $r_{\text{circ}} = 0,95 r_{\text{ind}}$, donde $r_{\text{ind}} = \max_i \sqrt{\text{PC1}_i^2 + \text{PC2}_i^2}$; esta re-escala es exclusivamente gráfica.

8.2.2 *RDA-Like* multivariado para variables cualitativas dicotómicas

Por la alta dimensionalidad de la matriz de SNP de los datos de arroz, para la selección óptima de hiperparámetros en el modelo de regresión logística multivariante con regularización *Elastic Net*, se aplicó una estrategia en dos fases. En la primera fase, se implementó un procedimiento de *screening* para reducir el espacio predictor de 186.833 a 10.000 SNP. Para ello, se calculó, para cada marcador, la

Grupo biológico	Variable	$\sqrt{\text{ECM}}$	R^2
Calidad del grano	Contenido de amilosa del grano	3,41	0,61
	Valor de dispersión alcalina	0,85	0,36
	Contenido de proteína en grano sin cáscara	0,95	0,12
Fenología	Días a floración	9,31	0,45
Morfología vegetativa	Altura de planta	14,10	0,53
	Longitud de hoja bandera	4,75	0,29
	Ancho de hoja bandera	0,16	0,60
Morfología reproductiva (inflorescencia)	Número de panículas por planta	6,62	0,66
	Longitud de panícula	2,53	0,45
	Número de ramas primarias	1,41	0,33
	Total de flores por panícula	41,74	0,31
Componentes del rendimiento	Número de granos vacíos	18,67	0,29
	Número de granos llenos	35,53	0,27
	Peso de 100 granos	0,34	0,48
Anatómica / forma del grano	Longitud del grano con cáscara	0,58	0,63
	Ancho del grano con cáscara	0,22	0,67
	Relación largo/ancho del grano	0,30	0,68
	Volumen del grano con cáscara	1,46	0,62

Tabla 8.2: Raíz del error cuadrática medio ($\sqrt{\text{ECM}}$) y coeficiente de determinación (R^2) del modelo multiple multivariado *Elastic Net*.

magnitud del gradiente de la función de pérdida logística multivariante respecto de los coeficientes asociados, estimando la contribución promedio de cada predictor sobre todas las variables respuesta. Los SNPs se ordenaron según la norma euclídea del gradiente agregado ($s_j = \sqrt{\sum_k g_{jk}^2}$) y se retuvieron los 10 000 de mayor magnitud, considerados los más informativos. Sobre este subconjunto se evaluó un rango de valores de $\delta \in \{0,10; 0,15; 0,20; 0,25; 0,30; 0,70; 0,90\}$ mediante validación cruzada de tres *folds*, identificándose $\delta^* = 0,10$ como el valor que minimizó consistentemente la función de pérdida conjunta. Cabe destacar que el valor seleccionado $\delta^* = 0,10$ sitúa al modelo más próximo al régimen ridge que al LASSO, privilegiando la estabilidad de los coeficientes frente a la sparsidad, una propiedad deseable dada la alta colinealidad entre marcadores SNP.

En la segunda fase, utilizando la matriz completa de SNP (342 x 186.833), se fijó $\delta = 0,10$ y se generó una secuencia de ocho valores de λ , calculada a partir de $\lambda_{\max}$ derivado de las condiciones KKT (norma en grupo), distribuidos logarítmicamente entre $0,5\lambda_{\max}$ y $10^{-4}\lambda_{\max}$. La validación cruzada identificó $\lambda_{\min}$ como el valor óptimo, correspondiente al mínimo promedio de la función de pérdida entre *folds*. Esta estrategia en dos etapas permitió optimizar eficientemente δ sobre un subconjunto manejable y luego refinar λ utilizando toda la información disponible. Se observa en

la Figura 8.10 el valor óptimo de $\lambda = 0,089$ para $\delta = 0,10$

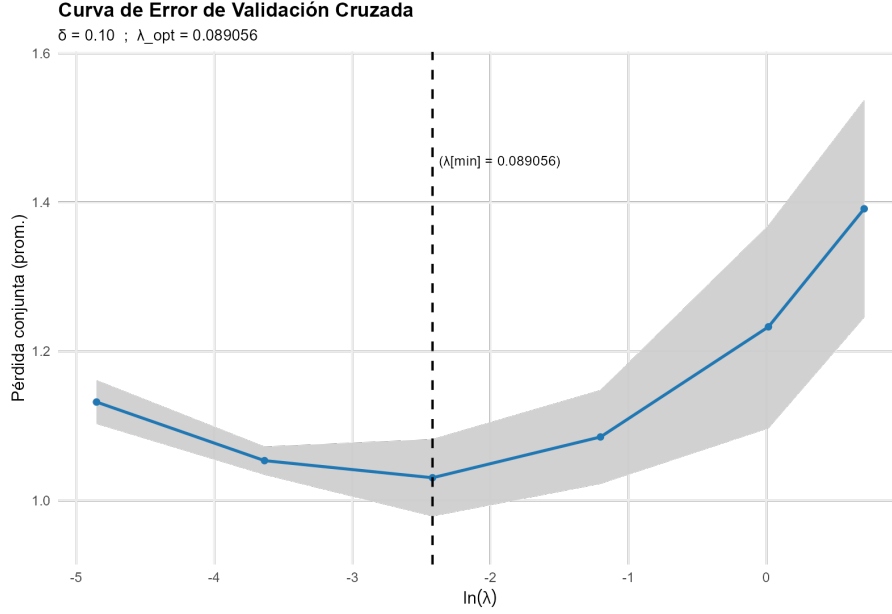


Figura 8.10: Curva de validación cruzada del modelo *Elastic Net* multivariado logístico correspondiente a $\delta = 0,10$. El eje de las ordenadas representa la pérdida conjunta promedio de validación cruzada, calculada a partir de la verosimilitud negativa sin penalizaciones. El eje de las abscisas representa $\ln(\lambda)$, donde $\ln$ denota el logaritmo neperiano.

Con esta configuración, el modelo estimó 11.889 celdas no nulas (2,12 %) en la matriz B , de un total de 560.499 posibles, y seleccionó 3.963 SNPs únicos (asociados al menos a un rasgo). La significación de las asociaciones se evaluó mediante un procedimiento de *bootstrap* con $B = 510$ réplicas, construyendo intervalos de confianza percentilares al 95 %. Este análisis permitió identificar 35 coeficientes significativamente distintos de cero, correspondientes a 20 SNP únicos de los 186.833 evaluados. En la representación de los resultados (Cuadro 8.3), se utilizaron colores para destacar los marcadores que aparecen asociados a más de un rasgo cualitativo, de modo que aquellos SNP repetidos entre diferentes rasgos comparten el mismo color.

El análisis *bootstrap* de los coeficientes de la matriz de precisión residual Θ permitió evaluar las asociaciones condicionales entre las variables cualitativas, una vez controlados los efectos de los predictores genéticos. En la Tabla 8.4 se presentan los intervalos de confianza de los parámetros θ_{ij} obtenidos a partir de 510 remuestreos.

Variables cualitativas	Snp _s Signifivcativos	Variable cualitativa	Snp _s Signifivcativos
Presencia de aristas	SNP-12.4192677	Pubescencia foliar	SNP-10.4091661
	SNP-2.20060954		SNP-12.4192677
	SNP-5.1066212		SNP-2.20060954
	SNP-5.990736		SNP-2.8607923
	SNP-6.5174235		SNP-2.8686776
	SNP-6.5179580		SNP-5.1066212
	SNP-6.5181054		SNP-5.1730850
	SNP-6.5182557		SNP-5.173474
	SNP-6.5188266		SNP-5.259180
	SNP-7.203722		SNP-5.266557
Desgrane espontáneo	SNP-10.4091661	Pubescencia foliar	SNP-5.287612
	SNP-2.20060954		SNP-5.87429
	SNP-5.173474		SNP-5.918878
	SNP-5.259180		SNP-5.990736
	SNP-5.87429		SNP-6.5174235
	SNP-5.990736		SNP-6.5179580
			SNP-6.5181054
			SNP-6.5182557
			SNP-6.5188266

Tabla 8.3: Marcadores significativos a partir de *bootraps* del modelo logístico *Elastic Net* multivariado de los datos de arroz.

Se consideraron significativas aquellas asociaciones cuyo intervalo de confianza no incluyó el valor cero. Bajo este criterio, se identificó un par de respuestas con dependencia residual significativa: desgrane espontáneo - Pubescencia foliar (IC: [0,299, 0,842]).

y1	y2	LI	LS	Significación
Presencia de aristas	Pubescencia foliar	-0,005	0,026	NS
Desgrane espontáneo	Presencia de aristas	-0,022	0,015	NS
Desgrane espontáneo	Pubescencia foliar	0,299	0,842	Significativo

Tabla 8.4: Intervalos de confianza de los parámetros θ_{ij} obtenidos a partir de 1.000 remuestros del modelo logístico *Elastic Net* multivariado.

La Figura 8.11 muestra la red de asociaciones residuales correspondiente al modelo ajustado con $\lambda_{\min}$, donde los nodos representan las variables de respuesta cualitativas y las aristas indican dependencias condicionales significativas luego de controlar por los efectos de los marcadores moleculares, derivadas de los coeficientes de Θ . Se consideraron significativas aquellas asociaciones cuyos intervalos de confianza bootstrap al (95 %) excluyeron al cero. El grosor de las aristas es proporcional a la magnitud absoluta de los coeficientes $|\Theta_{ij}|$, reflejando la intensidad de la asociación

residual entre pares de rasgos.

Las conexiones positivas (líneas verdes continuas) representan asociaciones directas entre los rasgos, mientras que las negativas (líneas naranjas discontinuas) indican relaciones inversas.

De acuerdo con los intervalos de confianza obtenidos por *bootstrap*, únicamente la asociación entre desgrane espontáneo y pubescencia foliar resultó estadísticamente significativa, ya que su intervalo de confianza no incluyó el cero. Las demás conexiones que aparecen en la red, incluyendo las relaciones entre desgrane espontáneo y presencia de aristas y entre pubescencia foliar y presencia de aristas, se muestran para completar la estructura de Θ , pero sus intervalos de confianza contienen el cero y, por lo tanto, no representan asociaciones residuales significativas.

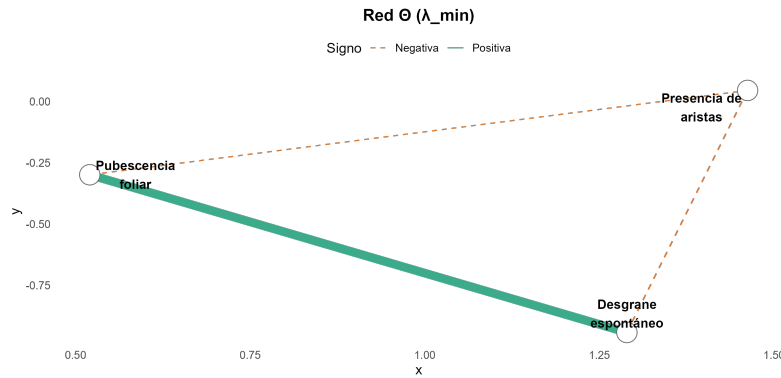


Figura 8.11: Red de asociaciones residuales entre variables cualitativas estimada a partir de los coeficientes de la matriz de precisión residual Θ del modelo multilogístico conjunto ajustado con $\lambda_{\min}$. El grosor de las aristas es proporcional a la magnitud absoluta de los coeficientes $|\Theta_{ij}|$. Las líneas verdes continuas corresponden a asociaciones positivas y las líneas naranjas discontinuas a asociaciones negativas.

Siguiendo el mismo enfoque empleado para los datos de quínoa, se llevó a cabo un análisis *RDA-like* en dos etapas. Primero, se calcularon probabilidades marginales predichas sin red ($\hat{Y}_{X\text{-only}} = \text{logit}^{-1}(XB + b_0)$) y luego, se aplicó un ACP sobre $\hat{Y}_{X\text{-only}}$ (estandarizado por rasgo). El triplot resultante integra: (i) individuos (accesiones) en el plano de las dos primeras componentes, (ii) flechas rojas para los rasgos cualitativos (Y), y (iii) flechas azules para los SNPs significativos por *bootstrap*. La Figura 8.12 muestra que las dos primeras componentes explican el 74,6 %

de la variabilidad en $\hat{Y}_{X\text{-only}}$. El eje 1 se asocia de manera positiva con la presencia de aristas, mientras que el desgrane espontáneo y la pubescencia foliar aportan a ambos ejes.

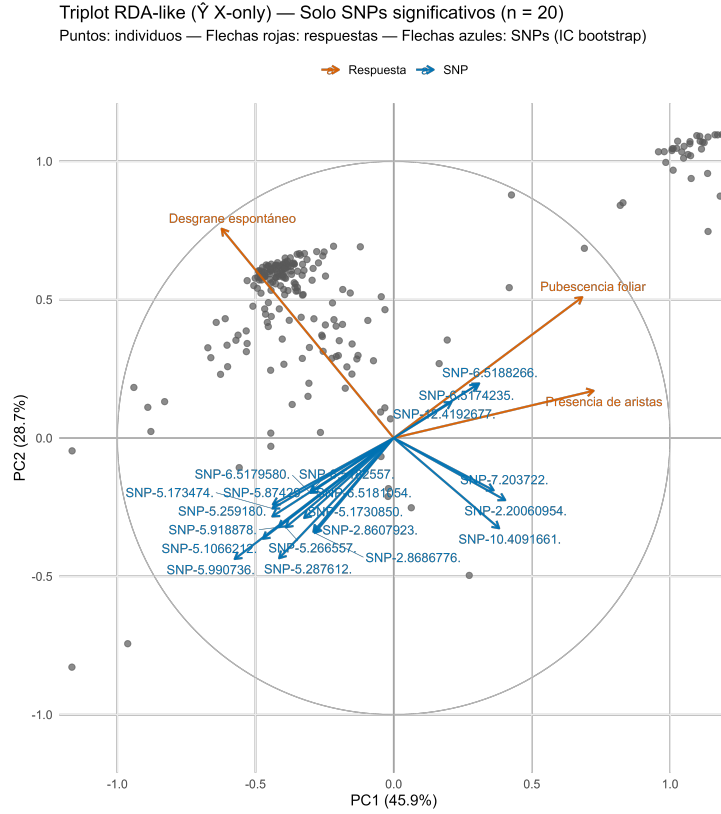


Figura 8.12: Triplot. Variables cualitativas de accesiones de arroz.

En conjunto, los resultados muestran una buena articulación entre simplicidad del modelo y capacidad interpretativa: se obtuvo un conjunto reducido pero robusto de marcadores con respaldo de *bootstrap*, una red residual Θ coherente y estable, y un análisis *RDA-like* que resume adecuadamente la variación fenotípica asociada a X .

Se distinguen distintas agrupaciones de accesiones: un grupo próximo al rasgo desgrane espontáneo, y otro grupo menor, ubicado en la dirección de los rasgos pubescencia foliar y presencia de aristas, en el cual se observan muy próximos los marcadores **SNP-6.5188266**, **SNP-6.5174235** y **SNP-12.4192677**.

8.2.3 *RDA-Like* Multivariado mixto

Con el mismo objetivo planteado para el análisis de quínoa, se llevó a cabo un análisis *RDA-like mixto*, aplicando un ACP sobre la matriz conjunta de respuestas predichas $[\hat{Y}_{\text{cuanti}}, \hat{Y}_{\text{cuali}}]$, derivada de los modelos *Elastic Net* multivariados —uno de tipo gaussiano para las 18 variables cuantitativas y otro logístico para las tres variables cualitativas. La Figura 8.13 muestra el triplot con respuesta mixta, correspondiente a las dos primeras componentes principales, que explicaron conjuntamente el 65,5 % de la variabilidad total. En el plano se observan las direcciones asociadas a los rasgos cuantitativos (flechas verdes), tales como porcentaje de amilosa, días a floración y relación largo/ancho del grano, y los rasgos cualitativos (flechas rojas) desgrane espontáneo, pubescencia foliar y presencia de aristas. Los marcadores moleculares significativos identificados por ambos modelos se representaron con flechas celestes violetas (cuantitativas) y turquesas (cualitativas). La proyección conjunta de las accesiones en este espacio evidencia patrones de covariación fenotípica y asociaciones consistentes entre los ejes genéticos y los rasgos explicados por ambos conjuntos de modelos. El eje 1 se asocia principalmente de manera positiva con la presencia de aristas, la dispersión alcalina y el ancho de grano con cáscara, mientras que muestra una asociación negativa con la longitud del grano con cáscara y el número de ramas.

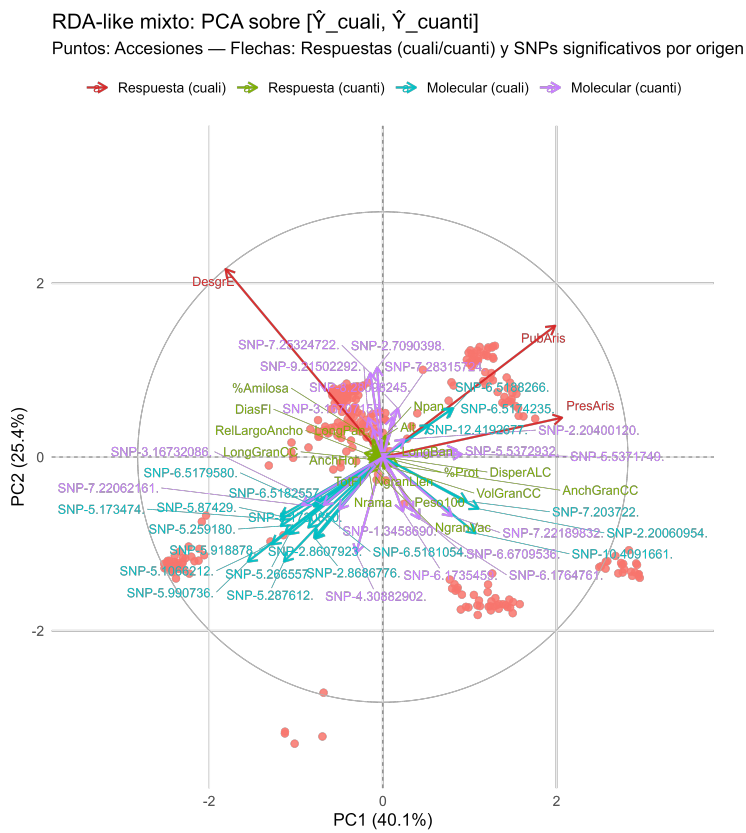


Figura 8.13: Triplot con respuesta mixta. Variables cuantitativas y cualitativas. Accesiones de arroz

Capítulo 9

Discusión

Estudios recientes han mostrado que el uso de múltiples rasgos en forma conjunta mejora sustancialmente la sensibilidad genética y la capacidad de detección de asociaciones respecto a los análisis univariados. Este trabajo se articula con lo planteado por Acharjee et al. (2013) y Boulesteix y Strimmer (2007), quienes muestran que, en contextos ómicos y de alta dimensionalidad, los enfoques multivariados y penalizados representan herramientas fundamentales para abordar miles de predictores, la fuerte correlación entre ellos y la dispersión de la señal en múltiples dimensiones. En particular, Acharjee et al. (2013) comparan distintos métodos de regresión regularizada (ridge, LASSO y *Elastic Net*) y evidencian cómo estas estrategias permiten estabilizar la estimación, mejorar la capacidad predictiva y realizar selección de variables de manera integrada. Por su parte, Boulesteix y Strimmer (2007) destacan la utilidad de técnicas como *PLS* para reducir la dimensionalidad y extraer componentes latentes que concentran la información relevante, facilitando el análisis de datos genómicos complejos.

Esta ventaja de integrar varios fenotipos correlacionados se confirma en aplicaciones concretas; por ejemplo, Hufnagel et al. (2024) aplican un multi-trait GWAS en sorgo para integrar rasgos de arquitectura radicular y eficiencia en fósforo, identificando loci pleiotrópicos que no emergen al analizar cada rasgo por separado. De manera similar, Udosen et al. (2023) utilizan un GWAS multivariado combinando presión arterial sistólica y diastólica. Su enfoque, que modela ambos rasgos correlacionados de manera conjunta, mediante un modelo lineal multivariado con meta-análisis de cohortes, permite detectar variantes genéticas que podrían pasarse por alto en GWAS univariados, aumentando la sensibilidad frente a loci con efectos

compartidos. Revisiones recientes (Shikha et al., 2021; Xiao et al., 2017) respaldan el uso de enfoques multivariantes y multi-ómicos en maíz, enfatizando que la integración de múltiples rasgos y capas de datos permite captar efectos distribuidos en múltiples loci, mejorando la sensibilidad de los análisis. En conjunto, estos trabajos subrayan que el uso de modelos multivariados con Penalización constituye un enfoque particularmente adecuado cuando los efectos biológicos se distribuyen en un gran número de variables y no pueden ser capturados eficazmente mediante métodos univariados tradicionales.

En este contexto, los modelos univariados (lineales y logísticos) ajustados inicialmente cumplieron un rol fundamental como punto de partida, al permitir identificar patrones básicos de asociación marcador molecular–rasgo y establecer un marco de contraste frente a los enfoques multivariados. La comparación con el modelo de regresión logística multivariada conjunto para respuestas binarias *Elastic Net* reveló un patrón coherente con la penalización ℓ_{21} : mientras los enfoques univariados detectaron asociaciones predominantemente específicas de cada rasgo, el modelo multivariado, al favorecer la selección conjunta de predictores, tendió a retener marcadores cuyos efectos se manifestaron de manera consistente en múltiples respuestas. La selección mediante *bootstrap* impuso, además, un criterio de estabilidad que reforzó esta preferencia por señales compartidas entre rasgos. En este marco, aunque ambos enfoques coincidieron en algunos marcadores clave, el modelo multivariado recuperó un conjunto más reducido y estructurado alrededor de efectos transversales que, en los análisis univariados, aparecían de manera fragmentada entre distintos modelos. Desde el punto de vista metodológico, resulta pertinente situar el modelo logístico multivariado conjunto para respuestas binarias, formulado como un Bernoulli multivariado condicionado y estimado mediante máxima verosimilitud penalizada *Elastic Net*, en relación con otros enfoques multivariados desarrollados en la literatura estadística. Si bien el modelo propuesto se vincula conceptualmente con los modelos de tipo Ising estudiados en el contexto de estadística gráfica, el enfoque adoptado en esta tesis difiere de los métodos clásicos de aprendizaje estructural en alta dimensión, como el propuesto por Ravikumar et al. (2010). En particular, mientras

dichos enfoques se basan en aproximaciones de pseudoverosimilitud y regresiones logísticas penalizadas nodo a nodo con el objetivo principal de identificar la estructura del grafo, el presente trabajo estima la verosimilitud conjunta completa del modelo Bernoulli multivariado condicionado, incorporando penalización *Elastic Net* para estabilizar la estimación y priorizar asociaciones marcador molecular–fenotipo relevantes. Esta elección resulta adecuada en escenarios con un número reducido de respuestas binarias, donde es factible evaluar explícitamente la función de partición.

Un comportamiento distinto se observó en el PLS multivariado para rasgos cuantitativos, que a diferencia del modelo logístico multivariado *Elastic Net*, identificó un número mayor de marcadores que sus contrapartes univariadas. Esta discrepancia puede atribuirse al carácter latente del PLS, que concentra la variación común en un espacio de componentes y permite que señales moderadas se acumulen a lo largo de dimensiones compartidas, incrementando su detectabilidad respecto de los modelos univariados. En este sentido, el desempeño del PLS responde a un criterio de representación y selección diferente del inducido por la penalización ℓ_{21} y por el esquema de estabilidad basado en *bootstrap*, lo que explica que ambos enfoques multivariados conduzcan a patrones de selección no necesariamente equivalentes.

Nuestros resultados, en coherencia con la formulación teórica, destacan así que la ventaja de los enfoques multivariados no es uniforme, sino que está modulada por la naturaleza del algoritmo empleado. Mientras la literatura a menudo reporta una ganancia global en ‘poder de detección’, nuestros resultados demuestran que este beneficio adopta expresiones metodológicas claramente diferenciadas, ya sea mediante una exploración sensible y exhaustiva que maximiza el número de señales potenciales (como lo hizo el modelo PLS), o a través de una selección robusta y conservadora que prioriza un conjunto reducido de asociaciones de alta confiabilidad y poder predictivo (como fue el caso del modelo *Elastic Net* multivariado).

Un resultado relevante se observó al comparar los ajustes multivariados *Elastic Net* realizados sobre subconjuntos de rasgos cuantitativos. Cuando el modelo se aplicó a cuatro respuestas, se identificó un número mayor de marcadores significativos que en el ajuste conjunto sobre las 17 respuestas cuantitativas. Este comportamiento

no implica pérdida de poder del modelo, sino que es consistente con la estructura de la penalización ℓ_{21} y con el criterio de estabilidad mediante *bootstrap*. Al incrementar la dimensionalidad de la respuesta, la selección conjunta exige que los efectos de un marcador se manifiesten de manera coherente en un conjunto más amplio de rasgos; en consecuencia, muchas asociaciones que resultan detectables en espacios de menor dimensión —por corresponder a señales parciales o específicas de algunos rasgos— dejan de superar el umbral de estabilidad en el análisis multivariado de mayor dimensión. Así, el modelo con 17 respuestas retiene un conjunto más reducido pero metodológicamente más robusto, privilegiando señales compartidas y patrones de asociación comunes a más de una respuesta. Este comportamiento es consistente con la teoría de los modelos multivariados penalizados con norma ℓ_{21} , en los cuales la selección conjunta induce un mecanismo de regularización que favorece la identificación de predictores relevantes de manera simultánea en múltiples rasgos y se vuelve más restrictivo a medida que aumenta la dimensionalidad del vector de respuestas (Liu et al., 2009; Obozinski et al., 2006). Bajo este esquema, efectos que resultan detectables en subconjuntos pequeños pueden no sostenerse cuando se requiere coherencia en un espacio de mayor dimensión. Asimismo, el criterio de estabilidad basado en *bootstrap* refuerza esta selectividad, descartando asociaciones débiles o no reproducibles (Meinshausen & Bühlmann, 2010; Shah & Samworth, 2013).

