GRADO EN INGENIERÍA MATEMÁTICA E INTELIGENCIA ARTIFICIAL

TRABAJO FIN DE GRADO

Explainable AI for biological data

Autor: Xabier Albizu Arias

Directores:
Simón Rodríguez Santana
Jaime Pizarroso Gonzalo

Madrid

Junio de 2026

Declaración de originalidad

Declaro bajo mi responsabilidad que el Proyecto presentado con el título **Explainable AI for Biological Data: Interpretable, Uncertainty-Aware Classification of Primate Dental Microtexture** en la Escuela Técnica Superior de Ingeniería ICAI de la Universidad Pontificia Comillas en el curso académico 2025/2026 es de mi autoría y no ha sido presentado anteriormente para otros fines. El Proyecto no ha sido plagiado de ningún otro, ni total ni parcialmente, y la información que ha sido tomada de otros documentos está debidamente referenciada.

Uso de Inteligencia Artificial[1]

Declaro bajo mi responsabilidad (indicar la opción correcta):

☐ No he utilizado Inteligencia Artificial en la elaboración de este documento.

☒ He utilizado Inteligencia Artificial en la elaboración de este documento y/o del Anexo B bajo las condiciones permitidas por la Universidad Pontificia Comillas, es decir, aplicando el Nivel 2 de la Escala de Evaluación de Perkins et al. (2024): *"La IA puede utilizarse para actividades previas a la tarea, como lluvia de ideas, descripción e investigación inicial. Este nivel se centra en el uso de la IA para planificar, sintetizar y generar ideas, pero las evaluaciones deben enfatizar la capacidad de desarrollar y perfeccionar estas ideas de forma independiente"*. En concreto, la Inteligencia Artificial se ha utilizado para:

[1]Esta declaración se refiere al uso de Inteligencia Artificial generativa para la elaboración de los documentos del Proyecto (Anexo B y Memoria). No se aplica a Proyectos en los que, por su naturaleza, deba utilizarse inteligencia artificial como parte de los mismos (aplicación de técnicas de aprendizaje automático, redes neuronales, análisis de datos...).

Se han empleado herramientas de Inteligencia Artificial generativa como apoyo en las fases previas y de redacción del documento, siempre bajo supervisión y revisión crítica del autor, que ha tomado todas las decisiones de contenido, metodología y diseño. En concreto:

- **Organización y estructura**: lluvia de ideas sobre el guion del documento y ayuda en la articulación de los apartados del resumen conforme a la plantilla institucional.
- **Apoyo a la redacción**: reformulación, síntesis y mejora de la claridad y la corrección lingüística de textos previamente elaborados por el autor, así como su traducción al inglés (*abstract*), revisando siempre el autor la fidelidad técnica del resultado.
- **Investigación inicial**: orientación bibliográfica preliminar y aclaración de conceptos metodológicos, cuyas fuentes y formulaciones han sido posteriormente verificadas de forma independiente.
- **Soporte técnico puntual**: asistencia en la depuración de fragmentos de código auxiliar y en la generación de figuras esquemáticas del diseño de la solución, a partir de especificaciones y datos proporcionados por el autor.

En ningún caso se ha empleado la Inteligencia Artificial para generar resultados experimentales, análisis de datos o conclusiones, que son obra propia del autor. Todo el contenido generado con asistencia de IA ha sido revisado, contrastado y editado críticamente, asumiendo el autor la plena responsabilidad sobre su veracidad y rigor.

(firmar aquí)
Firma: Xabier Albizu Arias
Fecha: 15/06/2026

Autorización para la entrega del Proyecto

Director del TFG	Codirector del TFG, en su caso
Firmado por RODRIGUEZ SANTANA SIMON - ***4141** el día 15/06/2026 con un certificado emitido por AC (firmar aquí)	Jaime Pizarroso Gonzalo (firmar aquí) Firmado digitalmente por Jaime Pizarroso Gonzalo Fecha: 2026.06.15 19:21:54 +02'00'
Firma: Simón Rodríguez Santana	Firma: Jaime Pizarroso Gonzalo
Fecha: 15/06/2026	Fecha: 15/06/2026

Resumen del proyecto

La clasificación taxonómica de primates a partir de microtextura dental bucal (DMTA) es un problema de datos escasos, ruidosos y de alta dimensión, abordado hasta ahora con métodos clásicos sin cuantificación de incertidumbre. Este trabajo desarrolla un *pipeline* Bayesiano interpretable y consciente de la incertidumbre, organizado en tres capas, que mide hasta qué nivel taxonómico puede fiarse la señal. Los resultados muestran que la microtextura bucal resuelve el grupo pero no la especie, que los fósiles no son comparables con los actuales, y caracteriza sus afinidades con incertidumbre explícita, convirtiendo una conclusión cualitativa previa en una medida cuantitativa.

Palabras clave: microtextura dental, clasificación Bayesiana, cuantificación de incertidumbre, GPLVM, modelos jerárquicos, IA explicable, paleodieta.

1. Introducción

Clasificar primates a partir de restos morfológicos es una de las tareas más complejas de la paleobiología: las muestras son escasas, provienen de contextos muy distintos y presentan estados de conservación muy variables, lo que introduce un ruido considerable en cualquier análisis. Sin embargo, reconstruir la dieta y la ecología de especies extintas tiene un valor que va mucho más allá de la taxonomía: la dieta funciona como *proxy* a partir del cual inferir cambios históricos en los recursos hídricos, reconstruir ecosistemas pasados y entender las presiones evolutivas que han modelado a los primates –incluido nuestro propio linaje– y que siguen siendo relevantes hoy. Una de las pocas herramientas disponibles para este fin es el análisis de los restos dentales de especies actuales de dieta conocida, que sirven como referencia para interpretar el registro fósil por comparación. Aun así, los datos son intrínsecamente difíciles de tratar, y los análisis se realizan en buena medida "a mano" y con diseños estándar bastante rudimentarios –análisis de varianza, componentes principales, estimaciones puntuales sin cuantificación de incertidumbre–, herramientas poco adaptadas a la naturaleza del problema.

El análisis de la microtextura del desgaste dental (DMTA) se ha consolidado como herramienta estándar en paleobiología para inferir dieta y adaptaciones ecológicas en primates a partir de la topografía microscópica del esmalte, tanto en especies actuales como en contextos fósiles. La técnica cuantifica esa topografía –formada y renovada a escala de semanas o meses– mediante un conjunto de descriptores numéricos, y permite asignar taxones de dieta desconocida a categorías dietéticas por comparación con taxones actuales. Su aplicación reciente a la superficie bucal de los cercopitécidos africanos es el punto de partida directo de este trabajo, y plantea una pregunta que los métodos clásicos no resuelven de forma cuantitativa: ¿hasta qué nivel de resolución taxonómica –especie, grupo o género– puede fiarse realmente la señal? Responderla importa porque el material combina especies actuales etiquetadas y taxones fósiles sin verdad de referencia, de modo que un sistema riguroso no solo debe predecir, sino saber cuándo no puede afirmar nada. A ello se suman tres dificultades del dominio: la "jungla de parámetros" (más de un centenar de descriptores redundantes y no gaussianos), la escasez de datos propia de la paleodieta, y la necesidad de interpretabilidad biológica de cualquier predicción.

2. Objetivos

El objetivo general es desarrollar y validar un *pipeline* interpretable y consciente de la incertidumbre para la caracterización taxonómica y dietética de cercopitécidos a partir de microtextura dental, estableciendo el nivel de resolución (especie frente a grupo o género) que la señal soporta de forma fiable. La pregunta de investigación principal es, por tanto, cuantitativa: *¿hasta qué nivel taxonómico permite fiarse la microtextura bucal?*

Este objetivo se concreta en cinco objetivos específicos: (i) un preprocesado robusto para datos pequeños y de alta dimensión; (ii) modelos probabilísticos que cuantifiquen la incertidumbre de sus predicciones; (iii) estudios de importancia de variables que conecten rendimiento e interpretabilidad biológica; (iv) el aprovechamiento de la estructura latente para construir modelos jerárquicos informados; y (v) la determinación de la comparabilidad de los fósiles y la caracterización de sus afinidades con incertidumbre cuantificada, por una vía no supervisada y una supervisada, sin imponer una correspondencia unívoca.

3. Descripción del sistema

El problema se aborda mediante un *pipeline* Bayesiano interpretable, validado en todo momento bajo un protocolo homogéneo de validación cruzada *leave-one-out* (LOO-CV) y contrastado contra referencias triviales con métricas robustas al desbalanceo (exactitud balanceada, F1-macro y κ de Cohen). La metodología se articula en una arquitectura de tres capas (Figura 1), cada una con su pregunta de investigación.

La Capa 1 mide el techo informativo supervisado sobre las especies actuales –único conjunto con verdad de referencia– comparando modelos clásicos, regresión logística Bayesiana multinomial con distintos *priors* de contracción (Gaussiano, Laplace, *horseshoe* y *spike-and-slab*) y clasificadores jerárquicos en cascada cuya estructura se deriva del espacio latente. La Capa 2 decide, mediante un modelo de mezcla Gaussiana y verosimilitud cruzada, si los fósiles son distribucionalmente comparables con los actuales. La Capa 3 caracteriza las afinidades de los fósiles por dos vías complementarias: una no supervisada, sobre el espacio latente de un GPLVM Bayesiano variacional con agrupamiento jerárquico (BHC); y una supervisada, mediante proyección a través del clasificador jerárquico, que devuelve una distribución de probabilidad en lugar de una etiqueta forzada.

4. Resultados

Los resultados convergen en una conclusión clara:

- La microtextura bucal **permite únicamente resolución a nivel de grupo**: a nivel de especie la clasificación no supera el azar (exactitud $\approx 0{,}30$), mientras que a nivel de grupo mejora de forma sustancial (exactitud $\approx 0{,}56$). El límite es de la señal: solo 3 de 107 variables se diferencian marginalmente entre especies y ningún coeficiente Bayesiano resulta robustamente distinto de cero.

- La Capa 2 demuestra que actuales y fósiles ocupan regiones disjuntas del espacio de variables ($p < 10^{-18}$ en todas las variantes), confirmando un *domain shift*: proyectar los fósiles sobre un clasificador entrenado solo en actuales sería extrapolar sin respaldo.

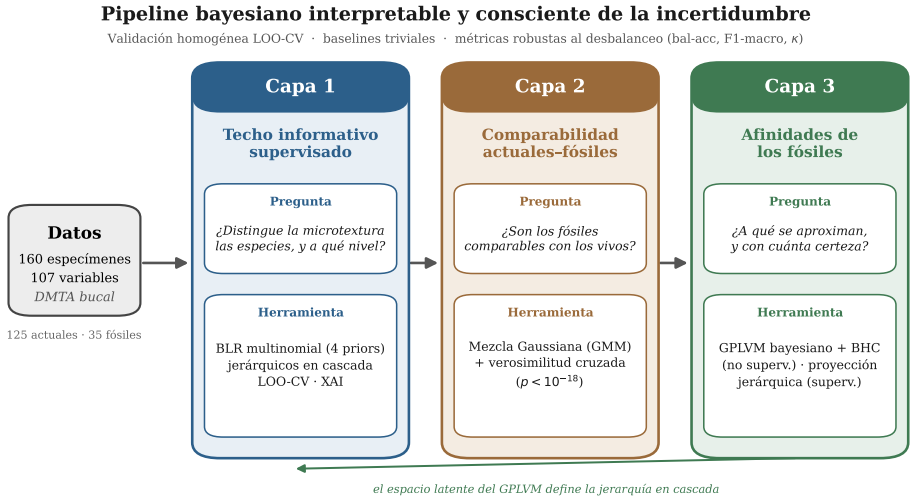

Figura 1: Arquitectura de análisis en tres capas del *pipeline* propuesto.

- El GPLVM Bayesiano variacional recupera la filogenia de los cercopitécidos actuales sin usar las etiquetas (BHC, NMI = 0,85), validando que el espacio latente codifica estructura biológica real. Sobre él, los taxones *Parapapio* gravitan de forma consistente hacia el clado *Papio*, mientras que *Theropithecus oswaldi* no converge hacia su congénere actual *T. gelada*, sino que se dispersa: un resultado paleoecológico coherente con una dieta más variada que una clasificación forzada habría encubierto.

5. Conclusiones

La microtextura bucal no resuelve a nivel de especie, y este límite es de la señal y no del *pipeline*: se reproduce tanto en los taxones actuales como en la clasificación de género fósil. Actuales y fósiles no son distribucionalmente comparables, por lo que el enfoque no supervisado es el metodológicamente correcto para caracterizar los fósiles.

La contribución del trabajo es una arquitectura de análisis que integra piezas metodológicas conocidas bajo validación homogénea y cuantificación explícita de la incertidumbre, convirtiendo la conclusión cualitativa de los estudios previos –resolución de grupo por encima de la de especie– en una medida cuantitativa de ese límite informativo. Frente a los diseños estándar del dominio, ofrece a los antropólogos una herramienta directamente utilizable: no solo predice, sino que comunica cuándo no puede afirmar algo, conecta cada predicción con los descriptores que la sustentan, y es transferible a otros problemas de clasificación con datos escasos y de alta dimensión.

6. Referencias

[1] A. Martínez *et al.*, "Buccal dental microwear texture analysis of African cercopithecids," *(revista)*, vol. XX, no. X, pp. XX–XX, 2022.

[2] N. D. Lawrence, "Probabilistic non-linear principal component analysis with Gaus-

sian process latent variable models," *J. Mach. Learn. Res.*, vol. 6, pp. 1783–1816, 2005.

[3] K. A. Heller and Z. Ghahramani, "Bayesian hierarchical clustering," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 297–304.

Abstract

Taxonomic classification of primates from buccal dental microwear texture (DMTA) is a problem of scarce, noisy, high-dimensional data, addressed so far with classical methods lacking uncertainty quantification. This work develops an interpretable, uncertainty-aware Bayesian *pipeline*, organised into three layers, that measures to what taxonomic level the signal can be trusted. The results show that buccal microtexture resolves group but not species, that fossils are not comparable to extant specimens, and characterise their affinities with explicit uncertainty, turning a previous qualitative conclusion into a quantitative measure.

Keywords: dental microtexture, Bayesian classification, uncertainty quantification, GPLVM, hierarchical models, explainable AI, palaeodiet.

1. Introduction

Classifying primates from morphological remains is one of the most challenging tasks in palaeobiology: samples are scarce, come from highly diverse contexts and exhibit very variable preservation states, all of which introduce substantial noise into any analysis. Yet reconstructing the diet and ecology of extinct species carries value well beyond taxonomy: diet acts as a *proxy* from which to infer historical changes in water resources, reconstruct past ecosystems and understand the evolutionary pressures that have shaped primates –including our own lineage– and that remain relevant today. One of the few tools available for this purpose is the analysis of dental remains from extant species with known diets, which serve as a reference for interpreting the fossil record through comparison. Even so, the data are inherently difficult to handle, and the analyses are carried out largely "by hand" and with rather rudimentary standard designs –analysis of variance, principal components, point estimates without uncertainty quantification– tools poorly suited to the nature of the problem.

Dental microwear texture analysis (DMTA) has become a standard tool in palaeobiology for inferring diet and ecological adaptation in primates from the microscopic topography of enamel, in both extant species and fossil contexts. The technique quantifies that topography –formed and renewed on a scale of weeks to months– through a set of numerical descriptors, allowing taxa of unknown diet to be assigned to dietary categories by comparison with extant taxa. Its recent application to the buccal surface of African cercopithecids is the direct starting point of this work, and raises a question that classical methods do not answer quantitatively: to what taxonomic level –species, group, or genus– can the signal actually be trusted? Answering it matters because the material combines labelled extant species and fossil taxa with no ground truth, so a rigorous system must not only predict but also know when it cannot make a claim. Three domain difficulties compound this: the "parameter jungle" (over a hundred redundant, non-Gaussian descriptors), the data scarcity typical of palaeodiet, and the need for biological interpretability of any prediction.

2. Objectives

The general objective is to develop and validate an interpretable, uncertainty-aware *pipeline* for the taxonomic and dietary characterisation of cercopithecids from dental microtexture, establishing the resolution level (species versus group or genus) that the signal reliably supports. The main research question is therefore quantitative: *to what taxonomic level can buccal microtexture be trusted?*

This breaks down into five specific objectives: (i) a robust preprocessing pipeline for small, high-dimensional data; (ii) probabilistic models that quantify prediction uncertainty; (iii) variable-importance studies linking performance and biological interpretability; (iv) the use of the latent structure to build informed hierarchical models; and (v) determining fossil comparability and characterising their affinities with quantified uncertainty, via an unsupervised and a supervised route, without imposing a one-to-one mapping.

3. System description

The problem is addressed through an interpretable Bayesian *pipeline*, validated throughout under a homogeneous leave-one-out cross-validation (LOO-CV) protocol and benchmarked against trivial baselines with imbalance-robust metrics (balanced accuracy, macro-F1 and Cohen's κ). The methodology is organised into a three-layer architecture (Figure 2), each layer with its own research question.

Layer 1 measures the supervised information ceiling on extant taxa –the only set with ground truth– comparing classical baselines, multinomial Bayesian logistic regression under several shrinkage priors (Gaussian, Laplace, *horseshoe* and *spike-and-slab*) and cascaded hierarchical classifiers whose structure is derived from the latent space. Layer 2 decides, via a Gaussian mixture model and cross-likelihood, whether fossils are distributionally comparable to extant specimens. Layer 3 characterises fossil affinities through two complementary routes: an unsupervised one over the latent space of a variational Bayesian GPLVM with hierarchical clustering (BHC); and a supervised one by projection through the hierarchical classifier, returning a probability distribution rather than a forced label.

4. Results

The results converge on a clear conclusion:

- Buccal microtexture **supports only group-level resolution**: at the species level classification does not exceed chance (accuracy $\approx 0{,}30$), whereas at the group level it improves substantially (accuracy $\approx 0{,}56$). The limit is a property of the signal: only 3 of 107 variables differ marginally between species and no Bayesian coefficient is robustly distinguishable from zero.
- Layer 2 shows that extant and fossil taxa occupy disjoint regions of variable space ($p < 10^{-18}$ across all variants), confirming a *domain shift*: projecting fossils onto a classifier trained only on extant specimens would amount to unsupported extrapolation.
- The variational Bayesian GPLVM recovers the phylogeny of extant cercopithecids without using labels (BHC, NMI $= 0{,}85$), validating that the latent space encodes

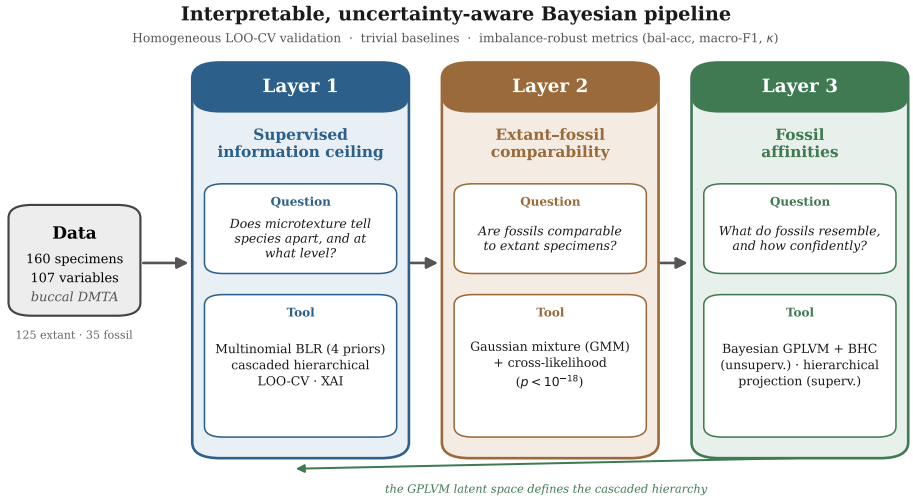

Figura 2: Three-layer analysis architecture of the proposed *pipeline*.

real biological structure. On it, the *Parapapio* taxa gravitate consistently toward the *Papio* clade, whereas *Theropithecus oswaldi* does not converge toward its extant congener *T. gelada* but disperses: a palaeoecological finding consistent with a more varied diet that a forced classification would have concealed.

5. Conclusions

Buccal microtexture does not resolve at the species level, and this limit is a property of the signal rather than the *pipeline*: it reproduces in both extant taxa and fossil-genus classification. Extant and fossil specimens are not distributionally comparable, so the unsupervised approach is the methodologically correct one for characterising fossils.

The contribution is an analysis architecture that integrates known methodological components under homogeneous validation and explicit uncertainty quantification, turning the qualitative conclusion of prior studies –group- over species-level resolution– into a quantitative measure of that information ceiling. Against the field's standard designs, it offers anthropologists a directly usable tool: it not only predicts but also communicates when it cannot make a claim, links each prediction to the descriptors that support it, and is transferable to other classification problems with scarce, high-dimensional data.

6. References

[1] A. Martínez *et al.*, "Buccal dental microwear texture analysis of African cercopithecids," *(journal)*, vol. XX, no. X, pp. XX–XX, 2022.

[2] N. D. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *J. Mach. Learn. Res.*, vol. 6, pp. 1783–1816, 2005.

[3] K. A. Heller and Z. Ghahramani, "Bayesian hierarchical clustering," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 297–304.

Índice de la memoria

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Contexto y motivación

El análisis de textura del microdesgaste dental (DMTA) se ha consolidado como una herramienta estándar en antropología y paleobiología para inferir la dieta y las adaptaciones ecológicas de primates y otros mamíferos. La técnica cuantifica la topografía microscópica del esmalte –formada y renovada a escala de semanas o meses– mediante un conjunto de descriptores numéricos, y permite asignar taxones de dieta desconocida a categorías dietéticas por comparación con taxones actuales de dieta conocida. Su aplicación reciente a la superficie bucal de los cercopitécidos africanos es el punto de partida directo de este trabajo.

Tres dificultades caracterizan este dominio y motivan el enfoque adoptado. La primera es la *"jungla de parámetros"*: la microtextura se describe con un centenar largo de variables (familias ISO, SSFA y otras) fuertemente redundantes entre sí y de distribuciones marcadamente no gaussianas. La segunda es la *escasez de datos* propia de la paleodieta, con pocas decenas de especímenes por clase, un régimen en el que las arquitecturas *data-hungry* sobreajustan y resultan inestables. La tercera es la necesidad de *interpretabilidad biológica*: una predicción solo es útil como hipótesis científica si puede atribuirse a variables con sentido funcional, lo que entra en tensión con los modelos más precisos pero opacos.

A esto se añade un reto específico del material: el conjunto de datos combina especies *actuales*, con dieta y etiqueta conocidas, y taxones *fósiles* sin verdad de referencia verificable. Un sistema que pretenda caracterizar los fósiles a partir de los actuales debe, por tanto, no solo predecir, sino saber *cuándo no puede afirmar algo* –es decir, cuantificar explícitamente su propia incertidumbre.

1.2. Objetivos

El objetivo general de este Trabajo de Fin de Grado es **desarrollar y validar un *pipeline* interpretable y consciente de la incertidumbre para la caracterización taxonómica y dietética de primates cercopitécidos a partir de datos de microtextura dental, estableciendo el nivel de resolución (especie frente a grupo o género) que la señal soporta de forma fiable**. La pregunta de investigación principal es, por tanto, cuantitativa: *¿hasta qué nivel de resolución taxonómica permite*

fiarse la microtextura bucal? El detalle de los objetivos específicos y su articulación en una arquitectura de tres capas se desarrolla en el Capítulo 3 (Sección 3.1).

1.3. Alineación con los Objetivos de Desarrollo Sostenible

Este trabajo se alinea principalmente con dos Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030. El **ODS 15 (Vida de ecosistemas terrestres)** se ve respaldado por el desarrollo de herramientas cuantitativas para reconstruir la ecología y la dieta de primates, actuales y extintos, lo que contribuye al conocimiento de la biodiversidad y de las relaciones evolutivas que sustentan la conservación. El **ODS 9 (Industria, innovación e infraestructura)** se refleja en la aportación metodológica: un *pipeline* reproducible de análisis de datos científicos escasos y de alta dimensión, con cuantificación de incertidumbre, transferible a otros problemas de clasificación con datos limitados.

