



MOX-Report No. 68/2026

Random mixtures in Bayes Hilbert spaces

Patanè, G.; Greven, S.; Menafoglio, A.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it

<https://mox.polimi.it>

Random mixtures in Bayes Hilbert spaces

Giulia Patanè^{*1}, Sonja Greven², and Alessandra Menafoglio¹

¹MOX, Department of Mathematics, Politecnico di Milano, Via Edoardo Bonardi 9, 20133 Milan, Italy

²Chair of Statistics, School of Business and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10117 Berlin, Germany

Abstract

We present a framework for the analysis and unmixing of random density mixtures in the Bayes Hilbert space $B^2(I)$. General identifiability results for mixtures in Hilbert spaces are established and applied to the Bayes Hilbert space setting. Building on these results, we propose a penalised maximum likelihood approach for the unmixing of Bayes Hilbert mixtures aimed at recovering the statistically space-efficient representation, together with a computationally efficient coordinate-wise maximisation algorithm for its implementation. The methodology is illustrated through a hyperspectral data application, where observations can be naturally embedded in the Bayes Hilbert space and analyzed in terms of distributional shape rather than amplitude. A complementary simulation study demonstrates the interpretability and practical performance of the proposed approach.

Keywords: Mixture models, Compositional data analysis, Convex geometry, Hyperspectral imaging, Functional data analysis

1 Introduction

Probability densities are central to statistics because they describe how values of a variable are distributed across a population. In many modern applications, the unit of analysis is not a single individual but an entire society or complex system. In these cases, the overall distribution can reveal important patterns that simple summaries, like the mean or variance, cannot. For example, two completely different distributions can have the same mean and variance, yet behave in fundamentally different ways. Capturing such differences requires looking at the full density, rather than just a few summary statistics.

In this work, we introduce and explore mixtures of densities, decomposing an observed density into interpretable constituent parts with compositional proportions. We prove the existence and uniqueness of the pair of vertices and proportions that maximise the statistical spread of the data over the space of compositional weights – i.e. positive weights summing to 1. Moreover, we propose a penalised maximum likelihood unmixing strategy and validate the model through simulations and on a case study – using data from AVIRIS Indian Pines [3].

Working with densities is not straightforward because they are constrained objects: they cannot be negative, and their total mass must always sum to one. Classical approaches in functional data analysis (FDA, [30]) generally assume that the data can vary freely, which is not true for densities. Applying these methods directly to densities can lead to misleading or invalid results.

To overcome this problem, researchers developed the concept of the *Bayes Hilbert space* [39], a mathematical framework specifically designed for densities. In this space, densities can be combined and manipulated in ways that respect their natural constraints. Indeed, transformations exist that allow traditional functional data analysis techniques to be applied safely, while making the study of densities simpler.

Embedding densities in the Bayes Hilbert space has already led to many practical advances. Researchers have used it to handle higher dimensional domains [11, 13], to work with non-Lebesgue reference measures [37], to model spatially dependent densities [23], to analyze situations where only sparse observations are available [20, 35], and where densities are the predictors [14] or the responses [18, 20, 21] of regression models.

Yet, one fundamental problem remains unaddressed within this framework: the unmixing of densities. Mixtures arise naturally in many fields, representing populations composed of multiple subgroups or overlapping latent processes. Consider, for instance, the income distribution of a country composed of distinct socioeconomic

^{*}Corresponding author: giulia.patane@polimi.it

groups: the observed density is a mixture of the group-specific densities, and unmixing it reveals the underlying subpopulation structure. Addressing this problem within the Bayes Hilbert space framework, however, is not straightforward. While simplicial PCA [12, 35] offers the closest existing tool for decomposing densities in the Bayes Hilbert (BH) space, it does so along orthogonal directions of variation – not as a mixture of components with compositional weights. Moreover, such a decomposition cannot be equated with classical probabilistic mixtures, which are linear combinations in the traditional sense and do not respect the geometric structure of the Bayes Hilbert space. This distinction matters whenever the population is heterogeneous and the goal is to recover latent subgroups or overlapping processes. As we discuss in Section 3, BH mixtures induce a genuinely different notion of mixing, and their interpretation requires care. Indeed, classical mixtures induce non-linear structures in Bayes Hilbert spaces. Conversely, mixtures defined within Bayes Hilbert spaces are given by convex combinations of distributions with respect to the underlying vector space operations, and therefore do not, in general, coincide with classical mixtures. This discrepancy has been partially addressed in the literature on α -mixtures, which introduce alternative, generally non-linear, combination rules [2, 31–33]. These mixtures have been investigated in several applied areas, including survival analysis and reliability modeling, stochastic ordering and comparison of heterogeneous populations, and Bayesian inference for lifetime models, where they provide flexible interpolation schemes between arithmetic – for $\alpha = 1$ – and geometric – $\alpha \rightarrow 0^+$ – aggregation of probability models. However, this line of work does not provide a practical tool for nonparametric estimation of the vertices and of the proportions’ distribution. In this work, we focus instead on exploring Bayes Hilbert mixtures as a powerful alternative grounded in the literature on compositional data analysis [11, 13, 14, 18, 20, 21, 35–37, 39].

A particularly natural application of this framework arises in hyperspectral imaging, where each pixel records a superposition of pure-material signatures. Hyperspectral data consist of reflectance – or radiance – curves measured across a dense grid of wavelengths, forming a high-dimensional functional signature for each pixel of the recorded scene. In this context, it is not the absolute magnitude of reflectance that is most informative, but rather the relative distribution of energy across wavelengths – that is, the overall shape of the spectral curve. Subtle variations in peaks, slopes, and absorption features reveal the presence of specific materials, surface properties, or ongoing chemical processes [22]. For this reason, hyperspectral reflectance curves are usually analysed after normalisation, encoding the relative distribution of energy across wavelengths rather than absolute intensity, and can thus be treated as densities within the Bayes Hilbert space framework – making unmixing a problem of recovering compositional proportions over a simplex of endmember densities. More broadly, because densities are positive, compactly supported, and continuous, they can represent any functional object whose key feature is its shape rather than its amplitude, making this framework widely applicable beyond classical probabilistic settings.