Por otro lado, no olvidemos que el objetivo del mapeo asociativo sobre datos existente en bancos de germoplasma, además de constituir un aprovechamiento de esa información, es un paso preliminar de búsqueda de relaciones marcador molecular–fenotipo candidatas para continuar con costosos ensayos de progenie para determinar con precisión la porción del cromosoma que controla un determinado carácter fenotípico.

La pertinencia de estas herramientas quedó particularmente evidenciada al aplicar el marco metodológico del Capítulo 7 a dos conjuntos de datos con estructuras radicalmente diferentes: por un lado, un panel de quinoa con un tamaño muestral reducido (25 accesiones) y un número moderado de predictores (313 microsatélites),

y por otro, un conjunto de arroz de escala genómica amplia (342 accesiones y 186.833 SNP). El hecho de que las mismas estrategias multivariadas penalizadas pudieran aplicarse de manera estable y coherente en ambos escenarios —uno pequeño y relativamente denso, el otro masivo y ultra-dimensional— refuerza la generalidad y robustez del enfoque. A ello se suma que los procedimientos permitieron abordar tanto rasgos cuantitativos como binarios mediante modelos específicos para cada tipo de respuesta, y que la integración posterior mediante el enfoque RDA-*like* permitió representar ambos bloques de información en un mismo espacio multivariado, preservando la coherencia estructural de los patrones de asociación. En el análisis aplicado a los datos de arroz, los marcadores SNP-3.16732086 y SNP-5.5371749 resultaron significativos en el modelo de regresión multivariada *Elastic Net* (luego del procedimiento de *bootstrap*) para la longitud del grano, y en el triplot RDA-*like* para variables cuantitativas el marcador SNP-3.16732086 se ubicó espacialmente próximo a dicha característica, lo cual coincide con lo reportado por McCouch et al. (2016), quienes identificaron esos mismos SNP como el polimorfismo funcional dentro del gen *GS3* y *GW5* asociados a la variación en longitud de grano en arroz y al ancho del grano respectivamente, utilizando un modelo lineal mixto (EMMAX). Este enfoque de síntesis, propuesto en esta tesis, no sólo facilitó la visualización integrada de variables moleculares y fenotípicas de distinta naturaleza, sino que además constituye —hasta donde alcanza nuestra revisión bibliográfica— una contribución metodológica sin antecedentes análogos en la literatura, al ofrecer un marco unificado, explícito y visual para explorar asociaciones marcador molecular-fenotipo en contextos de alta dimensionalidad.

Es importante señalar que, mientras para los rasgos cuantitativos los modelos multivariados pudieron implementarse directamente utilizando la librería *glmnet* de R, para el caso de respuestas binarias múltiples no encontramos una alternativa de análisis disponible en ningún paquete de R que considerara varias respuestas binarias de forma conjunta. Como paso innovador dentro de este marco metodológico, se desarrolló una función personalizada en R para un modelo logístico *Elastic Net* multivariado (ver Capítulo 7, modelo logístico multivariado con penalización *Elastic*

Net). La posibilidad de estimar la matriz de dependencias entre rasgos (Θ) constituye un avance frente a los modelos que asumen independencia entre respuestas, ofreciendo una descripción más realista de la estructura fenotípica.

Tal como se desarrolla en el capítulo 7, la combinación entre modelos multivariados penalizados —capaces de operar en contextos de alta dimensionalidad y de modelar simultáneamente múltiples rasgos— y técnicas multivariadas clásicas de síntesis, como el enfoque *RDA-like* aplicado a las predicciones $\hat{Y}$, permite abordar de manera conjunta aspectos que, por separado, resultarían inabordables. En particular, los métodos penalizados posibilitan la identificación de señales asociativas en escenarios donde $p \gg n$, mientras que la etapa de reducción multivariada facilita la integración y la representación conjunta de individuos, rasgos y predictores en un espacio canónico tipo triplot, de acuerdo con la tradición geométrica de la RDA en estudios de asociación multivariada (Capblancq & Forester, 2021). Esta articulación metodológica refuerza la robustez de los hallazgos y evidencia el valor de enfoques híbridos en genética estadística moderna.

Cabe señalar, sin embargo, que el desempeño de estos modelos multivariados penalizados depende en parte del tamaño muestral y del equilibrio entre las categorías fenotípicas, lo que puede limitar la potencia estadística para rasgos de baja frecuencia. Asimismo, el costo computacional asociado a la estimación conjunta de $\mathbf{B}$ y Θ se incrementa con el número de respuestas, lo que plantea la conveniencia de explorar optimizaciones futuras o estrategias de reducción dimensional previas. Desde una perspectiva aplicada, la identificación de conjuntos de marcadores asociados simultáneamente a múltiples rasgos permite delinear regiones genómicas de interés común, potencialmente vinculadas a mecanismos regulatorios compartidos. Esta evidencia es particularmente relevante para los programas de mejoramiento genético, donde la selección simultánea de atributos agronómicos y morfológicos es una prioridad.

Estos hallazgos abren la posibilidad de aplicar el marco metodológico a conjuntos de datos aún más complejos, como transcriptomas o metabolomas, e incluso a escenarios multi-especie dentro de bancos de germoplasma. Por otra parte, la

implementación desarrollada en R en el marco de esta tesis constituye una base reproducible que permitirá incorporar, en trabajos futuros, módulos de validación cruzada externa y análisis de estabilidad de asociaciones.

Los resultados obtenidos permiten sostener la hipótesis planteada de que la información contenida en los bancos de germoplasma constituye una base sólida para modelar asociaciones entre rasgos fenotípicos mixtos y su caracterización molecular. Esta conclusión refuerza la pertinencia de los enfoques multivariados penalizados en estudios de diversidad genética y mapeo asociativo con datos de alta dimensionalidad.

Capítulo 10

Conclusiones

El presente trabajo de tesis doctoral tuvo como objetivo principal abordar el problema del mapeo asociativo en contextos de alta dimensionalidad, considerando respuestas fenotípicas de naturaleza mixta y estructura multivariada. Desde una perspectiva estadística, la tesis se desarrolló en tres ejes principales, un análisis crítico de técnicas clásicas, el desarrollo de una propuesta metodológica integradora y su implementación en datos reales de interés en genética. En este marco, las principales conclusiones que se derivan de esta tesis, se presentan organizadas en dichos ejes.

1) Análisis de técnicas estadísticas existentes: En el marco del mapeo asociativo, cuyo último objetivo es la identificación de asociaciones confiables que justifiquen la realización de estudios posteriores de progenie, habitualmente costosos y asociados a un elevado requerimiento de recursos, los resultados de esta tesis permiten destacar la conveniencia de enfoques más conservadores. En este sentido, para rasgos cuantitativos, y desde una perspectiva comparativa, los resultados obtenidos permiten concluir que los enfoques multivariados penalizados presentan comportamientos diferenciados y complementarios. Si bien, el modelo PLS multivariado se destacó por su mayor capacidad exploratoria, al identificar un número más amplio de marcadores moleculares potencialmente asociados a los rasgos cuantitativos analizados, el modelo multivariado múltiple con penalización *Elastic Net* evidenció un comportamiento más conservador, priorizando un conjunto reducido de asociaciones, todas ellas coincidentes con las detectadas por el modelo PLS. Además, presentó resultados consistentes con los obtenidos mediante la técnica univariada tradicional y ampliamente difundida de mapeo asociativo con corrección por multiplicidad ba-

sada en el método de Li y Ji, en particular con el modelo *Naive* o ingenuo, que no incorpora la estructura poblacional subyacente. Por estas razones, este modelo se presenta como una herramienta adecuada para priorizar marcadores moleculares con mayor potencial de validación biológica.

Para caracteres cualitativos, el mapeo asociativo se encuentra marcadamente menos desarrollado y, si bien se han propuesto algunos modelos logísticos para variables binarias u ordinales, al momento de la realización de este trabajo no existe un marco multivariado que permita analizar de forma conjunta caracteres cualitativos en función de múltiples marcadores moleculares. Desde esta perspectiva, el análisis de las técnicas clásicas se limitó al empleo de modelos logísticos univariados, lo que puso de manifiesto la existencia de una brecha metodológica susceptible de ser explorada.

2) Propuesta metodológica desarrollada: A partir de las conclusiones derivadas del análisis comparativo de técnicas existentes, y frente a la inaplicabilidad de los modelos clásicos en escenarios de alta dimensionalidad, esta tesis formuló una propuesta metodológica original —con extensiones específicas para variables cuantitativas, cualitativas binarias y mixtas— basada en el modelo multivariado múltiple con penalización *Elastic Net*, denominada enfoque *RDA-like*. El marco propuesto permite recuperar e interpretar la estructura multivariada inducida por los predictores moleculares, integrando en un mismo esquema conceptual la selección de variables, la regularización y herramientas de análisis multivariado orientadas a la interpretación.

Para los rasgos cuantitativos, la implementación del modelo multivariado penalizado *Elastic Net* con distribución gaussiana resultó directa y operativamente viable, lo que refuerza la aplicabilidad práctica del enfoque en contextos reales de alta dimensionalidad. En este caso, la propuesta *RDA-like* utilizó las respuestas estimadas por el modelo como insumo para la recuperación de la estructura multivariada subyacente, dando lugar a una representación tipo triplot que integra individuos, variables fenotípicas y marcadores moleculares. Esta estrategia facilitó una lectura conjunta y coherente de las asociaciones detectadas, constituyéndose como una herramienta complementaria para el análisis e interpretación de los resultados del

mapeo asociativo.

En el caso de los rasgos cualitativos binarios, la contribución de la tesis adquiere un carácter principalmente conceptual y computacional. La extensión natural del modelo multivariado múltiple penalizado conduce a un modelo logístico multivariado conjunto con penalización *Elastic Net*, cuya implementación no se encuentra actualmente disponible en librerías estándar de R. Frente a esta limitación, el aporte central consistió en la formulación explícita del modelo teórico, integrando el enfoque de respuesta Bernoulli multivariada con predictores comunes propuesto por Dai et al. (2013) con un esquema de regularización Elastic Net para la estimación conjunta de los parámetros, cuya implementación computacional fue desarrollada específicamente en R para esta tesis.

Sobre la base de dicha formulación, se desarrollaron funciones específicas en R que permiten la estimación del modelo, la aplicación de penalización conjunta y su incorporación dentro del enfoque *RDA-like*. Este desarrollo habilita la extensión del marco metodológico al análisis conjunto de variables cualitativas binarias en escenarios de alta dimensionalidad, aun en ausencia de implementaciones preexistentes. De este modo, la tesis sienta las bases teóricas y computacionales para el análisis multivariado penalizado de caracteres cualitativos en estudios de mapeo asociativo, y constituye un punto de partida para futuras extensiones metodológicas y desarrollos computacionales.

Por último, se desarrolló el enfoque *RDA-like* para respuestas mixtas que permitió integrar en un único análisis variables cuantitativas y cualitativas binarias, mediante la construcción de una matriz de respuestas estimadas $\hat{\mathbf{Y}}$, conformada por los valores ajustados del modelo gaussiano multivariado *Elastic Net* y por las probabilidades estimadas del modelo logístico multivariado conjunto *Elastic Net*. A partir de esta matriz común, se obtuvo una representación multivariada conjunta que permitió la visualización simultánea de rasgos de naturaleza mixta y su relación con los marcadores moleculares en un mismo plano, mediante un gráfico tipo triplot para respuestas mixtas que integra información correspondiente a rasgos cuantitativos, cualitativos binarios, individuos y marcadores moleculares. Esta extensión refuerza

la flexibilidad del enfoque *RDA-like* para respuestas mixtas en estudios de mapeo asociativo, al permitir un uso integrado y eficiente de la información fenotípica disponible en bancos de germoplasma, contribuyendo al reaprovechamiento de datos de distinta naturaleza dentro de un marco metodológico unificado.

3) Resultados de implementar la propuesta en estudios de caso reales:

La propuesta metodológica desarrollada fue aplicada a dos conjuntos de datos reales de naturaleza contrastante, tanto en términos del número de individuos como de la dimensionalidad del espacio de predictores moleculares. En particular, se analizaron accesiones de quínoa del Noroeste Argentino y del *Rice Diversity Panel* 1 de arroz, los cuales difieren sustancialmente en el tamaño muestral, el número y tipo de marcadores moleculares y la complejidad de la estructura fenotípica. La posibilidad de abordar ambos estudios de caso dentro de un mismo marco metodológico pone de manifiesto la flexibilidad y generalidad del enfoque propuesto en escenarios caracterizados por relaciones $n \ll p$ marcadamente diferentes.

En ambos casos, esta metodología permitió identificar asociaciones marcador molecular–fenotipo consistentes desde el punto de vista estadístico y biológico, así como recuperar la estructura multivariada inducida por los predictores moleculares sobre el conjunto de respuestas. Para los rasgos cuantitativos, la combinación de modelos multivariados penalizados con el enfoque *RDA-like* facilitó una interpretación integrada de los resultados, posibilitando la visualización conjunta de individuos, caracteres fenotípicos y marcadores moleculares seleccionados, y aportando una lectura multivariada de las asociaciones que no resulta accesible mediante enfoques univariados clásicos.

Asimismo, la incorporación de procedimientos de remuestreo *bootstrap* permitió evaluar la estabilidad de las asociaciones identificadas y priorizar un subconjunto reducido de marcadores con mayor consistencia, reforzando el carácter conservador del enfoque en contextos de alta dimensionalidad. En el caso de los rasgos cualitativos binarios, la implementación del modelo logístico multivariado conjunto penalizado y su integración dentro del esquema *RDA-like* posibilitó la exploración conjunta de caracteres cualitativos en función de múltiples marcadores moleculares, constituyendo

una herramienta novedosa para el análisis de este tipo de respuestas.

De manera particular, el estudio de caso correspondiente al *Rice Diversity Panel* 1 implicó desafíos computacionales significativos asociados a la elevada dimensionalidad del conjunto de marcadores moleculares y a la necesidad de evaluar la estabilidad de los coeficientes estimados mediante procedimientos de remuestreo intensivos. Para hacer viable este análisis, fue necesario desarrollar una estrategia computacional específica basada en la partición del conjunto de predictores en bloques o *chunks*, lo que permitió ejecutar el *bootstrap* de forma secuencial y controlada. Esta solución posibilitó completar el análisis en un escenario que, de otro modo, resultaría computacionalmente intratable, garantizando al mismo tiempo la coherencia estadística del procedimiento y la reproducibilidad de los resultados.

Desde el punto de vista empírico, algunos resultados mostraron que la aplicación integrada de modelos multivariados penalizados y del enfoque *RDA-like* en el estudio de caso de quinoa del Noroeste Argentino permitió identificar asociaciones marcador molecular–fenotipo tanto para rasgos cuantitativos como cualitativos —destacándose, en el caso cuantitativo, los marcadores Q026-176, Q084-165 y QGA02-186 con efectos consistentes sobre múltiples caracteres. En el caso cualitativo, mientras los modelos univariados permiten identificar señales específicas para cada rasgo de manera independiente, el modelo multivariado incorpora explícitamente la covariación entre respuestas y, en consecuencia, facilita la identificación de patrones compartidos entre rasgos cualitativos, compatibles con la presencia de posibles efectos pleiotrópicos. En este sentido, el marcador Q0502-204 mostró asociación con los rasgos presencia de estría, forma de la panoja y tipo de la panoja, así como entre el marcador Q112-185 y los rasgos presencia de estría y presencia de rama. La proyección conjunta de las accesiones en el espacio *RDA-like* mixto, mediante el gráfico triplot para respuestas mixtas, evidenció patrones coherentes de asociación entre rasgos cuantitativos y cualitativos binarios, con un primer eje vinculado positivamente a la longitud de los hipocotilo, presencia de axilas y al pigmento de cotiledones y negativamente a rasgos de tamaño vegetativo, y un segundo eje dominado por la presencia de ramas, el tipo de panoja y en menor medida al número de ramas.

En el *Rice Diversity Panel* 1 de arroz, el enfoque propuesto aplicado a 18 rasgos cuantitativos identificó 17 SNP significativos, entre los cuales se destacan SNP-3.16732086 y SNP-5.5371749, ambos asociados a la longitud del grano. Cabe destacar que algunos de los SNP seleccionados por el modelo multivariado penalizado corresponden a regiones previamente asociadas a los genes *GS3* y *GW5*, identificados mediante enfoques GWAS como EMMAX (McCouch et al., 2016), reforzando la evidencia empírica a favor del enfoque propuesto. Para los rasgos cualitativos, se destacaron los SNP SNP-2.20060954 y SNP-5.990736, ambos asociados simultáneamente a los tres caracteres cualitativos analizados. El análisis *RDA-like* mixto en arroz evidenció además la presencia de grupos de accesiones claramente diferenciados, distinguiéndose un conjunto concentrado en el cuadrante II, cercano a rasgos como el porcentaje de amilosa, los días a floración y caracteres de grano, en concordancia con variables cualitativas como el desgrane espontáneo y SNP asociados.

Los resultados obtenidos en ambos conjuntos de datos reales ponen de manifiesto que la propuesta metodológica es aplicable a escenarios de diversa escala y complejidad, y que permite extraer información relevante tanto a nivel de asociaciones individuales como de estructura multivariada global. Estos hallazgos respaldan la utilidad práctica del enfoque propuesto como una alternativa flexible y operativamente viable para el análisis de mapeo asociativo en presencia de respuestas fenotípicas mixtas y alta dimensionalidad.

10.1 Limitaciones y futuras líneas de investigación

Si bien la propuesta metodológica demostró ser robusta y generalizable, es importante reconocer algunas limitaciones. En primer lugar, la naturaleza exploratoria de los análisis, aunque potenciada por el uso de técnicas multivariadas, no permite establecer relaciones causales. Estudios futuros que integren información complementaria a nivel genómico podrían contribuir a la validación de marcadores candidatos identificados de manera consistente en múltiples modelos y asociados tanto a rasgos cuantitativos como cualitativos, como Q078-185 o Q106-300.

En segundo lugar, el Modelo logístico *Elastic Net* multivariado para variables cualitativas se implementó aquí únicamente para respuestas binarias, quedando pendiente el desarrollo de su extensión al caso multinomial, lo cual constituye una línea de investigación futura directa y necesaria. Un aspecto adicional a considerar es la complejidad computacional asociada al análisis de datos de muy alta dimensionalidad, especialmente en el caso del panel de arroz, que incluyó 186.833 marcadores. El ajuste de modelos penalizados multivariados y la obtención de predicciones para procedimientos derivados (como los análisis *RDA-like*) demandaron tiempos de cómputo prolongados, así como un uso intensivo de memoria. Si bien estas dificultades se mitigaron mediante estrategias específicas —como el procesamiento por bloques, la paralelización y la optimización del entorno de trabajo en R—, la escalabilidad del método sigue siendo un desafío cuando se proyecta su aplicación a paneles genómicos aún más densos o a poblaciones más amplias. El desarrollo de algoritmos más eficientes o implementaciones en entornos de cómputo de alto rendimiento constituye, por tanto, una línea de mejora futura.

Finalmente, el poder de detección estuvo condicionado por el tamaño muestral y la densidad de marcadores disponibles; la incorporación de poblaciones más numerosas y paneles de mayor densidad (por ejemplo, derivados de secuenciación completa del genoma) permitiría alcanzar una caracterización más fina y estable de la arquitectura genética subyacente.