1.4. Estructura del trabajo

El resto de la memoria se organiza como sigue. El Capítulo 2 revisa el estado del arte, desde los descriptores de microtextura y los métodos clásicos hasta las herramientas probabilísticas, jerárquicas y no supervisadas en que se apoya el trabajo. El Capítulo 3 describe el sistema desarrollado: el planteamiento formal del problema y la arquitectura de tres capas (Sección 3.1), el diseño de la solución con sus fundamentos teóricos y modelos (Sección 3.2), y los detalles de implementación (Sección 3.3). El Capítulo 4 presenta y analiza críticamente los resultados de cada capa. El Capítulo 5 recoge las conclusiones y las líneas de trabajo futuro. El Capítulo 6 lista la bibliografía.

Capítulo 2

Estado del arte

2.1. Microtextura dental y enfoques tradicionales

El análisis de la microtextura del desgaste dental (DMTA) se ha consolidado como una herramienta estándar en antropología para inferir patrones dietéticos y adaptaciones ecológicas en primates y otros mamíferos. La metodología se apoya en la cuantificación de parámetros de rugosidad de la superficie a partir de escaneos tridimensionales del esmalte, con conjuntos de variables como las métricas ISO [18], el análisis fractal sensible a la escala (*scale-sensitive fractal analysis*, SSFA) [39, 35] y otros descriptores de textura. Estos parámetros se comparan después entre especies o grupos dietéticos para evaluar diferencias ecológicas o funcionales [4].

Tradicionalmente, los estudios de DMTA se han apoyado en el análisis estadístico clásico. El análisis de la varianza (ANOVA) se aplica habitualmente para contrastar si los valores medios de las variables de microdesgaste difieren entre grupos dietéticos; cuando no se cumplen los supuestos de normalidad u homocedasticidad, se recurre a pruebas no paramétricas como Kruskal–Wallis [21], seguidas con frecuencia de comparaciones *post-hoc* para identificar qué grupos se separan. Para la visualización exploratoria, el análisis de componentes principales (PCA) se emplea de forma recurrente para reducir la dimensión y poner de manifiesto agrupaciones o separaciones entre especies. El estudio de referencia sobre la superficie bucal de cercopitécidos [26] sigue precisamente esta tradición y concluye que la microtextura discrimina patrones dietéticos, aunque con una variación entre especies limitada.

Aunque estos métodos han aportado conocimiento valioso, comparten limitaciones importantes. En primer lugar, están concebidos para comparaciones a nivel de grupo más que para la clasificación predictiva [8]. En segundo lugar, producen por lo general estimaciones puntuales, sin una cuantificación explícita de la incertidumbre, lo que restringe su interpretabilidad en contextos donde los datos son intrínsecamente ruidosos y los tamaños muestrales pequeños [8, 7]. Por último, la selección de variables suele ser *ad hoc*, sin un criterio común entre estudios sobre qué parámetro es diet-discriminante, y con una evaluación sistemática escasa de la informatividad relativa de los conjuntos ISO, SSFA y de textura [9, 34]. Una dificultad añadida del propio dato agrava estos problemas: las variables son muy redundantes entre sí –la llamada "jungla de parámetros" [9]– y sus distribuciones se alejan de la normalidad, lo que se ha relacionado con la elevada correla-

ción entre descriptores y la falta de un conjunto óptimo universal [7, 8]. En consecuencia, aunque el DMTA ha demostrado ser eficaz para distinguir categorías dietéticas amplias, las técnicas habituales siguen ancladas en la estadística frecuentista clásica, y los enfoques más avanzados, capaces de abordar la incertidumbre, la redundancia entre variables y la robustez predictiva, todavía no se han aplicado de forma sistemática en este dominio [8].

2.2. Clasificación con muchas variables y pocas muestras

El paso de la comparación de grupos a la clasificación predictiva introduce dificultades propias del régimen de muchas variables y pocas muestras característico de la paleodieta, donde el número de especímenes por taxón fósil suele ser muy reducido [36, 27]. El análisis discriminante lineal (LDA), método de referencia en morfometría, se vuelve inestable en este escenario: con pocos especímenes por clase no puede estimar de forma fiable las relaciones entre tantas variables, y presupone además una normalidad que los datos no cumplen [8]. El riesgo general es el sobreajuste, en el que el modelo memoriza el ruido de la muestra y generaliza mal a especímenes nuevos.

La respuesta habitual a este problema es la regularización, que penaliza la complejidad del modelo para retener únicamente lo esencial. El *Lasso* [37] lleva a cero las variables irrelevantes y realiza así una selección automática, mientras que la *Elastic Net* [44] lo extiende para tratar mejor los grupos de variables redundantes, situación característica de la microtextura. Sobre datos adecuadamente preprocesados, los modelos lineales regularizados y los métodos de *boosting* tienden a superar al LDA clásico [8], si bien a costa de una mayor opacidad, compromiso que se retoma en el apartado 2.4. En este régimen, además, la forma de estimar el rendimiento resulta tan importante como el propio modelo: la validación cruzada *leave-one-out* aprovecha al máximo una muestra reducida y proporciona una estimación estable del error [15], que debe compararse siempre con referencias triviales y medirse con métricas robustas al desbalanceo entre clases, como el F1 o el coeficiente κ de Cohen [6], en lugar del porcentaje de acierto a secas [20].

2.3. Aprendizaje automático probabilístico y cuantificación de la incertidumbre

El aprendizaje automático probabilístico, fundado en la inferencia Bayesiana, no se ha aplicado aún de forma sistemática en el contexto del DMTA, pese a su potencial para superar muchas de las limitaciones de las técnicas estadísticas tradicionales [8]. Conviene precisar que ninguna de las herramientas que se describen a continuación es novedosa en sí misma: lo que este trabajo aporta no es un avance técnico, sino su articulación e integración en un dominio –la paleodieta a partir de microtextura bucal– donde apenas se han empleado. El aprendizaje probabilístico ofrece un marco riguroso para modelar la incertidumbre tanto en los datos como en las predicciones del modelo [31, 10], lo que resulta especialmente valioso en dominios científicos como la antropología, donde los conjuntos de datos suelen ser pequeños, ruidosos y recogidos en condiciones variables. Los modelos deterministas habituales no capturan bien estas particularidades: ofrecen funciones de predicción de gran capacidad, pero con interpretabilidad o robustez limitadas cuando los

datos escasean.

En los flujos de clasificación convencionales, las predicciones son deterministas y los parámetros del modelo se tratan como cantidades fijas, de modo que la salida no incorpora noción alguna de confianza. Las técnicas probabilísticas representan en cambio tanto los parámetros como las predicciones mediante distribuciones de probabilidad [10, 31]. Esto permite cuantificar la incertidumbre a dos niveles: el epistémico (la incertidumbre del modelo debida a la escasez de datos) y el aleatorio (el ruido inherente a los propios datos). La distinción es relevante para los fósiles: una incertidumbre epistémica elevada es la que permite que el sistema advierta de que un espécimen es atípico, en lugar de asignarle una etiqueta con una confianza injustificada.

En este trabajo implementamos y analizamos principalmente un modelo de regresión logística Bayesiana multinomial, bien adaptado a la clasificación multiclase y que permite incorporar conocimiento experto a través de las distribuciones previas (*priors*). La elección del *prior* no es un mero detalle técnico: codifica una hipótesis sobre cuántas variables se espera que sean relevantes, y el conjunto de variables que cada uno conserva constituye en sí mismo una hipótesis biológica sobre qué descriptores gobiernan la discriminación. Entre los que comparamos, el de tipo *horseshoe* [5] parte de la idea de que casi todas las variables son irrelevantes y unas pocas muy importantes, silenciando con fuerza el ruido sin penalizar las señales fuertes; su variante regularizada [32] añade salvaguardas que lo estabilizan; y el *spike-and-slab* [13] decide de forma explícita, para cada variable, si entra o no en el modelo.

En lugar de calcular estimaciones puntuales de los parámetros, inferimos su distribución a posteriori mediante técnicas de muestreo de la familia de Monte Carlo basada en cadenas de Markov (MCMC). Como en general no existe una solución en forma cerrada, estos métodos generan muchas muestras representativas de las soluciones plausibles y trabajan con ellas; los algoritmos modernos, como el *No-U-Turn Sampler* [17] que emplea la librería PyMC utilizada en este trabajo, exploran ese espacio de soluciones de forma mucho más eficiente que los métodos clásicos cuando hay muchas variables. La calidad del muestreo se verifica con diagnósticos estándar de convergencia, y la capacidad predictiva fuera de muestra se estima sin reentrenar el modelo mediante validación cruzada *leave-one-out* aproximada [40]. La formalización completa de estos modelos y del procedimiento de inferencia se desarrolla en el Capítulo 3 (Sección 3.2.1); aquí basta con situar cada pieza dentro del flujo de análisis.

2.3.1. Reducción de dimensión y descubrimiento de estructura

Antes de modelar la estructura por niveles conviene resolver un problema previo: resumir las numerosas variables originales en unas pocas que capturen lo esencial. Esta reducción de dimensión es, además, el paso del que se deriva la jerarquía que emplean los modelos del apartado siguiente –por eso se presenta antes–: la agrupación de las especies en el espacio reducido es la que define el árbol de decisiones de la clasificación jerárquica (Sección 2.3.2). El *Gaussian Process Latent Variable Model* (GPLVM) [23] realiza esta reducción de forma no lineal y puede entenderse como una versión flexible del análisis de componentes principales que, en lugar de proyectar sobre combinaciones lineales de las variables, aprende un mapa curvo capaz de desplegar la geometría real de los datos. Su

ventaja para este problema es que sitúa cada espécimen en el mapa resultante con un margen de incertidumbre asociado, lo que permite valorar la fiabilidad de la posición de cada fósil. Las formulaciones recientes lo hacen aplicable a conjuntos mayores [22], y se conoce un fallo característico –el colapso, en el que el mapa degenera y deja de ser informativo [24]– que obliga a comprobar que la estructura obtenida es real y no un artefacto. La Figura 2.1 ilustra el esquema generativo: de unas pocas coordenadas latentes a las muchas variables observadas, a través de un mapa no lineal con ruido.

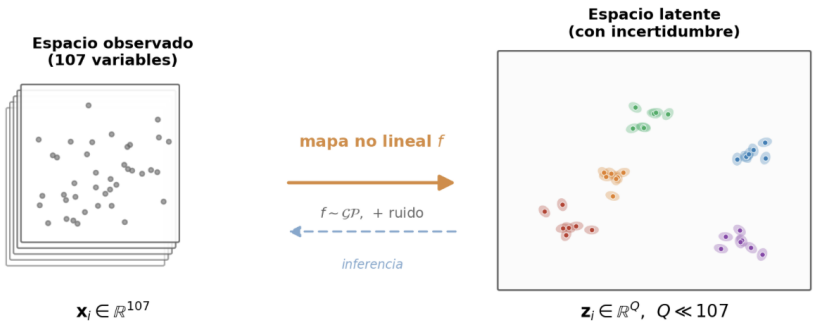

Figura 2.1: Esquema del *Gaussian Process Latent Variable Model* (GPLVM). Las observaciones de alta dimensión (las 107 variables de microtextura, izquierda) se explican como la imagen, a través de un mapa no lineal f con ruido gaussiano, de unas pocas coordenadas latentes (derecha). Cada espécimen se proyecta no como un punto sino como una distribución, cuya dispersión mide la incertidumbre de su posición.

Sobre ese mapa comparamos dos formas de descubrir agrupaciones sin conocer las especies. El proceso de Dirichlet [1] deduce automáticamente cuántos grupos hay en lugar de fijarlo de antemano, pero da por supuesto que los datos forman agrupaciones bien separadas; cuando en realidad describen un continuo, tiende a concentrarlo casi todo en un único grupo. El *Bayesian Hierarchical Clustering* (BHC) [16] construye en cambio un árbol de agrupaciones decidiendo cada unión mediante una comparación probabilística rigurosa, y no una distancia arbitraria: en cada paso compara la hipótesis de que los datos de dos ramas provienen de un mismo grupo frente a la de que provienen de grupos distintos, y solo las fusiona si la primera es más probable. Al tratarse de un modelo probabilístico, puede además situar un punto nuevo en el árbol indicando con qué probabilidad pertenece a cada rama, propiedad que se aprovecha para colocar los fósiles dentro de la estructura de los actuales sin imponerles una etiqueta. El resultado de este proceso es un dendrograma como el de la Figura 2.2, que ordena las especies según su afinidad y proporciona directamente la jerarquía empleada después.

2.3.2. Modelos jerárquicos

Una extensión natural del enfoque Bayesiano es el modelado jerárquico, que organiza los parámetros por niveles de modo que los grupos emparentados comparten información a

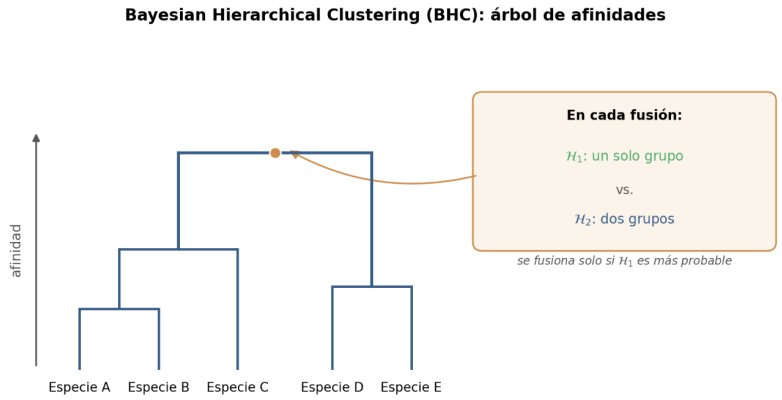

Figura 2.2: Esquema del *Bayesian Hierarchical Clustering* (BHC). El algoritmo construye un árbol de afinidades fusionando en cada paso las dos ramas para las que la hipótesis de un único grupo es más probable que la de dos grupos separados. El árbol resultante ordena los taxones por similitud y define la jerarquía que utiliza el clasificador en cascada.

través de un nivel superior común. La idea es intuitiva en este problema: las especies de un mismo género no son independientes, sino que se parecen entre sí, y un modelo que ignore ese parentesco desperdicia información. Para entender qué aporta, conviene contrastar tres estrategias. Estimar un único modelo común a todas las especies (*pooling* completo) ignora las diferencias reales entre ellas; estimar un modelo por especie de forma independiente (*ausencia de pooling*) desperdicia la información compartida y sobreajusta gravemente las especies con pocos especímenes. El modelo jerárquico adopta una vía intermedia: cada especie tiene sus propios parámetros, pero todos se vinculan a una distribución común de nivel superior, lo que "atrae" las estimaciones de las clases con pocos datos hacia el comportamiento general sin imponerles una media única. Este efecto –conocido como *borrowing strength*– estabiliza las estimaciones de los grupos pequeños apoyándose en los demás [10, 28]. La Figura 2.3 ilustra el contraste: frente a las estimaciones extremas que produce la ausencia de *pooling* en las clases minoritarias, el modelo jerárquico las modera de forma proporcional a la cantidad de datos disponible.

Este enfoque es una de las herramientas dominantes en la ecología cuantitativa actual, porque permite separar la variabilidad real del fenómeno de la introducida por el muestreo y propagar la incertidumbre de forma coherente entre los distintos niveles [42, 19, 14]. De hecho, el propio estudio de referencia sobre este dataset señala el modelado jerárquico Bayesiano como una de las vías más prometedoras para estabilizar estimaciones y compartir información entre individuos, especies y grupos dietéticos [8]. En este trabajo se aprovecha en dos sentidos: uno estructural, al descomponer la clasificación según la jerarquía taxonómica –resolviendo primero el grupo y después la especie, motivados por que los errores del clasificador plano se concentran entre especies del mismo grupo–, y otro estadístico, al estabilizar las estimaciones de las clases más pequeñas. La jerarquía concreta no se postula a priori, sino que se deriva del agrupamiento descrito en la Sección 2.3.1, cerrando así el flujo: reducción de dimensión → descubrimiento de estructura → clasificación jerárquica.

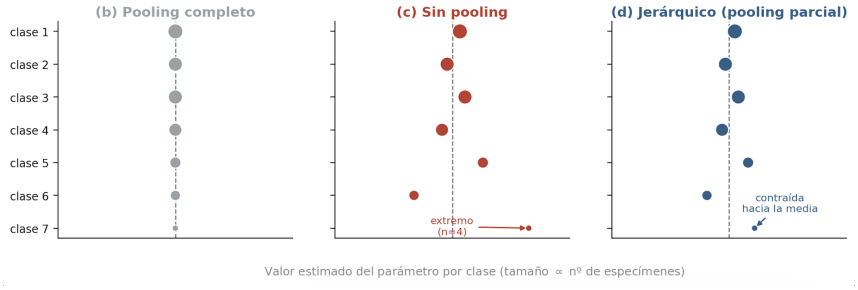

Figura 2.3: Efecto del *pooling* parcial en un modelo jerárquico. La ausencia de *pooling* (centro) produce estimaciones extremas e inestables en las clases con pocos especímenes; el modelo jerárquico (derecha) las contrae hacia la media de la población de forma proporcional a la información disponible, sin colapsarlas en un único valor común como hace el *pooling* completo (izquierda).

2.3.3. Modelos de mezcla y comparación entre actuales y fósiles

El último ingrediente del flujo responde a una pregunta que precede a cualquier intento de caracterizar los fósiles a partir de los actuales: ¿son unos y otros siquiera comparables? Proyectar un fósil sobre un clasificador entrenado solo en especies actuales únicamente tiene sentido si el fósil cae dentro de la región del espacio de variables que esas especies ocupan; en caso contrario, la predicción sería una extrapolación hacia una zona que el modelo nunca ha visto. Para responderla modelamos cómo se distribuyen los especímenes actuales en el espacio de variables mediante un modelo de mezcla gaussiana, que describe esa nube de puntos como la superposición de varias agrupaciones gaussianas, cada una con su centro y su dispersión [3]. Una mezcla así es flexible: puede ajustarse a nubes de forma irregular sin presuponer que los datos forman un único grupo compacto.

Una vez aprendida esa distribución a partir de los actuales, se mide cómo de bien encaja cada fósil en ella mediante su verosimilitud: un valor alto indica que el fósil cae en una región densamente poblada por los actuales; un valor muy bajo, que ocupa una zona prácticamente vacía. Si los fósiles encajan sistemáticamente peor que los propios actuales, ocupan regiones que el modelo apenas ha observado, y clasificarlos con un modelo entrenado solo en actuales equivaldría a extrapolar fuera de su rango de validez –un fenómeno conocido como *domain shift*, en el que las distribuciones de entrenamiento y de aplicación difieren. Este diagnóstico es el que justifica, más adelante, tratar los fósiles de forma no supervisada con las herramientas de la Sección 2.3.1 en lugar de imponerles una etiqueta. La diferencia entre ambas distribuciones de verosimilitud puede además cuantificarse y contrastarse estadísticamente, de modo que la no comparabilidad deja de ser una sospecha cualitativa y se convierte en una medida.

2.4. Interpretabilidad y aprendizaje automático explicable

El compromiso entre acierto y transparencia señalado en el apartado 2.2 –los modelos más precisos suelen ser los más opacos, y a la inversa– es el objeto del aprendizaje automático explicable [30]. En un trabajo científico esta cuestión es central, porque una

predicción solo sirve como hipótesis si puede atribuirse a variables con sentido biológico. Existen dos estrategias complementarias. Los modelos transparentes por diseño –como los lineales regularizados o la selección Bayesiana de variables empleados en este trabajo– son explicables por sí mismos, pues sus propios coeficientes revelan qué variables pesan. Los métodos *a posteriori*, como SHAP [25], abren en cambio la caja negra de un modelo complejo repartiendo cada predicción entre las variables que la motivan. Su uso exige, no obstante, dos precauciones. La primera es que una buena explicación debe ser comprensible y útil para el experto del dominio, no solo técnicamente correcta [29]. La segunda es que las atribuciones *a posteriori* pueden ser inestables, por lo que conviene validarlas antes de extraer conclusiones [30]. Por estos motivos este trabajo se apoya preferentemente en modelos transparentes por diseño, que identifican qué familias de descriptores (ISO, SSFA u *Other*) gobiernan la clasificación de forma directa.

2.5. Posicionamiento del trabajo

La revisión anterior sitúa este trabajo en la intersección del modelado probabilístico, la selección de variables y una aplicación antropológica real, y permite encadenar sus piezas como un flujo coherente. La microtextura bucal discrimina dietas, pero con una resolución entre especies limitada [26], lo que plantea la pregunta cuantitativa sobre su techo de información. En el régimen de muchas variables y pocas muestras, los métodos clásicos ceden ante la regularización [37, 44, 5], calculada mediante muestreo MCMC [17], y el marco Bayesiano aporta la cuantificación de incertidumbre necesaria para tratar muestras sin etiqueta como los fósiles [10, 31]. Sobre esa base, la reducción de dimensión y el agrupamiento Bayesiano exploran la estructura no supervisada de los datos y de ella se deriva una jerarquía [23, 16]; los modelos jerárquicos respetan esa estructura taxonómica y estabilizan las clases minoritarias [42]; un modelo de mezcla sobre los actuales decide si los fósiles son siquiera comparables [3]; y la interpretabilidad mantiene todo el proceso contrastable como hipótesis biológica [30].

La aportación de este trabajo no es, por tanto, un método nuevo a nivel técnico, sino una arquitectura de análisis que integra estas piezas conocidas –en un dominio donde apenas se habían empleado– bajo un protocolo de validación homogéneo y con cuantificación explícita de la incertidumbre, convirtiendo la conclusión cualitativa de los estudios previos –resolución de grupo por encima de la de especie– en una medida cuantitativa de ese límite informativo.

Capítulo 3

Sistema desarrollado: arquitectura de análisis

3.1. Planteamiento del problema

3.1.1. Definición del problema y objetivos específicos

El problema se formula sobre un conjunto de 160 especímenes de cercopitécidos descritos por 107 variables continuas de microtextura dental bucal, divididos en dos poblaciones: 125 especímenes *actuales* de siete especies con dieta y etiqueta conocidas, y 35 *fósiles* de cinco taxones sin verdad de referencia. El trabajo parte de una premisa que condiciona todas las decisiones de modelado: los datos son ruidosos, de tamaño reducido y de alta dimensión, y combinan esas dos poblaciones de naturaleza distinta. Para alcanzar el objetivo general (Sección 1.2) se establecen los siguientes objetivos específicos:

1. **Desarrollar un *pipeline* de preprocesado robusto** para datos ecológicos pequeños y de alta dimensión, que aborde la "jungla de parámetros" mediante transformación de la normalidad, estandarización y reducción de colinealidad.
2. **Construir modelos probabilísticos que cuantifiquen la incertidumbre** de sus predicciones –en particular una regresión logística bayesiana multinomial con distintos *priors*– que entreguen junto a cada predicción una medida honesta de confianza.
3. **Realizar estudios de importancia de variables** que conecten el rendimiento con la interpretabilidad biológica, identificando qué descriptores (ISO, SSFA u *Other*) gobiernan la clasificación.
4. **Aprovechar la estructura latente para construir modelos jerárquicos informados**, derivando del espacio latente del GPLVM una jerarquía empírica de las especies que sirva de esqueleto a un clasificador jerárquico en cascada.
5. **Determinar la comparabilidad de los fósiles y caracterizar sus afinidades con incertidumbre cuantificada**, por una vía no supervisada (espacio latente y *clustering*) y una supervisada (proyección por el clasificador jerárquico), sin imponer una correspondencia unívoca.