In hyperspectral data, pure material exhibits a characteristic signature – either in reflectance or radiance – allowing its remote identification within a scene, namely a spatially extended domain such as a planetary surface portion [40, 41]. When a scene is captured by remote sensors, every pixel of the resulting image corresponds to a hyperspectral observation. However, hyperspectral measurements rarely coincide exactly with pure-material signatures, as sensing technologies introduce spatial blurring and mixing effects due to neighbouring light sources and limited spatial resolution. For this reason, a dedicated research area known as *Hyperspectral Unmixing* (HU) has emerged, aiming to develop methods that estimate the proportions – or, in some formulations, the probabilities of occurrence – of each material contributing to the spectrum recorded at a given pixel. Moreover, since the set of materials present in a scene is not always known *a priori*, the pure-material signatures – endmembers – are often estimated in a fully unsupervised manner. Some approaches, such as Vertex Component Analysis (VCA, [25]), adopt the so-called *pure-pixel assumption*, which postulates that at least one pixel in the scene corresponds to each pure material; however, this assumption may fail in practice, potentially leading to inaccurate or non-identifiable endmember estimates. Non-negative Matrix Factorisation (NMF, [15]), which identifies the minimum-volume simplex enclosing the data, and Minimum-volume HU [17] relax this assumption, but remain sensitive to outliers: extreme observations can artificially inflate the enclosing simplex and distort the estimated endmembers. Our approach occupies a middle ground: rather than requiring pure pixels to be present in the dataset, we assume that the endmembers have positive probability density – that is, they need not be observed, but must be plausible under the data-generating distribution. Under this assumption, we identify the vertices that maximise the statistical spread of the data over the simplex, a criterion grounded in the covariance structure of the compositional weights and less sensitive to extreme observations than minimum-volume alternatives.

The article is organised as follows. In Section 3 we formally define and study random density mixtures within the Bayes Hilbert space, and discuss the difference with respect to probabilistic mixtures. Section 4 investigates the identifiability conditions of the corresponding unmixing problem in generic Hilbert spaces; Section 5 presents the proposed maximum likelihood estimation procedure in Bayes Hilbert spaces. The practical performance and applicability of the proposed methodology are demonstrated through an extensive simulation study (Section 6) – with synthetic probability density data – and a real-data application to hyperspectral imagery (Section 7).

2 Theoretical background

2.1 Bayes Hilbert Space

A Bayes Hilbert space on a compact interval $I \subset \mathbb{R}$, denoted by $B^2(I)$, is an infinite-dimensional Hilbert space specifically designed for the analysis of density functions [39]. Formally, $B^2(I)$ can be defined as the quotient space

$$L_+^2(I) / \sim_{B^2},$$

where

$$L_+^2(I) = \{f : I \rightarrow \mathbb{R}^+ \text{ s.t. } \log(f) \in L^2\},$$

and $\sim_{B^2}$ is the equivalence relation given by

$$f \sim_{B^2} g \text{ if and only if } f = \eta g \text{ a.e. on } I, \text{ for some constant } \eta > 0.$$

Thus, elements of $B^2(I)$ are equivalence classes of positive functions whose logarithm is square integrable, that differ only by a positive multiplicative constant. In other words, functions are identified up to scale, reflecting the fact that only their relative structure – or shape – is relevant. By convention, in the following, each equivalence class is represented, if possible, by its normalised element, i.e. the density integrating to one over I . As a Hilbert space, it is equipped with a sum $\oplus$ (a.k.a. perturbation), a multiplication by scalar $\odot$ (a.k.a. powering), and a scalar product, with respect to which $B^2(I)$ is complete. For $f, g \in B^2(I)$ and $\alpha \in \mathbb{R}$, the operations are defined as

$$\begin{cases} f \oplus g = \frac{f \cdot g}{\int_I f(s) \cdot g(s) ds}, \\ \alpha \odot f = \frac{f^\alpha}{\int_I f(s)^\alpha ds}. \end{cases} \quad (1)$$

To define a scalar product on $B^2(I)$, [39] introduced the *centered log-ratio (clr) transformation*, which maps densities into elements of

$$L_0^2(I) = \left\{ f \in L^2(I) \mid \int_I f(s) ds = 0 \right\}.$$

The clr-transformation is defined as

$$\text{clr}(f)(t) = \log(f(t)) - \frac{1}{\lambda(I)} \int_I \log(f(s)) ds, \quad (2)$$

where λ denotes the Lebesgue measure on I . By construction, with $\langle f, g \rangle_{B^2(I)} = \langle \text{clr}(f), \text{clr}(g) \rangle_{L^2(I)}$, the clr-transformation is an isometry between $B^2(I)$ and $L_0^2(I)$, and it preserves the Hilbert space structure. Moreover, $B^2(I)$ is separable [39], a property that is particularly important in statistical and computational contexts, where data must be efficiently represented and recorded.

This formulation is here considered for ease of exposition, although a wide range of extensions exists, including generalizations to more complex or unbounded supports and non-Lebesgue reference measures [13, 37, 39].

2.2 Linear mixtures

Let $\{h_j\}_{j=1}^m \subset B^2(I)$ be a set of densities. A linear mixture is defined as [10]

$$g = \sum_{j=1}^m p_j h_j, \quad \sum_{j=1}^m p_j = 1, \quad p_j \geq 0 \quad \forall j.$$

If we denote the vector of proportions $\mathbf{p} = (p_j)_j$, then by construction $\mathbf{p} \in S^m$, where $S^m \subset \mathbb{R}^m$ is the $(m-1)$ dimensional simplex. Thus, g is a linear convex combination of the densities $\{h_j\}_{j=1}^m$, or equivalently a weighted arithmetic mean, where the weights sum to 1. Let $X \sim g$. Then, X has the same distribution as the mixture Y , defined conditionally as $Y|\{J=j\} \sim h_j$ for all $j \in \{1, \dots, m\}$, with mixing index $J \sim \mathbf{p}$ supported on $\{1, \dots, m\}$.

Consequently, the linear mixture provides an appropriate model when each observation is assumed to be generated by first selecting a component density h_j with probability p_j , and then sampling from that density. In this sense, the linear mixture represents a purely *probabilistic* mixture, as it explicitly models heterogeneity through a random selection among distinct components.