Apéndice A

Código en R

Índice de códigos

A.1. Programa multivariado logístico <i>Elastic Net</i>	195
A.2. RDA-like multivariado <i>Elastic Net</i>	241
A.3. RDA-like Mixto	284

```
1 # =====
2 # MULTILOGIT CONJUNTO (multivariado) con Elastic Net
3 #   - Penalización en B: "l1" (independiente por coeficiente) o "l21" (
      compartida por predictor)
4 #   - Penalización en theta: L1 (sparse en dependencias entre respuestas)
5 #   - Optimización: FISTA con backtracking + aceleración de Nesterov
6 #   - Selección de  $\lambda$ : CV k-fold (reglas  $\lambda_{\min}$  y  $\lambda_{1SE}$ )
7 #   - Predicción: probabilidades marginales, modo conjunto y probabilidades
      conjuntas
8 #   - Exportación: coeficientes e interacciones a Excel
9 #   - Complejidad  $O(2^K)$  por enumerar estados (K relativamente pequeño)
10 # =====
11
12 rm(list = ls())
13
14 ## -----
15 ## Reproducibilidad y configuración de paralelo
16 ## -----
17 library(future.apply)
18 library(future)
19
20 seed_val <- 123
21 set.seed(seed_val)
22 RNGkind("L'Ecuyer-CMRG")
23
24 n_workers <- max(1, parallel::detectCores() - 1)
25 plan(multisession, workers = n_workers)
26 on.exit(plan(sequential), add = TRUE)
27
28 options(future.scheduling = 0)
29 options(future.globals.maxSize = 1500 * 1024^2)
```

```
30
31 if (requireNamespace("RhpcBLASctl", quietly = TRUE)) {
32   RhpcBLASctl::blas_set_num_threads(1)
33   RhpcBLASctl::omp_set_num_threads(1)
34 } else {
35   Sys.setenv(OMP_NUM_THREADS = "1")
36   Sys.setenv(OPENBLAS_NUM_THREADS = "1")
37   Sys.setenv(MKL_NUM_THREADS = "1")
38 }
39
40 options(future.rng.onMisuse = "error")
41
42 ## Número de réplicas bootstrap (múltiplo de n_workers)
43 B_reps <- n_workers * 30
44
45 ## -----
46 ## Rutas de entrada/salida (a completar por el usuario)
47 ## -----
48 path_excel <- "path/to/Datoscompletos.xlsx" # archivo con X e Y
49
50 SRC <- "path/to/src_objects" # carpeta con objetos RDS opcionales
51 DEST <- "path/to/results" # carpeta de salida para resultados
52
53 if (!dir.exists(SRC)) {
54   stop("Directorio SRC no existe: ", SRC)
55 }
56 if (!dir.exists(DEST)) {
57   dir.create(DEST, recursive = TRUE)
58 }
59
60 if (!file.exists(path_excel)) {
```

```
61  stop("Archivo Excel no encontrado: ", path_excel)
62  }
63
64  ## -----
65  ##  Helpers de I/O (usan SRC/DEST)
66  ## -----
67  safe_stamp <- function(path) {
68    if (file.exists(path)) {
69      sub("\\.([A-Za-z0-9]+)$",
70        paste0("_", format(Sys.time(), "%Y%m%d-%H%M%S"), ".\\1"),
71        path)
72    } else path
73  }
74
75  read_rds_from <- function(name, required = TRUE) {
76    f <- file.path(SRC, name)
77    if (!file.exists(f)) {
78      if (required) stop("No existe en SRC: ", f)
79      return(NULL)
80    }
81    readRDS(f)
82  }
83
84  write_rds_to <- function(obj, name, stamp = FALSE) {
85    f <- file.path(DEST, name)
86    if (stamp) f <- safe_stamp(f)
87    dir.create(dirname(f), recursive = TRUE, showWarnings = FALSE)
88    saveRDS(obj, f)
89    invisible(f)
90  }
91
```



```
92 write_xlsx_to <- function(sheets, name, stamp = TRUE) {
93   f <- file.path(DEST, name)
94   if (stamp) f <- safe_stamp(f)
95   dir.create(dirname(f), recursive = TRUE, showWarnings = FALSE)
96   writexl::write_xlsx(sheets, f)
97   invisible(f)
98 }
99
100 ggsave_to <- function(plot, name, ...) {
101   f <- file.path(DEST, name)
102   dir.create(dirname(f), recursive = TRUE, showWarnings = FALSE)
103   ggplot2::ggsave(filename = f, plot = plot, ...)
104   invisible(f)
105 }
106
107 png_to <- function(name, ...) { png (file.path(DEST, name), ...) }
108 pdf_to <- function(name, ...) { pdf (file.path(DEST, name), ...) }
109 tiff_to <- function(name, ...) { tiff(file.path(DEST, name), ...) }
110 jpeg_to <- function(name, ...) { jpeg(file.path(DEST, name), ...) }
111
112 ## -----
113 ##  Parámetros generales
114 ## -----
115
116 x_cols      <- 36:348      # columnas de X en el Excel
117 alpha_val   <- 0.40        # parámetro de Elastic Net
118 nlambdas    <- 60
119 lambda_min_ratio <- 0.01
120 nfolds      <- 5
121 penalty_mode <- "l21"      # "l21" (agrupada) o "l1" (coeficiente a
                             # coeficiente)
```

```
122
123 excel_path_main <- sprintf("Resultados_Logit_MultilogitConjunto_EN_%.xlsx",
    penalty_mode)
124
125 ## -----
126 ## Paquetes requeridos
127 ## -----
128 req_pkgs <- c("readxl", "writexl")
129 for (p in req_pkgs) {
130   if (!requireNamespace(p, quietly = TRUE)) {
131     stop(sprintf("Paquete requerido: '%s'.", p))
132   }
133 }
134 suppressPackageStartupMessages({
135   library(dplyr)
136   library(tidyr)
137 })
138
139 ## -----
140 ## Lectura de datos y construcción de X e Y_bin
141 ## -----
142 library(readxl)
143 dat <- readxl::read_excel(path_excel)
144
145 X <- as.matrix(dat[, x_cols, drop = FALSE])
146 storage.mode(X) <- "double"
147
148 ## Transformación robusta de variables dicotómicas a 0/1
149 to_bin_if_dichot <- function(v) {
150   out_fail <- list(bin = NULL, ok = FALSE, msg = "no dicotómica")
151   if (is.logical(v)) {
```

```
152   return(list(bin = as.integer(v), ok = TRUE, msg = "logical"))
153 }
154 if (is.factor(v)) v <- as.character(v)
155 if (is.numeric(v)) {
156   uv <- sort(unique(v[is.finite(v)]))
157   if (length(uv) == 2) {
158     bin <- ifelse(v == max(uv), 1L,
159                   ifelse(v == min(uv), 0L, NA_integer_))
160     return(list(bin = bin, ok = TRUE, msg = "numeric 2-level"))
161   }
162   if (all(v %in% c(0,1))) {
163     return(list(bin = as.integer(v), ok = TRUE, msg = "numeric 0/1"))
164   }
165   return(out_fail)
166 }
167 vv <- tolower(trimws(as.character(v)))
168 pos <- c("1","si","sí","true","t","yes","y","ap","presente","con","positivo")
169 neg <- c("0","no","false","f","ausente","sin","negativo")
170 map <- rep(NA_integer_, length(vv))
171 map[vv %in% pos] <- 1L
172 map[vv %in% neg] <- 0L
173 ambi <- which(is.na(map) & grepl("si|sí|no", vv))
174 if (length(ambi)) {
175   map[ambi] <- ifelse(grepl("si|sí", vv[ambi]), 1L,
176                       ifelse(grepl("\\bno\\b", vv[ambi]), 0L, NA_integer_))
177 }
178 uv <- unique(map[!is.na(map)])
179 if (length(uv) <= 2 && all(uv %in% c(0L,1L)) && length(uv) >= 1) {
180   return(list(bin = map, ok = TRUE, msg = "character 2-level"))
181 }
```

```
182   out_fail
183 }
184
185 ## Identificación automática de columnas dicotómicas
186 all_cols <- seq_along(dat)
187 cand_y   <- setdiff(all_cols, x_cols)
188
189 y_keep_names <- character(0)
190 Y_bin_list   <- list()
191
192 for (nm in names(dat)[cand_y]) {
193   res <- to_bin_if_dichot(dat[[nm]])
194   if (res$ok) {
195     Y_bin_list[[nm]] <- as.integer(res$bin)
196     y_keep_names     <- c(y_keep_names, nm)
197   }
198 }
199
200 if (length(Y_bin_list) == 0) stop("No se detectaron variables dicotómicas.")
201
202 Y_bin <- do.call(cbind, Y_bin_list)
203 colnames(Y_bin) <- y_keep_names
204 storage.mode(Y_bin) <- "double"
205
206 stopifnot(nrow(X) == nrow(Y_bin))
207 if (anyNA(Y_bin)) {
208   stop("Y_bin contiene valores faltantes.")
209 }
210
211 ## Conjunto oficial de respuestas (6 rasgos dicotómicos)
212 y_oficial <- c("AP", "Estrias", "RAMA", "FP", "PigCot", "TP")
```

```
213
214 y_present <- intersect(y_oficial, colnames(Y_bin))
215 if (length(y_present) < length(y_oficial)) {
216   faltan <- setdiff(y_oficial, y_present)
217   stop("Faltan respuestas en el set oficial: ",
218       paste(faltan, collapse = ", "))
219 }
220
221 niv <- sapply(y_present, function(v) {
222   uv <- sort(unique(na.omit(as.integer(Y_bin[, v]))))
223   length(uv)
224 })
225 if (any(niv != 2)) {
226   prob <- y_present[niv != 2]
227   stop("Las siguientes respuestas no son dicotómicas: ",
228       paste(prob, collapse = ", "))
229 }
230
231 Y_bin <- as.matrix(Y_bin[, y_oficial, drop = FALSE])
232
233 K <- ncol(Y_bin)
234 if (K > 8) stop("K demasiado grande para enumerar estados (2^K).")
235
236 stopifnot(nrow(X) == nrow(Y_bin))
237 stopifnot(all(colSums(is.na(X)) == 0))
238 stopifnot(all(colSums(is.na(Y_bin)) == 0))
239
240 ## -----
241 ## Utilidades para el modelo multilogit conjunto
242 ## -----
243
```

```
244 enumerate_states <- function(K) {
245   as.matrix(do.call(expand.grid, rep(list(0:1), K)))[, K:1, drop = FALSE]
246 }
247
248 mvb_moments <- function(eta, Theta, STATES = NULL, tol = 1e-12) {
249   K <- length(eta)
250   if (is.null(STATES)) STATES <- enumerate_states(K)
251   quad <- rowSums((STATES %*% Theta) * STATES) / 2
252   lin <- as.numeric(STATES %*% eta)
253   logw <- lin + quad
254   m <- max(logw)
255   w <- exp(logw - m)
256   Z <- sum(w)
257   if (Z <= tol || !is.finite(Z)) {
258     p <- rep(1 / nrow(STATES), nrow(STATES))
259     logZ <- log(nrow(STATES))
260   } else {
261     p <- w / Z
262     logZ <- m + log(Z)
263   }
264   if (any(!is.finite(p)) || abs(sum(p) - 1) > tol) {
265     p <- rep(1 / nrow(STATES), nrow(STATES))
266     logZ <- log(nrow(STATES))
267   }
268   Ey <- as.numeric(colSums(STATES * p))
269   Eyy <- t(STATES) %*% (STATES * p)
270   Eyy <- (Eyy + t(Eyy)) / 2
271   diag(Eyy) <- Ey
272   list(logZ = logZ, Ey = Ey, Eyy = Eyy)
273 }
274
```

```

275 obj_fun_joint <- function(Xs, Y, B, b0, Theta,
276                           alpha, lambda,
277                           penalty = c("l1", "l2l"),
278                           lambda_theta = 0.0, alpha_theta = 1.0,
279                           STATES = NULL) {
280   penalty <- match.arg(penalty)
281   Xs <- as.matrix(Xs); Y <- as.matrix(Y)
282   n <- nrow(Xs); p <- ncol(Xs); K <- ncol(Y)
283   if (!all(dim(B) == c(p, K))) B <- matrix(B, p, K)
284   b0 <- as.numeric(b0); if (length(b0) != K) b0 <- rep(b0[1], K)
285   Theta <- as.matrix(Theta); Theta <- (Theta + t(Theta)) / 2; diag(Theta) <-
    0
286
287   nll <- 0
288   for (i in 1:n) {
289     eta_i <- as.numeric(Xs[i,] %*% B + b0)
290     mm <- mvb_moments(eta_i, Theta, STATES)
291     nll <- nll +
292       ( - sum(Y[i,] * eta_i) -
293         sum(Theta[upper.tri(Theta)] *
294           (Y[i,] %o% Y[i,])[upper.tri(Y[i,] %o% Y[i,])]) +
295         mm$logZ )
296   }
297   nll <- nll / n
298
299   ridge_B <- (1 - alpha) * lambda * 0.5 * sum(B * B)
300   penB <- if (alpha > 0 && lambda > 0) {
301     if (penalty == "l1") {
302       alpha * lambda * sum(abs(B))
303     } else {
304       alpha * lambda * sum(sqrt(rowSums(B * B)))

```

```

305   }
306 } else 0
307
308 penTheta_ridge <- (1 - alpha_theta) * lambda_theta * 0.5 * sum(Theta *
    Theta)
309 penTheta_l1    <- if (alpha_theta > 0 && lambda_theta > 0) {
310   alpha_theta * lambda_theta * sum(abs(Theta[upper.tri(Theta)]))
311 } else 0
312
313 nll + ridge_B + penB + penTheta_ridge + penTheta_l1
314 }
315
316 smooth_obj_joint <- function(Xs, Y, B, b0, Theta,
317                             alpha, lambda, lambda_theta, alpha_theta,
318                             STATES = NULL) {
319   Xs <- as.matrix(Xs); Y <- as.matrix(Y)
320   n <- nrow(Xs); K <- ncol(Y)
321   if (is.null(STATES)) STATES <- enumerate_states(K)
322   Theta <- (Theta + t(Theta)) / 2; diag(Theta) <- 0
323
324   nll <- 0
325   for (i in 1:n) {
326     eta_i <- as.numeric(Xs[i,] %*% B + b0)
327     mm <- mvb_moments(eta_i, Theta, STATES)
328     nll <- nll +
329       ( - sum(Y[i,] * eta_i) -
330         sum(Theta[upper.tri(Theta)] *
331           (Y[i,] %o% Y[i,])[upper.tri(Y[i,] %o% Y[i,])]) +
332         mm$logZ )
333   }
334   nll <- nll / n

```



```

335 ridge_B <- (1 - alpha)      * lambda      * 0.5 * sum(B * B)
336 ridge_T <- (1 - alpha_theta) * lambda_theta * 0.5 * sum(Theta * Theta)
337 nll + ridge_B + ridge_T
338 }
339
340 ## -----
341 ## FISTA conjunto (B,b0,theta)
342 ## -----
343 fista_multilogit_joint <- function(X, Y,
344                                   alpha = 0.5, lambda = 0.01,
345                                   penalty = c("l1","l2l"),
346                                   lambda_theta = 0.01, alpha_theta = 1.0,
347                                   maxit = 800, tol = 1e-6,
348                                   standardize = TRUE, backtracking = TRUE,
349                                   t_init = 1e-3, verbose = FALSE,
350                                   B_init = NULL, b0_init = NULL, Theta_init
351                                   = NULL) {
352   penalty <- match.arg(penalty)
353   X <- as.matrix(X); Y <- as.matrix(Y)
354   storage.mode(X) <- "double"; storage.mode(Y) <- "double"
355   n <- nrow(X); p <- ncol(X); K <- ncol(Y)
356
357   x_means <- rep(0, p); x_sds <- rep(1, p); Xs <- X
358   if (standardize) {
359     x_means <- colMeans(X)
360     Xc <- sweep(X, 2, x_means, "-")
361     x_sds <- sqrt(colSums(Xc^2) / n); x_sds[x_sds == 0] <- 1
362     Xs <- sweep(Xc, 2, x_sds, "/")
363   }
364   B <- if (!is.null(B_init)) as.matrix(B_init) else matrix(0, p, K)

```

```

365 eps <- 1e-6; Ybar <- colMeans(Y)
366 b0_default <- log(pmin(pmax(Ybar, eps), 1 - eps) /
367                   pmax(1 - pmin(pmax(Ybar, eps), 1 - eps), eps))
368 b0 <- if (!is.null(b0_init)) as.numeric(b0_init) else b0_default
369
370 Theta <- if (!is.null(Theta_init)) {
371   M <- as.matrix(Theta_init); M <- (M + t(M)) / 2; diag(M) <- 0; M
372 } else matrix(0, K, K)
373
374 ZB <- B; Zb0 <- b0; ZTh <- Theta
375 t <- t_init; theta_acc <- 1
376 hist <- numeric()
377 STATES <- enumerate_states(K)
378 if (verbose) start_time <- Sys.time()
379
380 for (iter in 1:maxit) {
381   Ey_mat <- matrix(0, n, K)
382   Eyy_acc <- matrix(0, K, K)
383   for (i in 1:n) {
384     eta_i <- as.numeric(Xs[i,] %*% ZB + Zb0)
385     mm <- mvb_moments(eta_i, ZTh, STATES)
386     Ey_mat[i,] <- mm$Ey
387     Eyy_acc <- Eyy_acc + (mm$Eyy - (Y[i,] %o% Y[i,]))
388   }
389   Eyy_acc <- Eyy_acc / n
390
391   grad_b0 <- colMeans(Ey_mat - Y)
392   grad_B <- t(Xs) %*% (Ey_mat - Y) / n + (1 - alpha) * lambda * ZB
393   grad_Th <- Eyy_acc + (1 - alpha_theta) * lambda_theta * ZTh
394   grad_Th <- (grad_Th + t(grad_Th)) / 2
395   diag(grad_Th) <- 0

```

```

396
397 t_bt <- t
398 repeat {
399     B_new <- ZB - t_bt * grad_B
400     b0_new <- Zb0 - t_bt * grad_b0
401     Th_new <- ZTh - t_bt * grad_Th
402
403     if (alpha > 0 && lambda > 0) {
404         if (penalty == "l1") {
405             B_new <- sign(B_new) * pmax(abs(B_new) - t_bt * alpha * lambda, 0)
406         } else {
407             rn <- sqrt(rowSums(B_new * B_new))
408             s <- pmax(0, 1 - (t_bt * alpha * lambda) / pmax(rn, 1e-15))
409             B_new <- B_new * s
410         }
411     }
412
413     if (alpha_theta > 0 && lambda_theta > 0) {
414         U <- Th_new
415         ut <- upper.tri(U)
416         v <- U[ut]
417         v <- sign(v) * pmax(abs(v) - t_bt * alpha_theta * lambda_theta, 0)
418         U_new <- matrix(0, K, K)
419         U_new[ut] <- v
420         U_new <- U_new + t(U_new)
421         diag(U_new) <- 0
422         Th_new <- U_new
423     } else {
424         Th_new <- (Th_new + t(Th_new)) / 2; diag(Th_new) <- 0
425     }
426

```

```

427     lhs <- smooth_obj_joint(Xs, Y, B_new, b0_new, Th_new,
428                             alpha, lambda, lambda_theta, alpha_theta,
STATES)
429     rhs <- smooth_obj_joint(Xs, Y, ZB, Zb0, ZTh,
430                             alpha, lambda, lambda_theta, alpha_theta,
STATES) +
431     sum(grad_B * (B_new - ZB )) +
432     sum(grad_b0 * (b0_new - Zb0)) +
433     sum(grad_Th * (Th_new - ZTh)) +
434     (1 / (2 * t_bt)) *
435     ( sum((B_new - ZB )^2) +
436       sum((b0_new - Zb0)^2) +
437       sum((Th_new - ZTh)^2) )
438     if (t_bt < 1e-8) break
439     if (!backtracking || lhs <= rhs) break
440     t_bt <- t_bt * 0.7
441     if (t_bt < 1e-12) break
442   }
443
444   theta_new <- (1 + sqrt(1 + 4 * theta_acc^2)) / 2
445   gamma <- (theta_acc - 1) / theta_new
446
447   B_prev <- B; b0_prev <- b0; Th_prev <- Theta
448   B <- B_new; b0 <- b0_new; Theta <- Th_new
449
450   ZB <- B      + gamma * (B      - B_prev)
451   Zb0 <- b0    + gamma * (b0    - b0_prev)
452   ZTh <- Theta + gamma * (Theta - Th_prev)
453   theta_acc <- theta_new
454   t <- t_bt
455

```

```
456     fval <- obj_fun_joint(Xs, Y, B, b0, Theta,
457                           alpha, lambda, penalty,
458                           lambda_theta, alpha_theta, STATES)
459     if (length(hist) >= 1 && fval > tail(hist, 1)) {
460       theta_acc <- 1
461       ZB <- B; Zb0 <- b0; ZTh <- Theta
462     }
463     hist <- c(hist, fval)
464
465     if (length(hist) > 50) {
466       rel <- abs(hist[length(hist)] - hist[length(hist)-1]) /
467             max(1, abs(hist[length(hist)-1]))
468       if (rel < tol) break
469     }
470
471     if (verbose && (iter %% 25 == 0)) {
472       elapsed <- as.numeric(difftime(Sys.time(), start_time, units = "secs"))
473       cat(sprintf("iter=%4d, obj=%.6f, step=%.2e, time=%.1fs\n",
474                   iter, fval, t, elapsed))
475     }
476   }
477
478   if (standardize) {
479     B_orig <- sweep(B, 1, x_sds, "/")
480     shift <- as.numeric(x_means / x_sds)
481     b0_orig <- b0 - as.numeric(t(shift) %*% B)
482   } else {
483     B_orig <- B; b0_orig <- b0
484   }
485
486   list(
```

```
487     B = B_orig, b0 = b0_orig, Theta = Theta,
488     B_std = B, b0_std = b0,
489     history = hist, stepsize = t, iters = length(hist),
490     x_means = x_means, x_sds = x_sds
491 )
492 }
493
494 ## -----
495 ## Validación cruzada ( $\lambda$ -min y  $\lambda$ -1SE)
496 ## -----
497 cv_multilogit_joint <- function(X, Y,
498                               alpha = 0.5,
499                               nlambda = 50,
500                               lambda_min_ratio = 0.05,
501                               nfolds = 5,
502                               seed = 123,
503                               penalty = c("l1", "l21"),
504                               lambda_theta = 0.01, alpha_theta = 1.0,
505                               standardize = TRUE,
506                               verbose = FALSE) {
507   penalty <- match.arg(penalty)
508   set.seed(seed)
509   n <- nrow(X)
510   folds <- sample(rep(1:nfolds, length.out = n))
511
512   if (standardize) {
513     x_means_cv <- colMeans(X)
514     Xc_cv <- sweep(X, 2, x_means_cv, "-")
515     x_sds_cv <- sqrt(colSums(Xc_cv^2) / n); x_sds_cv[x_sds_cv == 0] <- 1
516     Xs_cv <- sweep(Xc_cv, 2, x_sds_cv, "/")
517   } else {
```

```
518   Xs_cv <- X
519 }
520
521 eps <- 1e-6
522 Ybar <- colMeans(Y)
523 pbar <- pmin(pmax(Ybar, eps), 1 - eps)
524 Pmat <- matrix(rep(pbar, each = nrow(X)), nrow = nrow(X))
525 G <- t(Xs_cv) %*% (Pmat - Y) / nrow(Xs_cv)
526
527 if (penalty == "l1") {
528   lam_max <- max(abs(G))
529 } else {
530   lam_max <- max(sqrt(rowSums(G^2)))
531 }
532 lam_min <- lam_max * lambda_min_ratio
533 lambda_seq <- exp(seq(log(lam_max), log(lam_min), length.out = nlambdas))
534
535 cvloss <- matrix(NA_real_, nlambdas, nfolds)
536 STATES_cv <- enumerate_states(ncol(Y))
537
538 for (f in 1:nfolds) {
539   tr <- which(folds != f); va <- which(folds == f)
540   Xtr <- X[tr, , drop = FALSE]; Ytr <- Y[tr, , drop = FALSE]
541   Xva <- X[va, , drop = FALSE]; Yva <- Y[va, , drop = FALSE]
542
543   B0 <- NULL; b00 <- NULL; Th0 <- NULL
544
545   for (li in seq_along(lambda_seq)) {
546     fit <- fista_multilogit_joint(
547       Xtr, Ytr,
548       alpha = alpha, lambda = lambda_seq[li],
```

```
549     penalty = penalty,
550     lambda_theta = lambda_theta, alpha_theta = alpha_theta,
551     standardize = standardize, verbose = FALSE,
552     B_init = B0, b0_init = b00, Theta_init = Th0
553   )
554   B0 <- fit$B_std
555   b00 <- fit$b0_std
556   Th0 <- fit$Theta
557
558   Xs_va <- if (standardize) {
559     sweep(sweep(Xva, 2, fit$x_means, "-"), 2, fit$x_sds, "/")
560   } else Xva
561
562   loss_va <- obj_fun_joint(
563     Xs_va, Yva, fit$B_std, fit$b0_std, fit$Theta,
564     alpha = 0, lambda = 0, penalty = penalty,
565     lambda_theta = 0, alpha_theta = 0, STATES = STATES_cv
566   )
567   cvloss[li, f] <- loss_va
568 }
569 if (verbose) cat("fold", f, "terminado\n")
570 }
571
572 mean_cvloss <- rowMeans(cvloss)
573 sd_cvloss <- apply(cvloss, 1, sd)
574 i_min <- which.min(mean_cvloss)
575
576 thresh <- mean_cvloss[i_min] + sd_cvloss[i_min]
577 eligible <- which(mean_cvloss <= thresh)
578 i_1se <- min(eligible)
579
```



```
580 list(  
581   lambda_seq = lambda_seq,  
582   cvloss      = cvloss,  
583   cv_mean     = mean_cvloss,  
584   cv_sd       = sd_cvloss,  
585   i_min       = i_min,  
586   lambda_min  = lambda_seq[i_min],  
587   i_1se      = i_1se,  
588   lambda_1se  = lambda_seq[i_1se],  
589   alpha       = alpha,  
590   penalty     = penalty,  
591   lambda_theta = lambda_theta,  
592   alpha_theta  = alpha_theta,  
593   standardize  = standardize,  
594   nfolds      = nfolds,  
595   seed        = seed  
596 )  
597 }  
598  
599 refit_multilogit_joint <- function(X, Y, cv, which = c("min", "1se"), verbose  
   = FALSE) {  
600   which <- match.arg(which)  
601   lambda_star <- if (which == "min") cv$lambda_min else cv$lambda_1se  
602  
603   fista_multilogit_joint(  
604     X, Y,  
605     alpha = cv$alpha, lambda = lambda_star,  
606     penalty = cv$penalty,  
607     lambda_theta = cv$lambda_theta, alpha_theta = cv$alpha_theta,  
608     standardize = cv$standardize, verbose = verbose  
609   )
```

```

610 }
611
612 ## -----
613 ## Predicción
614 ## -----
615 predict_multilogit_joint <- function(fit, X_new,
616                                     type = c("marginal", "joint_mode", "joint_
        probs"),
617                                     standardize = TRUE) {
618   type <- match.arg(type)
619   X_new <- as.matrix(X_new)
620   n <- nrow(X_new); K <- length(fit$b0_std)
621
622   Xs <- if (standardize) {
623     sweep(sweep(X_new, 2, fit$x_means, "-"), 2, fit$x_sds, "/")
624   } else X_new
625
626   STATES <- enumerate_states(K)
627
628   if (type == "marginal") {
629     Pm <- matrix(NA_real_, n, K)
630     for (i in 1:n) {
631       eta_i <- as.numeric(Xs[i,] %*% fit$B_std + fit$b0_std)
632       mm <- mvb_moments(eta_i, fit$Theta, STATES)
633       Pm[i,] <- mm$Ey
634     }
635     colnames(Pm) <- paste0("P(Y", seq_len(K), "=1)")
636     return(Pm)
637   }
638
639   if (type == "joint_mode") {

```

```

640  mode_idx <- integer(n)
641  mode_vec <- matrix(0, n, K)
642  for (i in 1:n) {
643    eta_i <- as.numeric(Xs[i,] %*% fit$B_std + fit$b0_std)
644    quad <- rowSums((STATES %*% fit$Theta) * STATES) / 2
645    lin <- as.numeric(STATES %*% eta_i)
646    logw <- lin + quad
647    mode_idx[i] <- which.max(logw)
648    mode_vec[i,] <- STATES[mode_idx[i], ]
649  }
650  colnames(mode_vec) <- paste0("Y", seq_len(K))
651  return(list(mode_pattern = mode_vec, mode_index = mode_idx))
652 }
653
654 if (type == "joint_probs") {
655   probs_list <- vector("list", n)
656   for (i in 1:n) {
657     eta_i <- as.numeric(Xs[i,] %*% fit$B_std + fit$b0_std)
658     quad <- rowSums((STATES %*% fit$Theta) * STATES) / 2
659     lin <- as.numeric(STATES %*% eta_i)
660     logw <- lin + quad
661     m <- max(logw)
662     w <- exp(logw - m)
663     s <- sum(w)
664     if (!is.finite(s) || s <= 0) {
665       p <- rep(1 / nrow(STATES), nrow(STATES))
666     } else {
667       p <- w / s
668     }
669     probs_list[[i]] <- p
670   }

```

```
671   attr(probs_list, "states") <- STATES
672   return(probs_list)
673 }
674 }
675
676 ## -----
677 ## Curva de CV
678 ## -----
679 plot_cv_multilogit_joint <- function(cv, log_x = TRUE,
680                                     main = "CV multilogit conjunto",
681                                     xlab = expression(lambda),
682                                     ylab = "Pérdida conjunta (prom.)") {
683   lam <- cv$lambda_seq
684   mu <- cv$cv_mean
685   sd <- cv$cv_sd
686   x <- if (log_x) log(lam) else lam
687   plot(x, mu, type = "b", pch = 16, xlab = xlab, ylab = ylab, main = main)
688   lines(x, mu + sd, lty = 2)
689   lines(x, mu - sd, lty = 2)
690   abline(v = x[cv$i_min], lty = 3)
691   abline(v = x[cv$i_1se], lty = 3)
692   legend("topright",
693         legend = c(
694           bquote(lambda[min] == .(format(cv$lambda_min, digits = 3))),
695           bquote(lambda[1*SE] == .(format(cv$lambda_1se, digits = 3))),
696         ),
697         bty = "n")
698 }
699
700 save_cv_curve <- function(cv, base_name = "CV_multilogit_conjunto",
701                           log_x = TRUE, width_in = 7, height_in = 4, dpi =
```

```

600) {
702 stamp   <- format(Sys.time(), "%Y%m%d-%H%M%S")
703 out_png  <- file.path(DEST, sprintf("%s_%s.png", base_name, stamp))
704 out_pdf  <- file.path(DEST, sprintf("%s_%s.pdf", base_name, stamp))
705 op <- par(no.readonly = TRUE); on.exit(par(op), add = TRUE)
706 png(out_png, width = width_in, height = height_in, units = "in", res = dpi)
707 par(mar = c(5,4,2,1))
708 plot_cv_multilogit_joint(cv, log_x = log_x, xlab = expression(log(lambda)))
709 dev.off()
710 pdf(out_pdf, width = width_in, height = height_in)
711 par(mar = c(5,4,2,1))
712 plot_cv_multilogit_joint(cv, log_x = log_x, xlab = expression(log(lambda)))
713 dev.off()
714 invisible(sub("\\.png$", "", out_png))
715 }
716
717 library(ggplot2)
718
719 plot_cv_pretty <- function(cv,
720                             opt = c("min", "1se"),
721                             show_bands = TRUE,
722                             title = "Curva de error de validación cruzada",
723                             subtitle_left = NULL) {
724   opt <- match.arg(opt)
725   df <- data.frame(
726     x = log(cv$lambda_seq),
727     mu = cv$cv_mean,
728     sd = cv$cv_sd
729   )
730   x_opt_min <- log(cv$lambda_min)
731   x_opt_1se <- log(cv$lambda_1se)