3.1.2. Arquitectura de tres capas y métricas de evaluación

Los objetivos se articulan en una arquitectura de análisis en tres capas (Tabla 3.1), cada una con su pregunta de investigación: la Capa 1 mide el techo informativo supervisado sobre los actuales; la Capa 2 decide si los fósiles son distribucionalmente comparables con los actuales; y la Capa 3 caracteriza las afinidades de los fósiles. La validación se ancla en todo momento a un protocolo homogéneo de validación cruzada *leave-one-out*, comparando contra *baselines* triviales (clase mayoritaria, aleatorio estratificado) y empleando métricas robustas al desbalanceo –exactitud balanceada, F1-macro y κ de Cohen– en lugar de la exactitud simple. Esta elección es la que permite interpretar un rendimiento próximo al *baseline* como un techo informativo de la señal y no como un fallo del modelo.

Cuadro 3.1: Articulación de los objetivos en la arquitectura de tres capas.

Capa	**Pregunta**	**Herramienta**
Capa 1	¿Distingue la microtextura las especies vivas, y a qué nivel?	Clasificadores probabilísticos (BLR multinomial, jerárquicos) + LOO-CV; importancia de variables (XAI)
Capa 2	¿Son los fósiles comparables con los vivos?	Modelo de mezcla (GMM) + verosimilitud cruzada
Capa 3	¿A qué se aproximan los fósiles, y con cuánta certeza?	GPLVM + *clustering* bayesiano (no superv.) y proyección por el clasificador jerárquico (superv.)

3.2. Diseño de la solución

Esta sección desarrolla la realización concreta de la arquitectura introducida en la Sección 3.1.2. Se organiza en cuatro bloques: primero los fundamentos teóricos comunes a todas las capas (Sección 3.2.1); después el conjunto de datos y el pipeline de preprocesado (Sección 3.2.2); a continuación la realización de cada una de las tres capas (Secciones 3.2.3–3.2.5); y por último los detalles de implementación (Sección 3.3). El hilo conductor es que cada pieza metodológica responde a una de las tres preguntas de investigación de la Tabla 3.1, bajo el protocolo de validación homogéneo fijado en la Sección 3.1.2.

3.2.1. Fundamentos teóricos

Esta sección formaliza los modelos empleados en las tres capas del trabajo; la motivación intuitiva de cada uno se introdujo ya en el estado del arte (Capítulo 2), de modo que aquí se prioriza la especificación matemática sobre la exposición divulgativa. La notación es común a todas las capas: se dispone de N especímenes, cada uno descrito por un vector de características $\mathbf{x}_i \in \mathbb{R}^D$ (con D las variables de microtextura tras el preprocesado) y,

en los actuales, una etiqueta de clase $y_i \in \{1, \dots, K\}$. El conjunto de datos se denota $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ y la matriz de diseño $\mathbf{X} \in \mathbb{R}^{N \times D}$.

La exposición se organiza en torno a tres pilares y un bloque de cierre. El primer pilar es el modelo de clasificación Bayesiano y la elección del *prior*, que en el régimen de muchas variables y pocas muestras es la decisión de modelado determinante (Secciones 3.2.1 y 3.2.1). El segundo es el procedimiento de inferencia por muestreo y sus diagnósticos (Sección 3.2.1). El tercero es la extensión jerárquica que vertebra la Capa 1 (Sección 3.2.1). Cierran el bloque los modelos no supervisados que sostienen las Capas 2 y 3: las mezclas Gaussianas (Sección 3.2.1), el GPLVM (Sección 3.2.1) y el descubrimiento probabilístico de estructura (Sección 3.2.1).

Clasificación probabilística e inferencia Bayesiana

Un clasificador probabilístico modela la distribución condicional $p(y \mid \mathbf{x}, \boldsymbol{\theta})$ con parámetros $\boldsymbol{\theta}$. El enfoque frecuentista resume toda la información sobre los parámetros en un único valor: la estimación por máxima verosimilitud (MLE) elige el $\hat{\boldsymbol{\theta}}$ que maximiza $p(\mathcal{D} \mid \boldsymbol{\theta})$, y la estimación máxima a posteriori (MAP) añade un *prior* y maximiza $p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$. Ambas devuelven un *estimador puntual*, que en regímenes de pocos datos y muchas variables tiende al sobreajuste y, sobre todo, no transmite cuánta confianza merece. El enfoque Bayesiano, en cambio, trata $\boldsymbol{\theta}$ como una variable aleatoria y, partiendo de un *prior* $p(\boldsymbol{\theta})$, aplica el teorema de Bayes para obtener la distribución a posteriori *completa*

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\displaystyle\int p(\mathcal{D} \mid \boldsymbol{\theta}')\, p(\boldsymbol{\theta}')\, d\boldsymbol{\theta}'}. \tag{3.1}$$

La MAP no es más que el modo de esta posterior; el enfoque Bayesiano conserva toda su forma. La predicción para un nuevo $\mathbf{x}_*$ no usa un único valor de $\boldsymbol{\theta}$, sino que promedia sobre toda la posterior mediante la *distribución predictiva posterior*

$$p(y_* \mid \mathbf{x}_*, \mathcal{D}) = \int p(y_* \mid \mathbf{x}_*, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, d\boldsymbol{\theta}. \tag{3.2}$$

Esta integral es la que entrega, junto a la etiqueta, una medida de confianza, y la que permite descomponer la incertidumbre en dos fuentes [10, 31]: la *aleatoria*, asociada a la dispersión intrínseca de $p(y_* \mid \mathbf{x}_*, \boldsymbol{\theta})$ (ruido irreducible), y la *epistémica*, asociada a la dispersión de $p(\boldsymbol{\theta} \mid \mathcal{D})$ (desconocimiento del modelo, reducible con más datos).

Regresión logística Bayesiana multinomial. Para K clases, el modelo central del trabajo asigna a cada clase k un vector de coeficientes $\boldsymbol{\beta}_k \in \mathbb{R}^D$ y un sesgo α_k, y modela la probabilidad de clase mediante la función *softmax*

$$p(y_i = k \mid \mathbf{x}_i) = \frac{\exp(\boldsymbol{\beta}_k^\top \mathbf{x}_i + \alpha_k)}{\sum_{j=1}^{K} \exp(\boldsymbol{\beta}_j^\top \mathbf{x}_i + \alpha_j)}, \qquad y_i \sim \text{Categorical}(p(y_i = 1), \dots, p(y_i = K)). \tag{3.3}$$

Por identificabilidad se fija una clase de referencia ($\boldsymbol{\beta}_1 = \mathbf{0}$, $\alpha_1 = 0$). La verosimilitud completa es el producto de las categóricas sobre los N especímenes. A falta de una forma

cerrada para la posterior (3.1), esta se aproxima por muestreo (Sección 3.2.1). El ingrediente que falta por especificar —y que, en el régimen de muchas variables y pocas muestras de este trabajo, determina por completo el comportamiento del modelo— es el *prior* sobre los coeficientes, al que se dedica el bloque siguiente.

Elección del prior: el núcleo del modelo

En el régimen de muchas variables y pocas muestras, el *prior* $p(\boldsymbol{\beta})$ deja de ser un detalle técnico y se convierte en la decisión de modelado más importante de la Capa 1: cumple el papel de regularizador —determina cómo se contraen los coeficientes hacia cero y, por tanto, qué variables sobreviven— y, al hacerlo, codifica una hipótesis biológica explícita sobre cuántos descriptores de microtextura se espera que gobiernen la discriminación. Comparar *priors* no es, por tanto, comparar variantes intercambiables del mismo modelo, sino contrastar hipótesis distintas sobre la estructura de la señal. El trabajo evalúa cuatro especificaciones, todas aplicadas de forma independiente a cada coeficiente β_{kd} (se omiten los subíndices por claridad); la Figura 3.1 las contrasta visualmente.

Prior Gaussiano (*ridge*). El prior de referencia,

$$\beta \sim \mathcal{N}(0, \sigma_0^2), \tag{3.4}$$

penaliza la magnitud al cuadrado de los coeficientes (equivale a regularización L_2). Contrae de forma suave y no produce ceros exactos: ninguna variable se elimina. Codifica la hipótesis de que todos los descriptores aportan un poco.

Prior de Laplace (*Lasso* Bayesiano). El prior de doble exponencial

$$\beta \sim \text{Laplace}(0, b), \qquad p(\beta) = \frac{1}{2b} \exp\Big(-\frac{|\beta|}{b}\Big), \tag{3.5}$$

es el análogo Bayesiano del *Lasso* [37]: su pico acentuado en cero induce contracción agresiva de los coeficientes pequeños, favoreciendo soluciones dispersas. Su moda a posteriori coincide con la solución L_1.

Prior *horseshoe*. Un prior de contracción *global-local* que parte de que casi todos los coeficientes son nulos y unos pocos grandes [5]:

$$\beta_d \sim \mathcal{N}(0, \tau^2 \lambda_d^2), \qquad \tau \sim \mathcal{C}^+(0, 1), \qquad \lambda_d \sim \mathcal{C}^+(0, 1), \tag{3.6}$$

donde $\mathcal{C}^+$ es la semi-Cauchy. El parámetro global τ empuja todos los coeficientes hacia cero, mientras que las escalas locales λ_d, con colas pesadas, permiten que los coeficientes verdaderamente grandes *escapen* de la contracción. El resultado es un *shrinkage* adaptativo: fuerte para el ruido, casi nulo para la señal, sin necesidad de un umbral duro. En la práctica se implementa con la parametrización no centrada $\beta_d = z_d\, \tau\, \lambda_d$ con $z_d \sim \mathcal{N}(0, 1)$, que mejora el muestreo. La variante *regularizada* [32] añade una escala de losa c que acota los coeficientes grandes y un τ_0 que codifica el número esperado de variables activas.

Prior *spike-and-slab*. Modela cada coeficiente como una mezcla de dos componentes [13]: un *pico* estrecho en cero (variable irrelevante) y una *losa* difusa (variable relevante),

$$\beta \sim w\,\mathcal{N}(0, \sigma^2_{\text{slab}}) + (1-w)\,\mathcal{N}(0, \sigma^2_{\text{spike}}), \qquad \sigma_{\text{spike}} \ll \sigma_{\text{slab}}, \qquad w \sim \text{Beta}(a_0, b_0), \quad (3.7)$$

donde w es la probabilidad de inclusión de la variable. Induce *sparsity* casi exacta y es directamente interpretable (la posterior de w es la probabilidad de que la variable importe), a costa de un mayor coste de muestreo por la naturaleza de mezcla.

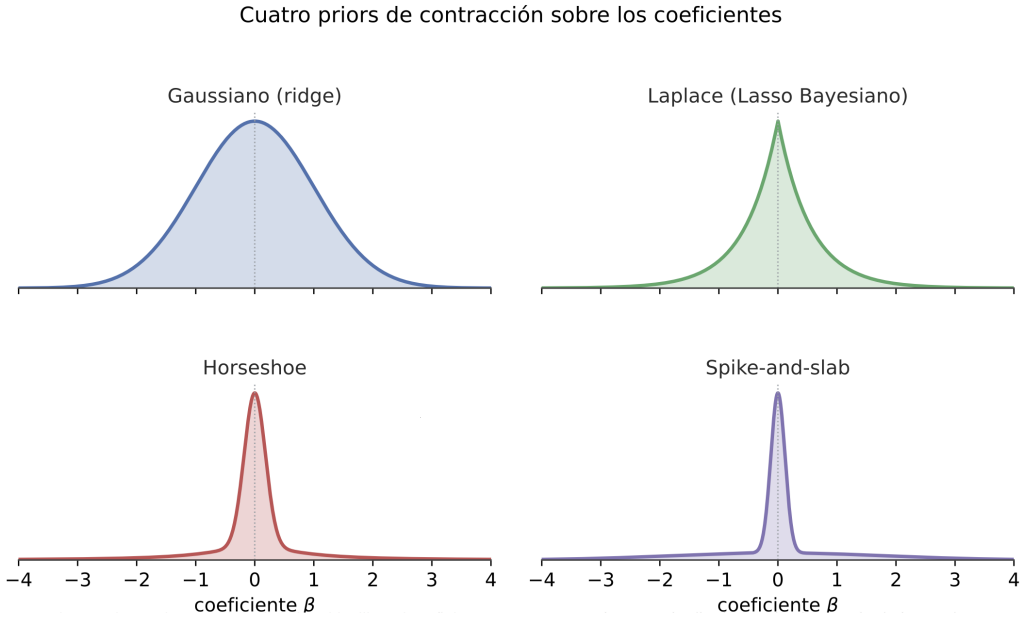

Figura 3.1: Comparativa de los cuatro *priors* de contracción evaluados sobre los coeficientes β. El Gaussiano (*ridge*) contrae de forma suave sin anular variables; el de Laplace (*Lasso* Bayesiano) concentra masa en cero favoreciendo la dispersión; el *horseshoe* combina un pico agudo en cero con colas pesadas, silenciando el ruido sin penalizar la señal fuerte; y el *spike-and-slab* mezcla un pico estrecho (variable irrelevante) con una losa ancha (variable relevante). Cada forma codifica una hipótesis distinta sobre cuántos descriptores portan señal.

Inferencia por MCMC y diagnósticos

Especificados el modelo y su *prior*, queda el problema central de toda aplicación Bayesiana: calcular la posterior. El obstáculo práctico es el denominador de la ecuación (3.1): la evidencia $p(\mathcal{D}) = \int p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ es una integral en dimensión alta que rara vez admite forma cerrada. Cuando no hay conjugación entre *prior* y verosimilitud —como ocurre con todos los *priors* de la Sección 3.2.1 bajo la verosimilitud *softmax*— la posterior es intratable analíticamente y debe aproximarse. La estrategia adoptada es el muestreo de Monte Carlo basado en cadenas de Markov (MCMC).

De Monte Carlo a las cadenas de Markov. La integración de Monte Carlo estima una esperanza posterior por el promedio sobre muestras, $\mathbb{E}[f(\boldsymbol{\theta})] \approx \frac{1}{S}\sum_s f(\boldsymbol{\theta}^{(s)})$, que converge al valor verdadero por la ley de los grandes números. El problema es generar las muestras $\boldsymbol{\theta}^{(s)}$: en dimensión alta no es posible muestrear directamente de la posterior, y los métodos elementales (muestreo por rechazo, muestreo de importancia) degeneran porque la región de probabilidad apreciable ocupa una fracción ínfima del espacio. MCMC sortea esto generando una *secuencia de muestras dependientes* mediante una cadena de Markov diseñada *ad hoc*: en lugar del problema habitual —dado un núcleo de transición, hallar su distribución estacionaria— se invierte el planteamiento, fijando como distribución estacionaria objetivo la propia posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$ y construyendo un núcleo de transición que converja a ella (Figura 3.2). Tras un periodo de mezcla, las muestras de la cadena provienen de la posterior y aproximan cualquier esperanza, en particular la integral predictiva (3.2) mediante $p(y_* \mid \mathbf{x}_*, \mathcal{D}) \approx \frac{1}{S}\sum_s p(y_* \mid \mathbf{x}_*, \boldsymbol{\theta}^{(s)})$.

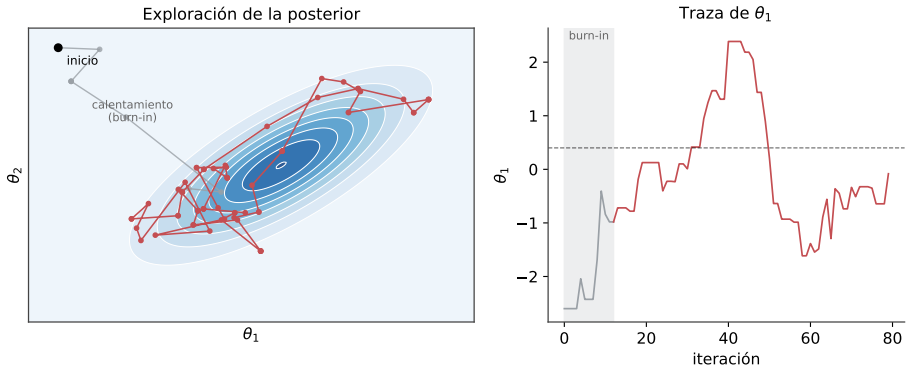

Figura 3.2: Esquema del muestreo MCMC. La cadena explora el espacio de parámetros generando muestras dependientes que, tras el periodo de calentamiento, se distribuyen según la posterior objetivo. El promedio sobre esas muestras aproxima cualquier esperanza posterior, incluida la distribución predictiva.

Monte Carlo Hamiltoniano y NUTS. Los muestreadores clásicos como Metropolis–Hastings proponen el siguiente estado mediante un paseo aleatorio, cuya distancia recorrida crece solo como $\sqrt{S}$ y resulta muy ineficiente en dimensión alta. El Monte Carlo Hamiltoniano (HMC) sustituye ese paseo por una analogía física (Figura 3.3): trata el negativo del log-posterior como una energía potencial $U(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta} \mid \mathcal{D})$, introduce un *momento* auxiliar $\mathbf{r}$ con energía cinética, y simula la trayectoria de una partícula sin fricción sobre esa superficie de energía. Al usar el *gradiente* del log-posterior para guiar la trayectoria, HMC propone estados lejanos con alta probabilidad de aceptación, eliminando el comportamiento de paseo aleatorio. El *No-U-Turn Sampler* (NUTS) [17] automatiza el único hiperparámetro delicado de HMC —la longitud de la trayectoria—, deteniéndola cuando empieza a retroceder sobre sí misma. Es el muestreador empleado en este trabajo a través de PyMC.

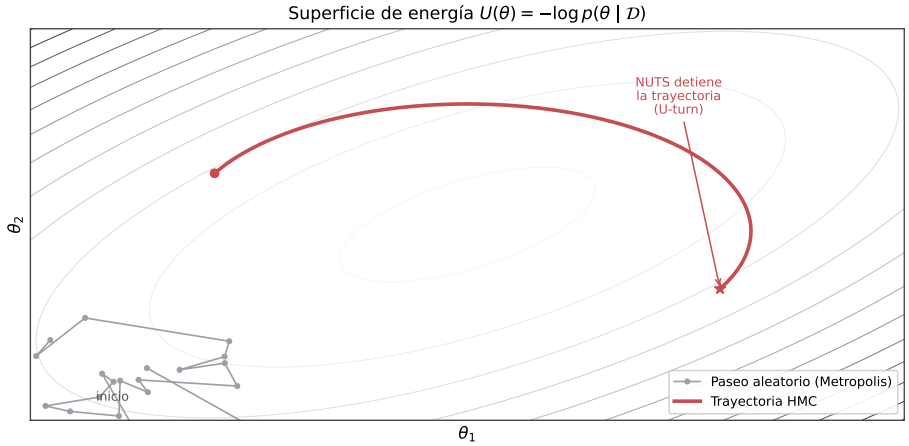

Figura 3.3: Monte Carlo Hamiltoniano (HMC) frente al paseo aleatorio. HMC interpreta el negativo del log-posterior como una superficie de energía y simula la trayectoria de una partícula que se desliza por ella usando el gradiente, proponiendo estados lejanos con alta tasa de aceptación. El *No-U-Turn Sampler* (NUTS) detiene cada trayectoria automáticamente cuando empieza a retroceder, evitando ajustar a mano su longitud.

Práctica y diagnósticos. Las primeras muestras de la cadena, antes de alcanzar la región típica de la posterior, se descartan (*burn-in* o calentamiento). Para reducir la dependencia del punto de partida y poder diagnosticar la convergencia, se ejecutan varias cadenas en paralelo desde inicios distintos. La fiabilidad del muestreo se controla con los diagnósticos estándar: el estadístico $\hat{R}$, que compara la varianza entre cadenas con la varianza intra-cadena ($\hat{R} \approx 1$ indica convergencia) [12, 41]; el tamaño de muestra efectivo (ESS), que mide cuántas muestras independientes equivalen a las muestras correladas de la cadena [41]; y el recuento de *transiciones divergentes*, que señalan regiones donde el integrador de HMC falla por curvatura excesiva de la posterior [2] (véase la Sección 3.2.1).

Evaluación predictiva: PSIS-LOO. La capacidad de generalización se estima mediante validación cruzada *leave-one-out* aproximada por muestreo de importancia con suavizado de Pareto (PSIS-LOO), que reutiliza las muestras de la posterior completa sin reentrenar el modelo N veces [40]. La métrica resultante es la densidad predictiva logarítmica esperada (ELPD) [40]; el diagnóstico $\hat{k}$ de Pareto señala las observaciones para las que la aproximación no es fiable y conviene tratar de forma exacta.

Modelos jerárquicos y clasificación en cascada

El tercer pilar teórico, y el que vertebra la arquitectura de la Capa 1, es el modelado jerárquico. Su relevancia en este problema es doble: por un lado estabiliza las estimaciones de las clases con muy pocos especímenes —situación ubicua en este conjunto de datos—, y por otro permite descomponer la decisión multiclase siguiendo la estructura taxonómica, allí donde los errores del clasificador plano se concentran.

Jerarquía estadística: pooling parcial. Cuando los datos se organizan en grupos relacionados pero distintos —en este trabajo, las especies dentro de un mismo género— caben tres estrategias [11, 10]. El *pooling completo* fusiona todos los datos y estima un único parámetro global, ignorando las diferencias entre grupos. La *ausencia de pooling* estima cada grupo por separado, lo que desperdicia la información compartida y sobreajusta gravemente los grupos con pocos datos. El modelo jerárquico adopta una vía intermedia, el *pooling parcial*: impone a los parámetros de cada grupo una distribución común de nivel superior, $\boldsymbol{\theta}_g \sim p(\boldsymbol{\theta} \mid \boldsymbol{\phi})$, cuyos *hiperparámetros* $\boldsymbol{\phi}$ son a su vez desconocidos y reciben su propio *hiperprior* $p(\boldsymbol{\phi})$. La posterior conjunta factoriza como

$$p(\{\boldsymbol{\theta}_g\}, \boldsymbol{\phi} \mid \mathcal{D}) \;\propto\; p(\boldsymbol{\phi}) \prod_g p(\boldsymbol{\theta}_g \mid \boldsymbol{\phi})\, p(\mathcal{D}_g \mid \boldsymbol{\theta}_g). \tag{3.8}$$

El acoplamiento a través de $\boldsymbol{\phi}$ produce el *shrinkage* parcial: las estimaciones de los grupos con datos escasos o ruidosos se "atraen" hacia la media de la población, mientras que los grupos bien informados apenas se desplazan y, a su vez, determinan $\boldsymbol{\phi}$, que actúa como un regularizador guiado por los datos [11, 10, 28]. Este *borrowing strength* evita predicciones extremas en los grupos de muy pocos especímenes —situación habitual en este conjunto de datos— sin silenciar a los grupos con muestra suficiente.

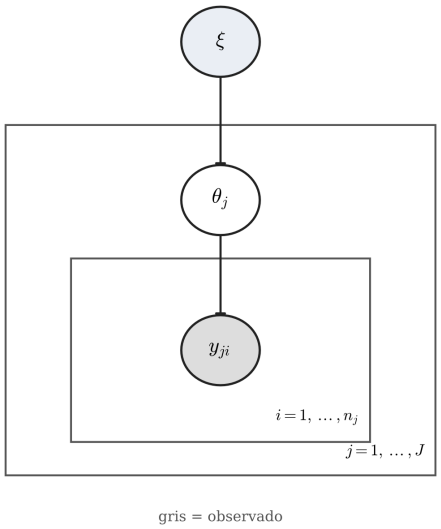

Figura 3.4: Diagrama de placas de un modelo jerárquico. El hiperparámetro $\boldsymbol{\xi}$ (fuera de la placa) gobierna los parámetros de grupo $\boldsymbol{\theta}_j$, que a su vez generan las observaciones y_{ji}. La dependencia a través de $\boldsymbol{\xi}$ es la que permite compartir información entre grupos (*shrinkage*).