In the next section, we introduce a framework for random Bayes Hilbert mixtures, highlighting the crucial differences with respect to the classical linear – probabilistic – mixture.

3 Bayes Hilbert mixtures

We now introduce the core theoretical developments of our work, beginning with the definition of Bayes Hilbert mixtures. Note that this construction does not depend on the specific geometry of Bayes spaces, but is valid in any Hilbert space. However, for the purposes of this work, we restrict our attention to mixtures defined within a Bayes Hilbert space and adopt the corresponding notation throughout. Let $\{h_j\}_{j=1}^m \subset B^2(I)$ be a set of linearly independent densities, where linearity has to be intended in the Bayes Hilbert sense, namely with respect to the operators $\oplus, \odot$ defined above. A *convex hull* in the Bayes Hilbert space $(B^2(I), \oplus, \odot)$ is a set

$$\mathcal{M}(\{h_j\}_{j=1}^m) = \left\{ g \in B^2(I) \mid g = \bigoplus_{j=1}^m p_j \odot h_j, \sum_j p_j = 1, p_j \geq 0 \forall j \right\}.$$

In the following, we name *Bayes Hilbert mixtures* the elements of $\mathcal{M}$. Note that, by projecting $\mathcal{M}$ through clr into the $L_0^2(I)$ space, the image is $\text{conv}(\text{clr}(h_1), \dots, \text{clr}(h_m))$, i.e. the classical (linear) convex hull of clr -transformed densities $\{h_j\}_{j=1}^m$.

Let

$$g = \bigoplus_{j=1}^m p_j \odot h_j = \frac{\prod_{j=1}^m h_j^{p_j}}{\int_I \prod_{j=1}^m h_j^{p_j}(x) dx} =_{B^2} \prod_{j=1}^m h_j^{p_j},$$

for a given set of $\{h_j\}_{j=1}^m$ in $B^2(I)$ and $\mathbf{p} = (p_j)_j$ in S^m . By construction, the resulting density g corresponds to the weighted geometric mean of the densities $\{h_j\}_{j=1}^m$. Consequently, the Bayes Hilbert mixture represents a form of aggregation that differs fundamentally from probabilistic mixing: rather than describing a population composed of distinct subpopulations, it yields a single distribution that integrates the characteristics of the original densities. In this sense, the Bayes Hilbert mixture is more appropriately viewed as a *fuzzy* mixture than as a probabilistic one.

The concept of a fuzzy mixture is widely recognized [16]: in general, it describes a framework in which each observation does not belong exclusively to a single component, but rather to all of them simultaneously, with varying degrees of membership. Just as *grey* is not simply a mixture of white and black pixels, but a shade in its own right – where white and black represent the two extremes of a continuum – the Bayes Hilbert mixture yields a genuinely new density that interpolates smoothly between the original components. It is strictly related to the problem of soft-clustering, as it models data belonging to more than one cluster at the same time, with different proportions. This seems appropriate e.g. in our application on hyperspectral unmixing, where pixels usually contain more than one material at the same time.

Figure 1 (top) displays an example with Beta distributions. The two continuous curves represent two densities d_1, d_2 in the Bayes Hilbert space $B^2([0, 1])$. The dashed curve represents the result of a linear mixture, while the dotted one represents the result of a Bayes Hilbert mixture. In both cases, $\mathbf{p} = (0.5, 0.5)$. The linear mixture is bimodal, with modes that are near the modes of the two original densities. By contrast, the Bayes Hilbert mixture is unimodal, with the mode in 0.5: this highlights that the two kinds of mixtures can be strongly different. For example, let us consider the densities and mixtures of Figure 1 (top) having a domain equal to a color map from white, 0, to black, 1. Figure 1 (bottom) displays pixels that are i.i.d sampled from d_1 (first from left, mostly light), d_2 (last, mostly dark), their linear mixture (second, heterogeneous mixture of light and dark pixels) and Bayes Hilbert mixture (third, mostly *grey*). This illustrates that linear mixtures are preferable when the data are believed to arise from separable subpopulations (here light and dark pixels), whereas Bayes Hilbert mixtures define a population, where mixing takes place within each observation (here within pixels, leading to differently grey pixels). The choice between the two mixing paradigms is thus driven by the setting and inferential goal.

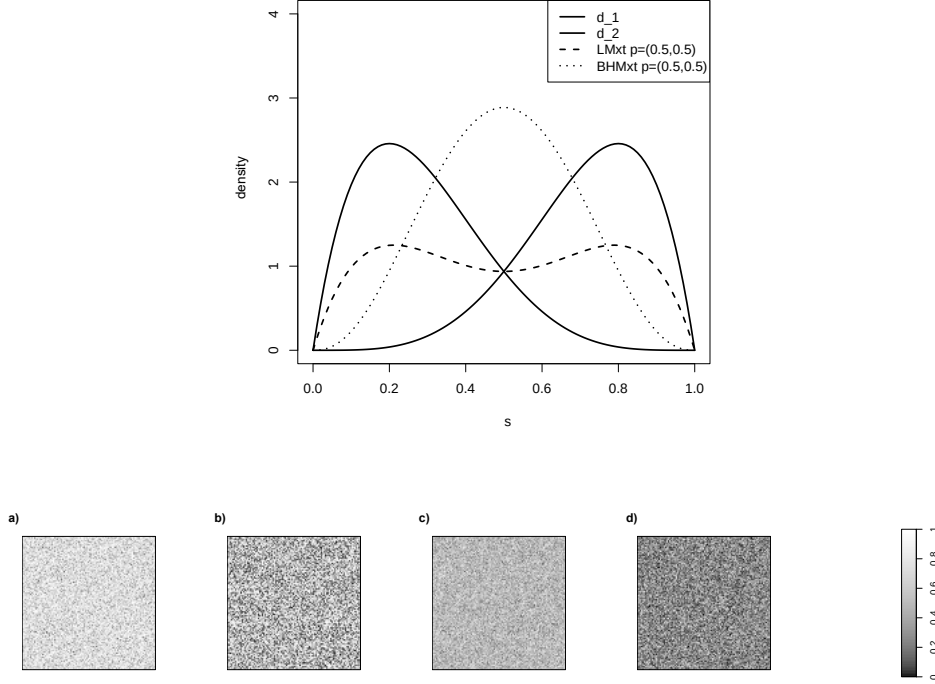


Figure 1: (Top) Density curves of Beta(2, 5) and Beta(5, 2) distributions (d_1 and d_2), their linear (uniform) mixture (dashed), and their Bayes Hilbert (uniform) mixture (dotted). (Bottom) Independent samples from a) d_1 , b) linear uniform mixture, c) BH uniform mixture, and d) d_2 , stored in a matrix for ease of visualization.