```

```
732 x_opt      <- if (opt == "min") x_opt_min else x_opt_1se
733
734 if (is.null(subtitle_left)) {
735   subtitle_left_final <- paste0("\u03B4 = ",
736                                 formatC(cv$alpha, format = "f", digits = 2)
737   )
738 } else {
739   subtitle_left_final <- subtitle_left
740   subtitle_left_final <- gsub("\u03B1", "\u03B4", subtitle_left_final,
741   fixed = TRUE)
742   subtitle_left_final <- gsub("alpha", "\u03B4", subtitle_left_final,
743   ignore.case = TRUE)
744 }
745
746 p <- ggplot(df, aes(x = x, y = mu)) +
747   (if (show_bands)
748     geom_ribbon(aes(ymin = mu - sd, ymax = mu + sd),
749               inherit.aes = TRUE, fill = "grey90") else NULL) +
750   geom_line(size = 0.8) +
751   geom_point(size = 1.6) +
752   geom_vline(xintercept = x_opt, linetype = "dashed") +
753   geom_vline(xintercept = if (opt == "min") x_opt_1se else x_opt_min,
754             linetype = "dotted") +
755   labs(
756     title = title,
757     subtitle = paste0(
758       subtitle_left_final,
759       " ; \u03BB_opt = ",
760       signif(if (opt == "min") cv$lambda_min else cv$lambda_1se, 6)
761     ),
762     x = "ln( $\lambda$ )",
```

```

760     y = "Pérdida conjunta (prom.)"
761   ) +
762   scale_x_continuous(expand = expansion(mult = c(0.02, 0.12))) +
763   coord_cartesian(clip = "off") +
764   theme_minimal(base_size = 13) +
765   theme(
766     panel.grid.minor = element_blank(),
767     plot.title = element_text(face = "bold", hjust = 0.5),
768     plot.subtitle = element_text(hjust = 0.5),
769     plot.margin = margin(10, 50, 10, 10)
770   )
771   print(p)
772   invisible(p)
773 }
774
775 save_cv_pretty <- function(cv, filename = "CV_multilogit_conjunto_pretty.png"
776   ,
777   opt = c("min", "1se"), width = 9, height = 4.8, dpi
778   = 300) {
779   opt <- match.arg(opt)
780   p <- plot_cv_pretty(cv, opt = opt)
781   is_plain_name <- dirname(filename) %in% c(".", "")
782   out_path <- if (is_plain_name) file.path(DEST, filename) else filename
783   dir.create(dirname(out_path), recursive = TRUE, showWarnings = FALSE)
784   ggsave(out_path, p, width = width, height = height, dpi = dpi)
785   invisible(out_path)
786 }
787
788 ## -----
789 ## Resumen del ajuste y exportación a Excel
790 ## -----

```

```
789 summary_multilogit_joint <- function(fit,
790                                     feature_names = NULL,
791                                     response_names = NULL,
792                                     top_pairs = 20,
793                                     threshold_abs_theta = 0) {
794   B <- fit$B
795   b0 <- fit$b0
796   Th <- fit$Theta
797   p <- nrow(B); K <- ncol(B)
798
799   if (is.null(feature_names)) feature_names <- paste0("X", seq_len(p))
800   if (is.null(response_names)) response_names <- paste0("Y", seq_len(K))
801
802   rownames(B) <- feature_names
803   colnames(B) <- response_names
804   names(b0) <- response_names
805   rownames(Th) <- response_names
806   colnames(Th) <- response_names
807
808   B_long <- do.call(rbind, lapply(seq_len(K), function(k) {
809     data.frame(
810       predictor = feature_names,
811       response = response_names[k],
812       beta = B[, k],
813       row.names = NULL,
814       check.names = FALSE
815     )
816   })))
817
818   ut <- which(upper.tri(Th), arr.ind = TRUE)
819   Th_tab <- data.frame(
```



```

820   y1 = response_names[ut[, 1]],
821   y2 = response_names[ut[, 2]],
822   theta = Th[ut],
823   abs_theta = abs(Th[ut]),
824   row.names = NULL, check.names = FALSE
825 )
826 if (threshold_abs_theta > 0)
827   Th_tab <- subset(Th_tab, abs_theta >= threshold_abs_theta)
828 Th_tab <- Th_tab[order(-Th_tab$abs_theta), ]
829
830 list(
831   intercepts = b0,
832   B_table = B_long[order(B_long$response, -abs(B_long$beta)), ],
833   Theta_table = head(Th_tab, top_pairs),
834   Theta_full = Th
835 )
836 }
837
838 export_multilogit_joint_excel <- function(fit, path_xlsx,
839                                           feature_names = NULL,
840                                           response_names = NULL,
841                                           add_cv = NULL,
842                                           threshold_abs_theta = 0) {
843   stopifnot(requireNamespace("writexl", quietly = TRUE))
844   sumo <- summary_multilogit_joint(fit,
845                                     feature_names = feature_names,
846                                     response_names = response_names,
847                                     top_pairs = Inf,
848                                     threshold_abs_theta = threshold_abs_theta)
849
850   sheet_intercepts <- data.frame(

```

```
851     response = names(sumo$intercepts),
852     intercept = as.numeric(sumo$intercepts),
853     check.names = FALSE
854 )
855 B_tab <- sumo$B_table
856 B_tab <- B_tab[order(B_tab$response, -abs(B_tab$beta)), ]
857
858 Th_full <- sumo$Theta_full
859 ut <- which(upper.tri(Th_full), arr.ind = TRUE)
860 sheet_theta <- data.frame(
861   y1 = rownames(Th_full)[ut[, "row"]],
862   y2 = colnames(Th_full)[ut[, "col"]],
863   theta = Th_full[ut],
864   abs_theta = abs(Th_full[ut]),
865   check.names = FALSE
866 )
867 if (threshold_abs_theta > 0)
868   sheet_theta <- subset(sheet_theta, abs_theta >= threshold_abs_theta)
869 sheet_theta <- sheet_theta[order(-sheet_theta$abs_theta), ]
870
871 sheets <- list(
872   "Interceptos" = sheet_intercepts,
873   "Coef_B"      = B_tab,
874   "Theta_pairs" = sheet_theta
875 )
876
877 if (!is.null(add_cv)) {
878   cvloss <- add_cv$cvloss
879   lambda_seq <- add_cv$lambda_seq
880   sheet_cv_mean <- data.frame(
881     lambda = lambda_seq,
```

```
882     cv_mean_loss = rowMeans(cvloss),
883     cv_sd_loss   = apply(cvloss, 1, sd),
884     check.names = FALSE
885   )
886   sheet_cv_folds <- as.data.frame(cvloss)
887   colnames(sheet_cv_folds) <- paste0("fold_", seq_len(ncol(sheet_cv_folds))
888   )
889   sheet_cv_folds$lambda <- lambda_seq
890   sheets[["CV_mean"]] <- sheet_cv_mean
891   sheets[["CV_folds"]] <- sheet_cv_folds
892 }
893 writexl::write_xlsx(sheets, path_xlsx)
894 invisible(TRUE)
895 }
896
897 safe_xlsx <- function(name) {
898   p <- file.path(DEST, name)
899   if (file.exists(p)) {
900     sub("\\.xlsx$", paste0("_", format(Sys.time(), "%Y%m%d-%H%M%S"), ".xlsx
901     "), p)
902   } else p
903 }
904
905 joint_loss_on <- function(fit, X, Y, standardize = TRUE) {
906   X <- as.matrix(X); Y <- as.matrix(Y)
907   Xs <- if (standardize) {
908     sweep(sweep(X, 2, fit$x_means, "-"), 2, fit$x_sds, "/")
909   } else X
910   STATES <- enumerate_states(ncol(Y))
911   obj_fun_joint(Xs, Y, fit$B_std, fit$b0_std, fit$Theta,
```

```

911         alpha = 0, lambda = 0, penalty = "l1",
912         lambda_theta = 0, alpha_theta = 0, STATES = STATES)
913     }
914
915     ## -----
916     ## Workflow principal: CV, refit, export
917     ## -----
918     delta_label <- function(val, digits = 2) {
919         bquote(delta == .(sprintf(paste0("%.\"", digits, "f"), val)))
920     }
921     lambda_opt_label <- function(cvj, opt = c("min", "l1se"), digits = 4) {
922         opt <- match.arg(opt)
923         lam <- if (opt == "min") cvj$lambda_min else cvj$lambda_l1se
924         bquote(lambda[.(opt)] == .(formatC(lam, format = "f", digits = digits)))
925     }
926     annotate_cv_header_delta <- function(cvj, opt = c("min", "l1se"),
927                                         side = 3, line = 0.5, adj = 0) {
928         opt <- match.arg(opt)
929         mtext(bquote(. (delta_label(cvj$alpha)) ~ "; " ~ .(lambda_opt_label(cvj, opt
930             )),
931             side = side, line = line, adj = adj)
932     }
933     ## 1) CV
934     cvj <- cv_multilogit_joint(
935         X, Y_bin,
936         alpha = alpha_val,
937         nlambda = nlambda, lambda_min_ratio = lambda_min_ratio,
938         nfolds = nfolds, seed = seed_val,
939         penalty = penalty_mode,
940         lambda_theta = 0.01, alpha_theta = 1.0,

```

```
941   standardize = TRUE, verbose = TRUE
942 )
943 saveRDS(cvj, file.path(DEST, "cv_multilogit_joint.rds"))
944 save_cv_curve(cvj, base_name = "CV_multilogit_conjunto", log_x = TRUE)
945
946 ## 2) Curva de CV
947 par(mar = c(5,4,2,1))
948 plot_cv_multilogit_joint(cvj, log_x = TRUE, xlab = expression(log(lambda)))
949 annotate_cv_header_delta(cvj, opt = "min")
950
951 ## 3) Refit en  $\lambda_{\min}$  y  $\lambda_{1SE}$ 
952 fit_min <- refit_multilogit_joint(X, Y_bin, cvj, which = "min", verbose =
    TRUE)
953 fit_1se <- refit_multilogit_joint(X, Y_bin, cvj, which = "1se", verbose =
    TRUE)
954 saveRDS(fit_min, file.path(DEST, "fit_min_multilogit.rds"))
955 saveRDS(fit_1se, file.path(DEST, "fit_1se_multilogit.rds"))
956
957 if (!exists("fit_min") && file.exists(file.path(SRC, "fit_min.RDS"))) {
958   fit_min <- read_rds_from("fit_min.RDS")
959 }
960 if (!exists("fit_1se") && file.exists(file.path(SRC, "fit_1se.RDS"))) {
961   fit_1se <- read_rds_from("fit_1se.RDS")
962 }
963 if (!exists("X") && file.exists(file.path(SRC, "X.RDS"))) {
964   X <- read_rds_from("X.RDS", required = FALSE)
965 }
966 if (!exists("dat") && file.exists(file.path(SRC, "dat.RDS"))) {
967   dat <- read_rds_from("dat.RDS", required = FALSE)
968 }
969
```

```
970 ## 4) Resúmenes básicos
971 nz_min <- sum(abs(fit_min$B) > 0)
972 nz_1se <- sum(abs(fit_1se$B) > 0)
973 cat(sprintf("Coeficientes no nulos -> min: %d | 1se: %d\n", nz_min, nz_1se))
974
975 cat("Pérdida conjunta (min): ", joint_loss_on(fit_min, X, Y_bin), "\n")
976 cat("Pérdida conjunta (1se): ", joint_loss_on(fit_1se, X, Y_bin), "\n")
977
978 nnz_Th_min <- sum(abs(fit_min$Theta[upper.tri(fit_min$Theta)]) > 1e-6)
979 nnz_Th_1se <- sum(abs(fit_1se$Theta[upper.tri(fit_1se$Theta)]) > 1e-6)
980 cat("Aristas en Theta (|theta| > 1e-6) -> min:", nnz_Th_min,
981     " | 1se:", nnz_Th_1se, "\n")
982
983 tab_Th_1se <- summary_multilogit_joint(
984   fit_1se,
985   feature_names = colnames(X),
986   response_names = colnames(Y_bin),
987   top_pairs = Inf)$Theta_table
988
989 tab_Th_min <- summary_multilogit_joint(
990   fit_min,
991   feature_names = colnames(X),
992   response_names = colnames(Y_bin),
993   top_pairs = Inf)$Theta_table
994
995 if (penalty_mode == "l21") {
996   cat("Predictores seleccionados (1se): ",
997       sum(rowSums(abs(fit_1se$B)) > 0), "\n")
998 }
999
1000 ## 5) Exportación de coeficientes y theta ( $\lambda_{\min}$  y  $\lambda_{1SE}$ )
```

```
1001 xlsx_1se <- safe_xlsx(sub("\\.xlsx$", "_1se.xlsx", excel_path_main))
1002 xlsx_min <- safe_xlsx(sub("\\.xlsx$", "_min.xlsx", excel_path_main))
1003
1004 export_multilogit_joint_excel(
1005   fit_1se, path_xlsx = xlsx_1se,
1006   feature_names = colnames(X),
1007   response_names = colnames(Y_bin),
1008   add_cv = cvj, threshold_abs_theta = 0
1009 )
1010
1011 export_multilogit_joint_excel(
1012   fit_min, path_xlsx = xlsx_min,
1013   feature_names = colnames(X),
1014   response_names = colnames(Y_bin),
1015   add_cv = cvj, threshold_abs_theta = 0
1016 )
1017
1018 stamp <- format(Sys.time(), "%Y%m%d-%H%M%S")
1019
1020 tiff(file.path(DEST, sprintf("CV_multilogit_conjunto_%s.tiff", stamp)),
1021      width = 7, height = 4, units = "in", res = 600, compression = "lzw")
1022 par(mar = c(5,4,2,1))
1023 plot_cv_multilogit_joint(cvj, log_x = TRUE, xlab = expression(log(lambda)))
1024 annotate_cv_header_delta(cvj, opt = "min")
1025 dev.off()
1026
1027 jpeg(file.path(DEST, sprintf("CV_multilogit_conjunto_%s.jpg", stamp)),
1028      width = 7, height = 4, units = "in", res = 600, quality = 100)
1029 par(mar = c(5,4,2,1))
1030 plot_cv_multilogit_joint(cvj, log_x = TRUE, xlab = expression(log(lambda)))
1031 annotate_cv_header_delta(cvj, opt = "min")
```

```

1032 dev.off()
1033
1034 plot_cv_pretty(cvj, opt = "min")
1035 save_cv_pretty(cvj,
1036               filename = file.path(DEST, "CV_multilogit_conjunto_pretty.png"
1037               ),
1038               opt = "min")
1039
1040 ## Resumen y tablas específicas para  $\lambda_{\min}$ 
1041 sumo_min <- summary_multilogit_joint(
1042   fit_min,
1043   feature_names = colnames(X),
1044   response_names = colnames(Y_bin),
1045   top_pairs = Inf
1046 )
1047 writexl::write_xlsx(list(
1048   "Interceptos_min" = data.frame(Resp = names(fit_min$b0), b0 = fit_min$b0),
1049   "Coef_B_min_top30" = head(sumo_min$B_table[order(-abs(sumo_min$B_table$beta
1050   )), ], 30),
1051   "Theta_pairs_min" = sumo_min$Theta_table
1052 ), safe_xlsx("Resumen_multilogit_min.xlsx"))
1053
1054 ## -----
1055 ##  Conteos y tablas para anexos ( $\lambda_{\min}$ )
1056 ##  -----
1057 nz_cells_min <- sum(abs(fit_min$B) > 0)
1058 snps_unicos_min <- sum(rowSums(abs(fit_min$B) > 0) > 0)
1059 snps_min <- rownames(fit_min$B)[rowSums(abs(fit_min$B) > 0) > 0]
1060
1061 snps_por_rasgo <- lapply(seq_len(ncol(Y_bin)), function(k) {

```



```
1061   rownames(fit_min$B)[abs(fit_min$B[, k]) > 0]
1062 })
1063 names(snps_por_rasgo) <- colnames(Y_bin)
1064
1065 cat("Celdas no nulas:", nz_cells_min, "\n",
1066     "SNPs únicos:",   snps_unicos_min, "\n")
1067
1068 B <- fit_min$B
1069 tab_coef <- do.call(rbind, lapply(seq_len(ncol(B)), function(k) {
1070   data.frame(
1071     Rasgo = colnames(B)[k],
1072     SNP   = rownames(B),
1073     Beta  = B[, k],
1074     OR    = exp(B[, k]),
1075     row.names = NULL, check.names = FALSE
1076   )
1077 })))
1078 tab_coef <- subset(tab_coef, abs(Beta) > 0)
1079 tab_coef <- tab_coef[order(tab_coef$Rasgo, -abs(tab_coef$Beta)), ]
1080
1081 incid <- (abs(B) > 0) * 1
1082 incid <- data.frame(SNP = rownames(B), incid, check.names = FALSE)
1083
1084 sumo_min <- summary_multilogit_joint(
1085   fit_min,
1086   feature_names = colnames(X),
1087   response_names = colnames(Y_bin),
1088   top_pairs = Inf
1089 )
1090
1091 writexl::write_xlsx(list(
```

```

1092   "Coef_no_nulos_min"      = tab_coef,
1093   "Incidencia_SNPxRasgo" = incid,
1094   "Theta_pairs_min"       = sumo_min$Theta_table
1095 ), safe_xlsx("Multilogit_min_resultados.xlsx"))
1096
1097 ## -----
1098 ## Bootstrap percentil de coeficientes y theta
1099 ## -----
1100 bootstrap_coef_ci_parallel <- function(X, Y,
1101                                     lambda_star,
1102                                     alpha = 0.5,
1103                                     penalty = c("l1", "l2l"),
1104                                     lambda_theta = 0.01, alpha_theta =
1105                                     1.0,
1106                                     B = B_reps, level = 0.95,
1107                                     standardize = TRUE, seed = 123,
1108                                     verbose = TRUE) {
1109   penalty <- match.arg(penalty)
1110   set.seed(seed)
1111   X <- as.matrix(X); Y <- as.matrix(Y)
1112   n <- nrow(X); p <- ncol(X); K <- ncol(Y)
1113
1114   boot_list <- future_lapply(
1115     seq_len(B), future.seed = TRUE,
1116     function(b) {
1117       idx <- sample.int(n, n, replace = TRUE)
1118       fit_b <- fista_multilogit_joint(
1119         X[idx, , drop = FALSE], Y[idx, , drop = FALSE],
1120         alpha = alpha, lambda = lambda_star,
1121         penalty = penalty,
1122         lambda_theta = lambda_theta, alpha_theta = alpha_theta,

```

```
1122     standardize = standardize, verbose = FALSE
1123   )
1124   list(B = fit_b$B, Theta = fit_b$Theta)
1125 }
1126 )
1127
1128 if (verbose) cat("Bootstrap paralelo terminado: ", B, "réplicas.\n")
1129
1130 B_arr <- array(NA_real_, dim = c(p, K, B))
1131 Th_arr <- array(NA_real_, dim = c(K, K, B))
1132 for (b in seq_len(B)) {
1133   B_arr[, , b] <- boot_list[[b]]$B
1134   Th_arr[, , b] <- boot_list[[b]]$Theta
1135 }
1136
1137 a <- (1 - level) / 2
1138 probs <- c(a, 0.5, 1 - a)
1139
1140 B_lower <- apply(B_arr, c(1,2), quantile, probs = probs[1], names = FALSE)
1141 B_med <- apply(B_arr, c(1,2), quantile, probs = probs[2], names = FALSE)
1142 B_upper <- apply(B_arr, c(1,2), quantile, probs = probs[3], names = FALSE)
1143 B_sig <- (B_lower > 0) | (B_upper < 0)
1144
1145 ut <- which(upper.tri(matrix(0, K, K)), arr.ind = TRUE)
1146 Th_lower <- matrix(0, K, K); Th_med <- matrix(0, K, K); Th_upper <- matrix
  (0, K, K)
1147 Th_sig <- matrix(FALSE, K, K)
1148 for (e in seq_len(nrow(ut))) {
1149   r <- ut[e,1]; c <- ut[e,2]
1150   vec <- Th_arr[r, c, ]
1151   qs <- stats::quantile(vec, probs = probs, names = FALSE)
```

```
1152   Th_lower[r,c] <- qs[1]; Th_med[r,c] <- qs[2]; Th_upper[r,c] <- qs[3]
1153   Th_lower[c,r] <- Th_lower[r,c]; Th_med[c,r] <- Th_med[r,c]; Th_upper[c,r]
      <- Th_upper[r,c]
1154   Th_sig[r,c] <- (qs[1] > 0) | (qs[3] < 0)
1155   Th_sig[c,r] <- Th_sig[r,c]
1156 }
1157 diag(Th_lower) <- diag(Th_med) <- diag(Th_upper) <- 0
1158
1159 feature_names <- colnames(X); if (is.null(feature_names)) feature_names <-
      paste0("X", seq_len(p))
1160 response_names <- colnames(Y); if (is.null(response_names)) response_names
      <- paste0("Y", seq_len(K))
1161 dimnames(B_lower) <- dimnames(B_med) <- dimnames(B_upper) <- dimnames(B_sig
      ) <-
1162   list(feature_names, response_names)
1163 dimnames(Th_lower) <- dimnames(Th_med) <- dimnames(Th_upper) <- dimnames(Th
      _sig) <-
1164   list(response_names, response_names)
1165
1166 B_long <- do.call(rbind, lapply(seq_len(K), function(k) {
1167   data.frame(
1168     SNP      = feature_names,
1169     Rasgo    = response_names[k],
1170     lower    = B_lower[, k],
1171     median   = B_med[, k],
1172     upper    = B_upper[, k],
1173     signif   = as.logical(B_sig[, k]),
1174     row.names = NULL, check.names = FALSE
1175   )
1176 })))
1177
```

```
1178 Th_tab <- data.frame(  
1179   y1 = response_names[ut[,1]],  
1180   y2 = response_names[ut[,2]],  
1181   lower = Th_lower[ut],  
1182   median = Th_med[ut],  
1183   upper = Th_upper[ut],  
1184   signif = as.logical(Th_sig[ut]),  
1185   row.names = NULL, check.names = FALSE  
1186 )  
1187  
1188 list(  
1189   B_ci = list(lower = B_lower, median = B_med, upper = B_upper,  
1190             signif = B_sig, long = B_long),  
1191   Theta_ci = list(lower = Th_lower, median = Th_med, upper = Th_upper,  
1192                 signif = Th_sig, long = Th_tab),  
1193   n_boot = B, level = level, lambda_star = lambda_star,  
1194   workers = future::nbrOfWorkers()  
1195 )  
1196 }  
1197  
1198 ## Bootstrap con  $\lambda$  seleccionado por CV (ejemplo:  $\lambda_{\min}$ )  
1199 lambda_star <- cvj$lambda_min  
1200  
1201 boot_ci <- bootstrap_coef_ci_parallel(  
1202   X, Y_bin,  
1203   lambda_star = lambda_star,  
1204   alpha = alpha_val,  
1205   penalty = penalty_mode,  
1206   lambda_theta = 0.01, alpha_theta = 1.0,  
1207   B = 1000, level = 0.95,  
1208   standardize = TRUE, seed = 123, verbose = TRUE
```

```
1209 )
1210 sig_B <- subset(boot_ci$B_ci$long, signif)
1211 nrow(sig_B)
1212 saveRDS(boot_ci, file.path(DEST, "boot_ci_multilogit.rds"))
1213
1214 ## -----
1215 ## Gráfico de coeficientes significativos por bootstrap
1216 ## -----
1217 plot_sig_coefs_boot <- function(boot_ci,
1218                                titulo = "Coeficientes significativos (
1219                                Bootstrap IC 95%)",
1220                                archivo_base = "EN_multilogit_coef_signif_
1221                                boot",
1222                                label_map = NULL,
1223                                ncol_facets = 2) {
1224   df <- boot_ci$B_ci$long
1225   df$SNP <- gsub("_", "-", df$SNP)
1226   df_sig <- subset(df, signif)
1227   if (nrow(df_sig) == 0) return(invisible(NULL))
1228
1229   if (!is.null(label_map)) {
1230     df_sig$Rasgo_lab <- ifelse(df_sig$Rasgo %in% names(label_map),
1231                               unname(label_map[df_sig$Rasgo]),
1232                               df_sig$Rasgo)
1233   } else {
1234     df_sig$Rasgo_lab <- df_sig$Rasgo
1235   }
1236   df_sig$Rasgo_lab <- factor(df_sig$Rasgo_lab, levels = unique(df_sig$Rasgo_
1237   lab))
1238
1239   suppressPackageStartupMessages({ library(dplyr); library(forcats) })
```

```
1237 df_sig <- df_sig %>%
1238   group_by(Rasgo_lab) %>%
1239   arrange(median, .by_group = TRUE) %>%
1240   mutate(SNP = fct_rev(factor(SNP, levels = unique(SNP)))) %>%
1241   ungroup()
1242
1243 library(ggplot2)
1244 p <- ggplot(df_sig, aes(x = median, y = SNP, fill = Rasgo_lab)) +
1245   geom_col(width = 0.75) +
1246   geom_errorbarh(aes(xmin = lower, xmax = upper), height = 0.25) +
1247   facet_wrap(~ Rasgo_lab, scales = "free_y", ncol = ncol_facets) +
1248   labs(title = titulo,
1249        x = "Beta (mediana bootstrap) - barras = IC 95%",
1250        y = "Predictor") +
1251   theme_minimal(base_size = 12) +
1252   theme(legend.position = "none",
1253        plot.title = element_text(hjust = 0.5, face = "bold"),
1254        strip.text = element_text(face = "bold"))
1255
1256 print(p)
1257
1258 stamp <- format(Sys.time(), "%Y%m%d-%H%M%S")
1259 tiff(sprintf("%s_%s.tiff", archivo_base, stamp),
1260      width = 7, height = 4.8, units = "in", res = 600, compression = "lzw")
1261 print(p); dev.off()
1262 jpeg(sprintf("%s_%s.jpg", archivo_base, stamp),
1263      width = 7, height = 4.8, units = "in", res = 600, quality = 100)
1264 print(p); dev.off()
1265 }
1266
1267 etiquetas_dep <- c(
```

```

1268   "AP"      = "Presencia de axilas",
1269   "Estrías" = "Presencia de estrías",
1270   "RAMA"    = "Presencia de ramas",
1271   "FP"      = "Forma de la panoja",
1272   "PigCot"  = "Pigmento de cotiledones",
1273   "TP"      = "Tipo de la panoja"
1274 )
1275
1276 plot_sig_coefs_boot(
1277   boot_ci,
1278   titulo      = "Coeficientes significativos (Bootstrap IC 95%) ( $\lambda_{\min}$ )",
1279   archivo_base = file.path(DEST, "EN_multilogit_coef_signif_boot"),
1280   label_map   = etiquetas_dep,
1281   ncol_facets = 2
1282 )
1283
1284 sig_B <- subset(boot_ci$B_ci$long, signif)
1285 sig_Th <- subset(boot_ci$Theta_ci$long, signif)
1286
1287 writexl::write_xlsx(list(
1288   "B_CI"      = boot_ci$B_ci$long[order(boot_ci$B_ci$long$Rasgo,
1289                                         boot_ci$B_ci$long$SNP,
1290                                         -abs(boot_ci$B_ci$long$median)),
1291   ],
1292   "B_signif"  = sig_B[order(sig_B$Rasgo, sig_B$SNP, -abs(sig_B$median)),
1293   ],
1294   "Theta_CI"  = boot_ci$Theta_ci$long[order(boot_ci$Theta_ci$long$y1,
1295                                             boot_ci$Theta_ci$long$y2,
1296                                             -abs(boot_ci$Theta_ci$long$
1297   median)), ],
1298   "Theta_signif" = sig_Th[order(sig_Th$y1, sig_Th$y2, -abs(sig_Th$median)), ]