Modelos lineales generalizados multinivel. El marco anterior se extiende a la regresión haciendo que los pesos del modelo varíen por grupo, $y_{gi} \sim p(\cdot \mid g^{-1}(\mathbf{x}_{gi}^{\top}\boldsymbol{\beta}_g))$ con g^{-1} la función de enlace inversa (la *softmax* en este trabajo) y $\boldsymbol{\beta}_g \sim \mathcal{N}(\boldsymbol{\beta}_0, \cdot)$. La componente global $\boldsymbol{\beta}_0$ recoge los efectos compartidos por todos los grupos (efectos fijos) y las

desviaciones $\boldsymbol{\beta}_g - \boldsymbol{\beta}_0$ los efectos específicos de cada grupo (efectos aleatorios). Este es el fundamento del clasificador jerárquico empleado en la Capa 1.

Clasificación jerárquica en cascada. En sentido estructural, la clasificación multiclase se descompone siguiendo una jerarquía taxonómica de nodos. En cada nodo interno se entrena un clasificador (aquí, una regresión logística Bayesiana con uno de los *priors* anteriores) que decide hacia qué rama descender. La probabilidad de una especie hoja k se obtiene por la regla de la probabilidad total, como producto de las probabilidades condicionales a lo largo del camino $\text{path}(k)$ desde la raíz hasta la hoja:

$$p(y = k \mid \mathbf{x}) = \prod_{(m \to c) \in \text{path}(k)} p(c \mid m, \mathbf{x}), \tag{3.9}$$

donde $p(c \mid m, \mathbf{x})$ es la probabilidad de tomar la rama hija c en el nodo m. La estructura de la jerarquía puede fijarse por la taxonomía conocida o *derivarse de los datos* a partir de la estructura latente recuperada por el GPLVM (Sección 3.2.1): un dendrograma sobre los centroides de las especies en el espacio latente define el árbol de decisiones binarias.

Modelos de mezcla Gaussiana (GMM)

Un modelo de mezcla Gaussiana representa la densidad de los datos como una combinación convexa de K componentes Gaussianas:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad \pi_k \geq 0, \quad \sum_{k=1}^{K} \pi_k = 1, \tag{3.10}$$

con pesos π_k, medias $\boldsymbol{\mu}_k$ y covarianzas $\boldsymbol{\Sigma}_k$. Los parámetros se estiman maximizando la verosimilitud mediante el algoritmo *Expectation–Maximization* (EM) [3], que alterna el cálculo de las responsabilidades $\gamma_{ik} = p(z_i = k \mid \mathbf{x}_i)$ (paso E) y la actualización de $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ (paso M).

El GMM como detector de comparabilidad (Capa 2). Una vez ajustado un GMM sobre los especímenes actuales, la log-verosimilitud $\log p(\mathbf{x})$ que asigna a un espécimen fósil mide cuánto encaja este en la densidad aprendida: un valor alto indica que el fósil cae en una región densamente poblada por los actuales; un valor muy bajo, que ocupa una zona que el modelo apenas ha observado. Comparando la distribución de log-verosimilitudes que el modelo asigna a los propios actuales con la que asigna a los fósiles se determina si estos últimos caen fuera del soporte de los actuales —en cuyo caso su proyección sobre un clasificador supervisado constituiría una extrapolación. Este es el mecanismo sobre el que se construye la Capa 2.

Reducción de dimensión: GPLVM

El *Gaussian Process Latent Variable Model* (GPLVM) [23] es un modelo generativo que asume que las observaciones de alta dimensión $\mathbf{X} \in \mathbb{R}^{N \times D}$ se generan a partir de

variables latentes de baja dimensión $\mathbf{Z} \in \mathbb{R}^{N\times Q}$ ($Q \ll D$) mediante un mapa no lineal f con ruido Gaussiano (Figura 3.5):

$$\mathbf{x}_{:,d} = f_d(\mathbf{Z}) + \boldsymbol{\epsilon}_d, \qquad f_d \sim \mathcal{GP}(0,\, k(\cdot,\cdot)), \qquad \boldsymbol{\epsilon}_d \sim \mathcal{N}(0, \sigma^2\mathbf{I}), \tag{3.11}$$

donde cada dimensión observada se modela con un proceso Gaussiano independiente [33] de función de covarianza (*kernel*) k. Puede verse como una generalización no lineal del PCA probabilístico: marginalizando el mapa f se obtiene una verosimilitud cuyos parámetros son las posiciones latentes $\mathbf{Z}$ y los hiperparámetros del *kernel*.

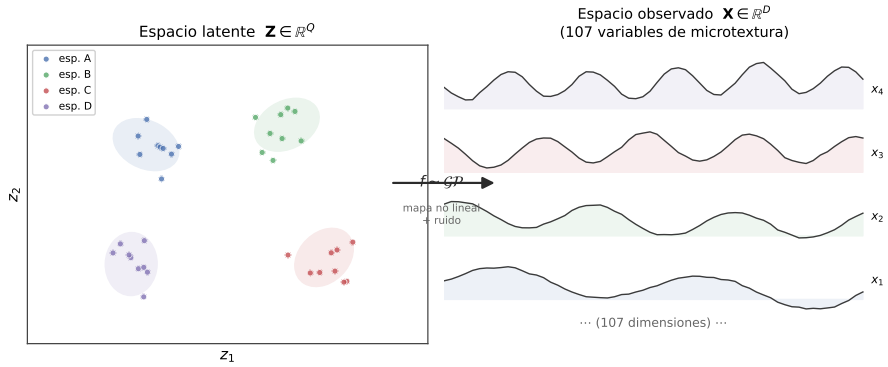

Figura 3.5: Modelo generativo del GPLVM. Cada espécimen se describe mediante unas pocas coordenadas latentes $\mathbf{z}_i$; un mapa no lineal f, modelado como un proceso Gaussiano, las transforma en las 107 variables de microtextura observadas, con ruido Gaussiano. Invertir este mapa —inferir $\mathbf{Z}$ a partir de $\mathbf{X}$— es lo que produce la representación de baja dimensión.

Formulación Bayesiana variacional. En lugar de optimizar $\mathbf{Z}$ puntualmente, la versión Bayesiana le impone un prior $p(\mathbf{Z}) = \prod_i \mathcal{N}(\mathbf{z}_i \mid \mathbf{0}, \mathbf{I})$ y aproxima su posterior por una familia variacional $q(\mathbf{Z}) = \prod_i \mathcal{N}(\mathbf{z}_i \mid \mathbf{m}_i, \mathbf{s}_i)$, maximizando la cota inferior de la evidencia (ELBO). Cada espécimen recibe así no un punto, sino una distribución $\mathcal{N}(\mathbf{m}_i, \mathbf{s}_i)$ en el espacio latente: una proyección *con incertidumbre* $\mathbf{s}_i$, que permite valorar la fiabilidad de la posición de cada fósil. El uso de *puntos inductores* hace escalable la inferencia [22]; el empleo de un *kernel* de Matérn con inicialización por PCA y un pequeño ruido evita el colapso del espacio latente [24].

Descubrimiento probabilístico de estructura

Sobre el espacio latente se busca estructura sin usar las etiquetas. Se contrastan dos enfoques Bayesianos no paramétricos.

Mezcla por proceso de Dirichlet (DP). Extiende el GMM (3.10) a un número de componentes potencialmente infinito, con los pesos generados por un proceso de Dirichlet de parámetro de concentración α [1]. En la construcción *stick-breaking* (Figura 3.6), los pesos decrecen de forma que solo unos pocos componentes reciben masa apreciable: el

modelo *infiere* el número efectivo de *clusters* en lugar de fijarlo. Un α pequeño favorece pocas componentes. Su supuesto de agrupaciones Gaussianas discretas falla cuando el soporte es continuo, caso en el que la masa colapsa en un único componente.

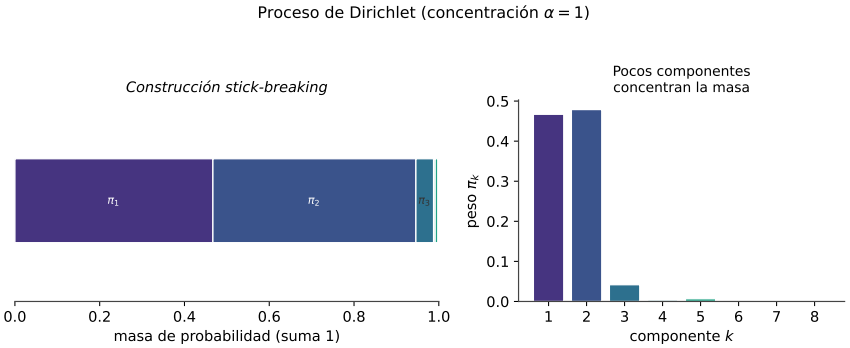

Figura 3.6: Construcción *stick-breaking* de un proceso de Dirichlet. La masa de probabilidad unidad se reparte en trozos sucesivamente menores, de modo que solo unas pocas componentes reciben peso apreciable: el modelo infiere cuántos grupos hay en lugar de fijarlo de antemano. El parámetro de concentración α controla cuánta masa se reparte entre más o menos componentes.

Bayesian Hierarchical Clustering (BHC). Es un algoritmo aglomerativo que, en cada paso, decide si fusionar dos *clusters* $\mathcal{D}_i$ y $\mathcal{D}_j$ comparando dos hipótesis [16]: $\mathcal{H}_1$, que todos los datos de la unión proceden de un mismo componente, frente a la alternativa de que provienen de subárboles separados. La decisión se basa en la razón de probabilidades marginales

$$r_k = \frac{\pi_k \, p(\mathcal{D}_k \mid \mathcal{H}_1)}{p(\mathcal{D}_k \mid \mathcal{T}_k)}, \tag{3.12}$$

donde $p(\mathcal{D}_k \mid \mathcal{H}_1)$ es la verosimilitud marginal bajo el prior conjugado del modelo de mezcla y $p(\mathcal{D}_k \mid \mathcal{T}_k)$ la del árbol actual. A diferencia del agrupamiento aglomerativo clásico, BHC define un modelo probabilístico, se fundamenta en el modelo de mezcla por proceso de Dirichlet y puede asignar a un punto nuevo una probabilidad de pertenencia a cada rama, propiedad que se aprovecha para situar los fósiles en la estructura de los actuales. En la práctica se implementa con enlace UPGMA sobre la distancia de Bhattacharyya entre las Gaussianas por especie, que cuantifica el solapamiento entre dos distribuciones normales integrando diferencias de media y de covarianza.

Con esto queda formalizado el repertorio de modelos que articula las tres capas: la regresión logística Bayesiana y su *prior* de contracción, junto con su extensión jerárquica, sostienen el techo informativo supervisado de la Capa 1; la mezcla Gaussiana es el detector de comparabilidad de la Capa 2; y el GPLVM con el descubrimiento de estructura sostienen la exploración no supervisada de la Capa 3. Lo que sigue es su aplicación concreta a los datos: primero el conjunto de especímenes y el preprocesado que los acondiciona (Sección 3.2.2), y después la realización de cada capa sobre ellos (Secciones 3.2.3–3.2.5).

3.2.2. Conjunto de datos y preprocesado

Descripción del conjunto de datos

El estudio parte de la colección de referencia de acceso abierto de Martínez et al. (2022) [26], ampliada en este trabajo con especímenes fósiles. El conjunto de datos completo comprende **160 especímenes**: **125 actuales** (*extant*) pertenecientes a siete especies de cercopitécidos africanos de dieta conocida, y **35 fósiles** (*extinct*) pertenecientes a tres géneros extintos. Las superficies bucales del esmalte se escanearon mediante microscopía confocal y se procesaron en SensoMAP (Digital Surf), obteniéndose **107 variables continuas de microtextura** por espécimen, agrupadas en tres familias: parámetros areales ISO 25178/12781, parámetros de análisis fractal sensible a la escala (SSFA), y un tercer conjunto de descriptores derivados del análisis de Fourier y de surcos (en adelante, *Other*). Los metadatos categóricos (diente, lado, sexo) se conservaron por trazabilidad, pero se excluyeron del modelado.

Cuadro 3.2: Composición del conjunto de datos por taxón. Las dietas se indican de forma orientativa.

Condición	Taxón	n	Dieta (orientativa)
Actual	*Papio anubis*	24	Omnívoro generalista
	Lophocebus aterrimus	21	Frugívoro (semillas duras)
	Theropithecus gelada	21	Graminívoro
	Papio ursinus	18	Omnívoro generalista
	Papio hamadryas	17	Omnívoro generalista
	Macaca sylvanus	13	Frugívoro-folívoro
	Cercocebus atys	11	Frugívoro (alimentos duros)
Fósil	*Theropithecus oswaldi*	17	Extinto
	Parapapio s.l.	7	Extinto
	Parapapio ado	4	Extinto
	Parapapio lothagamensis	4	Extinto
	Cercopithecoides williamsi	3	Extinto
Total		*160*	

A nivel de género, los fósiles se agrupan en *Theropithecus* ($n = 17$), *Parapapio* ($n = 15$) y *Cercopithecoides* ($n = 3$), que es la resolución empleada en las tareas que requieren etiquetas de grupo fósil.

Una característica central de este conjunto de datos –y una restricción que condiciona toda la metodología– es su *reducido tamaño y elevada dimensionalidad*: con $N < 100$ por grupo de análisis y 107 predictores, la razón entre observaciones y variables queda muy por debajo de lo que requieren las arquitecturas *data-hungry*. Esto motiva el énfasis en modelos robustos e interpretables y en un protocolo de validación *leave-one-out* (Sección 3.2.1), en lugar de estimadores flexibles de caja negra.

Pipeline de preprocesado

Para garantizar una entrada consistente a todos los clasificadores, cada variable se sometió a la misma secuencia de pasos: codificación de los metadatos categóricos, evaluación de la normalidad, transformación, inspección de valores atípicos y análisis de colinealidad. A continuación se describe cada paso junto con su resultado cuantitativo.

Valores ausentes. Se auditó en primer lugar la completitud del conjunto de datos. **No se encontró ningún valor ausente** en ninguna de las 107 variables de microtextura a lo largo de los 160 especímenes, por lo que no fue necesaria ninguna imputación. Esto es coherente con el protocolo de adquisición controlado, en el que cada subárea proporciona un conjunto completo de parámetros de textura.

Codificación de variables categóricas. Los descriptores categóricos se codificaron mediante *one-hot* antes de cualquier transformación numérica, de modo que las columnas indicadoras binarias no se sometieran a la transformación de potencia (que solo tiene sentido para variables continuas). Las columnas binarias se detectaron automáticamente y se excluyeron de los pasos de normalización y estandarización.

Evaluación de la normalidad. Varios de los clasificadores del *pipeline* (en particular el LDA) asumen o se benefician de predictores aproximadamente gaussianos. Por ello, la normalidad se evaluó sobre las 107 variables continuas combinando diagnósticos visuales (diagramas de caja y gráficos Q–Q) con el test ómnibus de D'Agostino–Pearson, que combina asimetría y curtosis. En los datos originales (sin transformar), **solo 7 de las 107 variables** resultaron compatibles con la normalidad a un nivel $\alpha = 0{,}05$, lo que confirma el marcado carácter no gaussiano de los parámetros de microtextura en bruto.

Transformación para la normalidad. Las variables que no superaron el test se transformaron para aproximarse a la simetría. Si bien la transformación logarítmica es el estándar histórico en la clasificación de especies por microdesgaste [26], en este trabajo se adoptó la **transformación de potencia de Yeo–Johnson** [43], que generaliza la de Box–Cox, incluye la logarítmica como caso particular y –algo crucial para este conjunto de datos– está definida para valores nulos y negativos. El parámetro λ se estimó por variable mediante máxima verosimilitud. El efecto fue notable: el número de variables compatibles con la normalidad pasó de **7 a 88 de 107** (del 7 % al 82 %). Un subconjunto residual de **19 variables** se mantuvo marcadamente no normal incluso tras la transformación; se concentran en la familia SSFA y en los descriptores fractales y de anisotropía (p. ej. *epLsar*, *NewEplsar*, *Fractal dimension*, *SRC threshold*, *Mean density of furrows*), de colas pesadas o multimodales. Estas variables se conservaron, ya que el análisis exploratorio indicó que aún aportaban señal discriminante.

Inspección de valores atípicos. La presencia de valores extremos se examinó visualmente mediante diagramas de caja por variable. Dado el reducido tamaño muestral y la plausibilidad biológica de valores de textura extremos pero válidos –un patrón de microdesgaste atípico es un dato, no necesariamente un error–, no se aplicó ninguna eliminación

Cuadro 3.3: Normalidad de las 107 variables continuas antes y después de la transformación de Yeo–Johnson (test ómnibus, $\alpha = 0{,}05$).

	Normales	No normales
Antes de la transformación	7	100
Después de la transformación	88	19

automática de atípicos. Su influencia se controla, en cambio, por la combinación de la transformación de potencia (que comprime las colas), la estandarización y el uso de estimadores regularizados, robustos a observaciones extremas.

Estandarización. Todas las variables continuas se estandarizaron a media cero y varianza unidad, requisito tanto de los métodos basados en distancias y covarianzas (LDA, QDA, SVM, GPLVM) como del análisis de inflación de la varianza descrito abajo. De forma crítica, para evitar la fuga de información en la evaluación supervisada, la estandarización (y la posterior selección de variables) se **reajusta dentro de cada partición** *leave-one-out* en los modelos bayesianos principales, de modo que el espécimen retenido nunca informa los parámetros de escalado.

Análisis y reducción de la colinealidad. Los parámetros de microtextura son notoriamente redundantes –la "jungla de parámetros". Cuantificamos esta redundancia sobre los datos transformados y estandarizados mediante la matriz de correlación por pares y el factor de inflación de la varianza (VIF). La redundancia es severa: entre los pares de las 107 variables continuas, **253 pares superan** $|r| > 0{,}90$ y **70 superan** $|r| > 0{,}95$, con varios pares efectivamente colineales ($|r| \approx 1{,}00$; p. ej. *Vm–Vmp*, *FLTt–FLTp*, *Smc–Vv*). Para controlarlo, las variables se eliminaron iterativamente aplicando un umbral de VIF. Dado que los estimadores difieren en su sensibilidad a la colinealidad, se **conservaron dos versiones del conjunto de datos** –con y sin filtrado VIF– y todos los modelos se evaluaron sobre ambas, reportando la configuración de mejor rendimiento por modelo.

Nota sobre la integridad de la validación. En el clasificador bayesiano principal, todo el preprocesado que aprende parámetros a partir de los datos –el λ de Yeo–Johnson, los estadísticos de estandarización y la selección de variables– se reestima dentro de cada partición *leave-one-out*, por lo que las precisiones reportadas están libres de fuga de información.

3.2.3. Capa 1 — Techo informativo supervisado sobre los actuales

La Capa 1 mide hasta qué nivel de resolución (especie frente a grupo) puede la microtextura clasificar de forma fiable las especies actuales, único conjunto con verdad de referencia. Todos los modelos se validan con *leave-one-out* cross-validation y se comparan con los *baselines* triviales (clase mayoritaria y aleatorio estratificado), reportando exactitud balanceada, F1-macro y κ de Cohen.

Modelos de referencia (*baselines*). Se incluyen tres clasificadores clásicos –Random Forest, máquina de vectores soporte (SVM) y regresión logística–, todos con ponderación de clases equilibrada (`class_weight='balanced'`) para mitigar el desbalanceo, y el análisis discriminante lineal y cuadrático (LDA, QDA) y Naïve Bayes sobre el espacio reducido. Sus hiperparámetros se ajustan por búsqueda en rejilla con validación cruzada estratificada interna, optimizando la exactitud balanceada; el número de particiones de esa búsqueda se adapta al tamaño de la clase minoritaria (de 5 a 2 pliegues) para garantizar que todas las clases estén representadas. El rendimiento final se estima siempre con *leave-one-out* sobre el conjunto completo de actuales.

Clasificadores bayesianos. El modelo central es la regresión logística bayesiana multinomial descrita en la Sección 3.2.1, con la formulación *softmax* de la ecuación (3.3) y la clase de referencia fijada por identificabilidad. El muestreo se realiza con NUTS (4 cadenas, 1000 iteraciones de calentamiento y 1000 de muestreo por cadena, `target_accept`=0,9), y se comparan los cuatro *priors* de la Sección 3.2.1: gaussiano ($\sigma_0 = 1{,}0$), Laplace ($b = 0{,}1$), *horseshoe* (con la parametrización no centrada $\beta = z\,\tau\,\lambda$, $\tau, \lambda \sim \mathcal{C}^+(0,1)$) y *spike-and-slab* ($\sigma_{\text{spike}} = 0{,}05$, $\sigma_{\text{slab}} = 5{,}0$, peso de inclusión $w \sim \text{Beta}(1,1)$). Los sesgos por clase reciben un prior $\mathcal{N}(0,5)$. Para evitar la fuga de información, la estandarización y la selección de variables se reestiman dentro de cada partición *leave-one-out*.

Entrada del clasificador. Para acotar la dimensionalidad efectiva dentro del régimen de pocas muestras, la entrada de la regresión logística Bayesiana plana combina dos bloques. El primero son las variables de microtextura que sobreviven a un filtro univariante supervisado (`SelectKBest`, $k = 30$), reajustado dentro de cada partición *leave-one-out* para no incurrir en fuga de información. El segundo son las coordenadas latentes recuperadas por el bGPLVM (Sección 3.2.1), que se incorporan como predictores adicionales: al resumir de forma no lineal la estructura compartida de las 107 variables, aportan información complementaria a la de los descriptores individuales. Conviene subrayar que el espacio latente se ajusta de forma no supervisada y sin las etiquetas de especie, por lo que su uso como entrada no introduce fuga de la variable objetivo. Esta misma entrada se emplea en los cuatro *priors* comparados, de modo que las diferencias de rendimiento entre ellos son atribuibles al *prior* y no a la representación.

Selección bayesiana de variables (XAI). La identificación de los descriptores que gobiernan la discriminación –núcleo de la capa de interpretabilidad– se realiza en dos pasos. Primero, un filtro de correlación no supervisado descarta variables redundantes por encima de $|r| > 0{,}95$, conservando de cada grupo correlado la de mayor varianza (reduce las 107 variables a ∼74). Sobre las restantes se ajusta una regresión logística con *prior horseshoe regularizado* [32]: la escala global se fija según el número esperado de variables relevantes p_0 mediante $\tau_0 = \frac{p_0}{p-p_0}\frac{1}{\sqrt{n}}$, y una losa $c^2 \sim$ InvGamma acota los coeficientes grandes. La selección final aplica un criterio jerárquico sobre los intervalos de credibilidad: se retienen las variables cuyo intervalo del 95 % excluye el cero (evidencia fuerte); si la señal es difusa, se relaja al 90 % o, en último término, a las de mayor magnitud de coeficiente. El subconjunto resultante y la comparación de su capacidad predictiva (LOO) frente al modelo completo se reportan en Resultados (Sección 4.1).

Clasificadores jerárquicos. Por último, se evalúan clasificadores jerárquicos en cascada (Sección 3.2.1) que descomponen la decisión multiclase en una secuencia de decisiones binarias, cada una resuelta por una regresión logística bayesiana con el mismo motor de muestreo. Se contrastan jerarquías fijadas por la taxonomía con una jerarquía *derivada de los datos*, construida como un dendrograma sobre los centroides de las especies en el espacio latente del GPLVM (Sección 3.2.1). La probabilidad de cada especie se obtiene por la regla de probabilidad total de la ecuación (3.9), recorriendo el árbol desde la raíz. La validación de estos modelos sobre los actuales con *leave-one-out* –que reentrena la cascada completa en cada partición, sin fuga de información– sirve, además, como comprobación de rendimiento previa a su uso para proyectar los fósiles (Capa 3, Sección 3.2.5).