Formally, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $G : \Omega \rightarrow B^2(I)$ be a random element in the Bayes Hilbert space $B^2(I)$.

Definition 3.1 (Bayes Hilbert random mixture). Let $\{h_j\}_{j=1}^m \subset B^2(I)$ be linearly independent in the Bayes Hilbert space. We say that G is a Bayes Hilbert random mixture (of $\{h_j\}_{j=1}^m$) if

$$G(\omega) \in \mathcal{M}(\{h_j\}_{j=1}^m) \quad \text{for all } \omega \in \Omega,$$

where $\mathcal{M}(\{h_j\}_{j=1}^m)$ denotes the associated convex hull.

So far, our formulation shows that if G is any random mixture, then it *can* be represented by a collection of vertices $\{h_j\}_{j=1}^m$ with an associated random proportion vector $\mathbf{p}$. To make this useful in practice, we may ask whether we also have the identifiability of any such G , i.e. whether it can be uniquely represented as a mixture. We discuss two variants of identifiability:

- (i) **Partial identifiability:** given $\{h_j\}_{j=1}^m$, the distribution of $\mathbf{p}$ is unique;
- (ii) **Total identifiability:** the representation of G is unique, namely, there exists a unique pair consisting of vertices and random proportions generating G .

The first property holds, as shown in Section B.1. The second property, however, does not hold in general. Given G , there may exist infinitely many pairs of vertices and corresponding random proportions that generate G as a random mixture.

In particular, larger convex hulls containing the support of (the distribution of) G , indicated as $\text{supp}(G)$, correspond to low-variance distributions of $\mathbf{p}$, whereas smaller convex hulls containing $\text{supp}(G)$ require more dispersed distributions of $\mathbf{p}$ – see Figure 6 for a schematic explanation. As the convex hull shrinks, its vertices approach $\text{supp}(G)$; equivalently, the vertices become increasingly likely to appear in a realization.

In real datasets, the vertices are typically latent and do not appear as pure realizations – see the discussion of the pure-pixel assumption in [41]. Nevertheless, one may assume that the vertices, even if not observed, could in principle occur. Formally, this amounts to seeking the representation whose vertices are closest to $\text{supp}(G)$, or equivalently, whose associated distribution of $\mathbf{p}$ is as spread out as possible. This is analogue to searching for the minimum-volume convex hull, as is standard in hyperspectral unmixing problems, but in a distributional perspective. We discuss this in more detail in B.4.

4 Identifiability of mixtures in Hilbert Spaces

We study the identifiability of the representation of a random mixture G through a set of vertices $\{h_j\}_{j=1}^m$ and the distribution of the random proportion. Throughout, G is assumed to be a random mixture supported on the convex hull of m fixed vertices.

Partial identifiability Given vertices $\{h_j\}_{j=1}^m$, Proposition 4 (proved in B) guarantees that the distribution of $\mathbf{p}$ is uniquely determined by G . This allows us to define the map

$$\phi_G: \mathcal{H}_G^m \rightarrow \mathfrak{D}(S^m), \quad \phi_G(\{h_j\}_{j=1}^m) = D \iff G \stackrel{d}{=} \bigoplus_{j=1}^m p_j \odot h_j, \quad \mathbf{p} \sim D,$$

where $\mathcal{H}_G^m$ is the set of m -tuples of linearly independent vertices whose convex hull contains $\text{supp}(G)$ almost surely (Definition B.1). Under suitable regularity conditions (see B), ϕ_G is continuous and, when G is *finite-symmetric* – meaning $\phi_G^{-1}(D)$ is finite for every D – ϕ_G is also a closed map, so $\text{Im}(\phi_G)$ is closed in $\mathfrak{D}(S^m)$.

Minimal representation. Let Σ_p denote the covariance matrix of the isometric log-ratio transformation ψ (ilr, [9]) of the proportion vector $\mathbf{p} \in S^m$, namely:

$$\Sigma_p = \text{Cov}(\mathbf{x}), \quad \mathbf{x} = \psi(\mathbf{p}), \quad \psi: S^m \rightarrow \mathbb{R}^{(m-1)} \text{ isomorphism} \quad (3)$$

We define a strict partial order $<_G$ on the set of valid representations of G . Since two m -tuples of vertices that differ only by a permutation yield the same convex hull, we work on the quotient space with respect to the equivalence relation $\sim_\pi: (h_1, \dots, h_m) \sim_\pi (h'_1, \dots, h'_m)$ if there exists a permutation π of $\{1, \dots, m\}$ such that $(h_1, \dots, h_m) = (h'_{\pi(1)}, \dots, h'_{\pi(m)})$. The set of valid representations is then

$$\mathcal{C}_G = \left\{ (\{h_j\}_j, \mathbf{p}) : G \stackrel{d}{=} \bigoplus_j p_j \odot h_j \right\} / \sim_\pi.$$

We set $(\{h_j\}_j, \mathbf{p}) <_G (\{\bar{h}_j\}_j, \bar{\mathbf{p}})$ whenever $\text{tr}(\Sigma_p) > \text{tr}(\Sigma_{\bar{p}})$. The quantity $\text{tr}(\Sigma_p)$ measures the total dispersion of $\mathbf{p}$: maximising it selects the representation in which the proportions spread out most within the convex hull, thereby utilising the available space efficiently. By contrast, directly minimising the volume of the convex hull tends to concentrate the vertices near the most extreme observations, which may distort the recovered statistical structure by overweighting atypical data points. Existence of a minimal element follows from Proposition 9 (see B).