```



```
1296 ), safe_xlsx("Bootstrap_IC_percentil.xlsx"))
1297
1298 ## -----
1299 ## Gráfico de la red theta (entre rasgos)
1300 ## -----
1301 plot_theta_network_ggraph <- function(
1302   fit, label_map = NULL,
1303   thr = 0.05, seed = 123,
1304   title = "Red residual entre rasgos (theta)",
1305   response_names = NULL,
1306   wrap_width = 18,
1307   repel = TRUE,
1308   base_size = 12,
1309   file_out = NULL,
1310   width = 10, height = 5, dpi = 300) {
1311
1312   Th <- as.matrix(fit$Theta); K <- ncol(Th)
1313   rn <- response_names; if (is.null(rn)) rn <- colnames(fit$B)
1314   if (is.null(rn)) rn <- colnames(Th)
1315   if (is.null(rn)) rn <- paste0("Y", seq_len(K))
1316   rownames(Th) <- rn; colnames(Th) <- rn
1317   Th <- (Th + t(Th)) / 2; diag(Th) <- 0
1318
1319   ut <- which(upper.tri(Th), arr.ind = TRUE)
1320   if (nrow(ut) == 0) return(invisible(NULL))
1321
1322   edges <- data.frame(
1323     from = rownames(Th)[ut[,1]],
1324     to   = colnames(Th)[ut[,2]],
1325     w    = Th[ut],
1326     aw   = abs(Th[ut]),
```

```
1327   sign = ifelse(Th[ut] >= 0, "Positiva", "Negativa"),
1328   check.names = FALSE
1329 )
1330 edges <- subset(edges, aw > thr)
1331 if (nrow(edges) == 0) return(invisible(NULL))
1332
1333 nodes <- data.frame(name = rn, check.names = FALSE)
1334 nodes$label <- if (!is.null(label_map)) {
1335   ifelse(nodes$name %in% names(label_map),
1336     unname(label_map[nodes$name]), nodes$name)
1337 } else nodes$name
1338
1339 for (pkg in c("stringr","ggraph","igraph", if (repel) "ggrepel")) {
1340   if (!requireNamespace(pkg, quietly = TRUE)) {
1341     stop("Falta el paquete '", pkg, "'.")
1342   }
1343 }
1344
1345 suppressPackageStartupMessages({
1346   library(igraph); library(ggraph); library(ggplot2); library(stringr)
1347   if (repel) library(ggrepel)
1348 })
1349 nodes$label <- stringr::str_wrap(nodes$label, width = wrap_width)
1350
1351 g <- igraph::graph_from_data_frame(edges, vertices = nodes, directed =
  FALSE)
1352
1353 set.seed(seed)
1354 p <- ggraph(g, layout = "fr") +
1355   geom_edge_link(aes(width = aw, linetype = sign, color = sign), alpha =
    0.85) +
```

```
1356   geom_node_point(shape = 21, size = 8, fill = "white", color = "grey40") +
1357   geom_node_text(
1358     aes(label = label),
1359     repel = repel,
1360     point.padding = grid::unit(0.3, "lines"),
1361     box.padding    = grid::unit(0.25, "lines"),
1362     force = 1, max.overlaps = Inf,
1363     size = 4, fontface = "bold"
1364   ) +
1365   scale_edge_width(range = c(0.5, 3.5), guide = "none") +
1366   scale_edge_color_manual(
1367     name    = "Signo",
1368     values = c("Positiva" = "#1b9e77", "Negativa" = "#d95f02")
1369   ) +
1370   scale_edge_linetype_manual(
1371     name    = "Signo",
1372     values = c("Positiva" = "solid", "Negativa" = "dashed")
1373   ) +
1374   coord_cartesian(clip = "off") +
1375   theme_minimal(base_size = base_size) +
1376   theme(
1377     panel.grid = element_blank(),
1378     legend.position = "top",
1379     plot.title = element_text(hjust = 0.5, face = "bold"),
1380     plot.margin = margin(20, 60, 20, 60)
1381   ) +
1382   labs(title = title)
1383
1384 print(p)
1385
1386 if (!is.null(file_out)) {
```

```

1387     ggsave(filename = file.path(DEST, file_out), plot = p,
1388           width = width, height = height, dpi = dpi)
1389   }
1390   invisible(p)
1391 }
1392
1393 plot_theta_network_ggraph(
1394   fit_1se,
1395   label_map      = etiquetas_dep,
1396   thr            = 0.08,
1397   title          = "Red theta ( $\lambda_{1SE}$ )",
1398   response_names = colnames(Y_bin),
1399   wrap_width     = 16,
1400   file_out       = "Red_Theta_1se.png",
1401   width          = 11, height = 5, dpi = 300
1402 )
1403
1404 plot_theta_network_ggraph(
1405   fit_min,
1406   label_map      = etiquetas_dep,
1407   thr            = 0.08,
1408   title          = "Red theta ( $\lambda_{min}$ )",
1409   response_names = colnames(Y_bin),
1410   wrap_width     = 16,
1411   file_out       = "Red_Theta_lmin.png",
1412   width          = 11, height = 5, dpi = 300
1413 )

```

Código A.1: Programa multivariado logístico *Elastic Net*

```

1 #####
2 ## Elastic Net multivariado (mgaussian) + RDA-like sobre  $\hat{Y}$ 

```

```
3 ## - Selección de Alpha y lambda_min (MSE mínimo) vía CV k-fold
4 ## - Métricas de ajuste (OOF e in-sample)
5 ## - Bootstrap simple (Alpha y lambda fijos) e IC por coeficiente
6 ## - Triplot RDA-like sobre  $\hat{Y}$ 
7 ## - Forest plots de coeficientes significativos
8 #####
9
10 ## --- Limpieza ---
11 rm(list = ls())
12 graphics.off()
13 set.seed(2025)
14
15 ## --- Librerías ---
16 suppressPackageStartupMessages({
17   library(readxl)
18   library(glmnet)
19   library(boot)
20   library(openxlsx)
21   library(FactoMineR)
22   library(factoextra)
23   library(ggplot2)
24   library(dplyr)
25   library(stringr)
26   library(ggrepel)
27   library(RColorBrewer)
28   library(grid)
29   library(tidytext)
30   library(MVN)
31   library(patchwork)
32 })
33
```

```
34 ## --- Directorio de salida (relativo al directorio de trabajo) ---
35 ruta_salida_base <- file.path(getwd(), "Resultados_EN")
36 if (!dir.exists(ruta_salida_base)) {
37   dir.create(ruta_salida_base, recursive = TRUE)
38 }
39
40 ## --- Datos ---
41 ## Se asume que el archivo se encuentra en el directorio de trabajo
42 Datos <- read_excel("Datoscuanti_Mole_CP.xlsx")
43 colnames(Datos) <- gsub("\\s+", "_", colnames(Datos))
44
45 ## --- Definición de conjuntos ---
46 cand_y <- colnames(Datos)[2:18]
47 y_cols <- cand_y[sapply(Datos[, cand_y], drop = FALSE, is.numeric)]
48 if (length(y_cols) == 0) stop("No se encontraron columnas numéricas en Datos
    [,2:18].")
49
50 max_xcol <- min(ncol(Datos), 337)
51 if (max_xcol < 25) stop("No hay columnas de predictores moleculares a partir
    de la 25.")
52 x_cols <- colnames(Datos)[25:max_xcol]
53
54 ## --- Filtrado filas completas en Y y X ---
55 Datos_sub <- Datos[, c(y_cols, x_cols)]
56 cc <- complete.cases(Datos_sub)
57 Datos_cc <- Datos_sub[cc, , drop = FALSE]
58
59 ## --- Matrices crudas ---
60 X_raw <- as.matrix(Datos_cc[, x_cols, drop = FALSE])
61 Y_raw <- as.matrix(Datos_cc[, y_cols, drop = FALSE])
62
```

```
63 stopifnot(!anyDuplicated(colnames(X_raw)))
64
65 validate_data <- function(X, Y, tol = 1e-12) {
66   issues <- character()
67
68   if (nrow(X) != nrow(Y)) {
69     issues <- c(issues, sprintf("X (%d) e Y (%d) tienen diferente número de
70                               filas",
71                               nrow(X), nrow(Y)))
72   }
73   if (anyNA(X)) issues <- c(issues, "Hay valores NA en X")
74   if (anyNA(Y)) issues <- c(issues, "Hay valores NA en Y")
75
76   zero_var_x <- apply(X, 2, function(x) {
77     isTRUE(all.equal(var(x, na.rm = TRUE), 0, tolerance = tol))
78   })
79   zero_var_y <- apply(Y, 2, function(x) {
80     isTRUE(all.equal(var(x, na.rm = TRUE), 0, tolerance = tol))
81   })
82
83   if (any(zero_var_x)) {
84     issues <- c(issues, sprintf("%d predictores con varianza ~= 0", sum(zero_
85                               var_x)))
86   }
87   if (any(zero_var_y)) {
88     issues <- c(issues, sprintf("%d respuestas con varianza ~= 0", sum(zero_
89                               var_y)))
90   }
91
92   if (length(issues) > 0) {
93     warning(paste(issues, collapse = "; "))
94   }
95 }
```

```
91   } else {
92     message("Validación de X e Y sin problemas detectados.")
93   }
94
95   invisible(list(
96     issues = issues,
97     zero_var_x = which(zero_var_x),
98     zero_var_y = which(zero_var_y)
99   ))
100 }
101
102 data_issues <- validate_data(X_raw, Y_raw)
103
104 ## --- Validación crítica de datos ---
105 if (nrow(X_raw) < 10) stop("Muy pocas observaciones para CV (n < 10)")
106 if (ncol(X_raw) == 0) stop("No quedaron predictores después del filtrado")
107 if (any(is.na(X_raw))) stop("NAs en X después de filtrado - revisar datos")
108 if (any(is.na(Y_raw))) stop("NAs en Y después de filtrado - revisar datos")
109
110 message(sprintf("Dimensiones finales: X(%d x %d), Y(%d x %d)",
111               nrow(X_raw), ncol(X_raw), nrow(Y_raw), ncol(Y_raw)))
112
113 ## --- Normalidad multivariada de Y ---
114 normalidad_multivariada <- mvn(data = Y_raw, mvn_test = "mardia")
115
116 ## --- Diccionario de etiquetas cortas para respuestas ---
117 etiquetas_dep <- c(
118   "STD"      = "DiamTallo",
119   "BNUMB"    = "NroRamas",
120   "STL"      = "FormaHojaSu",
121   "TEENUMB"  = "NroDientes",
```



```
122 "PETLEN" = "LongPecíolo",
123 "LEAFLEN" = "LongHoja",
124 "LEAFWI" = "AnchoHoja",
125 "PANLEN" = "LongPanoja",
126 "GLOLEN" = "LongGlomer",
127 "COTLEN" = "LongCotil",
128 "HYPLEN" = "LongHipo",
129 "EMM" = "DiasSE",
130 "SBF" = "DiasSBF",
131 "SFL" = "DiasSF",
132 "SMF" = "DiasSMF",
133 "HEIGHT" = "AltPlanta",
134 "GRADIAM" = "DiamGrano",
135 "ICTallo" = "IntColTallo",
136 "ICFA" = "IntColFA",
137 "ICCos" = "IntColCosecha",
138 "Densidad" = "DensPanoja",
139 "ICCot" = "IntColCot",
140 "IPHipo" = "IntPigHipo"
141 )
142
143 etiquetas_dep_largo <- c(
144 "STD" = "Diámetro de tallo",
145 "BNUMB" = "Número de ramas",
146 "STL" = "Forma de hoja (sup.)",
147 "TEENUMB" = "Número de dientes",
148 "PETLEN" = "Longitud de pecíolo",
149 "LEAFLEN" = "Longitud de hoja",
150 "LEAFWI" = "Ancho de hoja",
151 "PANLEN" = "Longitud de panoja",
152 "GLOLEN" = "Longitud de glomérulo",
```

```
153 "COTLEN" = "Longitud de cotiledón",
154 "HYPLEN" = "Longitud de hipocótilo",
155 "EMM"    = "Días siembra-emergencia",
156 "SBF"    = "Días siembra-botón floral",
157 "SFL"    = "Días siembra-floración",
158 "SMF"    = "Días siembra-madurez fisiológica",
159 "HEIGHT" = "Altura de planta",
160 "GRADIAM" = "Diámetro de grano"
161 )
162
163 ## ===== Selección de Alpha y lambda por MSE mínimo (lambda_min)
164      =====
165 set.seed(124)
166
167 K <- 5
168
169 grid_alpha <- seq(0.1, 0.9, by = 0.1)
170
171 res_alpha <- lapply(grid_alpha, function(a) {
172   cv <- cv.glmnet(
173     X_raw, Y_raw,
174     family = "mgaussian",
175     alpha = a,
176     nfolds = K,
177     intercept = TRUE,
178     standardize = TRUE,
179     standardize.response = TRUE,
180     keep = TRUE,
181     type.measure = "mse"
182   )
183   lam_min <- cv$lambda.min
184   idx_min <- which.min(abs(cv$lambda - lam_min))
```

```
183 mse_min <- cv$cvm[idx_min]
184 sd_min  <- cv$cvsd[idx_min]
185
186 fit_min <- glmnet(
187   X_raw, Y_raw,
188   family = "mgaussian",
189   alpha = a,
190   lambda = lam_min,
191   intercept = TRUE,
192   standardize = TRUE,
193   standardize.response = TRUE
194 )
195 nz_min <- sum(vapply(coef(fit_min), function(M) {
196   MM <- as.matrix(M)
197   sum(MM[rownames(MM) != "(Intercept)", 1] != 0)
198 }, integer(1)))
199
200 list(
201   alpha      = a,
202   cv         = cv,
203   lambda_min = lam_min,
204   mse_min    = mse_min,
205   sd_min     = sd_min,
206   nz_min     = nz_min
207 )
208 })
209
210 tab_alpha <- do.call(
211   rbind,
212   lapply(res_alpha, function(z) {
213     data.frame(
```

```
214     alpha    = z$alpha,
215     mse_min  = z$mse_min,
216     sd_min   = z$sd_min,
217     nz_min   = z$nz_min
218   )
219 })
220 )
221
222 best_idx <- which.min(tab_alpha$mse_min)
223 best_obj <- res_alpha[[best_idx]]
224 tol_1sd  <- best_obj$sd_min
225 cands    <- subset(tab_alpha, mse_min <= min(tab_alpha$mse_min) + tol_1sd)
226 alpha_final <- cands$alpha[which.min(cands$nz_min)]
227
228 which_alpha <- which.min(abs(sapply(res_alpha, function(z) z$alpha) - alpha_
    final))
229 obj_final   <- res_alpha[[which_alpha]]
230 cv_final    <- obj_final$cv
231 lambda_sel  <- obj_final$lambda_min
232
233 cat(sprintf("\nAlpha_final = %.3f | lambda_min = %.6f | MSE(lambda_min) =
    %.4f | no-ceros = %d\n",
234         alpha_final, lambda_sel, obj_final$mse_min, obj_final$nz_min))
235
236 openxlsx::write.xlsx(
237   tab_alpha,
238   file.path(ruta_salida_base, "cv_resumen_por_alpha_lambda_min.xlsx"),
239   overwrite = TRUE
240 )
241
242 stopifnot(all.equal(as.numeric(lambda_sel), as.numeric(cv_final$lambda_min)))
```

```
243
244 df_cv <- data.frame(
245   lambda = cv_final$lambda,
246   cvm     = cv_final$cvm,
247   cvsd    = cv_final$cvsd
248 )
249 df_cv$cvup  <- df_cv$cvm + df_cv$cvsd
250 df_cv$cvlo  <- df_cv$cvm - df_cv$cvsd
251 df_cv$loglam <- log(df_cv$lambda)
252
253 delta  <- alpha_final
254 lam_opt <- lambda_sel
255
256 sub_lab <- bquote(delta == .(format(delta, nsmall = 1)) ~ ";" ~
257                   lambda[opt] == .(format(lam_opt, digits = 4, nsmall = 4))
258 )
259
260 p_cv <- ggplot(df_cv, aes(x = loglam, y = cvm)) +
261   geom_ribbon(aes(ymin = cvlo, ymax = cvup), alpha = .2) +
262   geom_line(linewidth = 1) +
263   geom_vline(xintercept = log(lambda_sel), colour = "red", linetype = "solid"
264     ") +
265   geom_vline(xintercept = log(cv_final$lambda.1se), colour = "blue", linetype
266     = "dashed") +
267   scale_x_reverse() +
268   labs(
269     title      = "Curva de error de validación cruzada",
270     subtitle   = sub_lab,
271     x = expression(ln(lambda)),
272     y = "MSE"
273 ) +
```

```
271 theme_minimal(base_size = 12) +
272 theme(
273   plot.title = element_text(face = "bold", size = 18),
274   plot.subtitle = element_text(size = 13)
275 )
276
277 ggsave(
278   file.path(ruta_salida_base, "cv_curve_alpha_auto_lambda_min.png"),
279   p_cv,
280   width = 9, height = 5.5, dpi = 300
281 )
282
283 ## --- Ajuste final con lambda_min ---
284 fit_final <- glmnet(
285   X_raw, Y_raw,
286   family = "mgaussian",
287   alpha = alpha_final,
288   lambda = lambda_sel,
289   intercept = TRUE,
290   standardize = TRUE,
291   standardize.response = TRUE
292 )
293
294 coefs <- coef(fit_final)
295
296 res_nz <- do.call(
297   rbind,
298   lapply(names(coefs), function(dep) {
299     M <- as.matrix(coefs[[dep]])
300     df <- data.frame(
301       dependiente = dep,
```

```
302     predictor    = rownames(M),
303     coeficiente = as.numeric(M[, 1])
304   )
305   subset(df, predictor != "(Intercept)" & coeficiente != 0)
306 })
307 )
308
309 res_all <- do.call(
310   rbind,
311   lapply(names(coefs), function(dep) {
312     M <- as.matrix(coefs[[dep]])
313     data.frame(
314       dependiente = dep,
315       predictor    = rownames(M),
316       coeficiente = as.numeric(M[, 1])
317     )
318   })
319 )
320
321 openxlsx::write.xlsx(
322   list(
323     coef_no_cero = res_nz,
324     coef_todos   = res_all
325   ),
326   file.path(ruta_salida_base, "coeficientes_modelo_final_lambda_min.xlsx"),
327   overwrite = TRUE
328 )
329
330 ## =====
331 ## Métricas comparables (Elastic Net vs otros modelos)
332 ## OOF (CV) e In-sample, usando exactamente lambda_min
```

```
333 ## =====
334
335 stopifnot(!is.null(cv_final$fit.preval))
336 preval <- cv_final$fit.preval
337 dd <- dim(preval)
338 stopifnot(length(dd) == 3L)
339
340 lambda_sel <- as.numeric(cv_final$lambda.min)
341 lam_vec    <- as.numeric(cv_final$lambda)
342 n_lam      <- length(lam_vec)
343
344 lam_id <- which(lam_vec == lambda_sel)
345 if (length(lam_id) == 0L) {
346   lam_id <- which.min(abs(log(lam_vec) - log(lambda_sel)))
347   message(sprintf(
348     "lambda_min exacto no está indexado en fit.preval; se usa el más cercano:
349     %.6g (idx=%d/%d).",
350     lam_vec[lam_id], lam_id, n_lam
351   ))
352 } else {
353   lam_id <- lam_id[1]
354 }
355 lambda_oof_usado <- lam_vec[lam_id]
356 message(sprintf("lambda_min (modelo / OOF) = %.6g | idx=%d/%d", lambda_oof_
357   usado, lam_id, n_lam))
358
359 stopifnot(identical(unname(lambda_sel), unname(cv_final$lambda.min)))
360
361 if (!isTRUE(all.equal(lambda_oof_usado, lambda_sel))) {
362   warning(sprintf(
363     "El OOF no usa el lambda exacto (no estaba en fit.preval). Usado: %.6g |
```



```
    lambda_min: %.6g",
362    lambda_oof_usado, lambda_sel
363  ))
364 }
365
366 n_obs <- nrow(Y_raw)
367 n_resp <- ncol(Y_raw)
368
369 lambda_dim <- which(dd == n_lam)[1]
370 q_dim      <- which(dd == n_resp)[1]
371 n_dim      <- which(dd == n_obs)[1]
372 stopifnot(!is.na(lambda_dim), !is.na(q_dim), !is.na(n_dim))
373
374 slice <- switch(
375   lambda_dim,
376   `1` = preval[lam_id, , , drop = FALSE],
377   `2` = preval[, lam_id, , drop = FALSE],
378   `3` = preval[, , lam_id, drop = FALSE]
379 )
380
381 mat <- drop(slice)
382 if (!is.matrix(mat)) mat <- matrix(mat, nrow = n_obs, ncol = n_resp)
383 if (nrow(mat) == n_resp && ncol(mat) == n_obs) mat <- t(mat)
384
385 stopifnot(nrow(mat) == n_obs, ncol(mat) == n_resp)
386 Y_oof <- mat
387 rownames(Y_oof) <- rownames(Y_raw)
388 colnames(Y_oof) <- colnames(Y_raw)
389
390 Y_hat <- predict(fit_final, newx = X_raw, s = lambda_sel, type = "response")
    [, , 1]
```

```
391 dimnames(Y_hat) <- dimnames(Y_raw)
392
393 .metricas_por_col <- function(y, yhat) {
394   sse <- sum((y - yhat)^2, na.rm = TRUE)
395   sst <- sum((y - mean(y, na.rm = TRUE))^2, na.rm = TRUE)
396   mse <- sse / sum(is.finite(y - yhat))
397   rmse <- sqrt(mse)
398   mae <- mean(abs(y - yhat), na.rm = TRUE)
399   r2Q^2 <- if (sst > 0) 1 - sse/sst else NA_real_
400   c(MSE = mse, RMSE = rmse, MAE = mae, R2_o_Q^2 = r2Q^2)
401 }
402
403 met_oof <- dplyr::bind_rows(lapply(seq_len(n_resp), function(j) {
404   out <- .metricas_por_col(Y_raw[, j], Y_oof[, j])
405   data.frame(respuesta = colnames(Y_raw)[j], t(out), check.names = FALSE)
406 })))
407
408 met_in <- dplyr::bind_rows(lapply(seq_len(n_resp), function(j) {
409   out <- .metricas_por_col(Y_raw[, j], Y_hat[, j])
410   data.frame(respuesta = colnames(Y_raw)[j], t(out), check.names = FALSE)
411 })))
412
413 sd_y <- apply(Y_raw, 2, sd, na.rm = TRUE)
414 safe_div <- function(x, d) ifelse(is.finite(d) & d > 0, x/d, NA_real_)
415
416 met_oof_norm <- met_oof %>%
417   dplyr::mutate(
418     sdY = sd_y[match(respuesta, colnames(Y_raw))],
419     NRMSE = safe_div(RMSE, sdY),
420     NMAE = safe_div(MAE, sdY)
421   ) %>%
```

```
422   dplyr::select(respuesta, MSE, RMSE, NRMSE, MAE, NMAE, R2_o_Q^2)
423
424 met_in_norm <- met_in %>%
425   dplyr::mutate(
426     sdY   = sd_y[match(respuesta, colnames(Y_raw))],
427     NRMSE = safe_div(RMSE, sdY),
428     NMAE  = safe_div(MAE,  sdY)
429   ) %>%
430   dplyr::select(respuesta, MSE, RMSE, NRMSE, MAE, NMAE, R2_o_Q^2)
431
432 resumen_global <- data.frame(
433   Conjunto   = c("OOF (CV)", "In-sample"),
434   MSE_prom   = c(mean(met_oof$MSE, na.rm = TRUE),
435                  mean(met_in$MSE,  na.rm = TRUE)),
436   RMSE_prom  = c(mean(met_oof$RMSE, na.rm = TRUE),
437                  mean(met_in$RMSE,  na.rm = TRUE)),
438   MAE_prom   = c(mean(met_oof$MAE,  na.rm = TRUE),
439                  mean(met_in$MAE,   na.rm = TRUE)),
440   R2oQ^2_prom = c(mean(met_oof$R2_o_Q^2, na.rm = TRUE),
441                  mean(met_in$R2_o_Q^2, na.rm = TRUE)),
442   NRMSE_prom = c(mean(met_oof_norm$NRMSE, na.rm = TRUE),
443                  mean(met_in_norm$NRMSE, na.rm = TRUE)),
444   NMAE_prom  = c(mean(met_oof_norm$NMAE, na.rm = TRUE),
445                  mean(met_in_norm$NMAE,  na.rm = TRUE))
446 )
447
448 openxlsx::write.xlsx(
449   list(
450     INFO_lambda      = data.frame(
451       lambda_min = lambda_sel,
452       lam_idx    = lam_id,
```

```
453     nlambda      = n_lam
454   ),
455   OOF_original      = met_oof,
456   OOF_normalizado   = met_oof_norm,
457   InSample_original = met_in,
458   InSample_normalizado = met_in_norm,
459   Resumen_global     = resumen_global
460 ),
461 file.path(ruta_salida_base, "metricas_EN_lambda_min_comparables.xlsx"),
462 overwrite = TRUE
463 )
464
465 message("OOF promedio | RMSE = ",
466         round(resumen_global$RMSE_prom[1], 4),
467         " | NRMSE = ", round(resumen_global$NRMSE_prom[1], 4),
468         " | Q^2 = ", round(resumen_global$R2oQ^2_prom[1], 4))
469
470 ## ===== Bootstrap con lambda_min =====
471
472 boot_coef_robust <- function(data, indices, alpha_opt, lambda_opt, resp_names
473   , pred_names) {
474   d <- data[indices, , drop = FALSE]
475   Xb <- as.matrix(d[, pred_names, drop = FALSE])
476   Yb <- as.matrix(d[, resp_names, drop = FALSE])
477   P <- length(pred_names)
478   Q <- length(resp_names)
479
480   fit <- tryCatch({
481     glmnet(
482       Xb, Yb,
483       family = "mgaussian",
```

```
483     alpha = alpha_opt,
484     lambda = lambda_opt,
485     standardize = TRUE,
486     standardize.response = TRUE,
487     thresh = 1e-5
488   )
489 }, error = function(e) NULL)
490
491 if (is.null(fit)) return(rep(0, P * Q))
492
493 coef_list <- coef(fit)
494 out <- numeric(P * Q)
495 k <- 0
496
497 for (r in resp_names) {
498   M <- as.matrix(coef_list[[r]])
499   beta_r <- M[setdiff(rownames(M), "(Intercept)"), 1, drop = TRUE]
500   beta_r <- as.numeric(beta_r)
501   names(beta_r) <- setdiff(rownames(M), "(Intercept)")
502
503   beta_r_aligned <- setNames(numeric(P), pred_names)
504   beta_r_aligned[names(beta_r)] <- beta_r
505
506   for (p in seq_len(P)) {
507     k <- k + 1
508     out[k] <- beta_r_aligned[p]
509   }
510 }
511 out
512 }
513
```

```
514 Datos_boot <- cbind(as.data.frame(Y_raw), as.data.frame(X_raw))
515 resp_names <- colnames(Y_raw)
516 pred_names <- colnames(X_raw)
517
518 run_bootstrap <- function(Datos_boot, R_boot, alpha_final, lambda_sel,
519                           resp_names, pred_names,
520                           parallel_ok = TRUE, max_cores = 6, seed = 2025) {
521   RNGkind("L'Ecuyer-CMRG")
522   set.seed(seed)
523
524   do_boot <- function(parallel_mode = FALSE, cl = NULL, ncpus = 1) {
525     boot(
526       data      = Datos_boot,
527       statistic = boot_coef_robust,
528       R         = R_boot,
529       alpha_opt = alpha_final,
530       lambda_opt = lambda_sel,
531       resp_names = resp_names,
532       pred_names = pred_names,
533       parallel   = if (parallel_mode) "snow" else "no",
534       ncpus      = ncpus,
535       cl         = cl
536     )
537   }
538
539   if (parallel_ok && R_boot >= 500) {
540     n_cores <- max(1, min(parallel::detectCores() - 1, max_cores))
541     cl <- parallel::makeCluster(n_cores, type = "PSOCK")
542     on.exit(try(parallel::stopCluster(cl), silent = TRUE), add = TRUE)
543
544     parallel::clusterEvalQ(cl, {
```

```
545     suppressPackageStartupMessages(library(glmnet))
546   })
547   parallel::clusterExport(
548     cl,
549     c("boot_coef_robust", "resp_names", "pred_names",
550       "alpha_final", "lambda_sel", "Datos_boot"),
551     envir = environment()
552   )
553   parallel::clusterSetRNGStream(cl, seed)
554
555   res <- try(do_boot(parallel_mode = TRUE, cl = cl, ncpus = n_cores),
556             silent = TRUE)
557   if (!inherits(res, "try-error")) return(res)
558   message("El modo paralelo no se completó correctamente; se usa modo
559   secuencial.")
560 }
561
562 do_boot(parallel_mode = FALSE, ncpus = 1)
563 }
564
565 R_boot <- 1000
566
567 boot_res <- run_bootstrap(
568   Datos_boot = Datos_boot,
569   R_boot      = R_boot,
570   alpha_final = alpha_final,
571   lambda_sel  = lambda_sel,
572   resp_names  = resp_names,
573   pred_names  = pred_names,
574   parallel_ok = TRUE,
575   max_cores   = 6,
```

```
574 seed = 2025
575 )
576
577 P <- length(pred_names)
578 Q <- length(resp_names)
579
580 coefs_min <- coef(fit_final)
581
582 coef_mat <- do.call(
583   cbind,
584   lapply(resp_names, function(r) {
585     M <- as.matrix(coefs_min[[r]])
586     v <- M[setdiff(rownames(M), "(Intercept)"), 1, drop = TRUE]
587     v_aligned <- setNames(numeric(P), pred_names)
588     v_aligned[names(v)] <- as.numeric(v)
589     as.numeric(v_aligned)
590   })
591 )
592 colnames(coef_mat) <- resp_names
593 rownames(coef_mat) <- pred_names
594
595 orig_coefs_vec <- as.numeric(coef_mat)
596
597 ci_lower <- numeric(P * Q)
598 ci_upper <- numeric(P * Q)
599
600 for (j in seq_len(P * Q)) {
601   if (all(boot_res$t[, j] == 0)) {
602     ci_lower[j] <- 0
603     ci_upper[j] <- 0
604   } else {
```



```
605     qs <- quantile(boot_res$t[, j], c(0.025, 0.975),
606                   na.rm = TRUE, names = FALSE)
607     ci_lower[j] <- qs[1]
608     ci_upper[j] <- qs[2]
609   }
610 }
611
612 mat_point <- matrix(
613   orig_coefs_vec,
614   nrow = P, ncol = Q, byrow = FALSE,
615   dimnames = list(pred_names, resp_names)
616 )
617
618 res_list <- vector("list", Q)
619 for (r in seq_len(Q)) {
620   idx_r <- ((r - 1) * P + 1):(r * P)
621   res_list[[r]] <- data.frame(
622     respuesta = resp_names[r],
623     predictor = pred_names,
624     estimate  = mat_point[, r],
625     CI_lower  = ci_lower[idx_r],
626     CI_upper  = ci_upper[idx_r],
627     signif    = (ci_lower[idx_r] * ci_upper[idx_r] > 0),
628     stringsAsFactors = FALSE
629   )
630 }
631
632 boot_simple_df <- dplyr::bind_rows(res_list)
633
634 sig_total <- sum(boot_simple_df$signif, na.rm = TRUE)
635 sig_por_Y <- boot_simple_df %>%
```

```
636   group_by(respuesta) %>%
637   summarise(signif_n = sum(signif, na.rm = TRUE), .groups = "drop")
638 sig_por_X <- boot_simple_df %>%
639   group_by(predictor) %>%
640   summarise(signif_en_respuestas = sum(signif, na.rm = TRUE), .groups = "drop
        ") %>%
641   arrange(desc(signif_en_respuestas))
642
643 openxlsx::write.xlsx(
644   list(
645     Coeficientes_bootstrap = boot_simple_df,
646     Resumen_significancia = data.frame(
647       Metricas = c("Lambda_used", "Coef_significativos", "Porcentaje_
        significativos"),
648       Valores = c(lambda_sel, sig_total, sig_total/(P * Q) * 100)
649     ),
650     Top_por_respuesta = sig_por_Y %>% arrange(desc(signif_n)),
651     Top_predictores = sig_por_X %>% head(20)
652   ),
653   file.path(ruta_salida_base, "bootstrap_resultados_lambda_min.xlsx"),
654   overwrite = TRUE
655 )
656
657 saveRDS(
658   list(
659     alpha_final = alpha_final,
660     lambda_sel = lambda_sel,
661     cv_final = cv_final,
662     fit_final = fit_final,
663     boot_results = boot_res,
664     boot_df = boot_simple_df
```

```
665   ),
666   file.path(ruta_salida_base, "objetos_modelo_final_lambda_min.rds")
667 )
668
669 message(sprintf("Bootstrap con lambda_min completado: %d coeficientes
    significativos", sig_total))
670
671 ## ===== Gráficos de coeficientes significativos =====
672
673 sig_coefs <- boot_simple_df %>%
674   filter(signif & predictor != "(Intercept)") %>%
675   mutate(
676     dependiente = dplyr::recode(respuesta, !!!etiquetas_dep_largo, .default =
        respuesta),
677     predictor    = as.character(predictor)
678   )
679
680 if (nrow(sig_coefs) > 0) {
681   sig_coefs <- sig_coefs %>%
682     mutate(predictor = reorder_within(predictor, estimate, dependiente))
683
684   g_coef_sig <- ggplot(sig_coefs, aes(x = predictor, y = estimate, fill =
        dependiente)) +
685     geom_col(show.legend = FALSE) +
686     geom_errorbar(aes(ymin = CI_lower, ymax = CI_upper), width = 0.2, alpha =
        0.7) +
687     facet_wrap(~ dependiente, scales = "free_y") +
688     scale_x_reordered() +
689     coord_flip() +
690     labs(
691       title    = "Coeficientes significativos del modelo Elastic Net
```

```
multivariado",
692   subtitle = "IC 95% por bootstrap (Alpha y lambda fijos en lambda_min)",
693   x         = "Marcador molecular",
694   y         = "Coeficiente"
695 ) +
696 theme_minimal(base_size = 12) +
697 theme(
698   plot.title    = element_text(face = "bold", size = 14, hjust = 0.5),
699   plot.subtitle = element_text(size = 12, hjust = 0.5),
700   strip.text    = element_text(face = "bold", size = 11),
701   axis.text.y   = element_text(size = 8)
702 )
703
704 ggsave(
705   file.path(ruta_salida_base, "coeficientes_significativos_bootstrap.png"),
706   plot = g_coef_sig, width = 12, height = 8, dpi = 300
707 )
708 }
709
710 top_predictors <- sig_por_X %>%
711   filter(signif_en_respuestas > 0) %>%
712   arrange(desc(signif_en_respuestas)) %>%
713   head(20) %>%
714   pull(predictor)
715
716 if (length(top_predictors) > 0) {
717   heatmap_data <- boot_simple_df %>%
718     filter(predictor %in% top_predictors) %>%
719     mutate(
720       dependiente = dplyr::recode(respuesta, !!!etiquetas_dep_largo, .default
       = respuesta)
```

```
721   )
722
723   g_heatmap <- ggplot(heatmap_data, aes(x = predictor, y = dependiente, fill
724     = estimate)) +
725     geom_tile() +
726     scale_fill_gradient2(low = "blue", high = "red", mid = "white",
727       midpoint = 0, name = "Coef.") +
728     labs(
729       title = "Heatmap de coeficientes - Top predictores",
730       x     = "Marcadores moleculares",
731       y     = "Variables de respuesta"
732     ) +
733     theme_minimal() +
734     theme(
735       axis.text.x = element_text(angle = 45, hjust = 1),
736       plot.title  = element_text(face = "bold", hjust = 0.5)
737     )
738
739   ggsave(
740     file.path(ruta_salida_base, "heatmap_coeficientes.png"),
741     plot = g_heatmap, width = 10, height = 6, dpi = 300
742   )
743 }
744
745 ## ===== Gráficos por respuesta usando etiquetas largas =====
746
747 label_es <- function(x) dplyr::recode(x, !!!etiquetas_dep, .default = x)
748 label_code <- (function() {
749   inv <- setNames(names(etiquetas_dep), etiquetas_dep)
750   function(x) dplyr::recode(x, !!!inv, .default = x)
751 })()
```

```
751
752 sig_coefs_all <- boot_simple_df %>%
753   filter(signif, predictor != "(Intercept)") %>%
754   mutate(
755     dependiente_es = label_es(respuesta),
756     predictor       = as.character(predictor)
757   )
758
759 plots_por_resp <- list()
760
761 if (nrow(sig_coefs_all) > 0) {
762   top_k <- Inf
763
764   sig_coefs_all <- sig_coefs_all %>%
765     group_by(dependiente_es) %>%
766     arrange(desc(abs(estimate)), .by_group = TRUE) %>%
767     slice_head(n = top_k) %>%
768     ungroup()
769
770   .slug <- function(x) gsub("[^A-Za-z0-9]+", "_", x)
771
772   for (dep_es in sort(unique(sig_coefs_all$dependiente_es))) {
773     d <- sig_coefs_all %>% filter(dependiente_es == dep_es)
774
775     d <- d %>%
776       mutate(predictor = factor(predictor, levels = d[order(estimate), "
777         predictor", drop = TRUE]))
778
779     ymin <- min(d$CI_lower, na.rm = TRUE)
780     ymax <- max(d$CI_upper, na.rm = TRUE)
781     if (!is.finite(ymin) || !is.finite(ymax)) {
```

```
781     ymin <- min(d$estimate)
782     ymax <- max(d$estimate)
783   }
784   rng <- ymax - ymin
785   pad <- if (is.finite(rng) && rng > 0) 0.08 * rng else 1
786   lims <- c(ymin - pad, ymax + pad)
787
788   g_dep <- ggplot(d, aes(x = predictor, y = estimate)) +
789     geom_col(fill = "#2ca02c", alpha = 0.75, width = 0.7) +
790     geom_errorbar(aes(ymin = CI_lower, ymax = CI_upper),
791                   width = 0.25, linewidth = 0.4) +
792     geom_hline(yintercept = 0, linetype = "dashed", linewidth = 0.35) +
793     coord_flip() +
794     scale_y_continuous(limits = lims, expand = expansion(mult = 0)) +
795     labs(
796       title = dep_es,
797       x      = "Marcador molecular",
798       y      = "Valor del coeficiente (escala original de Y)"
799     ) +
800     theme_minimal(base_size = 12) +
801     theme(
802       plot.title      = element_text(face = "bold", hjust = 0.5),
803       panel.grid.minor = element_blank(),
804       axis.text.y      = element_text(size = 9)
805     )
806
807   plots_por_resp[[dep_es]] <- g_dep
808   out_png <- file.path(ruta_salida_base, sprintf("coef_sig_%s.png", .slug(
809     dep_es)))
810   ggsave(out_png, plot = g_dep, width = 7.8, height = 5.2, dpi = 300)
811 }
```

```
811
812 combinar_4 <- function(vars, outfile, ncol = 2) {
813   vars_es <- label_es(label_code(vars))
814   plist <- lapply(vars_es, function(v) {
815     if (!v %in% names(plots_por_resp)) {
816       stop(sprintf("No existe gráfico para '%s'", v))
817     }
818     plots_por_resp[[v]]
819   })
820   while (length(plist) %% ncol != 0) plist <- c(plist, list(NULL))
821
822   combinado <- wrap_plots(plist, ncol = ncol, guides = "collect") +
823     plot_annotation(
824       title = "Coeficientes significativos del modelo Elastic Net
825       multivariado",
826       subtitle = "IC 95% por bootstrap",
827       theme = theme(
828         plot.title = element_text(face = "bold", size = 15, hjust = 0.5)
829       ,
830         plot.subtitle = element_text(size = 12, hjust = 0.5)
831       )
832     )
833   ggsave(outfile, plot = combinado, width = 12, height = 8, dpi = 300)
834   invisible(combinado)
835 }
836
837 ## =====
838 ## RDA-like: PCA sobre  $\hat{Y}$  del modelo EN multivariado final (lambda_min)
839 ## =====
```



```
840 id_filas <- if ("Accesion" %in% names(Datos)) {
841   Datos$Accesion[cc]
842 } else if ("Accession" %in% names(Datos)) {
843   Datos$Accession[cc]
844 } else {
845   seq_len(nrow(Y_raw))
846 }
847 stopifnot(length(id_filas) == nrow(Y_raw))
848
849 Yhat <- predict(fit_final, newx = X_raw, s = lambda_sel, type = "response")[,
      , 1]
850 dimnames(Yhat) <- dimnames(Y_raw)
851
852 openxlsx::write.xlsx(
853   list(
854     Y_original = as.data.frame(Y_raw),
855     Y_hat      = as.data.frame(Yhat)
856   ),
857   file.path(ruta_salida_base, "Y_y_Yhat_modelo_final_lambda_min.xlsx"),
858   overwrite = TRUE
859 )
860
861 pca_hat <- FactoMineR::PCA(as.data.frame(Yhat), scale.unit = TRUE, graph =
      FALSE)
862
863 scores_ind <- as.data.frame(pca_hat$ind$coord[, 1:2])
864 scores_ind$Accesion <- id_filas
865 colnames(scores_ind)[1:2] <- c("PC1", "PC2")
866
867 scores_var <- as.data.frame(pca_hat$var$coord[, 1:2])
868 colnames(scores_var) <- c("PC1", "PC2")
```

```
869 scores_var$var <- rownames(scores_var)
870 scores_var$var_es <- dplyr::recode(scores_var$var, !!!etiquetas_dep, .default
    = scores_var$var)
871
872 eig <- pca_hat$eig
873 pc1_var <- round(eig[1, 2], 1)
874 pc2_var <- round(eig[2, 2], 1)
875
876 openxlsx::write.xlsx(
877   list(
878     PCA_Yhat_eig      = as.data.frame(eig),
879     PCA_Yhat_varcoord = scores_var,
880     PCA_Yhat_indcoord = scores_ind
881   ),
882   file.path(ruta_salida_base, "PCA_Yhat_resumen.xlsx"),
883   overwrite = TRUE
884 )
885
886 if (!exists("boot_simple_df")) {
887   stop("No se encontró 'boot_simple_df'. Es necesario ejecutar antes el
    bloque de bootstrap.")
888 }
889
890 sig_markers <- boot_simple_df %>%
891   filter(signif, predictor != "(Intercept)") %>%
892   distinct(predictor) %>%
893   pull(predictor)
894
895 arrow_markers <- NULL
896 if (length(sig_markers) > 0) {
897   sig_markers <- intersect(sig_markers, colnames(X_raw))