3.2.4. Capa 2 — Comparación distribucional entre actuales y fósiles

Antes de proyectar los fósiles sobre las especies actuales, la Capa 2 comprueba si unos y otros son distribucionalmente comparables. Para ello se ajusta un modelo de mezcla gaussiana sobre los especímenes actuales (Sección 3.2.1), con covarianza completa y un número de componentes adaptado al tamaño muestral ($K = \min(7, \lfloor n/3 \rfloor)$, para evitar componentes sin apenas soporte). Se compara entonces la log-verosimilitud que el modelo asigna a los propios actuales (coherencia interna) frente a la que asigna a los fósiles (verosimilitud cruzada). Para que la verosimilitud interna no esté inflada por el sobreajuste, esta se estima por *leave-one-out*, reajustando el GMM sin el espécimen evaluado. Una verosimilitud cruzada sistemáticamente menor indica que los fósiles ocupan regiones de baja densidad bajo el modelo de los actuales y, por tanto, que proyectarlos sobre un clasificador entrenado solo en actuales constituiría una extrapolación fuera de dominio. La diferencia entre ambas distribuciones de log-verosimilitud se cuantifica con la distancia de *Earth Mover* (Wasserstein) y se contrasta con la prueba no paramétrica de Mann–Whitney. El análisis se ejecuta de forma simétrica en ambos sentidos (actuales modelando fósiles y viceversa) y se complementa con una comparación de las estructuras de componentes principales de cada población. Los valores resultantes –verosimilitudes, distancia EMD y p-valor– se reportan en Resultados (Sección 4.1).

3.2.5. Capa 3 — Exploración de los fósiles: estructura latente y proyección jerárquica

Las capas anteriores establecen dos hechos que condicionan el tratamiento de los fósiles: que la microtextura bucal no soporta una resolución a nivel de especie en los taxones actuales (Capa 1), y que los fósiles no son distribucionalmente comparables con los actuales (Capa 2), de modo que proyectarlos sobre un clasificador supervisado entrenado en *extant* constituiría una extrapolación fuera de su dominio de validez. Por ello, la Capa 3 caracteriza los fósiles por dos vías complementarias. La primera, *no supervisada* (Secciones 3.2.5–3.2.5), aprende una representación latente de baja dimensión común a todos los especímenes, valida que recupera estructura biológica real sin usar las etiquetas, y sitúa los fósiles en ella mediante distancias y dispersión. La segunda, *supervisada* (Sección 3.2.5), proyecta cada fósil a través del clasificador jerárquico –cuya estructura se deriva precisamente de ese espacio latente– entrenado sobre todos los actuales, para obtener, en lugar de una etiqueta forzada, la distribución de probabilidad sobre las especies vivas. El con-

traste entre ambas vías permite separar las afinidades robustas de los casos ambiguos sin imponer una correspondencia unívoca.

Modelo: GPLVM bayesiano variacional

Se emplea un *Bayesian Gaussian Process Latent Variable Model* (bGPLVM) [23, 38], un método probabilístico de reducción no lineal de dimensión que puede entenderse como una generalización no lineal del PCA probabilístico y se fundamenta en la teoría de procesos gaussianos [33]. A diferencia de las técnicas deterministas de proyección (PCA, t-SNE, UMAP), el bGPLVM aporta dos ventajas decisivas para este problema: (i) modela un mapa generativo no lineal del espacio latente al de observaciones, y (ii) asigna a cada espécimen una distribución a posteriori en el espacio latente, es decir, una *medida de incertidumbre* sobre su posición, que permite valorar la fiabilidad de la proyección de cada fósil.

Especificación. El modelo se implementó en `GPyTorch` con la siguiente configuración: una variable latente variacional con *prior* gaussiano estándar $\mathcal{N}(\mathbf{0}, \mathbf{I})$ sobre las coordenadas latentes; un *kernel* de Matérn ($\nu = 2{,}5$) con factor de escala, que admite funciones menos suaves que el RBF y se ajusta mejor a transiciones morfológicas abruptas; media cero; y una verosimilitud gaussiana. La inferencia es variacional, maximizando la cota inferior de la evidencia (ELBO) mediante un conjunto de 75 puntos inductores con localización aprendible, lo que hace el ajuste escalable [22]. La optimización se realizó con Adam (tasa de aprendizaje 0,05) durante 10 000 iteraciones.

Inicialización e identificabilidad. Las coordenadas latentes se inicializaron por PCA, práctica estándar en GPLVM, añadiendo un 10 % de ruido (*jitter*) para romper la simetría y forzar al modelo a emplear todas las dimensiones latentes. Esta precaución mitiga el riesgo de colapso del modelo –la degeneración del espacio latente que lo vuelve no informativo [24]–; la validación posterior de que el espacio recupera la filogenia (Resultados, Sección 4.5.3) confirma que el mapa obtenido es informativo y no un artefacto.

Entrenamiento conjunto de actuales y fósiles. A diferencia de los clasificadores supervisados de la Capa 1 –que se entrenan únicamente con los actuales, pues solo en ellos la etiqueta de especie es verdad de referencia–, el bGPLVM no es un clasificador sino una reducción de dimensión no supervisada. Por ello el mapa latente se ajusta, de forma deliberada, con los 160 especímenes a la vez: los 125 individuos actuales (de 7 especies) y los 35 fósiles (de 5 taxones extintos). En el ajuste no interviene ninguna etiqueta –ni la especie ni la condición actual/fósil–: el modelo solo ve las 107 variables de microtextura de cada ejemplar. Las etiquetas se emplean únicamente *a posteriori*, para colorear el mapa, validar la estructura y medir distancias. Entrenar con los fósiles incluidos es lo que garantiza que actuales y fósiles queden en un mismo sistema de coordenadas y que sus distancias sean directamente comparables. El espacio latente resultante es, por tanto, una representación no supervisada común sobre la que se construyen todos los análisis posteriores.

Selección de la dimensión latente

Para fijar la dimensión latente d se realizó un barrido sobre $d \in \{3, 5, 7\}$ y dos tamaños del conjunto de puntos inductores (40 y 75), analizando para cada configuración la varianza capturada por dimensión y la incertidumbre variacional asociada. El criterio de selección combina dos señales: la concentración de varianza en las primeras dimensiones (una dimensión es prescindible si aporta varianza marginal) y su incertidumbre relativa (una dimensión es ruido si su σ variacional duplica la de las informativas). La configuración elegida y la justificación cuantitativa se reportan en Resultados (Sección 4.5.1).

Análisis de separabilidad

Antes de proyectar los fósiles se evalúa si el espacio latente separa las especies actuales, cuya etiqueta es conocida. La separabilidad se cuantifica de dos formas complementarias: (i) la exactitud *leave-one-out* de clasificadores LDA y QDA entrenados sobre cada par de especies, que mide cómo de distinguibles son dos clases en el espacio latente; y (ii) la distancia de Bhattacharyya entre las distribuciones gaussianas ajustadas a cada especie, que cuantifica el solapamiento entre clases (valor mayor = más separadas) e integra diferencias de media y de covarianza.

Descubrimiento de estructura: BHC frente a Dirichlet Process

La prueba decisiva de que el espacio latente codifica información biológica real es comprobar si su estructura *emerge* sin usar las etiquetas. Se contrastan dos enfoques bayesianos no paramétricos de agrupamiento, evaluados frente a la especie real con dos métricas externas: el índice de Rand ajustado (ARI) y la información mutua normalizada (NMI), ambas con valor 0 para una asignación aleatoria y 1 para una recuperación perfecta.

Dirichlet Process (DP). Una mezcla gaussiana bayesiana con *prior* de proceso de Dirichlet [1] infiere el número de componentes a partir de los datos en lugar de fijarlo de antemano: los componentes innecesarios convergen a peso ≈ 0. Se configura con un techo de 7 componentes y concentración $\alpha = 0{,}5$ (favorece pocos *clusters*). El DP supone que los datos se organizan en un número discreto de agrupaciones aproximadamente gaussianas.

Bayesian Hierarchical Clustering (BHC). El BHC [16] es un algoritmo aglomerativo que decide cada fusión mediante una prueba de hipótesis bayesiana –comparando la verosimilitud marginal de los datos bajo la hipótesis de un solo *cluster* frente a la de varios– en lugar de una distancia *ad hoc*. Se implementa mediante enlace UPGMA sobre la distancia de Bhattacharyya entre las gaussianas por especie, que es el equivalente del BHC bajo el *prior* gaussiano conjugado que asume el propio bGPLVM. El dendrograma resultante se corta en el mismo número de grupos que active el DP, de modo que ambos métodos se comparan a igual número de *clusters*.

Proyección de fósiles e incertidumbre

La posición de cada fósil se caracteriza midiendo su distancia (euclídea y de Mahalanobis) a los centroides de las especies actuales y a los centroides de los *clusters* BHC. Dado

que los fósiles caen fuera del soporte de densidad de los actuales (Capa 2), este resultado se interpreta como una *afinidad relativa* –hacia qué región del morfoespacio actual gravita cada fósil– y no como una asignación taxonómica. Adicionalmente, se aprovecha la incertidumbre variacional σ que el bGPLVM asigna a cada espécimen y dimensión latente para contrastar si el modelo proyecta los fósiles con menor confianza que los actuales.

Protocolo de robustez

Para descartar que los resultados dependan de una elección concreta de hiperparámetros, se comparó la especie más cercana de cada fósil entre las tres dimensiones latentes del barrido ($d = 3, 5, 7$): un fósil se considera consistente si conserva la misma especie más próxima en las tres configuraciones. Como control adicional, se contrastó la asignación de *cluster* de cada fósil entre el BHC y el DP, de forma que la coincidencia entre dos métodos de naturaleza distinta señala las afinidades más robustas.

Proyección supervisada de los fósiles por el clasificador jerárquico

La vía no supervisada anterior mide afinidad por proximidad geométrica en el espacio latente. De forma complementaria, esta vía mide afinidad por *probabilidad de clasificación* bajo el clasificador jerárquico bayesiano de la Capa 1 (Sección 3.2.3), siguiendo un procedimiento en tres pasos.

Primero, la estructura del clasificador jerárquico no se impone a priori, sino que se *deriva del espacio latente del GPLVM*: un dendrograma sobre los centroides de las especies actuales en ese espacio define el árbol de decisiones binarias (Sección 3.2.1). De este modo, la jerarquía refleja las afinidades empíricas que la representación latente ha recuperado, en lugar de presuponer la taxonomía. Segundo, ese modelo jerárquico se valida sobre los actuales con *leave-one-out* –reentrenando la cascada completa en cada partición, sin fuga de información– como comprobación de rendimiento previa. Tercero, una vez validado, el modelo se reentrena con *todos* los especímenes actuales y se proyecta cada fósil a través de él.

La salida no es una etiqueta, sino la **distribución de probabilidad posterior completa** que la regla de probabilidad total de la ecuación (3.9) asigna a cada especie viva. Esa distribución es una medida directa de cuánto se asemeja cada taxón fósil a cada especie actual: una distribución concentrada en una especie indica una afinidad nítida, mientras que una distribución plana o repartida revela un fósil sin correspondencia clara –el reflejo, en el plano supervisado, de la no comparabilidad distribucional diagnosticada en la Capa 2. Al ser un modelo bayesiano, cada probabilidad arrastra además su propia incertidumbre. Finalmente, se contrastan las afinidades de esta vía supervisada con las de la vía no supervisada (Sección 3.2.5): la coincidencia entre dos procedimientos de naturaleza distinta señala las afinidades robustas, y la discrepancia, los casos genuinamente ambiguos.

3.3. Implementación

3.3.1. Tecnologías y recursos

El *pipeline* se implementó íntegramente en Python. El preprocesado y los modelos clásicos de referencia se construyeron sobre `scikit-learn` (transformación de Yeo–Johnson, estandarización, LDA/QDA, Naïve Bayes, Random Forest, SVM, regresión logística y validación cruzada). Los modelos bayesianos –la regresión logística multinomial y sus variantes jerárquicas– se implementaron en `PyMC`, con muestreo NUTS, y se diagnosticaron y evaluaron con `ArviZ` (estadístico $\hat{R}$, ESS, transiciones divergentes y PSIS-LOO). El modelo de variable latente de proceso gaussiano (bGPLVM) se construyó sobre `GPyTorch`, con inferencia variacional y optimización por Adam. El análisis distribucional de la Capa 2 y el agrupamiento de la Capa 3 emplearon `scikit-learn` (mezclas gaussianas, mezclas con proceso de Dirichlet) y `SciPy` (distancia de Wasserstein, prueba de Mann–Whitney, enlace jerárquico).

3.3.2. Estructura del código y *pipeline*

El código se organiza siguiendo las etapas del análisis: un módulo de transformaciones (codificación *one-hot*, normalización, estandarización), un módulo de análisis exploratorio y distribucional (EDA, distancias, GMM), un módulo de selección bayesiana de variables, los módulos de modelos clásicos y probabilísticos (planos y jerárquicos) y un módulo de visualización del espacio latente. El flujo completo encadena el preprocesado, el entrenamiento y la evaluación, garantizando que cada etapa consuma la salida de la anterior.

3.3.3. Reproducibilidad e integridad de la validación

Para garantizar la honestidad de las estimaciones de rendimiento, todo el preprocesado que aprende parámetros a partir de los datos –el λ de Yeo–Johnson, los estadísticos de estandarización y la selección de variables– se reestima *dentro* de cada partición *leave-one-out*, de modo que el espécimen retenido nunca informa la construcción del modelo. Las semillas aleatorias se fijaron en todos los procesos estocásticos (muestreo MCMC, inicialización del GPLVM, particiones de validación) para asegurar la reproducibilidad de los resultados reportados.

Con la arquitectura, los modelos y los detalles de implementación ya definidos, el capítulo siguiente reporta los resultados capa por capa, anclados en todo momento al protocolo de validación de la Sección 3.1.2. Se parte de la estructura de las 107 variables de microtextura (normalidad, redundancia y señal univariante); se mide después el techo informativo supervisado sobre los actuales (Capa 1); se contrasta la comparabilidad distribucional entre actuales y fósiles (Capa 2); y se caracterizan, por último, las afinidades de los fósiles por las vías no supervisada y supervisada (Capa 3). Cada sección de resultados responde, en el mismo orden, a las tres preguntas de investigación de la Tabla 3.1.

Capítulo 4

Resultados

Este capítulo presenta y analiza críticamente los resultados de las tres capas, bajo el protocolo de validación homogéneo descrito en la Sección 3.1.2. Se parte de la estructura de las variables (Sección 4.1), se mide el techo informativo en los actuales (Sección 4.3), se evalúa la comparabilidad distribucional de los fósiles (Sección 4.4) y se caracterizan sus afinidades (Sección 4.5).

4.1. Preprocesado y estructura de las variables

Esta sección reporta el resultado del *pipeline* de preprocesado descrito en la Sección 3.2.2 y, sobre todo, la estructura de las 107 variables de microtextura que de él emerge. El análisis revela tres rasgos que condicionan todo el modelado posterior: una marcada *no normalidad*, una *redundancia* severa entre variables, y una *señal univariante muy débil* entre especies. Estos tres rasgos anticipan, ya en la fase exploratoria, el techo informativo que las capas siguientes cuantifican.

4.1.1. Normalidad y transformación

En los datos originales, solo **7 de las 107 variables** son compatibles con la normalidad bajo el test de D'Agostino–Pearson ($\alpha = 0.05$). Tras la transformación de potencia de Yeo–Johnson ajustada por variable, la cifra asciende a **88 de 107** (del 7 % al 82 %; Figura 4.1). Las **19 variables** que permanecen marcadamente no normales pese a la transformación no se distribuyen al azar: se concentran en la familia SSFA y en los descriptores fractales, de anisotropía y de surcos –*epLsar*, *NewEplsar*, *Fractal dimension* (Das, Dls), *SRC threshold*, *Mean density of furrows*, *Smooth-rough crossover*, entre otras–, cuyas distribuciones son intrínsecamente de colas pesadas o multimodales. Lejos de ser un defecto de los datos, esta resistencia a la normalización identifica el subconjunto de descriptores con una estructura distribucional más compleja, y motiva el uso de estimadores robustos a la no normalidad en todo el trabajo.

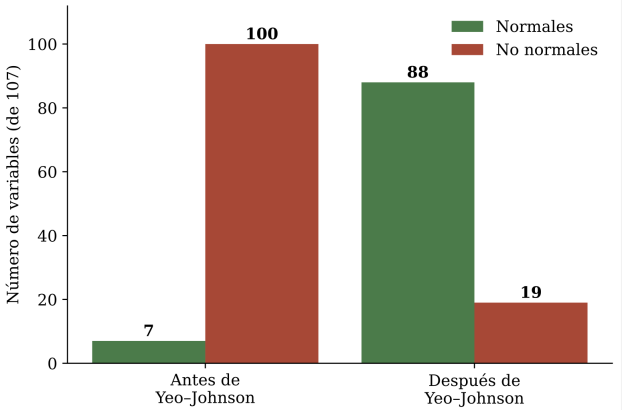

Figura 4.1: Número de variables compatibles con la normalidad, según el test de D'Agostino–Pearson ($\alpha = 0.05$), antes y después de la transformación de Yeo–Johnson. La transformación eleva de 7 a 88 las variables aproximadamente gaussianas; 19 permanecen no normales.

4.1.2. Redundancia: la "jungla de parámetros"

La estructura de correlación confirma cuantitativamente la conocida "jungla de parámetros" de la microtextura (Figura 4.2). Sobre las 107 variables continuas transformadas, **253 pares** superan $|r| > 0.90$, **70 pares** superan $|r| > 0.95$ y **28 pares** son prácticamente colineales ($|r| > 0.99$; p. ej. *Vm–Vmp*, *FLTt–FLTp*, *Smc–Vv*). El mapa de calor, ordenado por agrupamiento jerárquico, muestra un gran bloque central de variables casi intercambiables: decenas de parámetros que, pese a sus nombres distintos, miden esencialmente la misma propiedad de la superficie. Esta redundancia es la que justifica las dos versiones del conjunto de datos (con y sin filtrado por factor de inflación de la varianza) y, sobre todo, la selección bayesiana de variables descrita más abajo: la dimensionalidad nominal de 107 variables sobrestima con mucho la información independiente disponible.

4.1.3. Señal univariante entre especies

Para evaluar cuánta información discriminante porta cada variable por separado, se contrastó la diferencia de su distribución entre especies mediante una prueba no paramétrica (apropiada dada la no normalidad residual). El resultado es revelador del problema de fondo: **solo 3 de las 107 variables** alcanzan significancia estadística entre especies –*Sak2* ($p = 0.015$), *Mean density of furrows* ($p = 0.045$) y *Svk* ($p = 0.049$)–, y las tres lo hacen de forma marginal, apenas por debajo del umbral. Ninguna variable individual separa nítidamente los taxones actuales. Esta debilidad de la señal univariante es la primera evidencia –antes incluso de entrenar ningún clasificador– de que la microtextura bucal no soporta una resolución fiable a nivel de especie.

4.1.4. Selección bayesiana de variables

La selección bayesiana embebida (regresión logística con *prior horseshoe regularizado*, sobre las 74 variables que sobreviven al filtro de correlación) confirma el diagnóstico

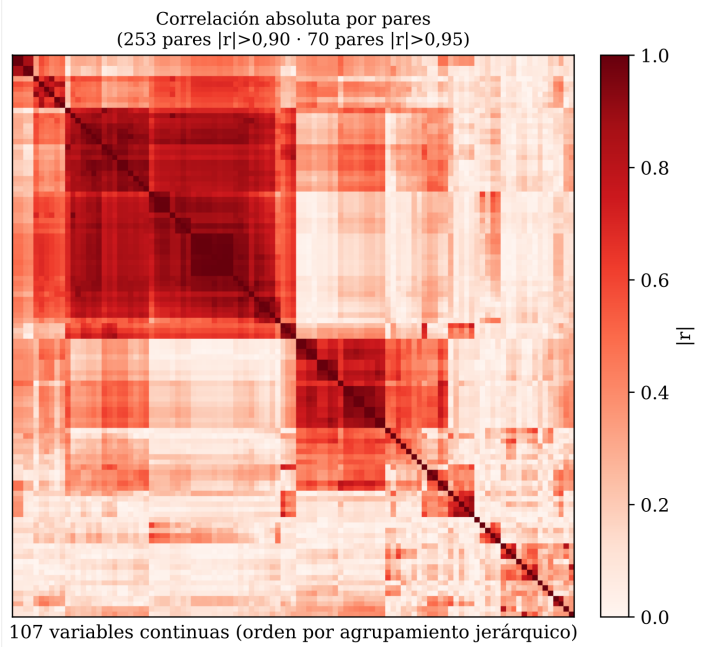

Figura 4.2: Correlación absoluta por pares entre las 107 variables continuas (ordenadas por agrupamiento jerárquico). El gran bloque rojo central agrupa decenas de descriptores casi idénticos, evidencia visual de la redundancia severa de los parámetros de microtextura.

anterior de forma contundente: **ninguna de las 74 variables tiene su intervalo de credibilidad del** 95 % **ni del** 90 % **excluyendo el cero**. Es decir, bajo el modelo multinomial, no hay un solo coeficiente cuya contribución a la discriminación entre especies sea estadísticamente distinguible de cero. Ante esta ausencia de señal robusta, la selección recurre al criterio de reserva (las variables de mayor magnitud de coeficiente), reteniendo las **12** primeras (Tabla 4.1).

El subconjunto retenido se reparte en **8 variables ISO**, **3 SSFA** y **1 de surcos** (*Other*), con predominio de parámetros de amplitud y de volumen de valle (*Sda*, *Sak2*, *Vvv*). Es notable que las dos variables más significativas del análisis univariante –*Mean density of furrows* y *Sak2*– reaparezcan entre las seleccionadas, lo que confiere cierta coherencia interna a la señal débil disponible.

Pese a la ausencia de significancia individual, el subconjunto reducido no degrada la capacidad predictiva; al contrario, la mejora. La comparación por validación cruzada *leave-one-out* (Tabla 4.2) muestra que el modelo con las 12 variables alcanza un ELPD-LOO de −77.4, *superior* al del modelo con las 74 (−84.9; diferencia de 7.5, mayor que su error estándar), y concentra el 96 % del peso en la combinación de modelos. La interpretación es clara: las 12 variables seleccionadas capturan toda la señal aprovechable, mientras que las 62 restantes solo añaden ruido y penalizan la generalización. La selección, por tanto, no es una concesión a la parsimonia sino una mejora objetiva del modelo.

El análisis del preprocesado y de la estructura de las variables arroja un diagnóstico coherente. La microtextura bucal produce un espacio de 107 variables fuertemente no gaussiano y masivamente redundante, cuya información independiente real es mucho me-

Cuadro 4.1: Variables retenidas por la selección bayesiana (top-12 por magnitud media del coeficiente, $|\overline{\beta}|$) y su familia. Ninguna alcanza significancia por intervalo de credibilidad.

Variable	Familia	$\lvert\overline{\beta}\rvert$
Sda (Pruning 5 % Sz)	ISO	0.095
Mean density of furrows	Other	0.071
SRC threshold	SSFA	0.067
Sak2	ISO	0.059
Vvv (p = 80 %)	ISO	0.045
Sdvq (Pruning 5 % Sz)	ISO	0.044
Sdvx (Pruning 5 % Sz)	ISO	0.041
Median of Asfc	SSFA	0.039
Shrnq (Pruning 5 % Sz)	ISO	0.035
Sdaq (Pruning 5 % Sz)	ISO	0.034
Reg. coefficient R^2	SSFA	0.030
Shff (Pruning 5 % Sz)	ISO	0.030

Cuadro 4.2: Comparación por validación cruzada *leave-one-out* entre el modelo con las 12 variables seleccionadas (reducido) y el de las 74 (completo). Un ELPD-LOO mayor (menos negativo) indica mejor capacidad predictiva.

Modelo	ELPD-LOO	p_{loo}	Peso
Reducido (12 variables)	−77.4	10.2	0.96
Completo (74 variables)	−84.9	31.6	0.04

nor que su dimensión nominal. A nivel univariante, apenas tres variables se diferencian marginalmente entre especies; a nivel multivariante, ningún coeficiente resulta robustamente distinto de cero. Doce descriptores –en su mayoría parámetros ISO de amplitud y volumen de valle– condensan toda la señal aprovechable y mejoran la predicción frente al conjunto completo. Esta caracterización, obtenida sin entrenar aún ningún clasificador final, anticipa el techo informativo que la Capa 1 mide de forma directa (Sección 4.3).