Uniqueness requires an additional assumption. We say that G satisfies the *Probabilistic Pure Pixel* (3P) assumption if each vertex of the minimal representation can be approached with positive probability. Formally

Definition 4.1. (Probabilistic Pure Pixel assumption) We say that G satisfies the Probabilistic Pure Pixel (3P) assumption if there exists $(\{h_1^*, \dots, h_m^*\}, \mathbf{p})$ in the minimal set, such that $\forall j = 1, \dots, m$ and $\forall \epsilon > 0$, $\mathbb{P}(G \in D_\epsilon(h_j^*)) > 0$, where $D_\epsilon(h_j^*)$ is the ball centered in h_j^* and with radius ϵ .

This differs from the classical pure pixel assumption in Hyperspectral Unmixing, which requires each vertex to appear as a pure observation almost surely; the 3P assumption only requires this event to have positive probability.

Theorem 1 (Minimal representation). *If G is finite-symmetric and satisfies the 3P assumption, the minimal element of $(\mathcal{C}_G, <_G)$ is unique. The unique minimal representation is given by the vertices $\{h_j^*\}_{j=1}^m$ paired with $\mathbf{p}^* \sim D^* = \phi_G(\{h_j^*\}_{j=1}^m)$.*

Corollary 2. *The vertices $\{h_j^*\}_{j=1}^m$ in Definition 4.1, paired with the corresponding random proportion*

$$\mathbf{p}^* \sim D^* = \phi_G(\{h_j^*\}_{j=1}^m)$$

are the unique minimal representation in Theorem 1.

By forcing $\mathbf{p}$ to be spread out while estimating the representation of G , we obtain the minimal representation, which corresponds to the representation for which the vertices belong to the closure of G 's support. This fact is crucial for the interpretability of our model. All proofs and auxiliary results are collected in B.

5 Bayes Hilbert unmixing via penalised maximum likelihood

In real datasets consisting of multiple realizations of a random mixture G , the assumption that G is a noiseless mixture is generally unrealistic. Without an additive noise term absorbing existing residual variability in $B^2(I)$, every observed realization would require its own vertex to be represented exactly, leading to an unbounded growth of m and to overfitting of the vertices. We therefore incorporate a noise component and fix the number of components m (the choice of this hyperparameter is discussed in 5.5). We model G as

$$G = G_{\mathcal{M}} \oplus \epsilon_{B^2} \quad (4)$$

where $G_{\mathcal{M}}$ is a noiseless mixture for which the 3P assumption holds (see Definition 4.1), and $\epsilon_{B^2} \in B^2(I)$ is a density-valued noise term with mean equal to the uniform distribution $[0]_{B^2}$, the neutral element of the BH space, and a covariance operator Γ_{ϵ} . The perturbation term ϵ_{B^2} allows G to deviate, with probability one, from the convex hull on which $G_{\mathcal{M}}$ lies. More precisely,

$$G_{\mathcal{M}} := \bigoplus_{j=1}^m p_j \odot h_j, \quad (5)$$

where $\{h_j\}_{j=1}^m \subset \mathcal{H}_G^m$ are the vertices and $\mathbf{p} := (p_j)_{j=1}^m \sim D \in \mathfrak{D}(S^m)$ are random proportions. The pair

$$(\{h_j\}_{j=1}^m, \mathbf{p}) \in \mathcal{H}_G^m \times \mathfrak{D}(S^m)$$

constitutes the minimal representation of $G_{\mathcal{M}}$. This representation is unique up to permutations of the indices, according to the identifiability results established in Section 4. In this section, we propose a strategy for estimating both the vertices and the distribution of the random proportions that generate $G_{\mathcal{M}}$, based solely on a sample of n independent and identically distributed (i.i.d.) realizations of the noisy mixture G .

5.1 Modelling the proportions' distribution

In Equation (5), $(p_j)_{j=1}^m = \mathbf{p} \in S^m$ denotes the vector of proportions. We assume that the isometric log-ratio (ilr, [9]) transformation of $\mathbf{p}$ follows a Gaussian distribution, namely

$$\mathbf{p} \sim \psi^{-1}(\mathcal{N}_{m-1}(\boldsymbol{\mu}_p, \Sigma_p)),$$

for some $\boldsymbol{\mu}_p \in \mathbb{R}^{m-1}$, $\Sigma_p \in \mathbb{R}^{(m-1) \times (m-1)}$ positive semidefinite, where $\psi : S^m \rightarrow \mathbb{R}^{m-1}$ denotes the discrete ilr transformation. The ilr transformation maps the compositional vector $\mathbf{p}$ from the simplex S^m to the real Euclidean space $\mathbb{R}^{m-1}$ by projecting log-ratio coordinates onto an orthonormal basis, thereby preserving distances and endowing the simplex with a standard (finite-dimensional) Hilbert space structure. Equivalently, in ilr coordinates, the model reads

$$\mathbf{z} = \psi(\mathbf{p}) \sim \mathcal{N}_{m-1}(\boldsymbol{\mu}_p, \Sigma_p),$$

and the composition is recovered through the inverse mapping

$$\mathbf{p} = \psi^{-1}(\mathbf{z}).$$

This formulation makes explicit that the Gaussian assumption is imposed in Euclidean space, while the induced distribution on the simplex is ilr-normal. While other log-ratio transformations, such as the clr, are possible, they differ from the ilr by a linear transformation and thus do not affect normality.

5.2 Modelling the density objects

The model detailed in Equation (4) and (5) includes different density objects: the noisy realizations $\{g_i\}_{i=1}^n$, which are observable, the latent vertices $\{h_j\}_{j=1}^m$ and the noise ϵ_{B^2} .

In our work, we represent each density object through their clr-transformation, which is discretised – applying truncation – via finite Karhunen-Loève decomposition (i.e., simplicial functional principal component analysis, SFPCA, [12]) of G . This yields, for a generic density g :

$$g(x) = \text{clr}^{-1} \left(\sum_{l=1}^k c_l \phi_l(x) \right) \quad (6)$$

where $\{\phi_l\}_{l=1}^k \in L_0^2(I)$ are the first k simplicial principal components (SFPCs) – which form a k -dimensional orthonormal basis – and $c_j \in \mathbb{R}$ are the scores of $g(x)$ over the SFPCs – i.e., $c_j = \langle g, \phi_j \rangle_{B^2}$, with $\langle \cdot, \cdot \rangle_{B^2}$ the inner product in $B^2(I)$.