```

```
898   if (length(sig_markers) > 0) {
899     Xsig_std <- scale(X_raw[, sig_markers, drop = FALSE])
900     S_ind    <- scale(as.matrix(scores_ind[, c("PC1", "PC2")]))
901     R <- cor(Xsig_std, S_ind, use = "pairwise.complete.obs")
902     arrow_markers <- as.data.frame(R)
903     arrow_markers$predictor <- rownames(arrow_markers)
904     rownames(arrow_markers) <- NULL
905     colnames(arrow_markers)[1:2] <- c("PC1", "PC2")
906     arrow_markers$norma <- sqrt(arrow_markers$PC1^2 + arrow_markers$PC2^2)
907     scale_mark <- 0.9
908     arrow_markers$PC1 <- arrow_markers$PC1 * scale_mark
909     arrow_markers$PC2 <- arrow_markers$PC2 * scale_mark
910     if (nrow(arrow_markers) > 25) {
911       arrow_markers <- arrow_markers %>%
912         arrange(desc(norma)) %>%
913         slice(1:25)
914     }
915   }
916 }
917
918 scores_var$PC1_s <- scores_var$PC1
919 scores_var$PC2_s <- scores_var$PC2
920
921 g_triplot <- ggplot() +
922   geom_point(data = scores_ind, aes(PC1, PC2), size = 2, alpha = 0.7) +
923   ggrepel::geom_text_repel(
924     data = scores_ind,
925     aes(PC1, PC2, label = Accesion),
926     size = 3, max.overlaps = 80, min.segment.length = 0
927   ) +
928   geom_segment(
```

```
929   data = scores_var,
930   aes(x = 0, y = 0, xend = PC1_s, yend = PC2_s),
931   arrow = arrow(length = unit(0.18, "cm")),
932   linewidth = 0.6,
933   alpha = 0.9
934 ) +
935 ggrepel::geom_text_repel(
936   data = scores_var,
937   aes(x = PC1_s, y = PC2_s, label = var_es),
938   size = 3.2, fontface = "bold", color = "grey20",
939   max.overlaps = Inf, min.segment.length = 0
940 ) +
941 {
942   if (!is.null(arrow_markers) && nrow(arrow_markers) > 0) {
943     geom_segment(
944       data = arrow_markers,
945       aes(x = 0, y = 0, xend = PC1, yend = PC2),
946       arrow = arrow(length = unit(0.14, "cm")),
947       linewidth = 0.4,
948       alpha = 0.7,
949       linetype = "dotted"
950     )
951   }
952 } +
953 {
954   if (!is.null(arrow_markers) && nrow(arrow_markers) > 0) {
955     ggrepel::geom_text_repel(
956       data = arrow_markers,
957       aes(x = PC1, y = PC2, label = predictor),
958       size = 2.7, color = "grey25",
959       max.overlaps = 60, min.segment.length = 0
```

```
960     )
961   }
962 } +
963 labs(
964   title = "RDA-like: PCA sobre  $\hat{Y}$  (Elastic Net multivariado)",
965   subtitle = sprintf("PC1 = %.1f%%, PC2 = %.1f%% | Marcadores
966   significativos (bootstrap): %d",
967     pc1_var, pc2_var,
968     ifelse(!is.null(arrow_markers), nrow(arrow_markers),
969     0)),
970   x = sprintf("PC1 (%.1f%% var)", pc1_var),
971   y = sprintf("PC2 (%.1f%% var)", pc2_var)
972 ) +
973 coord_equal() +
974 theme_minimal(base_size = 12) +
975 theme(
976   plot.title = element_text(face = "bold"),
977   panel.grid.minor = element_blank()
978 )
979
980 ggsave(
981   file.path(ruta_salida_base, "tripplot_RDA_like_Yhat_lambda_min.png"),
982   plot = g_tripplot, width = 10.5, height = 8.0, dpi = 300
983 )
984
985 colores <- c(
986   "Individuos" = "grey30",
987   "Variables respuesta" = "#1f77b4",
988   "Marcadores (sig.)" = "#d62728"
```

```
989 g_triplot2 <- ggplot() +
990   geom_point(
991     data = scores_ind,
992     aes(PC1, PC2, color = "Individuos"),
993     size = 2, alpha = 0.75, show.legend = TRUE
994   ) +
995   ggrepel::geom_text_repel(
996     data = scores_ind,
997     aes(PC1, PC2, label = Accesion, color = "Individuos"),
998     size = 3, max.overlaps = 80, min.segment.length = 0,
999     show.legend = FALSE
1000   ) +
1001   geom_segment(
1002     data = scores_var,
1003     aes(x = 0, y = 0, xend = PC1_s, yend = PC2_s, color = "Variables
1004 respuesta"),
1005     arrow = arrow(length = unit(0.18, "cm")),
1006     linewidth = 0.7, alpha = 0.95, show.legend = TRUE
1007   ) +
1008   ggrepel::geom_text_repel(
1009     data = scores_var,
1010     aes(x = PC1_s, y = PC2_s, label = var_es, color = "Variables respuesta"),
1011     size = 3.2, fontface = "bold", max.overlaps = Inf,
1012     min.segment.length = 0, show.legend = FALSE
1013   ) +
1014   {
1015     if (!is.null(arrow_markers) && nrow(arrow_markers) > 0) {
1016       geom_segment(
1017         data = arrow_markers,
1018         aes(x = 0, y = 0, xend = PC1, yend = PC2, color = "Marcadores (sig.)"
1019       ),
```

```
1018     arrow = arrow(length = unit(0.14, "cm")),
1019     linewidth = 0.5, alpha = 0.8, linetype = "dotted", show.legend = TRUE
1020   )
1021 }
1022 } +
1023 {
1024   if (!is.null(arrow_markers) && nrow(arrow_markers) > 0) {
1025     ggrepel::geom_text_repel(
1026       data = arrow_markers,
1027       aes(x = PC1, y = PC2, label = predictor, color = "Marcadores (sig.)")
1028     ,
1029     size = 2.7, max.overlaps = 60, min.segment.length = 0,
1030     show.legend = FALSE
1031   )
1032 } +
1033 scale_color_manual(name = NULL, values = colores) +
1034 labs(
1035   title = "RDA-like: PCA sobre  $\hat{Y}$  (Elastic Net multivariado)",
1036   subtitle = sprintf("PC1 = %.1f%%, PC2 = %.1f%% | Marcadores
1037     significativos (bootstrap): %d",
1038     pc1_var, pc2_var,
1039     ifelse(!is.null(arrow_markers), nrow(arrow_markers),
1040     0)),
1041   x = sprintf("PC1 (%.1f%% var)", pc1_var),
1042   y = sprintf("PC2 (%.1f%% var)", pc2_var)
1043 ) +
1044 coord_equal() +
1045 theme_minimal(base_size = 12) +
1046 theme(
1047   plot.title      = element_text(face = "bold"),
```

```
1046   legend.position = "top",
1047   panel.grid.minor = element_blank()
1048 )
1049
1050 ggsave(
1051   file.path(ruta_salida_base, "tripplot_RDA_like_Yhat_lambda_min_con_leyenda.
1052             png"),
1053   plot = g_tripplot2, width = 10.5, height = 8.0, dpi = 300
1054 )
1055 ## ===== Selección de todos los marcadores significativos (bootstrap) =====
1056
1057 marcadores_vec <- boot_simple_df %>%
1058   dplyr::filter(signif, predictor != "(Intercept)") %>%
1059   dplyr::distinct(predictor) %>%
1060   dplyr::pull(predictor)
1061
1062 if (length(marcadores_vec) == 0L) {
1063   warning("No se encontraron marcadores significativos por bootstrap; se
1064           omitirán los forest plot.")
1065 } else {
1066   message("Marcadores significativos seleccionados: ",
1067           paste(marcadores_vec, collapse = ", "))
1068 }
1069
1070 sig_por_X <- boot_simple_df %>%
1071   dplyr::filter(signif, predictor != "(Intercept)") %>%
1072   dplyr::group_by(predictor) %>%
1073   summarise(
1074     signif_en_respuestas = sum(signif, na.rm = TRUE),
1075     max_abs_est           = max(abs(estimate), na.rm = TRUE),
```



```
1075   .groups = "drop"
1076 ) %>%
1077 dplyr::arrange(
1078   dplyr::desc(signif_en_respuestas),
1079   dplyr::desc(max_abs_est)
1080 )
1081
1082 marcadores_vec <- intersect(sig_por_X$predictor, marcadores_vec)
1083
1084 ## ===== Forest plot por marcador (coeficientes estandarizados) =====
1085
1086 sd_y <- apply(Y_raw, 2, sd, na.rm = TRUE)
1087 sd_tab <- data.frame(respuesta = names(sd_y), sdY = as.numeric(sd_y))
1088
1089 df_std <- boot_simple_df %>%
1090   inner_join(sd_tab, by = "respuesta") %>%
1091   filter(predictor %in% marcadores_vec,
1092          predictor != "(Intercept)") %>%
1093   mutate(
1094     resp_es = dplyr::recode(respuesta, !!!etiquetas_dep_largo, .default =
1095       respuesta),
1096     est_std = estimate / sdY,
1097     CIlo_std = CI_lower / sdY,
1098     CIup_std = CI_upper / sdY,
1099     predictor = factor(predictor, levels = marcadores_vec)
1100   )
1101
1102 orden_resp <- df_std %>%
1103   group_by(resp_es) %>%
1104   summarise(score = max(abs(est_std), na.rm = TRUE), .groups = "drop") %>%
1105   arrange(desc(score)) %>%
```

```
1105   pull(resp_es)
1106
1107 df_std <- df_std %>%
1108   mutate(resp_es = factor(resp_es, levels = rev(orden_resp)))
1109
1110 lims <- range(c(df_std$CIlo_std, df_std$CIup_std), na.rm = TRUE)
1111 pad <- ifelse(diff(lims) > 0, 0.08 * diff(lims), 1)
1112 lims <- c(lims[1] - pad, lims[2] + pad)
1113
1114 g_forest_std <- ggplot(
1115   df_std,
1116   aes(y = resp_es, x = est_std, xmin = CIlo_std, xmax = CIup_std)
1117 ) +
1118   geom_vline(xintercept = 0, linetype = "dashed",
1119             linewidth = 0.4, color = "grey30") +
1120   geom_errorbarh(height = 0.22, alpha = 0.9) +
1121   geom_point(size = 2.7) +
1122   facet_grid(. ~ predictor, scales = "fixed", space = "fixed") +
1123   scale_x_continuous(limits = lims, expand = expansion(mult = 0.02)) +
1124   labs(
1125     title      = "Forest plot por marcador (coeficientes estandarizados por sd(
1126     Y))",
1127     subtitle = "Comparables entre respuestas",
1128     x          = "Coeficiente por 1 sd de Y",
1129     y          = "Respuesta"
1130   ) +
1131   theme_minimal(base_size = 12) +
1132   theme(
1133     plot.title      = element_text(face = "bold", hjust = 0.5),
1134     strip.text.x     = element_text(face = "bold"),
1135     panel.grid.minor = element_blank(),
```

```
1135     axis.text.y      = element_text(size = 9)
1136   )
1137
1138   ggsave(
1139     file.path(ruta_salida_base, "forestplot_estandarizado_por_marcador.png"),
1140     g_forest_std, width = 9.5, height = 6.5, dpi = 300
1141   )
1142
1143   ## ===== Forest plot por grupos biológicos =====
1144
1145   mapa_grupos <- c(
1146     "EMM"      = "Fenología",
1147     "SBF"      = "Fenología",
1148     "SFL"      = "Fenología",
1149     "SMF"      = "Fenología",
1150     "HEIGHT"   = "Morf. vegetativa",
1151     "LEAFLEN"  = "Morf. vegetativa",
1152     "LEAFWI"   = "Morf. vegetativa",
1153     "PETLEN"   = "Morf. vegetativa",
1154     "COTLEN"   = "Morf. vegetativa",
1155     "HYPLEN"   = "Morf. vegetativa",
1156     "PANLEN"   = "Morf. reproductiva",
1157     "GLOLEN"   = "Morf. reproductiva",
1158     "GRADIAM"  = "Morf. reproductiva",
1159     "STD"      = "Anatómicas/forma",
1160     "STL"      = "Anatómicas/forma",
1161     "TEENUMB"  = "Anatómicas/forma"
1162   )
1163
1164   sd_y <- apply(Y_raw, 2, sd, na.rm = TRUE)
1165   sd_df <- data.frame(resp_code = names(sd_y), sdY = as.numeric(sd_y))
```

```
1166
1167 df_signif_std <- boot_simple_df %>%
1168   filter(signif, predictor != "(Intercept)", predictor %in% marcadores_vec)
1169   %>%
1170   mutate(resp_code = respuesta) %>%
1171   left_join(sd_df, by = "resp_code") %>%
1172   mutate(
1173     grupo      = dplyr::recode(resp_code, !!!mapa_grupos, .default = "Otros")
1174     ,
1175     resp_label = dplyr::recode(resp_code, !!!etiquetas_dep_largo, .default =
1176     resp_code),
1177     est_std    = estimate / sdY,
1178     lo_std     = CI_lower / sdY,
1179     hi_std     = CI_upper / sdY,
1180     signo      = ifelse(est_std >= 0, "positivo", "negativo")
1181   )
1182
1183 niv_grupo <- c(
1184   "Fenología",
1185   "Morf. vegetativa",
1186   "Morf. reproductiva",
1187   "Anatómicas/forma",
1188   "Otros"
1189 )
1190
1191 df_signif_std <- df_signif_std %>%
1192   mutate(
1193     grupo = factor(grupo, levels = niv_grupo),
1194     resp_label = factor(
1195       resp_label,
1196       levels = rev(unique(resp_label[order(grupo, resp_label)]))
1197     )
1198   )
```

```
1194   )
1195   )
1196
1197 colores_marc <- c(
1198   "QGA02_186" = "#1b9e77",
1199   "Q084_165"  = "#7570b3",
1200   "Q026_176"  = "#d95f02"
1201 )
1202
1203 lims <- range(c(df_signif_std$lo_std, df_signif_std$hi_std), na.rm = TRUE)
1204 pad  <- ifelse(diff(lims) > 0, 0.08 * diff(lims), 1)
1205 lims <- c(lims[1] - pad, lims[2] + pad)
1206
1207 g_bio <- ggplot(
1208   df_signif_std,
1209   aes(x = est_std, y = resp_label, color = predictor, shape = signo)
1210 ) +
1211   geom_errorbarh(
1212     aes(xmin = lo_std, xmax = hi_std),
1213     height = 0.0, alpha = 0.55
1214   ) +
1215   geom_point(size = 2.6) +
1216   geom_vline(xintercept = 0, linetype = "dashed", linewidth = 0.5) +
1217   scale_x_continuous(limits = lims, expand = expansion(mult = 0.02)) +
1218   scale_color_manual(name = "Marcador", values = colores_marc) +
1219   scale_shape_manual(name = "Signo", values = c(negativo = 17, positivo = 16)
1220   ) +
1221   labs(
1222     title      = "Forest plot por grupo biológico y marcador",
1223     subtitle   = "Coeficientes estandarizados por sd(Y)",
1224     x          = "Coeficiente por 1 sd de Y",
```

```
1224     y           = "Respuesta"
1225   ) +
1226   facet_grid(
1227     rows  = vars(grupo),
1228     cols  = vars(predictor),
1229     scales = "free_y",
1230     space  = "free_y",
1231     switch = "y",
1232     labeller = labeller(grupo = label_wrap_gen(12))
1233   ) +
1234   theme_minimal(base_size = 12) +
1235   theme(
1236     strip.placement  = "outside",
1237     strip.position   = "left",
1238     strip.text.y.left = element_text(
1239       angle = 0,
1240       face  = "bold",
1241       margin = margin(r = 6)
1242     ),
1243     strip.text.x      = element_text(face = "bold"),
1244     panel.spacing.y   = unit(0.8, "lines"),
1245     legend.position   = "top",
1246     panel.grid.minor  = element_blank()
1247   )
1248
1249 n_resp <- length(unique(df_signif_std$resp_label))
1250 alto   <- max(7, 5 + 0.20 * n_resp)
1251
1252 ggsave(
1253   file.path(ruta_salida_base, "forest_grupo_biologico_std.png"),
1254   g_bio, width = 14, height = alto, dpi = 300
```

```

1255 )
1256
1257 message("Análisis completado. Resultados en: ", ruta_salida_base)