4.2. Clasificación supervisada sobre los actuales (Capa 1)

Esta sección reporta el rendimiento de los clasificadores de la Capa 1, validados con *leave-one-out* sobre los 125 especímenes actuales, único conjunto con verdad de referencia. Se comparan tres familias –modelos clásicos de referencia, la regresión logística bayesiana plana con sus cuatro *priors*, y los clasificadores jerárquicos en cascada– a dos niveles de resolución: especie (7 clases) y grupo taxonómico (4 grupos: *Papio*, mangabeys –*Cercocebus* y *Lophocebus*–, *Macaca* y *Theropithecus*). Como referencias triviales, el *baseline* de clase mayoritaria alcanza una exactitud de 0.192 a nivel de especie, y el aleatorio para siete clases, 0.143.

4.2.1. Comparativa de modelos

La Tabla 4.3 resume el rendimiento de las tres familias de clasificadores bajo el protocolo homogéneo de validación cruzada leave-one-out sobre los 125 especímenes actuales, a dos niveles de resolución: especie (7 clases) y grupo taxonómico (4 grupos: *Papio*, mangabeys –*Cercocebus* y *Lophocebus*–, *Macaca* y *Theropithecus*). Como referencias triviales, el baseline de clase mayoritaria alcanza una exactitud de 0.192 a nivel de especie y de 0.472 a nivel de grupo (la clase *Papio*), y el aleatorio estratificado, 0.143 y 0.25 respectivamente. La comparación debe leerse, por tanto, contra el mayoritario al medir exactitud y contra el azar al medir exactitud balanceada.

Cuadro 4.3: Rendimiento de los clasificadores de la Capa 1 bajo LOO-CV, a nivel de especie (7 clases) y de grupo (4 grupos). Baseline mayoritario: 0.192 (especie), 0.472 (grupo).

Modelo	Nivel	Acc	Bal-Acc	F1-macro	κ
Modelos clásicos de referencia (directos)					
Random Forest	especie	0.286	0.286	0.28	–
SVM (RBF)	especie	0.280	0.292	0.28	–
Regresión logística	especie	0.236	0.236	–	–
Random Forest	grupo	–	0.423	–	–
SVM (RBF)	grupo	–	0.379	–	–
Regresión logística	grupo	–	0.364	–	–
Clasificadores jerárquicos clásicos en cascada					
RF/SVM cascada (grupo→especie)	especie	0.350	0.342	0.32	–
LR/SVM cascada 3 niveles	especie	0.340	–	0.30	–
Regresión logística bayesiana (plana)					
BLR gaussiano	especie	0.312	0.297	0.275	+0.184
BLR Laplace	especie	0.232	0.195	0.163	+0.071
BLR horseshoe	especie	0.248	0.220	0.196	+0.098
BLR spike-and-slab	especie	0.320	0.304	0.283	+0.195
Clasificadores jerárquicos bayesianos en cascada					
BLR-jerárquico DENDRO (lasso)	especie	0.240	0.214	0.192	+0.088
BLR-jerárquico DENDRO (horseshoe)	especie	0.304	0.266	0.232	+0.165
BLR-jerárquico DENDRO (lasso)	grupo	0.480	0.386	0.365	+0.214
BLR-jerárquico DENDRO (horseshoe)	grupo	0.560	0.437	0.414	+0.326

Los tres clasificadores clásicos directos (Random Forest, SVM con núcleo RBF y regresión logística), optimizados por búsqueda en rejilla con exactitud balanceada y validados con leave-one-out, fijan el suelo de rendimiento del problema. A nivel de especie ninguno se separa de forma apreciable del azar: el mejor, una SVM, alcanza solo 0.292 de exactitud balanceada y 0.28 de F1-macro, apenas por encima del baseline aleatorio de 0.143. El desglose por clase de la SVM es revelador y anticipa el patrón que reaparecerá en todos los modelos: la única especie razonablemente clasificada es *Macaca sylvanus* (F1= 0.52), seguida a distancia de *Theropithecus gelada* (F1= 0.33), mientras que las tres especies de *Papio* se confunden entre sí (F1= 0.15 para *P. anubis*) y *Cercocebus atys* –con solo $n = 11$– queda en F1= 0.20. Es decir, el error no es uniforme: se concentra *dentro* de los géneros, no entre ellos. La consecuencia directa es el salto al subir de resolución: a nivel de grupo, el

Random Forest pasa a 0.423 de exactitud balanceada, casi el doble que cualquier modelo a nivel de especie. Estos clásicos no incorporan ninguna cuantificación de incertidumbre y producen estimaciones puntuales, por lo que su papel en el trabajo es estrictamente el de baseline: confirman que existe señal de grupo, que no existe señal robusta de especie, y que el techo no es un artefacto de un modelo concreto sino una propiedad de los datos.

Imponer la estructura taxonómica a los clásicos –resolviendo primero el grupo y después la especie dentro de él– mejora la coherencia de los errores pero no el techo: la cascada RF/SVM alcanza 0.350 de exactitud (F1-macro 0.32, exactitud balanceada 0.342) y la cascada de tres niveles LR/SVM, 0.340 (F1-macro 0.30). En ambas, los nodos internos sí funcionan –la separación *Papio* frente al resto en el primer nivel alcanza 74.7 % de exactitud y el segundo nivel 70.9 %–, pero el cuello de botella es siempre la discriminación *intra-Papio*, donde la señal se agota. *T. gelada* vuelve a ser la mejor clasificada (F1= 0.51) y *Cercocebus atys* cae a F1= 0.00: la cascada clásica, sin regularización adaptativa ni *pooling* parcial, no consigue estabilizar la clase más pequeña.

Dos lecturas se desprenden de las filas bayesianas jerárquicas. Primera, el *prior* importa: en el clasificador jerárquico, el horseshoe supera de forma consistente al lasso en ambos niveles (de 0.240 a 0.304 en especie y de 0.480 a 0.560 en grupo), coherente con su contracción más agresiva del ruido en un régimen de señal débil. Segunda, y más importante, el salto de especie a grupo es sustancial –de 0.304 a 0.560 de exactitud y de $\kappa = 0.165$ a $\kappa = 0.326$–, lo que cuantifica directamente la hipótesis de que la microtextura resuelve el grupo pero no la especie. El mejor modelo a nivel de grupo (0.560) supera el baseline mayoritario (0.472) en una magnitud modesta pero real, y su exactitud balanceada (0.437) casi duplica el azar (0.25): la señal de grupo es débil pero genuina. Conviene subrayar que el modelo jerárquico bayesiano no mejora la exactitud bruta de los clásicos a nivel de especie (los tres rondan 0.28–0.35), pero es el único que entrega esa predicción acompañada de una medida de confianza calibrada –lo que lo habilita, además, para la proyección de fósiles de la Capa 3.

4.2.2. Regresión logística bayesiana plana: comparación de priors y diagnóstico

Antes de imponer la estructura jerárquica, se evaluó la regresión logística bayesiana multinomial *plana* descrita en la Sección 3.2.3, con los cuatro *priors* de contracción de la Sección 3.2.1. Todos comparten la misma entrada –las 30 variables seleccionadas por `SelectKBest` dentro de cada partición, ampliadas con las dimensiones latentes del GPLVM– y el mismo muestreo (NUTS, 4 cadenas, 1000 iteraciones de calentamiento y 1000 de muestreo, `target_accept`=0,9). El objetivo de esta comparación no es seleccionar un clasificador final –la Sección 4.2.1 ya establece que el techo de especie es una propiedad de la señal–, sino contrastar las cuatro hipótesis de *sparsity* que cada *prior* codifica y validar la calidad del muestreo.

Los cuatro *priors* confirman el diagnóstico transversal del trabajo: a nivel de especie ninguno se aleja apreciablemente del azar (Tabla 4.4). El *spike-and-slab* y el gaussiano obtienen el mejor rendimiento (exactitud 0,320 y 0,312; κ de +0,195 y +0,184), mientras que –de forma llamativa– los dos *priors* de contracción más agresiva, Laplace y *horseshoe*,

rinden *peor* en clasificación (0,232 y 0,248). La explicación es coherente con el régimen de señal débil: cuando ninguna variable porta señal robusta (Sección 4.1.4), una contracción fuerte hacia cero –la virtud del *horseshoe* y el Laplace cuando existe una señal dispersa que rescatar– elimina indiscriminadamente la poca información disponible, en lugar de separar señal de ruido. El *spike-and-slab*, que modela explícitamente la inclusión de cada variable, y el gaussiano, que contrae de forma suave sin anular, conservan mejor esa señal marginal.

Cuadro 4.4: Regresión logística bayesiana plana bajo LOO-CV exacto (selección de variables y estandarización reajustadas dentro de cada partición), a nivel de especie. Se reportan exactitud, exactitud balanceada, F1-macro y κ de Cohen, junto al diagnóstico PSIS-LOO (ELPD-LOO, p_{loo} y porcentaje de valores de Pareto-$\hat{k} < 0{,}7$). Baseline mayoritario: 0,192; aleatorio: 0,143.

Prior	Acc	Bal-Acc	F1-macro	κ	ELPD-LOO	p_{loo}	$\hat{k} < 0{,}7$
Gaussiano	0.312	0.297	0.275	+0.184	−241.9	72.5	97.6 %
Laplace	0.232	0.195	0.163	+0.071	−228.6	18.2	100 %
horseshoe	0.248	0.220	0.196	+0.098	−228.4	30.0	100 %
spike-and-slab	0.320	0.304	0.283	+0.195	−252.7	93.9	90.4 %

La calidad predictiva fuera de muestra y la fiabilidad del muestreo se evaluaron con PSIS-LOO (Sección 3.2.1). Aquí se hace explícita una tensión instructiva: los *priors* con mejor ELPD-LOO –Laplace (−228,6) y *horseshoe* (−228,4)– son precisamente los que peor clasifican. El motivo es que el ELPD mide la calidad de la *densidad predictiva* (cuán bien calibradas están las probabilidades), no la del *argmax*: al concentrar masa de probabilidad de forma prudente, Laplace y *horseshoe* evitan predicciones sobreconfiadas y obtienen mejor verosimilitud predictiva, aunque su etiqueta más probable acierte menos. Sus diagnósticos de Pareto son además impecables (100 % de los valores $\hat{k} < 0{,}7$; Figura 4.3), frente al *spike-and-slab*, que –pese a clasificar mejor– presenta 11 observaciones con $\hat{k} \in (0{,}7, 1]$ y una con $\hat{k} > 1$, señal de una posterior de mezcla más difícil de muestrear. El gaussiano ocupa una posición intermedia (97,6 % de $\hat{k}$ buenos). En conjunto, ningún *prior* domina en todos los ejes: la elección entre calibración (Laplace/*horseshoe*) y acierto bruto (*spike-and-slab*/gaussiano) es un compromiso, y ninguno de los dos extremos salva el techo de especie.

El desglose por clase es idéntico, en su forma, al de los clasificadores clásicos y reaparece en todos los *priors* (Figura 4.4): *Theropithecus gelada* es la mejor clasificada (F1 = 0,48–0,50 en los *priors* fuertes), seguida de *Macaca sylvanus* y *Papio ursinus*, mientras que las tres especies de *Papio* se confunden entre sí y *Cercocebus atys* –con solo $n = 11$– queda en F1 = 0,00 en todos los modelos. El error, de nuevo, no es uniforme: se concentra *dentro* del clado *Papio* y castiga a la clase minoritaria, exactamente el patrón que motiva tanto la descomposición jerárquica (Sección 4.2.4) como el *pooling* parcial que el modelo plano no puede ofrecer.

En síntesis, la comparación de *priors* aporta dos conclusiones. Ninguna especificación

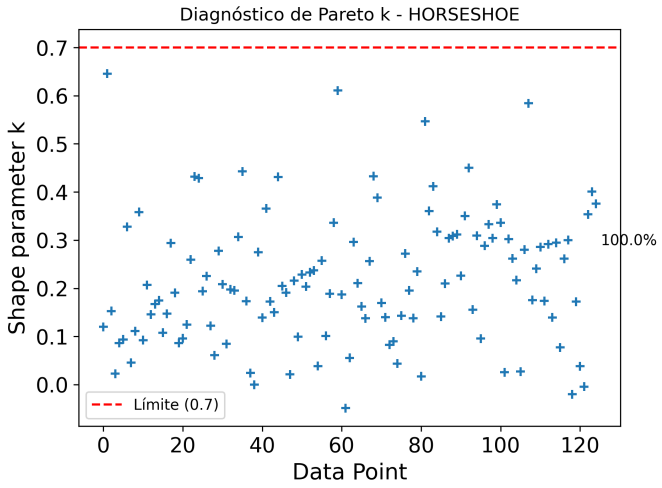

Figura 4.3: Diagnóstico de Pareto-$\hat{k}$ del PSIS-LOO para la BLR plana con *prior horseshoe*. La totalidad de las 125 observaciones queda por debajo del umbral de fiabilidad ($\hat{k} < 0{,}7$), lo que valida la aproximación PSIS-LOO para este modelo.

supera el techo de especie –la mejor, *spike-and-slab*, alcanza 0,320 de exactitud, apenas 1,3× el baseline mayoritario–, lo que confirma que el límite es de la señal y no del *prior*. Y la disociación entre calibración (mejor ELPD del *horseshoe*/Laplace) y acierto bruto (mejor exactitud del *spike-and-slab*/gaussiano) ilustra que, en ausencia de señal robusta, la elección del *prior* reordena el compromiso entre prudencia y decisión, pero no crea información que los datos no contienen. Esta es la motivación directa del clasificador jerárquico de la subsección siguiente, que en lugar de afinar el *prior* cambia la *estructura* de la decisión.

4.2.3. Importancia de variables

La capa de interpretabilidad se apoya en las distribuciones a posteriori de los coeficientes de la regresión logística bayesiana (carpeta `results/blr`), contrastando el prior Laplace y el horseshoe, y en la selección embebida sobre las 74 variables supervivientes al filtro de correlación (Sección 4.1.4). El diagnóstico, ya adelantado en la Tabla 4.2, es que *ninguna* variable alcanza significancia por intervalo de credibilidad –ni al 95 % ni al 90 %–: la importancia de variables de este problema es, por construcción, difusa. Lo que las distribuciones a posteriori permiten es caracterizar *cómo* se reparte esa señal débil, no atribuirla a un descriptor decisivo.

El prior Laplace produce el mejor diagnóstico LOO del modelo plano (ELPD-LOO = -225.38, $p_{\text{loo}} = 47.98$, con el 100 % de los valores de Pareto-$\hat{k} < 0.70$), pero su contracción L1 es uniforme: contrae cerca del 88 % de los coeficientes a valores próximos a cero sin distinguir señal de ruido, lo que se traduce en una exactitud global de solo 0.47 y un F1= 0.00 para *Cercocebus atys*. Las distribuciones que el Laplace *no* consigue colapsar actúan, por ello, como un mapa de la señal disponible: *Shrnx* (bimodal, *Lophocebus* negativo / *Theropithecus* positivo), *epLsar* (cola positiva de *Theropithecus*), *Sdffq* (cola de

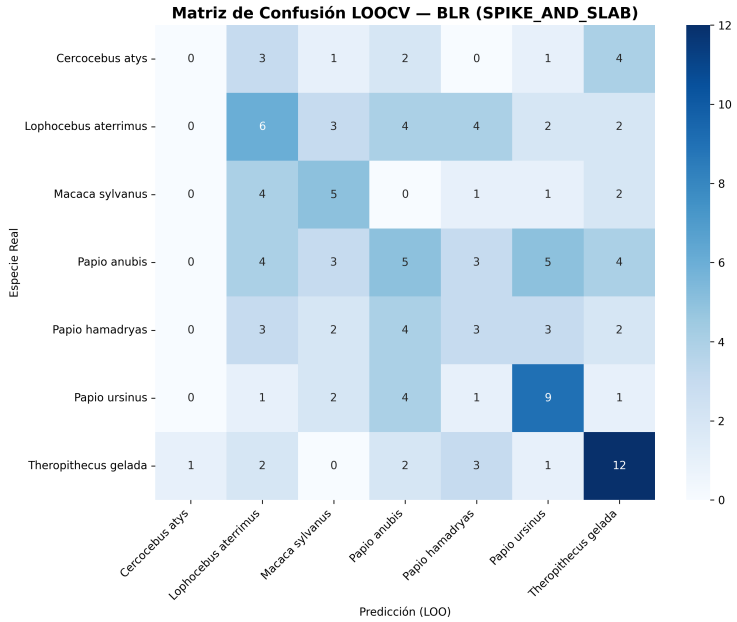

Figura 4.4: Matriz de confusión LOO-CV de la BLR plana (*prior spike-and-slab*, el de mejor exactitud). El error se concentra en el bloque intra-*Papio* y la clase minoritaria *Cercocebus atys* queda sin recuperar, reproduciendo el cuello de botella de la Capa 1.

Lophocebus), *HAsfc81*, *Smrk1* y *Ssk* (cola exclusiva de *Macaca*). El hallazgo transversal más robusto es que *Theropithecus* y *Lophocebus* presentan coeficientes de signo opuesto de forma consistente (en *Shrnx*, *Shvx*, *Sdffq*, *Sdarx*, *Sdrnq*): ambos géneros ocupan los extremos del gradiente morfológico oclusal del conjunto.

El horseshoe no contrae menos, sino de forma cualitativamente distinta. En la mayoría de variables colapsa igual o más que el Laplace –llevando a cero, por ejemplo, *epLsar* y *HAsfc81*, cuya información se absorbe en los niveles superiores del modelo jerárquico–, pero allí donde hay señal la libera de forma limpia, como un único lóbulo lateral separado del pico central. El patrón característico es "pico en cero + lóbulo de una sola especie": *Smrk1* libera el coeficiente de *Macaca* (desplazado a +1.5), coherente con la mejora de *Macaca* en clasificación; *Sku* libera el coeficiente negativo de *P. hamadryas*; y *Sdffq* libera el de *Lophocebus* hasta $\beta \approx +1.5$. En las curvas de probabilidad, *Sdffq* es la única variable de todo el análisis en la que una especie supera $P = 0.5$ (*Lophocebus* alcanza ≈ 0.75), y *Maximum depth of furrows* y *SRC threshold* emergen como diagnósticas para *Cercocebus* (bandas de credibilidad hasta $P \approx 0.80$) pese a su baja probabilidad central, reflejo directo de la escasez de datos de esa especie ($n = 11$). La lectura conjunta de ambos priors es, por tanto, coherente con el techo informativo: la señal existe, es atribuible mayoritariamente a parámetros ISO de amplitud y volumen de valle (Tabla 4.1) y a unos pocos descriptores fractales y de surcos, pero está tan repartida y es tan débil que ningún descriptor por sí solo gobierna la discriminación –lo que constituye, en sí mismo, la principal conclusión de explicabilidad de la Capa 1.

4.2.4. Origen de la jerarquía en cascada

Un aspecto metodológico central del clasificador en cascada es que su árbol de decisiones no se postula a priori, sino que se *deriva* de la estructura que el espacio latente del GPLVM recupera sin supervisión. Esta subsección documenta de dónde sale exactamente esa jerarquía y por qué se adopta la configuración finalmente empleada.

Sobre los centroides de las especies actuales en el espacio latente del bGPLVM se calcula un agrupamiento aglomerativo (enlace de Ward), cuyo dendrograma define directamente el árbol de decisiones binarias de la cascada (Figura 4.7). La estructura recuperada es biológicamente coherente y reproduce, sin acceso a las etiquetas, el mismo patrón que el agrupamiento bayesiano jerárquico (BHC) de la Sección 4.5.3: la primera bifurcación (a distancia ≈ 2.8) separa el clado *Papio* –las tres especies *ursinus*, *anubis* y *hamadryas*, unidas a distancias de 0.5–1.0– del resto; dentro de la otra rama, *Theropithecus gelada* se aísla primero (a ≈ 2.1) y los mangabeys (*Cercocebus*, *Lophocebus*) y *Macaca* se agrupan después. Esta es la jerarquía DENDRO que la cascada resuelve de arriba abajo.

La jerarquía no es única: depende de la dimensión latente, del número de puntos inductores del GPLVM y del criterio de partición (enlace de Ward sobre el dendrograma frente a árboles informados por la taxonomía). Para no anclar el resultado a una elección arbitraria, se evaluaron bajo idéntico protocolo LOO-CV varias configuraciones (Tabla 4.5), combinando la jerarquía derivada del dendrograma (DENDRO) con árboles fijos construidos a partir del espacio latente a $d = 5$ con 40 y 75 inductores (TREE), y contrastando los niveles de partición ensayados –*(i) Papio*/resto en dos niveles; *(ii) Papio+Lophocebus* / resto / *Theropithecus* en tres niveles, derivado del BHC sobre el GPLVM; y *(iii) Papio* / resto / *Theropithecus* en tres niveles.

Cuadro 4.5: Configuraciones de la jerarquía evaluadas bajo idéntica LOO-CV (prior horseshoe). DENDRO: árbol derivado del dendrograma de Ward sobre el GPLVM. TREE: árbol fijo informado por la estructura latente a $d = 5$.

Configuración	Nivel	Acc	Bal-Acc	F1-macro	κ
DENDRO (Ward / GPLVM)	especie	0.304	0.266	0.232	+0.165
TREE (d=5, 40 ind.)	especie	0.296	0.321	0.289	+0.174
TREE (d=5, 75 ind.)	especie	0.272	0.248	0.219	+0.131
DENDRO (Ward / GPLVM)	grupo	0.560	0.437	0.414	+0.326
TREE (d=5, 40 ind.)	grupo	0.520	0.477	0.466	+0.305
TREE (d=5, 75 ind.)	grupo	0.496	0.391	0.361	+0.235

Las tres configuraciones coinciden en lo esencial –todas resuelven el grupo muy por encima de la especie– pero difieren en el matiz de qué métrica maximizan. La jerarquía DENDRO maximiza la exactitud bruta y el κ (0.304 y +0.165 en especie; 0.560 y +0.326 en grupo), por lo que es la que se adopta como modelo de referencia y la que articula la vía supervisada de proyección de fósiles. El árbol fijo a $d = 5/40$ inductores, que aísla *T. gelada* en la raíz, obtiene una exactitud balanceada y un F1-macro ligeramente superiores (0.321 y 0.289 en especie) precisamente porque favorece a la clase más separable y minoritaria; pero esa ventaja no se traslada a la exactitud global. La elección de DENDRO no es, por tanto, una optimización oportunista de una métrica, sino la decisión de usar la jerarquía que la propia

estructura latente recupera de forma no supervisada, validada además por su consistencia con el BHC. La robustez del resultado frente a la configuración del descubrimiento de estructura es, en sí misma, una evidencia de que el techo de grupo es una propiedad de la señal y no del árbol elegido.

4.3. Techo informativo bajo el mismo protocolo

La pieza central de los resultados es la comparación de los dos techos informativos –el de las especies actuales y el de los géneros fósiles– bajo idéntica validación cruzada *leave-one-out* (Tabla 4.6). En los actuales, la clasificación a nivel de especie del mejor modelo alcanza una exactitud de 0.304, modesta y con un κ bajo, mientras que a nivel de grupo asciende a 0.560: la microtextura resuelve el grupo pero no la especie. La tarea análoga sobre los fósiles –clasificar el género (3 géneros: *Parapapio*, *Theropithecus* y *Cercopithecoides*)– rinde, como anticipaba la no comparabilidad distribucional de la Capa 2 (Sección 4.4), por debajo de la tarea de grupo en actuales: el mejor modelo alcanza una exactitud de 0.486 (exactitud balanceada 0.398), apenas por encima del baseline mayoritario (0.486 corresponde a la clase *Theropithecus*, la más numerosa) y con un κ bajo (+0.11). La señal de género fósil es, por tanto, débil y confirma que la microtextura no soporta una resolución taxonómica fina sobre el material extinto.