The clr -transformation of the density ϵ_{B^2} is assumed to be distributed as a Gaussian process (GP, [6]), namely $\epsilon_{B^2} = \text{clr}^{-1}(\epsilon)$ with

$$\epsilon \sim GP_{L_0^2}(0, \Gamma_\epsilon).$$

Analogous to (6), we can expand $\epsilon = \text{clr}(\epsilon_{B^2})$ in the basis of the first k eigenfunctions (SFPCs) of G , with coefficients then following a multivariate Gaussian distribution of dimension k with mean $\mu_\epsilon = \mathbf{0}$ and covariance matrix Σ_ϵ . The equivalence is shown in [35]. Hereafter, the symbol $\mathbf{c}_i \in \mathbb{R}^k$ will denote the vector of coefficients of $\text{clr}(g_i)$, associated with the k elements of the basis $\{\phi_1, \dots, \phi_k\}$. Analogously, the matrix $\mathbb{H} \in \mathbb{R}^{k \times m}$ will store the coefficients of h_j , $j \in \{1, \dots, m\}$ column-wise. For clarity, the complete model can be summarised as follows. Each observed density g_i is represented by its clr -coefficient vector $\mathbf{c}_i \in \mathbb{R}^k$, which satisfies

$$\mathbf{c}_i = \mathbb{H}\mathbf{p}_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}_k(\mathbf{0}, \Sigma_\epsilon), \quad \mathbf{p}_i = \psi^{-1}(\mathbf{z}_i), \quad \mathbf{z}_i \sim \mathcal{N}_{m-1}(\boldsymbol{\mu}_p, \Sigma_p). \quad (7)$$

5.3 Problem formulation

The parameters $(\mathbb{H}, \boldsymbol{\mu}_p, \Sigma_p, \Sigma_\epsilon)$ are estimated by maximising the marginal likelihood of the observed data $\{g_i\}_{i=1}^n$, obtained by integrating out the latent proportion vectors $\{\mathbf{p}_i\}$. Since this integral is intractable, we resort to the MCEM algorithm: in the E-step, the current conditional expectation of the complete-data log-likelihood is approximated via Monte Carlo; in the M-step, this approximation is maximised with respect to the parameters. Observing only $\{g_i\}_{i=1}^n \subset B^2(I)$ – represented by $\{\mathbf{c}_i\}_{i=1}^n \subset \mathbb{R}^{n \times k}$ –, independent realizations of G , we seek to estimate via maximum likelihood the vertices $h_1, \dots, h_m \in B^2(I)$ through the coefficients' matrix $\mathbb{H} \in \mathbb{R}^{k \times m}$, the residuals' covariance operator $\Gamma_\epsilon \in \mathbb{R}^{k \times k}$ through Σ_ϵ , the mean ilr -proportion $\boldsymbol{\mu}_p \in \mathbb{R}^{(m-1)}$ and its covariance $\Sigma_p \in \mathbb{R}^{(m-1) \times (m-1)}$.

Throughout this section, $f_{\mathbf{p}_i}$ denotes the model density of $\mathbf{p}_i$ induced by the ilr -normal assumption, and $f_{\mathbf{p}_i|\mathbf{c}_i}$ its conditional density given the observed coefficient vector $\mathbf{c}_i$. For given $\mathbf{p}_i$, $i \in \{1, \dots, n\}$, the log-likelihood would be:

$$l(\mathbb{H}, \Sigma_\epsilon | \mathbf{p}_1, \dots, \mathbf{p}_n) = \sum_{i=1}^n l_i(\mathbb{H}, \mathbf{p}_i, \Sigma_\epsilon) = -\frac{1}{2} \sum_{i=1}^n (\mathbf{c}_i - \mathbb{H}\mathbf{p}_i)' \Sigma_\epsilon^{-1} (\mathbf{c}_i - \mathbb{H}\mathbf{p}_i) - \frac{n}{2} \log(\det(\Sigma_\epsilon)) + \text{const}. \quad (8)$$

However, in our scenario the proportions $\mathbf{p}_i$ are latent variables. Following the *Monte Carlo Expectation–Maximization (MCEM)* framework, in each M-step we maximise the expected complete-data log-likelihood (the Q -function from the E-step) with respect to the parameters, namely:

$$l(\mathbb{H}, \boldsymbol{\mu}_p, \Sigma_p, \Sigma_\epsilon) = \sum_{i=1}^n \mathbb{E}_{\mathbf{p}_i|\mathbf{c}_i} [l_i(\mathbb{H}, \mathbf{p}_i, \Sigma_\epsilon) + \log f_{\mathbf{p}_i|\mathbf{c}_i}(\mathbf{p}_i)], \quad (9)$$

where the expectation is based on the conditional distribution of $\mathbf{p}_i|\mathbf{c}_i$ with the current values of the parameters. In each M-step, we aim to find the set of parameters $(\mathbb{H}, \boldsymbol{\mu}_p, \Sigma_p, \Sigma_\epsilon)$ which maximizes the expected log-likelihood in Equation (9).

The conditional expectation in (9) is not analytically available. The $\mathbf{p}_i$ are obtained through a nonlinear transformation ψ of a Gaussian random variable, but are not Gaussian themselves. As a consequence, the model density and the likelihood of $\mathbf{p}_i$ are not conjugate, and the resulting conditional distribution is not Gaussian. Nevertheless, expectations with respect to $\mathbf{p}_i$ can be approximated via Monte Carlo simulation. Specifically, at each iteration t the current parameter estimates $(\mathbb{H}^{(t)}, \boldsymbol{\mu}_p^{(t)}, \Sigma_p^{(t)}, \Sigma_\epsilon^{(t)})$ are held fixed; the latent variables $\mathbf{p}_i$ are then sampled from the conditional $f_{\mathbf{p}_i|\mathbf{c}_i}$ (E-step), the expectation in (9) is approximated by Monte Carlo, and the parameters are updated by maximising this approximation (M-step). MCEM is well established in the literature [6] and has already been employed in the Bayes Hilbert space setting [35].