```

Código A.2: RDA-like multivariado *Elastic Net*

```

1 #####
2 ## RDA_mixto.R - PCA (RDA-like) sobre [ $\hat{Y}$ _cuali,  $\hat{Y}$ _cuanti]
3 ## - Entrada: X, Yhat_cuali, Yhat_cuanti
4 ## - SNPs significativos por bloque: sig_cuali, sig_cuanti
5 ## - Opcional: dat con columnas Accesion/Accession y (si existe) Grupo
6 ## - Salidas: Triplot PNG/TIFF + Excel con eig, scores, loadings y
   correlaciones SNP-PC
7 #####
8
9 rm(list = ls())
10 graphics.off()
11
12 suppressPackageStartupMessages({
13   library(FactoMineR)
14   library(ggplot2)
15   library(ggrepel)
16   library(ggnewscale)
17   library(openxlsx)
18 })
19
20 ## ----- Helpers -----
21 safe_scale <- function(M, center = TRUE, scale = TRUE, eps = 1e-8) {
22   M <- as.matrix(M)
23   if (center) M <- sweep(M, 2, colMeans(M, na.rm = TRUE), "-")
24   if (scale) {
25     sds <- apply(M, 2, sd)

```

```
26     sds[is.na(sds)] <- 0
27     sds <- ifelse(sds < eps, 1, sds)
28     M <- sweep(M, 2, sds, "/")
29   }
30   M
31 }
32
33 corr_con_scores <- function(M_vars_std, scores_2pc) {
34   R <- cor(M_vars_std, scores_2pc, use = "pairwise.complete.obs")
35   out <- data.frame(
36     PC1 = R[, 1],
37     PC2 = R[, 2],
38     row.names = rownames(R),
39     check.names = FALSE
40   )
41   out$norma <- sqrt(out$PC1^2 + out$PC2^2)
42   out
43 }
44
45 circle_df <- function(r, n = 200) {
46   t <- seq(0, 2 * pi, length.out = n)
47   data.frame(x = r * cos(t), y = r * sin(t))
48 }
49
50 ## ----- Función principal -----
51 RDA_mixto <- function(
52   X,                      # n x p (SNPs)
53   Yhat_cuali,             # n x Kc (probabilidades / puntajes
54   cualitativos)           # n x Kq (predicciones cuantitativas)
55   Yhat_cuanti,            # n x Kq (predicciones cuantitativas)
56   sig_cuali = character(0), # SNPs significativos (cuali)
```



```
56   sig_cuanti = character(0), # SNPs significativos (cuanti)
57   dat = NULL,                # opcional: data.frame con Accesion(/Accession
                               ) y/o Grupo
58   etiquetas_cuali = NULL,    # named vector: etiquetas para columnas de
                               Yhat_cuali
59   etiquetas_cuanti = NULL,   # idem para Yhat_cuanti
60   DEST = getwd(),
61   balance_blocks = TRUE,     # ponderar cada bloque para contribución
                               comparable
62   top_snps_por_bloque = 40,  # máximo de SNPs por bloque en el gráfico
63   seed_val = 123
64 ) {
65   set.seed(seed_val)
66
67   stopifnot(nrow(X) == nrow(Yhat_cuali), nrow(X) == nrow(Yhat_cuanti))
68
69   ## --- Alineación de filas por rownames (si existen) ---
70   if (!is.null(rownames(X))) {
71     rn <- rownames(X)
72     if (!is.null(rownames(Yhat_cuali)))
73       Yhat_cuali <- Yhat_cuali[rn, , drop = FALSE]
74     if (!is.null(rownames(Yhat_cuanti)))
75       Yhat_cuanti <- Yhat_cuanti[rn, , drop = FALSE]
76     if (!is.null(dat) && !is.null(rownames(dat)))
77       dat <- dat[rn, , drop = FALSE]
78   }
79
80   ## --- Filtrado de varianza y estandarización por columna ---
81   eps <- 1e-8
82   keep_cuali <- apply(Yhat_cuali, 2, sd) > eps
83   keep_cuanti <- apply(Yhat_cuanti, 2, sd) > eps
```

```
84
85 if (!any(keep_cuali)) stop("Todas las columnas de Yhat_cuali tienen var ~
    =0.")
86 if (!any(keep_cuanti)) stop("Todas las columnas de Yhat_cuanti tienen var ~
    =0.")
87
88 Yc <- safe_scale(Yhat_cuali[, keep_cuali, drop = FALSE], TRUE, TRUE, eps)
89 Ya <- safe_scale(Yhat_cuanti[, keep_cuanti, drop = FALSE], TRUE, TRUE, eps)
90
91 ## --- Balanceo por bloque (opcional) ---
92 if (balance_blocks) {
93   tr_c <- sum(apply(Yc, 2, var))
94   tr_a <- sum(apply(Ya, 2, var))
95   k <- ncol(Yc) + ncol(Ya)
96   w_c <- sqrt(k / (2 * ncol(Yc))) * ifelse(tr_c > eps, 1 / sqrt(tr_c), 1)
97   w_a <- sqrt(k / (2 * ncol(Ya))) * ifelse(tr_a > eps, 1 / sqrt(tr_a), 1)
98   Yc <- Yc * w_c
99   Ya <- Ya * w_a
100 }
101
102 ## --- Matriz combinada y PCA ---
103 Ymix <- cbind(Yc, Ya)
104 colnames(Ymix) <- c(colnames(Yc), colnames(Ya))
105
106 pca <- FactoMineR::PCA(Ymix, scale.unit = FALSE, graph = FALSE)
107
108 eig_tbl <- pca$eig
109 ind_coord <- as.data.frame(pca$ind$coord[, 1:2])
110 colnames(ind_coord) <- c("PC1", "PC2")
111 var_coord <- as.data.frame(pca$var$coord[, 1:2])
112 colnames(var_coord) <- c("PC1", "PC2")
```

```
113
114 pc1_lab <- paste0("PC1 (", round(eig_tbl[1, 2], 1), "%)")
115 pc2_lab <- paste0("PC2 (", round(eig_tbl[2, 2], 1), "%)")
116
117 ## --- Etiquetas de individuos (Accesión) y Grupo ---
118 scores <- ind_coord
119
120 if (!is.null(dat)) {
121   if ("Grupo" %in% names(dat)) {
122     scores$Grupo <- dat[rownames(scores), "Grupo", drop = TRUE]
123   } else {
124     scores$Grupo <- factor("Sin grupo")
125   }
126
127   cand <- c(
128     "Accesion", "Accession", "accesion", "accession",
129     "Genotipo", "genotype", "ID", "id", "Nombre", "name"
130   )
131   pos <- which(names(dat) %in% cand)
132
133   if (length(pos) >= 1) {
134     scores$ID <- as.character(dat[rownames(scores), pos[1], drop = TRUE])
135   } else {
136     scores$ID <- if (!is.null(rownames(X))) rownames(X) else as.character(
137       seq_len(nrow(X)))
138   }
139 } else {
140   scores$Grupo <- factor("Sin grupo")
141   scores$ID <- if (!is.null(rownames(X))) rownames(X) else as.character(
142     seq_len(nrow(X)))
143 }
```

```
142
143 ## --- Flechas de variables (respuestas) por origen ---
144 p_c <- ncol(Yc)
145 vars_cuali <- rownames(pca$var$coord)[1:p_c]
146 vars_cuanti <- rownames(pca$var$coord)[(p_c + 1):(p_c + ncol(Ya))]
147
148 arrows_resp_cuali <- data.frame(
149   Var    = vars_cuali,
150   xend   = var_coord$PC1[match(vars_cuali, rownames(var_coord))],
151   yend   = var_coord$PC2[match(vars_cuali, rownames(var_coord))],
152   Origen = "Respuesta (cuali)",
153   stringsAsFactors = FALSE
154 )
155
156 arrows_resp_cuanti <- data.frame(
157   Var    = vars_cuanti,
158   xend   = var_coord$PC1[match(vars_cuanti, rownames(var_coord))],
159   yend   = var_coord$PC2[match(vars_cuanti, rownames(var_coord))],
160   Origen = "Respuesta (cuanti)",
161   stringsAsFactors = FALSE
162 )
163
164 arrows_resp_cuali$Var_lab <- if (!is.null(etiquetas_cuali)) {
165   dplyr::recode(arrows_resp_cuali$Var, !!!etiquetas_cuali, .default =
166     arrows_resp_cuali$Var)
167 } else {
168   arrows_resp_cuali$Var
169 }
170
171 arrows_resp_cuanti$Var_lab <- if (!is.null(etiquetas_cuanti)) {
172   dplyr::recode(arrows_resp_cuanti$Var, !!!etiquetas_cuanti, .default =
```

```
    arrows_resp_cuanti$Var)
172 } else {
173     arrows_resp_cuanti$Var
174 }
175
176 ## --- Flechas moleculares por origen (correlación con scores de individuos
    ) ---
177 S <- as.matrix(ind_coord[, 1:2, drop = FALSE])
178
179 get_arrows_mol <- function(sig_vec, etiqueta_origen) {
180     sig_vec <- intersect(sig_vec, colnames(X))
181     if (length(sig_vec) == 0) return(NULL)
182     Xs <- scale(X[, sig_vec, drop = FALSE])
183     C <- corr_con_scores(Xs, S)
184     C$SNP <- rownames(C)
185     C$Origen <- etiqueta_origen
186     C[order(-C$norma), , drop = FALSE]
187 }
188
189 A_cuali <- get_arrows_mol(sig_cuali, "Molecular (cuali)")
190 A_cuanti <- get_arrows_mol(sig_cuanti, "Molecular (cuanti)")
191
192 if (!is.null(A_cuali) && nrow(A_cuali) > top_snps_por_bloque)
193     A_cuali <- A_cuali[1:top_snps_por_bloque, ]
194 if (!is.null(A_cuanti) && nrow(A_cuanti) > top_snps_por_bloque)
195     A_cuanti <- A_cuanti[1:top_snps_por_bloque, ]
196
197 arrows_mol <- NULL
198
199 if (!is.null(A_cuali)) {
200     arrows_mol <- rbind(
```

```
201     arrows_mol,
202     data.frame(
203       Var      = A_cuali$SNP,
204       xend     = A_cuali$PC1,
205       yend     = A_cuali$PC2,
206       Origen   = A_cuali$Origen,
207       Var_lab  = A_cuali$SNP,
208       check.names = FALSE
209     )
210   )
211 }
212
213 if (!is.null(A_cuanti)) {
214   arrows_mol <- rbind(
215     arrows_mol,
216     data.frame(
217       Var      = A_cuanti$SNP,
218       xend     = A_cuanti$PC1,
219       yend     = A_cuanti$PC2,
220       Origen   = A_cuanti$Origen,
221       Var_lab  = A_cuanti$SNP,
222       check.names = FALSE
223     )
224   )
225 }
226
227 ## --- Escalado conjunto de flechas y círculo de correlación ---
228 df_arrows <- rbind(
229   arrows_resp_cuali,
230   arrows_resp_cuanti,
231   if (!is.null(arrows_mol)) arrows_mol else NULL
```

```
232 )
233
234 arrow_mat <- as.matrix(df_arrows[, c("xend", "yend")])
235 max_arrow <- max(sqrt(rowSums(arrow_mat^2)))
236 max_scores <- max(abs(c(scores$PC1, scores$PC2)))
237 scale_fac <- 0.9 * max_scores / max_arrow
238
239 df_arrows$xend <- df_arrows$xend * scale_fac
240 df_arrows$yend <- df_arrows$yend * scale_fac
241 circ <- circle_df(scale_fac)
242
243 df_arrows$Origen <- factor(
244   df_arrows$Origen,
245   levels = c(
246     "Respuesta (cuali)",
247     "Respuesta (cuanti)",
248     "Molecular (cuali)",
249     "Molecular (cuanti)"
250   )
251 )
252
253 ## --- Limpieza numérica para ploteo robusto ---
254 is_fin <- function(x) is.finite(x) & !is.na(x)
255 scores <- subset(scores, is_fin(PC1) & is_fin(PC2))
256 df_arrows <- subset(df_arrows, is_fin(xend) & is_fin(yend))
257
258 shrink <- function(x, y, s = 0.95) {
259   r <- sqrt(x^2 + y^2)
260   r[r == 0] <- 1
261   list(x = x * s, y = y * s)
262 }
```

```
263 tmp <- shrink(df_arrows$xend, df_arrows$yend, s = 0.95)
264 df_arrows$xend <- tmp$x
265 df_arrows$yend <- tmp$y
266
267 R <- max(abs(c(scores$PC1, scores$PC2, df_arrows$xend, df_arrows$yend)))
268 if (!is.finite(R) || R <= 0) R <- 1
269 xlim_fix <- c(-1.05 * R, 1.05 * R)
270 ylim_fix <- c(-1.05 * R, 1.05 * R)
271
272 df_arrows$norma_plot <- sqrt(df_arrows$xend^2 + df_arrows$yend^2)
273 umbral_norma <- 0.02 * R
274 df_arrows$plot_label <- df_arrows$norma_plot >= umbral_norma
275
276 ## --- Triplot mixto ---
277 p_mix <- ggplot() +
278   geom_point(
279     data = scores,
280     aes(PC1, PC2, color = Grupo),
281     size = 2, alpha = 0.85
282   ) +
283   scale_color_discrete(name = "Grupo") +
284   ggnewscale::new_scale_color() +
285   geom_hline(yintercept = 0, linewidth = 0.3, linetype = 2, color = "grey40") +
286   geom_vline(xintercept = 0, linewidth = 0.3, linetype = 2, color = "grey40") +
287   geom_path(data = circ, aes(x, y), linewidth = 0.3, color = "grey70") +
288   geom_segment(
289     data = df_arrows,
290     aes(x = 0, y = 0, xend = xend, yend = yend, color = Origen),
291     arrow = arrow(length = unit(0.25, "cm")),
```



```
292     linewidth = 0.7, alpha = 0.95
293   ) +
294   ggrepel::geom_text_repel(
295     data = transform(df_arrows, lab = ifelse(is.na(Var_lab), Var, Var_lab))
296   ,
297     aes(x = xend, y = yend, label = lab, color = Origen),
298     size = 3.1, segment.size = 0.2,
299     max.overlaps = 120, seed = 1, box.padding = 0.28
300   ) +
301   coord_equal(xlim = xlim_fix, ylim = ylim_fix) +
302   labs(
303     title = "RDA-like mixto: PCA sobre [ $\hat{Y}_{cuali}$ ,  $\hat{Y}_{cuanti}$ ]",
304     subtitle = "Puntos: accesiones; flechas: respuestas (cuali/cuanti) y
305     SNPs significativos por origen",
306     x = pc1_lab, y = pc2_lab, color = NULL
307   ) +
308   theme_minimal(base_size = 12) +
309   theme(
310     legend.position = "top",
311     panel.grid.minor = element_blank()
312   )
313
314 p_mix <- p_mix +
315   ggrepel::geom_text_repel(
316     data = scores,
317     aes(PC1, PC2, label = ID),
318     size = 3, color = "grey20",
319     box.padding = 0.25, point.padding = 0.25,
320     segment.size = 0.25, max.overlaps = Inf, seed = 1
321   )
```

```
321 print(p_mix)
322
323 ## --- Guardado de resultados ---
324 if (!dir.exists(DEST)) dir.create(DEST, recursive = TRUE)
325 ggsave(
326   file.path(DEST, "Triplot_RDA_mixto.png"),
327   p_mix, width = 12, height = 8, dpi = 600
328 )
329 ggsave(
330   file.path(DEST, "Triplot_RDA_mixto.tiff"),
331   p_mix, width = 12, height = 8, dpi = 600, units = "in"
332 )
333
334 wb <- createWorkbook()
335 addWorksheet(wb, "eig")
336 writeData(wb, "eig", eig_tbl, rowNames = FALSE)
337
338 addWorksheet(wb, "ind_coord")
339 writeData(
340   wb, "ind_coord",
341   cbind(id = rownames(pca$ind$coord), pca$ind$coord),
342   rowNames = FALSE
343 )
344
345 addWorksheet(wb, "var_coord")
346 writeData(
347   wb, "var_coord",
348   cbind(variable = rownames(pca$var$coord), pca$var$coord),
349   rowNames = FALSE
350 )
351
```

```
352 addWorksheet(wb, "arrows")
353 writeData(wb, "arrows", df_arrows, rowNames = FALSE)
354
355 if (!is.null(A_cuali)) {
356   addWorksheet(wb, "corr_SNP_PC_cuali")
357   writeData(
358     wb, "corr_SNP_PC_cuali",
359     A_cuali[, c("SNP", "PC1", "PC2", "norma", "Origen")],
360     rowNames = FALSE
361   )
362 }
363 if (!is.null(A_cuanti)) {
364   addWorksheet(wb, "corr_SNP_PC_cuanti")
365   writeData(
366     wb, "corr_SNP_PC_cuanti",
367     A_cuanti[, c("SNP", "PC1", "PC2", "norma", "Origen")],
368     rowNames = FALSE
369   )
370 }
371
372 saveWorkbook(
373   wb,
374   file.path(DEST, "RDA_mixto_resultados.xlsx"),
375   overwrite = TRUE
376 )
377
378 invisible(list(
379   p      = p_mix,
380   eig    = eig_tbl,
381   scores = scores,
382   var_coord = var_coord,
```

```
383     arrows    = df_arrows  
384   )  
385 }
```