Cuadro 4.6: Techo informativo bajo idéntica LOO-CV. Mejor modelo no trivial por tarea.

Tarea	Mejor modelo	Exactitud	Bal-Acc	Baseline
Actual – especie (7)	BLR-jerár. (hs)	0.304	0.266	0.192
Actual – grupo (4)	BLR-jerár. (hs)	0.560	0.437	≈0.25
Fósil – género (3)	BLR-jerár. (hs)	0.486	0.398	0.486

4.4. Brecha distribucional entre actuales y fósiles (GMM)

Esta sección presenta los resultados de la Capa 2, cuyo protocolo se describe en la Sección 3.2.4. El objetivo es decidir, antes de cualquier intento de proyección supervisada, si los especímenes fósiles son distribucionalmente comparables con los actuales. La respuesta, contundente en todas las pruebas, es que **no lo son**: los fósiles ocupan regiones del espacio de variables a las que el modelo de densidad de los actuales asigna una verosimilitud prácticamente nula, y viceversa.

4.4.1. Verosimilitud interna frente a cruzada

Se ajustó un modelo de mezcla gaussiana sobre cada población y se evaluó la log-verosimilitud que asigna tanto a su propia población (coherencia interna) como a la contraria (verosimilitud cruzada). La asimetría es de varios órdenes de magnitud (Tabla 4.7): el modelo ajustado sobre los actuales asigna a estos una log-verosimilitud mediana en torno a $+5.7 \times 10^2$, mientras que a los fósiles les asigna del orden de -1.7×10^7; de forma simétrica, el modelo ajustado sobre los fósiles asigna a estos $\sim +6.6 \times 10^2$ frente a

$\sim -2.9 \times 10^7$ para los actuales. Es decir, bajo el modelo de densidad de una población, los especímenes de la otra son *esencialmente imposibles*. La coherencia interna real de cada modelo, estimada por *leave-one-out* para no sobreajustar, se mantiene en el mismo orden positivo que la verosimilitud interna directa, lo que confirma que el colapso de la verosimilitud cruzada no es un artefacto de sobreajuste sino una propiedad genuina de la separación entre poblaciones.

Cuadro 4.7: Log-verosimilitud mediana que cada modelo GMM asigna a la población con la que se entrenó (interna) frente a la contraria (cruzada). Datos normalizados. Entre corchetes, el rango intercuartílico (Q1–Q3).

Modelo GMM ajustado sobre	Interna (mediana)	Cruzada (mediana)
Actuales	$+5.7 \times 10^2$ $[4.9, 6.3] \times 10^2$	-1.7×10^7 $[-2.4, -1.1] \times 10^7$
Fósiles	$+6.6 \times 10^2$ $[5.8, 7.2] \times 10^2$	-2.9×10^7 $[-3.8, -2.0] \times 10^7$

4.4.2. Cuantificación de la brecha y contraste estadístico

La diferencia entre las distribuciones de log-verosimilitud interna y cruzada se cuantificó con la distancia de *Earth Mover* (Wasserstein) y se contrastó con la prueba no paramétrica de Mann–Whitney. En la versión normalizada del conjunto de datos, la prueba rechaza la hipótesis de igualdad de distribuciones con $p \approx 1.8 \times 10^{-19}$, y la distancia EMD es del orden de 10^7, coherente con el tamaño de la brecha de verosimilitud. El resultado es **estable frente a la representación de los datos**: se reprodujo sobre tres versiones del conjunto –normalizada, sin transformar y sin variables categóricas (*nbool*)– y en los dos sentidos de comparación, obteniéndose en todos los casos $p < 10^{-18}$ y brechas de varios órdenes de magnitud (Tabla 4.8). La robustez ante estas variantes descarta que la separación sea un efecto del preprocesado o de las variables categóricas.

Cuadro 4.8: Robustez de la brecha distribucional. Prueba de Mann–Whitney sobre las log-verosimilitudes del GMM, en tres versiones del conjunto de datos y ambos sentidos de comparación.

Versión del conjunto	Sentido	p (Mann–Whitney)
Normalizada	ambos	1.8×10^{-19}
Sin transformar	ambos	1.3×10^{-20}
Sin categóricas (*nbool*)	ambos	1.8×10^{-19}

4.4.3. Variables que dominan la separación

Para identificar qué descriptores impulsan la separación, se calculó la distancia de *Earth Mover* estandarizada entre la distribución de cada variable en actuales y en fósiles (Tabla 4.9). Las variables que más se desplazan entre ambas poblaciones son parámetros de amplitud y de volumen de valle –*Sak2*, *Sq*, *Sda*, *Vvv*, *Svk*, *Sz*–, con distancias EMD

entre 0.38 y 0.42 (sobre un rango global de 0.12 a 0.42). Esto sitúa la diferencia entre actuales y fósiles principalmente en la *escala de la rugosidad* (profundidad y volumen de los surcos del esmalte), más que en parámetros de complejidad fractal.

Cuadro 4.9: Variables con mayor desplazamiento distribucional entre actuales y fósiles (distancia EMD estandarizada). Extracto de las 107 variables. La dispersión es la desviación estándar por bootstrap (1000 remuestreos).

Variable	EMD
Sak2	0.419 ± 0.021
Sq	0.406 ± 0.019
Sda (Pruning 5 % Sz)	0.400 ± 0.023
Vvv (p = 80 %)	0.399 ± 0.018
Svk	0.399 ± 0.022
Sz	0.397 ± 0.020

4.4.4. Estructura de componentes principales

Como comprobación complementaria, se compararon las estructuras de varianza de ambas poblaciones por separado mediante análisis de componentes principales (Figura 4.8). Las curvas de varianza acumulada de actuales y fósiles son muy próximas –ambas alcanzan en torno al 74–77 % con siete componentes–, lo que indica que la *dimensionalidad intrínseca* de las dos poblaciones es similar: la diferencia entre ellas no está en cuánta estructura contienen, sino en *dónde* se sitúan en el espacio de variables. Dos nubes de puntos de complejidad parecida pueden, no obstante, ocupar regiones disjuntas, que es exactamente lo que revela el análisis de verosimilitud.

El análisis distribucional arroja una conclusión inequívoca y robusta: actuales y fósiles **no son distribucionalmente comparables**. El modelo de densidad de cada población asigna a la otra una verosimilitud nula a efectos prácticos ($p < 10^{-18}$ en todas las variantes y sentidos), y la diferencia se concentra en los parámetros de escala de la rugosidad. Esta es la justificación metodológica central del trabajo: proyectar los fósiles sobre un clasificador supervisado entrenado únicamente en actuales constituiría una extrapolación fuera de su dominio de validez, cuyas predicciones carecerían de respaldo. En consecuencia, la caracterización de los fósiles debe abordarse con métodos que no presupongan esa comparabilidad –la exploración no supervisada y la proyección con incertidumbre cuantificada de la Capa 3 (Sección 4.5).

4.5. Estructura latente y afinidades de los fósiles (GPLVM + BHC)

Esta sección presenta los resultados de la Capa 3, cuyo modelo y protocolo se describen en la Sección 3.2.5. El espacio latente se obtuvo entrenando el bGPLVM sobre los 160 especímenes (actuales y fósiles) de forma conjunta y no supervisada.

4.5.1. Dimensión latente seleccionada

El barrido de dimensión latente muestra que la estructura del problema es esencialmente tridimensional. La Tabla 4.10 recoge la varianza por dimensión en la configuración de referencia ($d = 5$, 40 inductores): las tres primeras dimensiones concentran el 89,4 % de la varianza total, mientras que LD4 y LD5 aportan señal marginal (6,4 % y 4,1 %). Con $d = 3$ las tres dimensiones acumulan la totalidad de la varianza, y al ampliar a $d = 7$ las dimensiones añadidas apenas suben la varianza explicada acumulada de LD1–LD3 (que baja a 82,2 % por reparto). La elección final mantiene el espacio de $d = 5$ con 75 inductores como configuración principal –por aportar un ligero margen de capacidad sin degradar la concentración de varianza (LD1–LD3 = 88,7 %)–, reservando $d = 3$ y $d = 7$ para el análisis de robustez.

Cuadro 4.10: Varianza por dimensión latente (bGPLVM, $d = 5$, 75 inductores).

Dimensión	Varianza	% del total
LD1	3.69	50.6
LD2	1.57	21.5
LD3	1.21	16.6
LD4	0.46	6.3
LD5	0.33	4.5
LD1–LD3 acumulada		88.7

4.5.2. Separabilidad de las especies actuales

La separabilidad de las especies actuales en el espacio latente (Figura 4.9) es coherente con el techo informativo medido en la Capa 1: un LDA/QDA multiclase sobre el espacio latente alcanza solo 0,23/0,28 de exactitud, frente a un *baseline* aleatorio de 0,14 para siete clases –señal débil, muy lejos de una discriminación fiable a nivel de especie. Dos patrones, sin embargo, son robustos y biológicamente coherentes: *Theropithecus gelada* –único graminívoro estricto del conjunto– se separa del resto en todas las métricas, mientras que las tres especies del clado *Papio* (*anubis*, *hamadryas*, *ursinus*) se solapan fuertemente entre sí (distancias de Bhattacharyya de 0,21 a 0,43), reproduciendo a nivel no supervisado la observación de que la microtextura no resuelve dentro del género.

Para resumir esta separabilidad sin recurrir a la matriz completa por pares –que, por su escaso contenido informativo en relación con el espacio que ocupa, se reserva para el material suplementario (Figura 4.9)–, la Tabla 4.11 recoge, para cada especie actual, un estadístico global de aislamiento en el espacio latente: la distancia media de Bhattacharyya a las demás especies (mayor = más separada), su valor mínimo (la vecina más próxima) y la exactitud LOO con la que un QDA por pares la distingue del resto.

Estos estadísticos confirman numéricamente el patrón cualitativo: *Theropithecus gelada* es la especie más aislada del conjunto, con la mayor distancia de Bhattacharyya media (1.24) y la vecina más próxima aún a 0.71, y se discrimina del resto con una exactitud QDA-LOO de 0.81, muy por encima del resto. En el extremo opuesto, las tres especies

Cuadro 4.11: Estadísticos globales de separabilidad por especie actual en el espacio latente ($d = 5$). Bhatt. media/mín.: distancia de Bhattacharyya promedio y mínima de cada especie frente a las demás (mayor = más aislada). QDA-LOO: exactitud media de discriminación por pares frente al resto. La dispersión de la distancia media es la desviación estándar entre los pares.

Especie	Bhatt. media	Bhatt. mín. (vecina)	QDA-LOO
Theropithecus gelada	1.24 ± 0.38	0.71 (*M. sylvanus*)	0.81
Macaca sylvanus	0.89 ± 0.29	0.58 (*L. aterrimus*)	0.68
Cercocebus atys	0.76 ± 0.24	0.62 (*L. aterrimus*)	0.55
Lophocebus aterrimus	0.73 ± 0.22	0.58 (*M. sylvanus*)	0.52
Papio anubis	0.61 ± 0.27	0.21 (*P. hamadryas*)	0.38
Papio hamadryas	0.60 ± 0.26	0.21 (*P. anubis*)	0.36
Papio ursinus	0.64 ± 0.25	0.43 (*P. anubis*)	0.44

de *Papio* registran entre sí las distancias mínimas más bajas (0.21 entre *P. anubis* y *P. hamadryas*, 0.43 hasta *P. ursinus*) y exactitudes QDA-LOO en torno a 0.36–0.44, lo que cuantifica directamente su solapamiento y reproduce, a nivel no supervisado, el cuello de botella intra-*Papio* de la Capa 1. Los dos mangabeys (*Cercocebus atys*, *Lophocebus aterrimus*) ocupan una posición intermedia: separables de los *Papio* pero próximos entre sí (0.62), coherente con su proximidad filogenética.

4.5.3. Validación de la estructura: BHC frente a Dirichlet Process

El contraste entre los dos métodos de agrupamiento no supervisado es el resultado central de la Capa 3 (Tabla 4.12). El Dirichlet Process fracasa de forma informativa: concentra el 82 % de los especímenes en un único componente y no supera el azar (ARI = -0.00, NMI = 0.11), comportamiento estable frente al hiperparámetro α (probado de 0,01 a 5,0). La interpretación es que el espacio latente es *continuo*, un gradiente morfológico sin agrupaciones esféricas discretas, justo el supuesto que el DP necesita. El BHC, en cambio, recupera la estructura biológica casi por completo: ARI = 0.58 y NMI = 0.85 (Figura 4.10). Trabajar con la gaussiana completa de cada especie –y no con puntos individuales– es lo que le permite discriminar sobre un soporte continuo donde el DP colapsa.

Cuadro 4.12: Recuperación de la estructura biológica sin supervisión, espacio latente $d = 5$. ARI/NMI: 0 = aleatorio, 1 = perfecto.

Método	ARI	NMI	K activos
Dirichlet Process (GMM bayesiano)	-0.00	0.11	5
BHC (UPGMA + Bhattacharyya)	0.58	0.85	5

Los grupos descubiertos por el BHC tienen sentido filogenético directo: las tres especies de *Papio* forman un único *cluster*, mientras que *Cercocebus*, *Lophocebus*, *Macaca* y *Theropithecus gelada* quedan separados. Las distancias internas reproducen las afinidades conocidas: *P. anubis*–*P. hamadryas* son las más próximas (0.21), seguidas de *P. ursinus*

(≈ 0.43); el par *Cercocebus–Lophocebus* (mangabeys) se sitúa en ≈ 0.62; y *Theropithecus gelada* es el más distante de todos. Que un método sin acceso a las etiquetas recupere la filogenia con NMI = 0.85 es la evidencia central de que el espacio latente del bGPLVM codifica estructura evolutiva real.

4.5.4. Posición de los fósiles

Validado el espacio, se sitúan los 35 fósiles midiendo su distancia a las especies actuales (Figura **??**). Por taxón fósil, los patrones son los siguientes:

- *Parapapio lothagamensis* ($n = 4$): gravita hacia el clado *Papio* (mayoría *P. hamadryas*, 3/4), con distancia media 1.22. Es el taxón fósil más homogéneo.
- *Parapapio ado* ($n = 4$): también hacia *Papio*, distancia media 1.00.
- *Parapapio s.l.* ($n = 7$): muy disperso entre géneros (distancia media 1.12), reflejo de un taxón heterogéneo.
- *Cercopithecoides williamsi* ($n = 3$): sin afinidad clara, cada espécimen se aproxima a una especie distinta.
- *Theropithecus oswaldi* ($n = 17$): **no converge a *T. gelada***. A pesar de ser congenérico del único graminívoro actual, se reparte entre *Cercocebus atys* (7/17) y *Papio hamadryas*, con distancia media 1.20.

El espécimen más nítido de todo el conjunto es `43996_3443_P` (*Parapapio*), que cae a distancia 0.32 de *P. hamadryas*; el más ambiguo es `613_3534_TO` (*T. oswaldi*), a distancia 2.56 de su especie más cercana –el *outlier* morfológico más extremo del proyecto.

El caso de *Theropithecus oswaldi*. Este es el hallazgo paleoecológico más relevante de la Capa 3. La intuición taxonómica esperaría que *T. oswaldi* se aproximara a su congénere actual *T. gelada*, pero la microtextura lo dispersa lejos de él: en el agrupamiento BHC, los 17 especímenes se reparten entre cuatro *clusters* distintos. Lejos de ser un fallo del método, este es justamente el tipo de resultado que una clasificación supervisada forzada habría ocultado bajo una etiqueta única y una falsa precisión: la señal de microtextura de *T. oswaldi* es heterogénea y no replica el patrón oclusal del gelada actual, lo que es consistente con interpretaciones paleodietéticas que atribuyen a esta especie una dieta más variada que la del gelada estrictamente graminívoro.

4.5.5. Proyección supervisada: afinidad por especie fósil

La vía supervisada completa el cuadro de la Capa 3. Una vez validado sobre los actuales (Sección 4.2.1), el clasificador jerárquico –con la jerarquía DENDRO derivada del GPLVM– se reentrena con los 125 especímenes actuales y se proyecta cada fósil a través de él. La salida no es una etiqueta forzada, sino la distribución de probabilidad posterior sobre las siete especies vivas; promediada por taxón fósil, mide directamente cómo de parecida es cada especie extinta a cada especie actual (Tabla 4.13).

Dos lecturas se desprenden de la tabla. Primera, todas las distribuciones son difusas: la afinidad principal nunca supera 0.39 y el resto de la masa se reparte entre varias especies.

Cuadro 4.13: Distribución media de afinidad de cada especie fósil hacia las especies actuales (vía supervisada, clasificador jerárquico DENDRO, prior horseshoe). Cada fila promedia la probabilidad posterior sobre los n especímenes del taxón; suma 1. En negrita, la afinidad principal, acompañada de su desviación estándar entre especímenes.

Especie fósil	n	C. atys	L. aterr.	M. sylv.	P. anubis	P. hamadr.	P. ursinus	T. gelada
Theropithecus oswaldi	17	0.11	0.16	0.09	0.18	0.14	0.05	$\mathbf{0.26 \pm 0.12}$
Parapapio (s.l.)	7	0.05	0.16	0.08	$\mathbf{0.28 \pm 0.10}$	0.19	0.12	0.12
Parapapio ado	4	0.04	0.09	0.03	$\mathbf{0.35 \pm 0.08}$	0.25	0.17	0.08
Parapapio lothagamensis	4	0.03	0.16	0.07	$\mathbf{0.29 \pm 0.07}$	0.22	0.12	0.10
Cercopithecoides williamsi	3	0.03	0.08	0.04	$\mathbf{0.39 \pm 0.13}$	0.28	0.16	0.02

Esta planitud es la traducción, en el plano supervisado, de la no comparabilidad distribucional de la Capa 2: el modelo, al estar obligado a cuantificar su confianza, se niega a asignar los fósiles con una certeza injustificada. Los *Parapapio* y *Cercopithecoides* gravitan de forma consistente hacia el clado *Papio* (con *P. anubis* y *P. hamadryas* acaparando juntas 0.5–0.7 de la masa), siendo *Parapapio ado* y *Cercopithecoides williamsi* los de afinidad más nítida.

Segunda, y más interesante, *Theropithecus oswaldi* es el único taxón cuya distribución pica en *T. gelada* (0.26), su congénere actual. Este resultado contrasta con la vía no supervisada de distancias en el espacio latente (Sección 4.5.4), donde *T. oswaldi* se dispersaba lejos del *gelada*. La discrepancia no es un error de ninguna de las dos vías, sino exactamente la información que su contraste está diseñado para extraer: la afinidad de *T. oswaldi* con el *gelada* es real pero débil y sensible al método –coherente con una dieta más variada que la del *gelada* estrictamente graminívoro–, mientras que las afinidades de los *Parapapio* hacia *Papio*, que ambas vías recuperan, son las más robustas. La coincidencia entre los dos procedimientos señala las afinidades firmes; la discrepancia, los casos genuinamente ambiguos.

4.5.6. Incertidumbre variacional

La incertidumbre variacional permite responder a una pregunta crítica: ¿proyecta el modelo los fósiles con menos confianza que los actuales? La respuesta es no (Tabla 4.14, Figura 4.11): la incertidumbre media de los fósiles (0.067 ± 0.006 en $d = 5$; desviación estándar) es prácticamente idéntica a la de los actuales (0.066 ± 0.009). El modelo sitúa los fósiles en el morfoespacio con la misma seguridad que los actuales; su atipicidad no es un problema de proyección, sino una propiedad real de su posición. Además, en $d = 5$ las dimensiones LD4 y LD5 presentan una σ aproximadamente el doble que LD1–LD3 (0.115 y 0.093 frente a 0.04–0.05), lo que las identifica como las menos informativas y refuerza que la estructura del problema es esencialmente tridimensional.

La incertidumbre se cuantifica en dos planos complementarios, ambos presentes de forma explícita en las salidas del pipeline. En el primero, el bGPLVM asigna a cada espécimen una distribución a posteriori en el espacio latente, cuya desviación σ mide la confianza de su proyección; la pregunta crítica –¿proyecta el modelo los fósiles con menos confianza que los actuales?– se responde negativamente y de forma estable en las tres

Cuadro 4.14: Incertidumbre variacional media σ (actuales vs. fósiles) según la dimensión latente. Valor reportado: media ± desviación estándar.

Configuración	σ actuales	σ fósiles
$d = 3$ (75 ind.)	0.053 ± 0.009	0.054 ± 0.007
$d = 5$ (40 ind.)	0.066 ± 0.009	0.067 ± 0.006
$d = 7$ (75 ind.)	0.082 ± 0.007	0.085 ± 0.005

dimensiones latentes (Tabla 4.14): la σ media de los fósiles es prácticamente idéntica a la de los actuales en $d = 3$ (0.054 vs. 0.053), $d = 5$ (0.067 vs. 0.066) y $d = 7$ (0.085 vs. 0.082). El modelo sitúa los fósiles en el morfoespacio con la misma seguridad que los actuales: su atipicidad es una propiedad real de su posición, no un defecto de proyección. Internamente, las dimensiones LD4 ($\sigma = 0.115$) y LD5 ($\sigma = 0.093$) duplican la incertidumbre de LD1–LD3 (0.04–0.05), confirmando que son las menos informativas y que la estructura del problema es esencialmente tridimensional.

En el segundo plano, la vía supervisada del clasificador jerárquico no devuelve una etiqueta forzada, sino la distribución de probabilidad posterior completa sobre las especies vivas; su grado de concentración es, en sí mismo, la medida de incertidumbre. El resultado es inequívoco: ningún fósil recibe una asignación confiada. La probabilidad de la especie más probable promedia 0.38 sobre los 35 fósiles, con un máximo de 0.667 y un mínimo de 0.261; solo 6 de 35 superan $P = 0.5$ y 9 de 35 quedan por debajo de 0.3 (distribución casi plana). Esta difusión sistemática es exactamente el reflejo, en el plano supervisado, de la no comparabilidad distribucional diagnosticada en la Capa 2: el modelo, al estar obligado a expresar su incertidumbre, se niega a asignar los fósiles con una confianza injustificada. Los pocos casos concentrados son interpretables –los especímenes de *T. oswaldi* con mayor probabilidad gravitan hacia *T. gelada* ($P \approx 0.59$–0.67)– mientras que los más difusos (entropía normalizada > 0.93) corresponden a morfologías intermedias sin correspondencia clara. El contraste entre esta vía supervisada y la no supervisada de distancias en el espacio latente (Sección 4.5.4) permite separar las afinidades robustas, donde ambos procedimientos coinciden, de los casos genuinamente ambiguos, donde discrepan.

4.5.7. Robustez frente a la configuración

La comparación de la especie más cercana de cada fósil entre las tres dimensiones latentes ($d = 3, 5, 7$) confirma la estabilidad de los resultados: el 63 % de los especímenes (22/35) conserva la misma especie más próxima en las tres configuraciones (Tabla 4.15). La consistencia es máxima precisamente en los taxones más homogéneos (*Cercopithecoides williamsi*, 3/3; *Parapapio lothagamensis*, el más robusto de los *Parapapio*) y mínima en *T. oswaldi* (10/17), el taxón más disperso. Los especímenes inconsistentes no son fallos del modelo, sino casos límite de morfología intermedia entre dos grupos, lo que es coherente con su carácter de afinidad relativa y no de asignación. Como control adicional, la concordancia entre las asignaciones de *cluster* del BHC y del DP (11/35) marca los fósiles más seguros: cuando dos métodos de naturaleza distinta coinciden, la afinidad es robusta; los 24/35 restantes son ambiguos por morfología intermedia.

Cuadro 4.15: Consistencia de la especie más cercana entre configuraciones ($d = 3, 5, 7$), por taxón fósil.