The conditional distribution $f_{\mathbf{p}_i|\mathbf{c}_i}$ of $\mathbf{p}_i$ is such that:

$$f_{\mathbf{p}_i|\mathbf{c}_i}(\mathbf{p}_i) \propto \exp \left(-\frac{1}{2} (\psi(\mathbf{p}_i) - \boldsymbol{\mu}_p)' \Sigma_p^{-1} (\psi(\mathbf{p}_i) - \boldsymbol{\mu}_p) - \frac{1}{2} (\mathbf{c}_i - \mathbb{H}\mathbf{p}_i)' \Sigma_\epsilon^{-1} (\mathbf{c}_i - \mathbb{H}\mathbf{p}_i) \right). \quad (10)$$

For the E-step, we can sample from the conditional distribution without explicitly computing the normalizing constant, using methods that range from importance sampling to more advanced algorithms such as Hamiltonian Monte Carlo (HMC). Importance sampling is easy to implement and interpret, e.g. [4], but poorly scalable as m increases. Moreover, expectations are approximated by weighted averages, with weights given by the ratio between the target conditional density and the auxiliary density. Conversely, HMC is designed to efficiently explore high-dimensional conditional distributions [27], and in this case expectations are computed as simple sample averages, since all samples have unit weight and are already distributed according to the target conditional distribution. For this reason, we present both options.

In the E-step, the expectation in (9) is approximated by drawing B samples $\{\mathbf{p}_{b,i}\}_{b=1}^B$ from an auxiliary distribution π and computing importance-weighted averages. Here $\mathbf{p}_{b,i}$ denotes the b -th sample drawn from π for the i -th observation, with weight $w_{b,i} = \frac{\eta_{b,i}}{\sum_b \eta_{b,i}}$ and $\eta_{b,i} = \frac{f_{\mathbf{p}_i|\mathbf{c}_i}(\mathbf{p}_{b,i})}{\pi(\mathbf{p}_{b,i})}$. For HMC, $\pi = f_{\mathbf{p}_i|\mathbf{c}_i}$ and $w_{b,i} = 1/B$ for all b .

Notice that, together with the maximization of the expected log-likelihood with respect to the model parameters, we are also implicitly maximizing the expected log-density of $\mathbf{p}_i$. This term alone does not control the size of the induced convex hull $\mathcal{M}(\{h_j\}_{j=1}^m)$, since $\text{tr}(\Sigma_p)$ may become arbitrarily small. In such a case, the proportions concentrate around the mean $\boldsymbol{\mu}_p$, while the locations of the vertices remain essentially unconstrained. To address this issue, and to improve the identifiability of the vertices – see Section 4 –, we introduce two additional penalization terms: one controlling the size of the convex hull by penalising small dispersion, and one regularising the shape of the estimated vertices when exact pure observations are not available. The resulting objective function to be *maximised* is:

$$\hat{l}(\mathbb{H}, \boldsymbol{\mu}_p, \Sigma_p, \Sigma_\epsilon) = \sum_{b=1}^B \sum_{i=1}^n w_{b,i} (l(\mathbb{H}, \mathbf{p}_{b,i}, \Sigma_\epsilon) + \log f_{\mathbf{p}_i|\mathbf{c}_i}(\mathbf{p}_{b,i})) - \lambda_{tr} \text{tr}(\Sigma_p^{-1}) - \lambda_{sm} \sum_j \|D^2[\text{clr}(h_j)]\|_{L^2(I)}^2, \quad (11)$$

where D^2 indicates the second derivative operator, and choice of λ_{tr} and λ_{sm} is discussed in Section 5.5.

5.4 Implementation

In this section we describe the practical implementation of the E- and M-steps introduced in Section 5.3, showing that each parameter block admits a closed-form update within the coordinate-wise maximisation of the marginal likelihood. The resulting updates are reported below and form the building blocks of the iterative estimation procedure, as described in Algorithm 1.

Update of the vertices $\mathbb{H}$

Fixing $\Sigma_\epsilon, \boldsymbol{\mu}_p, \Sigma_p$, the maximization of $\hat{l}$ has a closed form solution $\mathbb{H}^*$, that is:

$$\text{vec}(\mathbb{H}^*) = (A_r' \otimes I_{k \times k} + I_{m \times m} \otimes A_l)^{-1} \text{vec}(C) \quad (12)$$

where $\otimes$ denotes the Kroeneker product, $I_{k \times k}$ and $I_{m \times m}$ are, respectively, the identity matrix of dimension $k \times k$ and $m \times m$, $\text{vec}(\cdot)$ is the operation that vectorises the matrices by column, and:

$$\begin{cases} A_l = (2\lambda_{sm} \cdot \Sigma_\epsilon \cdot D_2) \\ A_r = \sum_{b=1}^B \sum_{i=1}^n w_{b,i} \cdot (\mathbf{p}_{b,i} \mathbf{p}_{b,i}') \\ C = \sum_{b=1}^B \sum_{i=1}^n w_{b,i} \cdot \mathbf{c}_i \mathbf{p}_{b,i}' \end{cases} \quad (13)$$

where $D_2 \in \mathbb{R}^{k \times k}$ is the matrix representation of the second-derivative operator in the SFPC basis, i.e., $(D_2)_{kl} = \langle \phi_k'', \phi_l'' \rangle_{L^2(I)}$. When $\lambda_{sm} = 0$, the maximiser of $\hat{l}$ becomes $\mathbb{H}^* = C A_r^{-1}$. $\mathbb{H}^* = C A_r^{-1}$ can be interpreted as a weighted mean of the observations (encoded in C) corrected by A_r^{-1} , which adjusts for vertices' overlap. The diagonal entries of A_r reflect average squared conditional proportions per vertex, while the off-diagonal entries capture their joint co-occurrence. Applying A_r^{-1} thus mitigates vertex overlap. The derivation of (12) is provided in C.