Código A.3: RDA-*like* Mixto

Referencias

- Abdessemed, L., & Escofier, B. (1996). Analyse factorielle multiple de tableaux de fréquences: comparaison avec l'analyse canonique des correspondances. *Journal de la Société de Statistique de Paris*, 137, 3-18.
- Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS regression). *WIREs Computational Statistics*, 2, 97-106. <https://doi.org/10.1002/wics.51>
- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*, 2, 433-459. <https://doi.org/10.1002/wics.101>
- Acharjee, A., Finkers, R., Visser, R. G. F., & Maliepaard, C. (2013). Comparison of regularized regression methods for omics data. *Metabolomics: Open Access*, 3, 126. <https://doi.org/10.4172/2153-0769.1000126>
- Acquaah, G. (2012). *Plant breeding: Principles and methods* (3rd). Wiley-Blackwell.
- Agresti, A. (2002). *Categorical data analysis* (2nd). Wiley.
- Agresti, A. (2013). *Categorical Data Analysis* (3.^a ed.). Wiley.
- Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16, 125-127. <https://doi.org/10.1080/00401706.1974.10489157>
- Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3.^a ed.). Wiley.
- Auditoría General de la Nación. (2019). *Informe de auditoría N.º 64: Gestión de los recursos genéticos para la alimentación y la agricultura*. Auditoría General de la Nación. https://www.agn.gob.ar/sites/default/files/informes/informe_064_2019.pdf
- Bartlett, M. S. (1950). Tests of significance in factor analysis. *British journal of Statistical Psychology*, 3, 77-85. <https://doi.org/10.1111/j.2044-8317.1950.tb00285.x>

- Bartlett, M. S. (1951). The effect of standardization on a χ^2 approximation in factor analysis. *Biometrika*, 38, 337-344.
- Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on Imaging Sciences*, 2, 183-202. <https://doi.org/10.1137/080716542>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *journal of the Royal Statistical Society: Series B (Methodological)*, 57, 289-300.
- Benjamini, Y., & Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian journaltitle of Statistics*, 24, 407-418.
- Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29, 1165-1188. <https://doi.org/10.1214/aos/1013699998>
- Bergmann, B., & Hommel, G. (1988). Improvements of general multiple test procedures for redundant systems of hypotheses. En P. Bauer, G. Hommel & E. Sonnemann (Eds.), *Multiple Hypotheses Testing* (pp. 100-115). Springer.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *journal of the Royal Statistical Society: Series B*, 36, 192-236.
- Bi, W., Zhou, W., Dey, R., Mukherjee, B., Sampson, J. N., & Lee, S. (2021). Efficient mixed model approach for large-scale genome-wide association studies of ordinal categorical phenotypes. *The American journal of Human Genetics*, 108, 825-839. <https://doi.org/10.1016/j.ajhg.2021.03.019>
- Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. En *Studi in onore del professore Salvatore Ortu Carboni* (pp. 13-60).
- Boulesteix, A.-L., & Strimmer, K. (2007). Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8, 32-44. <https://doi.org/10.1093/bib/bbl016>
- Bramardi, S., Costa Tártara, S., & Curti, R. (2015). Mesa redonda: "Mejoramiento genético" [Mesa redonda desarrollada entre el 6 y el 9 de octubre]. *Actas del 1.º Congreso Argentino de Estadística, XLIII Coloquio de la Sociedad*

- Argentina de Estadística y XX Reunión Científica del Grupo Argentino de Biometría.*
- Bramardi, S., Costa Tártara, S., & Curti, R. (2017). Métodos de análisis multivariado a tres vías para identificación de asociaciones entre caracteres fenotípicos y marcadores moleculares [17–20 de octubre de 2017]. *Libro de resúmenes del XLV Coloquio Argentino de Estadística.*
- Bramardi, S., & Lavalle, A. (2021). *Apuntes del curso Métodos estadísticos para el análisis de datos en estructuras de tres vías* [Apuntes de curso, Doctorado en Estadística, Universidad Nacional de Rosario].
- Breseghello, F., & Sorrells, M. E. (2006). Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics*, 172, 1165-1177. <https://doi.org/10.1534/genetics.105.044586>
- Capblancq, T., & Forester, B. R. (2021). Redundancy analysis: A Swiss army knife for landscape genomics. *Methods in Ecology and Evolution*, 12, 2298-2309. <https://doi.org/10.1111/2041-210X.13722>
- Cheng, J., Levina, E., Wang, P., & Zhu, J. (2014). Sparse Ising models with covariates. *Biometrics*, 70, 943-953. <https://doi.org/10.1111/biom.12202>
- Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researchers* (2nd). Wiley.
- Cheverud, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, 87, 52-58.
- Christidis, A.-A., Lakshmanan, L., Smucler, E., & Zamar, R. (2020). Split regularized regression. *Technometrics*. <https://doi.org/10.1080/00401706.2019.1635533>
- Claros Díaz, M. G. (1998). Marcadores moleculares: Qué son, cómo se obtienen y para qué valen. *Encuentros en la Biología*, 49.
- Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., & Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, 142, 169-196. <https://doi.org/10.1007/s10681-005-1681-5>

- Copelli, S. B. (2010). *Genética: Desde la herencia a la manipulación de los genes*. Fundación Azara. Consultado el 11 de octubre de 2023, desde <https://www.fundacionazara.org.ar/img/libros/genetica.pdf>
- Costa Tártara, S. M., Manifesto, M. M., Bramardi, S. J., & Bertero, H. D. (2012). Genetic structure in cultivated quinoa (*Chenopodium quinoa* Willd.), a reflection of landscape structure in Northwest Argentina. *Conservation Genetics*, 13, 1027-1038. <https://doi.org/10.1007/s10592-012-0350-1>
- Cuadras, J. M. (2019). *Nuevos métodos de análisis multivariante*. Universidad de Barcelona.
- Curti, R. N., Andrade, A., Bramardi, S., Velásquez, B., & Bertero, H. D. (2012). Ecogeographic structure of phenotypic diversity in cultivated populations of quinoa from Northwest Argentina. *Annals of Applied Biology*, 160, 114-121. <https://doi.org/10.1111/j.1744-7348.2011.00524.x>
- Dai, B., Ding, S., & Wahba, G. (2013). Multivariate Bernoulli distribution. *Bernoulli*, 19, 1465-1483. <https://doi.org/10.3150/12-BEJSP10>
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511802843>
- Demidenko, E. (2004). *Mixed models: Theory and applications*. John Wiley & Sons.
- Dobson, A. J. (2002). *An Introduction to Generalized Linear Models* (2nd). Chapman & Hall/CRC.
- Dray, S., Dufour, A.-B., & Chessel, D. (2007). The ade4 package – II: Two-table and K-table methods. *R News*, 7, 47-52. https://cran.r-project.org/doc/Rnews/Rnews_2007-2.pdf
- Eckart, C., & Young, G. (1936). The Approximation of One Matrix by Another of Lower Rank. *Psychometrika*, 1, 211-218.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7, 1-26. <https://doi.org/10.1214/aos/1176344552>
- Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

- Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32, 407-499. <https://doi.org/10.1214/009053604000000067>
- Eizenga, G. C., Ali, M. L., Bryant, R. J., Yeater, K. M., McClung, A. M., & McCouch, S. R. (2014). Registration of the rice diversity panel 1 for genomewide association studies. *Journal of Plant Registrations*, 8(2), 109-116. <https://doi.org/10.3198/jpr2013.03.0013crmp>
- Enciclopedia Colaborativa Cubana (ECURED). (2020). *Selección asistida por marcadores*. https://www.ecured.cu/Selecci%C3%B3n_asistida_por_marcadores
- Engels, J. M. M., & Visser, L. (Eds.). (2003). *A guide to effective management of germplasm collections*. IPGRI.
- Escofier, B., & Pagès, J. (1992). *Analyses factorielles simples et multiples: Objectifs, méthodes et interprétation*. Dunod.
- Escofier, B., & Pagès, J. (2008). *Analyses factorielles simples et multiples: objectifs, méthodes et interprétation* (4.^a ed.). Dunod.
- Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, 29, 751-760. <https://www.jstor.org/stable/2529140>
- Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models* (2nd). Chapman; Hall/CRC.
- Flint-Garcia, S. A., Thornsberry, J. M., & Buckler, E. S. (2003). Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology*, 54, 357-374. <https://doi.org/10.1146/annurev.arplant.54.031902.134907>
- Food and Agriculture Organization of the United Nations. (2008). *Informe nacional sobre el estado de los recursos fitogenéticos para la agricultura y la alimentación: Argentina*. <https://www.fao.org/4/i1500e/argentina.pdf>
- Food and Agriculture Organization of the United Nations. (2010). *The second report on the state of the world's plant genetic resources for food and agriculture*. FAO. <https://www.fao.org/3/i1500e/i1500e00.htm>

- Food and Agriculture Organization of the United Nations. (2012). *Phenotypic characterization of animal genetic resources*. <https://www.fao.org/3/i2686e/i2686e00.htm>
- Foster, D. P., & Stine, R. A. (2008). α -investing: a procedure for sequential control of expected false discoveries. *journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 429-444.
- Friedman, J., Hastie, T., Hoefling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1, 302-332. <https://doi.org/10.1214/07-AOAS131>
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *journal of Statistical Software*, 33, 1-22. <https://doi.org/10.18637/jss.v033.i01>
- Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *journal of Computational and Graphical Statistics*, 7, 397-416. <https://doi.org/10.1080/10618600.1998.10474784>
- Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. *journal of the American Statistical Association*, 70, 320-328. <https://doi.org/10.1080/01621459.1975.10479865>
- Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185, 1-17. [https://doi.org/10.1016/0003-2670\(86\)80028-9](https://doi.org/10.1016/0003-2670(86)80028-9)
- Griffiths, A. J. F., Doebley, J., Peichel, C. L., & Wassarman, D. A. (2021). *Genética: Análisis y principios* (7.^a ed.). McGraw-Hill Education.
- Guindon, F. (2017). *Construcción de un mapa de ligamiento e identificación de QTLs (quantitative trait loci) para caracteres de importancia agronómica en arveja (Pisum sativum L.)* [Tesis doctoral]. Universidad Nacional de Rosario, Facultad de Ciencias Agrarias.
- Gujarati, D. N. (2009). *Econometría*. McGraw-Hill.

- Gupta, P. K., Rustgi, S., & Kulwal, P. L. (2005). Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Molecular Biology*, 57, 461-485. <https://doi.org/10.1007/s11103-005-0257-z>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd). Springer.
- Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38, 226-231. <https://doi.org/10.1007/BF01245622>
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75, 800-802. <https://doi.org/10.1093/biomet/75.4.800>
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67. <https://doi.org/10.1080/00401706.1970.10488634>
- Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian journal of Statistics*, 6, 65-70. <http://www.jstor.org/stable/4615733>
- Hommel, G. (1986). Multiple test procedures for arbitrary dependence structures. *Metrika*, 33, 321-336.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3.^a ed.). Wiley.
- Hovde Liland, K., Mevik, B.-H., Wehrens, R., & Hiemstra, P. (2024). *pls: Partial least squares and principal component regression* (Ver. 2.8-5). <https://cran.r-project.org/package=pls>
- Hufnagel, B., Bernardino, K. C., Malosetti, M., Sousa, S. M., Silva, L. A., Guimarães, C. T., Coelho, A. M., Santos, T. T., Viana, J. H. M., Schaffert, R. E., Kochian, L. V., Eeuwijk, F. A. v., & Magalhaes, J. V. d. (2024). Multi-trait association mapping for phosphorous efficiency reveals flexible root architectures in sorghum. *BMC Plant Biology*, 24, 562. <https://doi.org/10.1186/s12870-024-05183-5>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer.

- Jannink, J.-L., Bink, M. C. A. M., & Jansen, R. C. (2001). Using complex plant pedigrees to map valuable genes. *Trends in Plant Science*, 6, 337-342. [https://doi.org/10.1016/S1360-1385\(01\)02017-9](https://doi.org/10.1016/S1360-1385(01)02017-9)
- Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178, 1709-1723. <https://doi.org/10.1534/genetics.107.080101>
- Kearsey, M. J., & Pooni, H. S. (1996). *The genetical analysis of quantitative traits*. Chapman; Hall.
- Kiers, H. A. L. (1991). Hierarchical relations among three-way methods. *Psychometrika*, 56, 449-470. <https://doi.org/10.1007/BF02294485>
- Kroonenberg, P. M. (2008). *Applied Multiway Data Analysis*. Wiley.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
- Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *journal of Statistical Software*, 25, 1-18. <https://doi.org/10.18637/jss.v025.i01>
- Legendre, P., & Legendre, L. (2012). *Numerical ecology* (3rd English ed.). Elsevier.
- Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations: heterotic models. *Genetics*, 49, 49-67.
- Lewontin, R. C., & Hubby, J. L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. *Genetics*, 54, 577-594. <https://doi.org/10.1093/genetics/54.2.577>
- Li, A., & Barber, R. F. (2019). Multiple testing with the structure-adaptive Benjamini-Hochberg algorithm. *journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81, 45-74. <https://doi.org/10.1111/rssb.12298>
- Li, J., & Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95, 221-227. <https://doi.org/10.1038/sj.hdy.6800717>

- Liu, J., Ji, S., & Ye, J. (2009). Multi-task Feature Learning via Efficient $\ell_{2,1}$ -Norm Minimization. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 339-348.
- Lombardo, L., Vanzetti, L., & Helguera, M. (2010). Utilización de marcadores moleculares en el mejoramiento del trigo [Tomo 4, Vol. 2]. *Análisis de Semillas*, 4.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748. <https://doi.org/10.1093/jnci/22.4.719>
- Markert, C. L., & Møller, F. (1959). Multiple forms of enzymes: tissue, ontogenic, and species-specific patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 45, 753-763. <https://doi.org/10.1073/pnas.45.5.753>
- Martínez, M. C., Helguera, M., & Carrera, A. (2004). Marcadores moleculares. En *Biotecnología y mejoramiento vegetal II* (pp. 70-85). INTA.
- Maxam, A. M., & Gilbert, W. (1977). A New Method for Sequencing DNA. *Proceedings of the National Academy of Sciences*, 74 (2), 560-564. <https://doi.org/10.1073/pnas.74.2.560>
- McCouch, S. R., Wright, M. H., Tung, C.-W., Maron, L. G., McNally, K. L., Fitzgerald, M., Singh, N., DeClerck, G., Agosto-Pérez, F., Korniliev, P., Greenberg, A. J., Naredo, M. E. B., Mercado, S. M. Q., Harrington, S. E., Shi, Y., Branchini, D. A., Kuser-Falcão, P. R., Leung, H., Ebana, K., ... Mezey, J. (2016). Open access resources for genome-wide association mapping in rice. *Nature Communications*, 7, 10532. <https://doi.org/10.1038/ncomms10532>
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2.^a ed.). Chapman & Hall.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. En P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105-142). Academic Press.

- Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 417-473. <https://doi.org/10.1111/j.1467-9868.2010.00740.x>
- Moose, S. P., & Mumm, R. H. (2008). Molecular Plant Breeding as the Foundation for 21st Century Crop Improvement. *Plant Physiology*, 147, 969-977. <https://doi.org/10.1104/pp.108.118232>
- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *journal of the Royal Statistical Society, Series A (General)*, 135, 370-384. <https://doi.org/10.2307/2344614>
- Nikoloff, N., & Ruiz de Arcaute, C. (2021). Introducción a la citogenética. En *Estructura, función y comportamiento cromosómico* (pp. 48-68). Facultad de Ciencias Naturales y Museo, Universidad Nacional de La Plata. https://ri.conicet.gov.ar/bitstream/handle/11336/161294/CONICET_Digital_Nro.62c704c2-f676-4566-9d19-75d8ed3ba98c_B.pdf
- Obozinski, G., Taskar, B., & Jordan, M. I. (2006). *Multi-task feature selection* (Technical report). University of California, Berkeley. https://www.researchgate.net/publication/228666354_Multi-task_feature_selection
- OpenAI. (2025). *ChatGPT* (Ver. GPT-5). <https://chat.openai.com>
- Organización de las Naciones Unidas para la Agricultura y la Alimentación. (1989). *Revisión de la situación mundial de los recursos fitogenéticos* (technical report). FAO.
- Organización de las Naciones Unidas para la Alimentación y la Agricultura. (2021a). *Glosario: WIEWS - El sistema mundial de información y alerta sobre los recursos fitogenéticos para la alimentación y la agricultura*. Consultado el 11 de octubre de 2023, desde <https://www.fao.org/wiews/glossary/es/>
- Organización de las Naciones Unidas para la Alimentación y la Agricultura. (2021b). *Tercer informe nacional sobre el estado de los recursos fitogenéticos para la alimentación y la agricultura en Argentina* (technical report). FAO. <https://openknowledge.fao.org/handle/20.500.14283/cd4871es>

- Panahi, B., Mohammadzadeh Jalaly, H., & Hamid, R. (2024). Using next-generation sequencing approach for discovery and characterization of plant molecular markers. *Current Plant Biology*, 40, 100412. <https://doi.org/10.1016/j.cpb.2024.100412>
- Paredes, O., & Gepts, P. (1995). Extensive introgression of Middle American germplasm into Chilean common bean cultivars. *Genetic Resources and Crop Evolution*, 42, 29-41. <https://doi.org/10.1007/BF02439207>
- Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545-554. <https://doi.org/10.1093/biomet/58.3.545>
- Peña Malavera, A., Bruno, C., & Balzarini, M. (2018). Control de falsos descubrimientos en mapeo asociativo con poblaciones estructuradas. *Journal of Basic and Applied Genetics*, 28, 37-49.
- Peña Malavera, A. N. (2015). *Aproximaciones estadísticas para el mapeo asociativo en estudios genéticos* [Tesis doctoral]. Universidad Nacional de Córdoba. <https://rdu.unc.edu.ar/items/2d0c8d98-f7f9-4554-9570-9d085d1fa315>
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945-959. <https://doi.org/10.1093/genetics/155.2.945>
- Quenouille, M. H. (1949). Problems in plane sampling. *Annals of Mathematical Statistics*, 20, 355-375. <https://doi.org/10.1214/aoms/1177729989>
- Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353-360. <https://doi.org/10.1093/biomet/43.3-4.353>
- R Core Team. (2025). *R: A language and environment for statistical computing*. <https://www.r-project.org/>
- Ramdas, A. K., Barber, R. F., Wainwright, M. J., & Jordan, M. I. (2019). A unified treatment of multiple testing with prior knowledge using the p-filter. *Annals of Statistics*, 47, 2790-2821. <https://doi.org/10.1214/18-AOS1765>
- Rao, C. R. (1973). *Linear statistical inference and its applications*. Wiley.

- Ravikumar, P., Wainwright, M. J., & Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1 -regularized logistic regression. *The Annals of Statistics*, 38, 1287-1319. <https://doi.org/10.1214/09-AOS691>
- Rencher, A. C. (2002). *Methods of multivariate analysis* (2.^a ed.). Wiley.
- Revelle, W. (2025). *psych: Procedures for psychological, psychometric, and personality research* (Ver. 2.4.6). <https://CRAN.R-project.org/package=psych>
- Ripley, B., & Venables, W. (2023). *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models* (Ver. 7.3-19). <https://cran.r-project.org/package=nnet>
- Robert, P., & Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. *journal of the Royal Statistical Society: Series C (Applied Statistics)*, 25, 257-265. <https://doi.org/10.2307/2347233>
- Rogers, A. R., & Huff, E. (2009). Linkage disequilibrium between loci with unknown phase. *Genetics*, 182, 839-844. <https://doi.org/10.1534/genetics.108.093153>
- Rojas-Beltrán, J. A. (2007). *Tecnología de los marcadores moleculares*. Fundación PROINPA.
- Rosipal, R., & Trejo, L. J. (2001). Kernel partial least squares regression in reproducing kernel Hilbert space. *journal of Machine Learning Research*, 2, 97-123. <https://www.jmlr.org/papers/v2/rosipal01a.html>
- Rüger, B. (1978). Das maximale Signifikanzniveau des Tests Lehne H_0 ab, wenn k unter n gegebenen Tests zur Ablehnung führen. *Metrika*, 25, 171-178.
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA Sequencing with Chain-Terminating Inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), 5463-5467. <https://doi.org/10.1073/pnas.74.12.5463>
- Seijo, G., Lavia, G. I., Robledo, G., Fernández, A., & Solís Neffa, V. G. (2004). La citogenética molecular e inmunocitogenética en el estudio de los genomas vegetales. En *Biotecnología y mejoramiento vegetal II* (pp. 34-46). INTA. <https://exa.unne.edu.ar/biologia/fisiologia.vegetal/BiotecnologiayMejoramientovegetalII.pdf>

- Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *journal of the American Statistical Association*, 81, 826-831. <https://doi.org/10.1080/01621459.1986.10478341>
- Shah, R. D., & Samworth, R. J. (2013). Variable selection with error control: Another look at stability selection. *journal of the Royal Statistical Society: Series B*, 75, 55-80. <https://doi.org/10.1111/j.1467-9868.2011.01034.x>
- Sharma, S. K., Bryan, G. J., Bolser, D., de Boer, J., Sønderkær, M., Amoros, W., & Guzman, F. (2018). Linkage disequilibrium and evaluation of genome-wide association mapping models in tetraploid potato. *G3: Genes, Genomes, Genetics*, 8, 3185-3202. <https://doi.org/10.1534/g3.118.200377>
- Shikha, K., Shahi, J. P., Vinayan, M. T., Zaidi, P. H., Singh, A. K., & Sinha, B. (2021). Genome-wide association mapping in maize: status and prospects. *3 Biotech*, 11, 244. <https://doi.org/10.1007/s13205-021-02799-4>
- Stewart, D. W., & Love, W. A. (1968). A general canonical correlation index. *Psychological Bulletin*, 70, 160-163. <https://doi.org/10.1037/h0026143>
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *journal of the Royal Statistical Society: Series B*, 36, 111-147. <https://doi.org/10.1111/j.2517-6161.1974.tb00994.x>
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *journal of the Royal Statistical Society: Series B*, 58, 267-288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Tikhonov, A. N. (1943). On the stability of inverse problems. *Doklady Akademii Nauk SSSR*, 39, 195-198.
- Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics*, 29, 614-623. <https://doi.org/10.1214/aoms/1177706647>
- Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., Nguyen-Viet, T. A., Wedow, R., Zacher, M., Furlotte, N. A., 23andMe Research Team, Social Science Genetic Association Consortium, Magnusson, P., Oskarsson, S., Johannesson, M., Visscher, P. M., Laibson, D., Cesarini, D.,

- Neale, B. M., & Benjamin, D. J. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, 50, 229-237. <https://doi.org/10.1038/s41588-017-0009-4>
- Udosen, B., Soremekun, O., Kamiza, A., Machipisa, T., Cheickna, C., Omotuyi, O., Soliman, M., Wélé, M., Nashiru, O., Chikowore, T., & Fatumo, S. (2023). Meta-analysis and multivariate GWAS analyses in 80,950 individuals of African ancestry identify novel variants associated with blood pressure traits. *International journaltitle of Molecular Sciences*, 24, 2164. <https://doi.org/10.3390/ijms24032164>
- van den Wollenberg, A. L. (1977). Redundancy analysis: An alternative for canonical correlation analysis. *Psychometrika*, 42, 207-219. <https://doi.org/10.1007/BF02294050>
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4.^a ed.). Springer.
- Villegas Castagnasso, E., Posik, D. M., Crespi, J. A., & Catanesi, C. I. (2021). *Marcadores genéticos: Introducción al análisis y su aplicación en diversas áreas biológicas*. Editorial de la Universidad Nacional de La Plata (EDULP). <https://sedici.unlp.edu.ar/handle/10915/131500>
- Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1, 1-305. <https://doi.org/10.1561/2200000001>
- Williams, J. G. K., Kubelik, A. R., Livak, K. J., Rafalski, J. A., & Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research*, 18, 6531-6535. <https://doi.org/10.1093/nar/18.22.6531>
- Wold, H. (1982). Soft modeling: The basic design and some extensions [Part 2]. En K. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (pp. 1-54). North-Holland.

- Wold, H. O. A. (1966). Estimation of principal components and related models by iterative least squares. En P. R. Krishnaiah (Ed.), *Multivariate Analysis* (pp. 391-420). Academic Press.
- Wu, T. T., & Lange, K. (2008). Coordinate descent procedures for lasso penalized regression. *The Annals of Applied Statistics*, 2, 224-244. <https://doi.org/10.1214/07-AOAS147>
- Xiao, Y., Liu, H., Wu, L., Warburton, M., & Yan, J. (2017). Genome-wide association studies in maize: Praise and stargaze. *Molecular Plant*, 10, 359-374. <https://doi.org/10.1016/j.molp.2016.12.008>
- Xu, Y. (2010). *Molecular plant breeding*. CABI. <https://doi.org/10.1079/9781845933920.0000>
- Yang, Q., & Wang, Y. (2012). Methods for analyzing multivariate phenotypes in genetic association studies. *Journal of Probability and Statistics*, 2012, 652569. <https://doi.org/10.1155/2012/652569>
- Yu, J., & Buckler, E. S. (2006). Genetic association mapping and genome organization of maize. *Current Opinion in Biotechnology*, 17, 155-160. <https://doi.org/10.1016/j.copbio.2006.02.003>
- Zhu, C., Gore, M., Buckler, E. S., & Yu, J. (2008). Status and prospects of association mapping in plants. *The Plant Genome*, 1, 5-20. <https://doi.org/10.3835/plantgenome2008.02.0089>
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67, 301-320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>