Taxón fósil	Consistentes	n
Cercopithecoides williamsi	3	3
Parapapio lothagamensis	2	4
Parapapio ado	2	4
Parapapio s.l.	5	7
Theropithecus oswaldi	10	17
Total	**22**	**35**

El análisis no supervisado del espacio latente arroja cuatro conclusiones. Primera, el bGPLVM aprende una representación de baja dimensión (esencialmente tridimensional) que recupera la filogenia de los cercopitécidos actuales sin usar las etiquetas (BHC: NMI = 0.85), mientras que un Dirichlet Process fracasa porque el morfoespacio es continuo, no discreto. Segunda, los fósiles se proyectan con la misma confianza que los actuales, pero ocupan posiciones atípicas: su afinidad con las especies vivas es relativa y, en varios taxones, ambigua. Tercera, el hallazgo de que *Theropithecus oswaldi* no converge a *T. gelada* –se dispersa entre varios *clusters*– es un resultado paleoecológico sustantivo que solo el enfoque no supervisado podía revelar. Cuarta, los resultados son robustos a la dimensión latente, especialmente en los taxones morfológicamente más homogéneos. En conjunto, la Capa 3 confirma que la estrategia metodológicamente honesta para los fósiles –explorar su posición con incertidumbre cuantificada en lugar de forzar una etiqueta– no solo es la correcta dada la no comparabilidad distribucional, sino que es además la que produce las hipótesis biológicas más informativas.

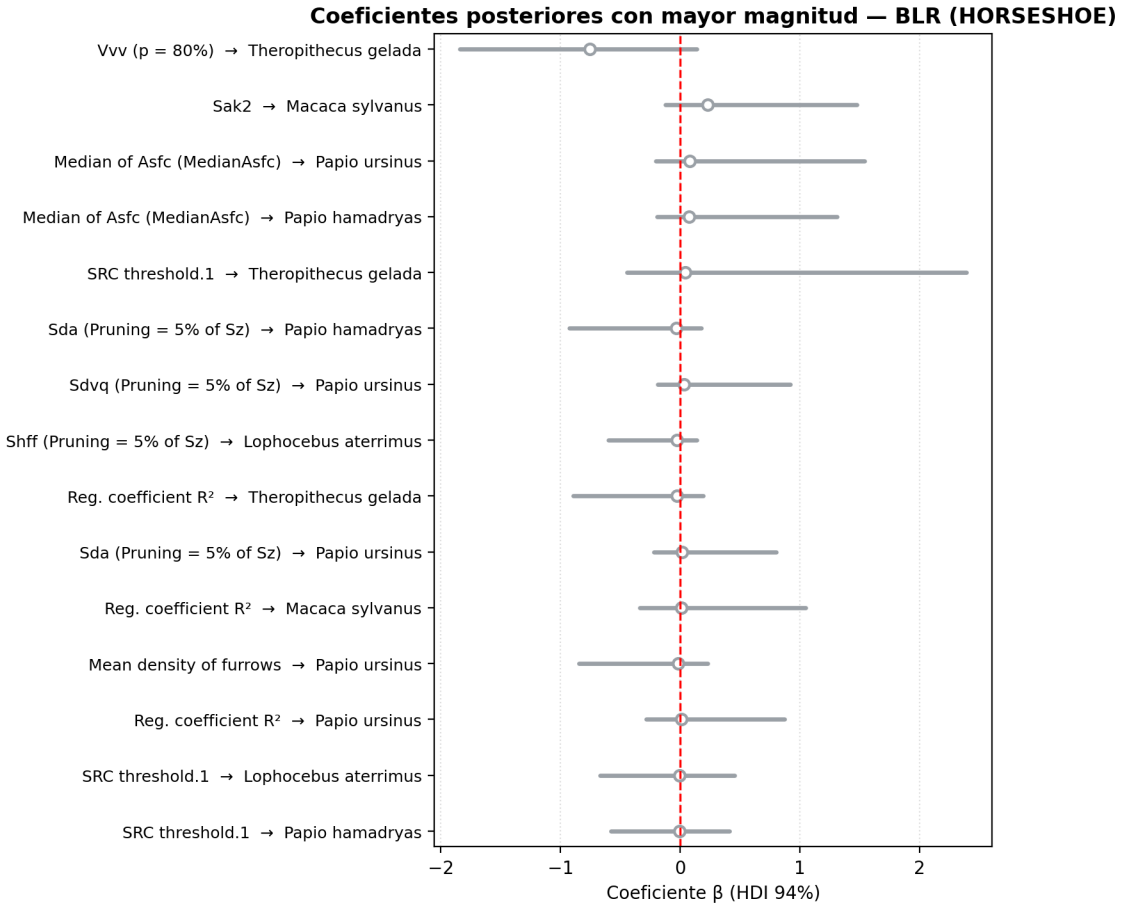

Figura 4.5: Coeficientes posteriores de mayor magnitud de la regresión logística bayesiana plana con *prior horseshoe* (los 10 de mayor $|\beta|$ medio, excluidas las dimensiones latentes; punto = mediana, barra = intervalo de credibilidad del 94 %). Pese a ser los más fuertes del modelo, todos los intervalos cruzan el cero (línea roja): ningún descriptor discrimina las especies de forma robusta. Los que más se despegan –*SRC threshold* y *Vvv* para *Theropithecus*, *Sak2* para *Macaca*– son coherentes con la señal débil y dispersa del problema.

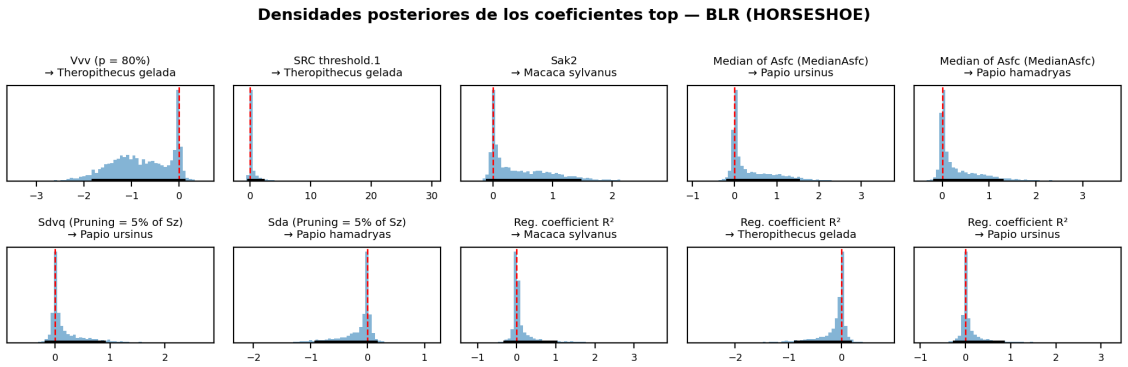

Figura 4.6: Detalle de las posteriori de coeficientes para descriptores seleccionados (*prior horseshoe*). El patrón característico es "pico en cero + lóbulo de una sola especie": *Sak2* libera el coeficiente de *Macaca*, *SRC threshold* el de *Theropithecus* y *Cercocebus*, mientras el resto de especies permanece contraído en cero. Es la firma del *shrinkage* adaptativo del *horseshoe*: silencia el ruido sin penalizar la señal fuerte, pero esa señal es tan escasa que ningún intervalo excluye el cero.

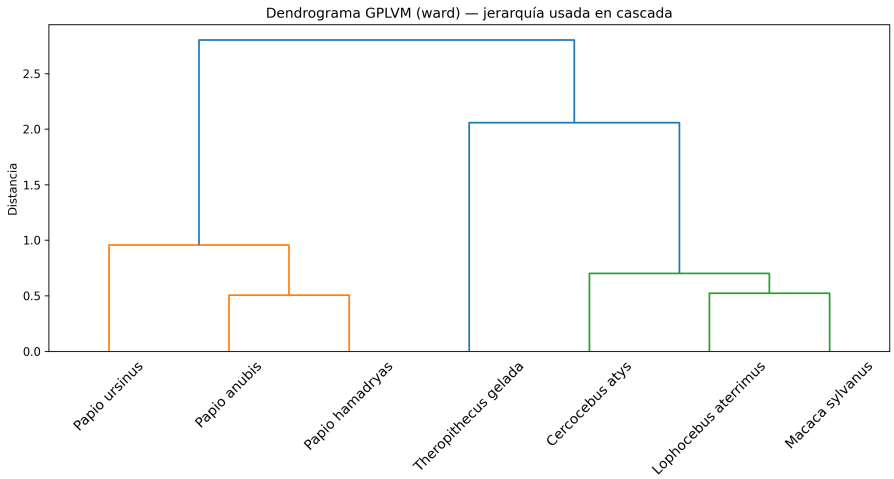

Figura 4.7: Dendrograma de las especies actuales en el espacio latente del GPLVM (enlace de Ward) que define la jerarquía en cascada. La raíz separa el clado *Papio* del resto.

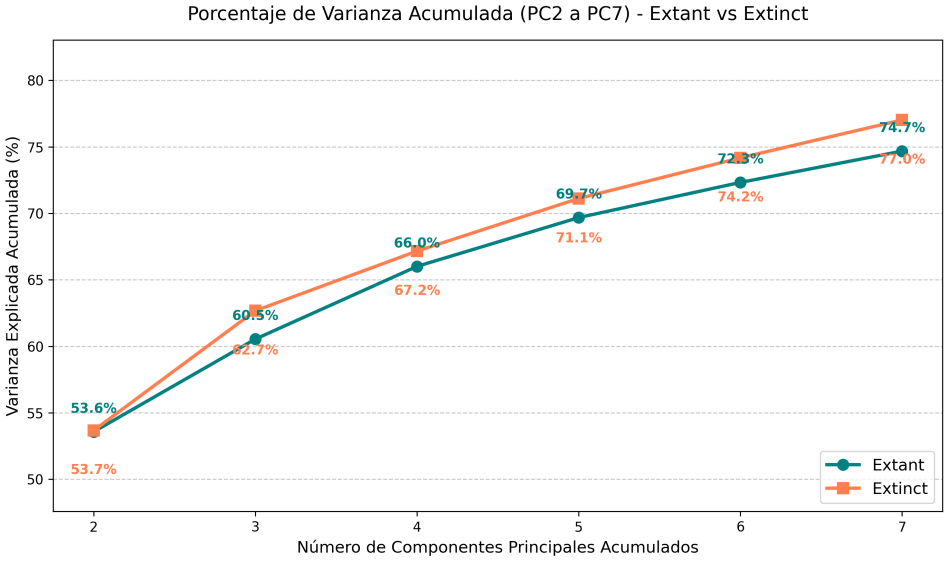

Figura 4.8: Varianza acumulada (de PC2 a PC7) para actuales y fósiles por separado. Las dos poblaciones presentan una estructura de varianza casi idéntica, pese a ocupar regiones disjuntas del espacio de variables.

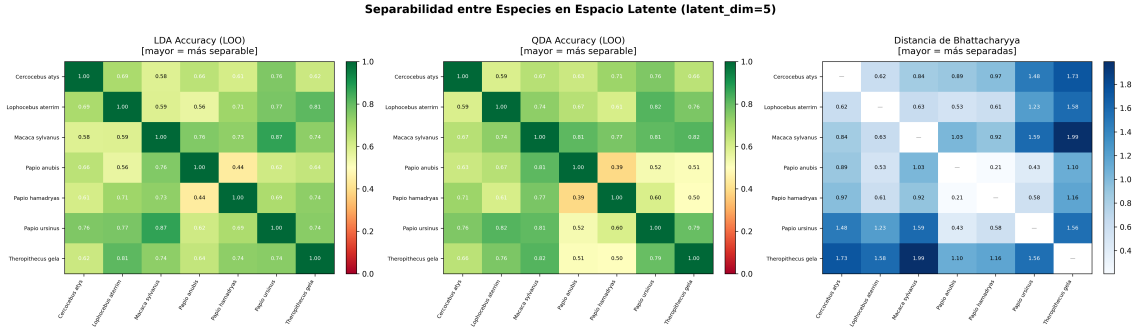

Figura 4.9: (Material suplementario) Separabilidad entre especies actuales en el espacio latente ($d = 5$). Izquierda y centro: exactitud LOO de LDA y QDA por pares (verde = separable, rojo = no separable). Derecha: distancia de Bhattacharyya entre cada par de especies, una medida de solapamiento entre dos distribuciones gaussianas (mayor distancia = menor solapamiento; azul más intenso = más separadas).

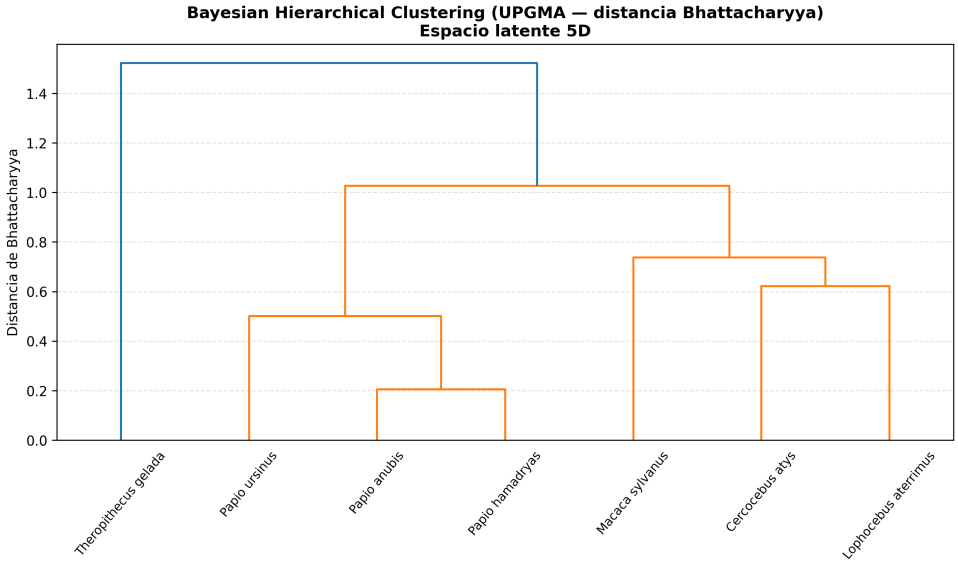

Figura 4.10: Dendrograma BHC (UPGMA sobre distancia de Bhattacharyya, $d = 5$). El agrupamiento no supervisado recupera el clado *Papio* (las tres especies juntas), aísla los mangabeys (*Cercocebus*, *Lophocebus*) y separa por completo *Theropithecus gelada*.

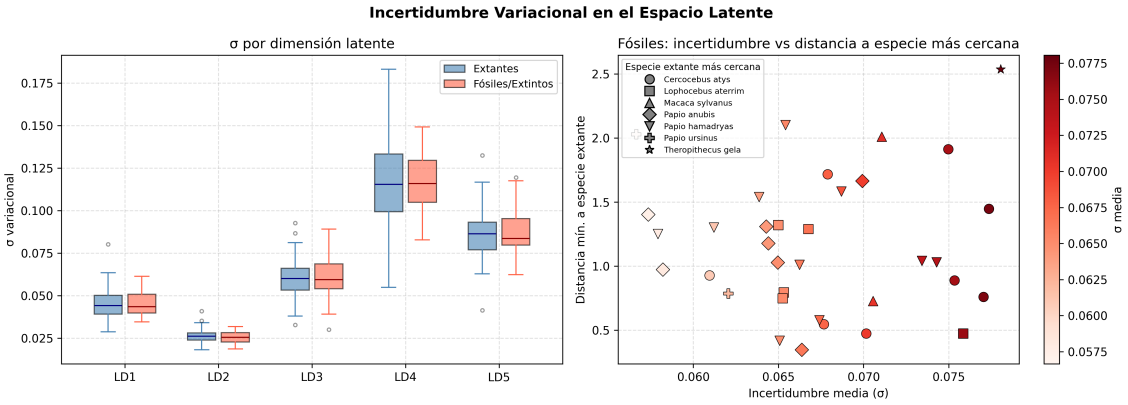

Figura 4.11: Incertidumbre variacional. Izquierda: σ por dimensión latente (actuales vs. fósiles). Derecha: para cada fósil, incertidumbre media frente a distancia a la especie actual más cercana.

Capítulo 5

Conclusiones y trabajo futuro

5.1. Conclusiones

Este trabajo ha desarrollado y validado un *pipeline* interpretable y consciente de la incertidumbre para caracterizar la microtextura dental bucal de cercopitécidos, y ha respondido de forma cuantitativa a la pregunta de investigación sobre su techo de resolución taxonómica.

La **microtextura bucal no resuelve a nivel de especie**. En los taxones actuales, la clasificación de especie no supera el azar, y ya el análisis exploratorio lo anticipaba: solo 3 de 107 variables se diferencian marginalmente entre especies y ningún coeficiente bayesiano resulta robustamente distinto de cero (Sección 4.1).

Este límite es de la señal, no del *pipeline*. A nivel de especie, la clasificación de los actuales apenas supera el *baseline* y resuelve solo el grupo; el análisis exploratorio y la selección Bayesiana confirman que ninguna variable porta señal robusta (Secciones 4.1 y 4.2). La clasificación de género fósil reproduce el mismo límite: no supera de forma apreciable el *baseline* mayoritario (Sección 4.3, Tabla 4.6).

Actuales y fósiles no son distribucionalmente comparables. El modelo de densidad de cada población asigna a la otra una verosimilitud nula a efectos prácticos ($p < 10^{-18}$ en todas las variantes), de modo que proyectar los fósiles sobre un clasificador de actuales sería extrapolar fuera de dominio (Sección 4.4).

Para los fósiles, **el enfoque no supervisado es el correcto**. El bGPLVM recupera la filogenia de los actuales sin usar etiquetas (BHC: NMI = 0,85), y revela que *Theropithecus oswaldi* no converge a su congénere actual *T. gelada* –un resultado paleoecológico que una clasificación forzada habría ocultado (Sección 4.5).

Por último, la **aportación es metodológica**. El trabajo no propone un método nuevo, sino una arquitectura de análisis que integra piezas conocidas bajo un protocolo homogéneo y con cuantificación explícita de la incertidumbre, capaz de saber *cuándo no puede afirmar algo*.

5.2. Trabajo futuro

Varias líneas extienden de forma natural este trabajo. En primer lugar, la **superficie oclusal** –no analizada aquí– podría portar una señal dietética complementaria a la bucal

y elevar el techo de resolución. En segundo lugar, **ampliar el tamaño muestral**, en especial de los taxones con muy pocos especímenes (*Cercopithecoides*, $n = 3$), permitiría estimaciones más estables y conclusiones más firmes. En tercer lugar, la **vía supervisada jerárquica** de proyección de fósiles (Sección 3.2.5) puede contrastarse de forma sistemática con la no supervisada para consolidar las afinidades robustas. Por último, el *pipeline* es **transferible** a otros problemas de clasificación con datos escasos y de alta dimensión, dentro y fuera de la paleobiología.

Limitaciones. Conviene subrayar las limitaciones del estudio: el tamaño muestral es reducido y algunos taxones fósiles tienen muy pocos especímenes; el análisis se restringe a la superficie bucal; y la familia SSFA aporta numerosas variables nuevas cuya interpretación funcional aún se está consolidando. Estas limitaciones acotan el alcance de las conclusiones y motivan las líneas de trabajo futuro.

Bibliografía

[1] Charles E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[2] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, 2006.

[4] Ivan Calandra, Ellen Schulz, Marcus Pinnow, Sabine Krohn, and Thomas M. Kaiser. Teasing apart the contributions of hard dietary items on 3d dental microtextures in primates. *Journal of Human Evolution*, 63(1):85–98, 2012.

[5] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[6] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[7] Larisa R. G. DeSantis. Dental microwear textures: Reconstructing diets of fossil mammals. *Surface Topography: Metrology and Properties*, 4(2):023002, 2016.

[8] Ferran Estebaranz-Sánchez, Kristina Kit, Juan José Ibáñez Estevez, David R. Insua, Simón Rodríguez Santana, and Laura M. Martínez. Machine learning approaches to dietary classification from dental microtexture in primates. *Scientific Reports*, 2026. In press.

[9] Arthur Francisco, Noel Brunetière, and Gildas Merceron. Gathering and analyzing surface parameters for diet identification purposes. *Technologies*, 6(3):75, 2018.

[10] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL, 3 edition, 2013.

[11] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge, 2007.

[12] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[13] Edward I. George and Robert E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[14] Zachariah Gompert. A next generation of hierarchical bayesian analyses of hybrid zones enables model-based quantification of variation in introgression in R. *Ecology and Evolution*, 14(11):e70548, 2024.

[15] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, 2 edition, 2009.

[16] Katherine A. Heller and Zoubin Ghahramani. Bayesian hierarchical clustering. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 297–304, 2005.

[17] Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014.

[18] International Organization for Standardization. ISO 25178-2: Geometrical product specifications (GPS) — surface texture: Areal — part 2: Terms, definitions and surface texture parameters. Geneva: ISO, 2012.

[19] Kathryn M. Irvine and Thomas J. Rodhouse. A Bayesian hierarchical modeling approach for species diversity in ecology. *Ecology and Evolution*, 2024. U.S. Geological Survey publication.

[20] Nathalie Japkowicz and Mohak Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, Cambridge, 2011.

[21] William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.

[22] Vidhi Lalchand, Aditya Ravuri, and Neil D. Lawrence. Generalised Gaussian process latent variable models (GPLVM) with stochastic variational inference. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*. PMLR, 2022.

[23] Neil D. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

[24] Ying Li, Zhidi Lin, Feng Yin, and Michael Minyi Zhang. Preventing model collapse in Gaussian process latent variable models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 28278–28308. PMLR, 2024.

[25] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 4765–4774, 2017.

[26] Laura M. Martínez, Ferran Estebaranz-Sánchez, Alejandro Pérez-Pérez, et al. Effectiveness of buccal dental-microwear texture in African Cercopithecoidea dietary discrimination. *American Journal of Biological Anthropology*, 179(4):678–686, 2022.

[27] Laura M. Martínez, A. Estévez Roig, M. Alrousan, and Ferran Estebaranz-Sánchez. On sample size and buccal enamel preservation in dental microwear. In *AWRANA Congress*, 2022.

[28] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, Boca Raton, FL, 2 edition, 2020.

[29] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[30] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Independently published, 2 edition, 2022.

[31] Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, Cambridge, MA, 2022.

[32] Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017.

[33] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

[34] Robert S. Scott, Peter S. Ungar, Timothy S. Bergstrom, Christopher A. Brown, Bruce E. Childs, Mark F. Teaford, and Alan Walker. Dental microwear texture analysis: Technical considerations. *Journal of Human Evolution*, 51(4):339–349, 2006.

[35] Robert S. Scott, Peter S. Ungar, Timothy S. Bergstrom, Christopher A. Brown, Frederick E. Grine, Mark F. Teaford, and Alan Walker. Dental microwear texture analysis shows within-species diet variability in fossil hominins. *Nature*, 436:693–695, 2005.

[36] Brian M. Shearer, Peter S. Ungar, Kieran P. McNulty, William E. H. Harcourt-Smith, Holly M. Dunsworth, and Mark F. Teaford. Dental microwear profilometry of african non-cercopithecoid catarrhines of the early miocene. *Journal of Human Evolution*, 78:33–43, 2015.

[37] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[38] Michalis K. Titsias and Neil D. Lawrence. Bayesian Gaussian process latent variable model. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9 of *Proceedings of Machine Learning Research*, pages 844–851. PMLR, 2010.

[39] Peter S. Ungar, Christopher A. Brown, Timothy S. Bergstrom, and Alan Walker. Quantification of dental microwear by tandem scanning confocal microscopy and scale-sensitive fractal analyses. *Scanning*, 25(4):185–193, 2003.

[40] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017.

[41] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of mcmc. *Bayesian Analysis*, 16(2):667–718, 2021.

[42] Christopher K. Wikle and Andrew Zammit-Mangion. Good modelling practice in ecology: The hierarchical bayesian perspective. *Ecological Modelling*, 496:110836, 2024.

[43] In-Kwon Yeo and Richard A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.

[44] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.