Update of the noise covariance Σ_ϵ

Conditional on $(\mathbb{H}, \boldsymbol{\mu}_p, \Sigma_p)$, the maximization of $\hat{l}$ with respect to Σ_ϵ reduces to the estimation of a weighted residual covariance matrix. The optimiser is given by

$$\Sigma_\epsilon^* = \frac{1}{\sum_b \sum_i w_{b,i}} \sum_{b=1}^B \sum_{i=1}^n w_{b,i} (\mathbf{c}_i - \mathbb{H} \mathbf{p}_{b,i}) (\mathbf{c}_i - \mathbb{H} \mathbf{p}_{b,i})' \quad (14)$$

that is, the conditional-weighted covariance of the reconstruction errors. Hence, Σ_ϵ captures the dispersion of the data around the current mixture representation.

Update of the proportion's parameters

Fixing $(\mathbb{H}, \Sigma_\epsilon, \Sigma_p)$, the update of $\boldsymbol{\mu}_p$ corresponds to a weighted average in the transformed coordinate system. Specifically,

$$\boldsymbol{\mu}_p^* = \frac{1}{\sum_b \sum_i w_{b,i}} \sum_b \sum_i w_{b,i} \boldsymbol{\psi}(\mathbf{p}_{b,i}), \quad (15)$$

that is, the conditional mean of the transformed proportions. This expression highlights that the Gaussian assumption is enforced in the Euclidean space induced by ψ .

Finally, fixing $(\mathbb{H}, \Sigma_\epsilon, \mu_p)$, the maximization with respect to Σ_p yields a regularised weighted covariance estimator:

$$\Sigma_p^* = \frac{1}{\sum_b \sum_i w_{b,i}} \left(\sum_b \sum_i w_{b,i} (\psi(p_{b,i}) - \mu_p)(\psi(p_{b,i}) - \mu_p)' + \lambda_{tr} I_{(m-1) \times (m-1)} \right), \quad (16)$$

where the trace-penalization term $\lambda_{tr} I$ ensures numerical stability and prevents degeneracy of the covariance matrix.

The MCEM algorithm [6], with a coordinate-wise optimization in the M-step whose steps all have a closed-form solution (as shown in Equations (12), (14), (15), (16)) is reported below. The algorithm proceeds as follows. At each iteration `it`, a simple random sample (SRS) – i.e., drawn uniformly without replacement – of $n_s = \lceil m\alpha_r \rceil$ observations is selected, where $\alpha_r \in (0, 1]$ is the batch fraction. SRS ensures that each observation has equal probability of being selected, avoiding systematic bias in the parameter updates. Setting $\alpha_r = 1$ recovers the full-batch version of the algorithm. For each parameter block, conditional samples $\{p_{b,i}\}$ are generated and the corresponding closed-form update is applied in sequence: first Σ_ϵ , then $\mathbb{H}$, then μ_p , and finally Σ_p . Convergence is assessed via the relative change in $\mathbb{H}$; the algorithm stops when this falls below a tolerance threshold `tol`. The algorithm supports a *stochastic-batch option*: rather than using all observations at each update step, a fresh SRS batch is drawn before each parameter update (lines 4, 6, 8, 10). This introduces additional stochasticity that can help escape local optima and reduces the computational cost per iteration, at the price of a noisier gradient signal. An alternative is to draw a single batch per full iteration. The tuning parameters λ_{tr} , λ_{sm} , and m are discussed in Section 5.5. The batch fraction α_r should be set according to the redundancy of the dataset: when observations are spatially or otherwise correlated, α_r can be reduced substantially without loss of information – in the case study of Section 7 we set $\alpha_r = 0.2$ – whereas for non-redundant datasets $\alpha_r = 1$ is recommended. For the convergence criterion, we set `tol` = 0.05 (i.e., a 5% relative change threshold on $\mathbb{H}$) and `max_it` = 50, which proved sufficient for convergence in all simulation and case study settings considered in this work. In absence of prior knowledge, we recommend to choose as initial values $\mu_p^{(0)} = 0$, $\Sigma_p^{(0)} = I_{(m-1) \times (m-1)}$. The initialisation of $\mathbb{H}^{(0)}$ and $\Sigma_\epsilon^{(0)}$ ensures that the algorithm starts aligned with the directions of maximal variation in the data: the initial vertices coincide with the first m simplicial principal components, while $\Sigma_\epsilon^{(0)}$ is calibrated to the variance explained by the first discarded component, as measured by the $(m+1)$ -th eigenvalue.

Algorithm 1 Algorithm for MCEM estimation – with stochastic-batch option

Require: $\{\mathbf{c}_i\}_{i=1}^n$, $\Sigma_\epsilon^{(0)}$, $\mathbb{H}^{(0)}$, $\mu_p^{(0)}$, $\Sigma_p^{(0)}$, D_2 , λ_{tr} , λ_{sm} , α_r , `max_it`, `tol`

Output: $(\mu_p, \Sigma_p, \mathbb{H}, \Sigma_\epsilon)$

```

1:  $m \leftarrow \text{ncol}(\mathbb{H}^{(0)})$ 
2:  $n_s \leftarrow \lceil m\alpha_r \rceil$ 
3: for it = 1 to max_it do
4:   SRS batch  $\{\mathbf{c}_{i_s}\}_{i_s=1}^{n_s}$ 
5:    $\Sigma_\epsilon^{(\text{it})} \leftarrow \text{update}_{\Sigma_\epsilon}(\cdot)$  (Eq. (14))
6:   SRS batch  $\{\mathbf{c}_{i_s}\}_{i_s=1}^{n_s}$ 
7:    $\mathbb{H}^{(\text{it})} \leftarrow \text{update}_{\mathbb{H}}(\cdot)$  (Eq. (12))
8:   SRS batch  $\{\mathbf{c}_{i_s}\}_{i_s=1}^{n_s}$ 
9:    $\mu_p^{(\text{it})} \leftarrow \text{update}_{\mu_p}(\cdot)$  (Eq. (15))
10:  SRS batch  $\{\mathbf{c}_{i_s}\}_{i_s=1}^{n_s}$ 
11:   $\Sigma_p^{(\text{it})} \leftarrow \text{update}_{\Sigma_p}(\cdot)$  (Eq. (16))
12:   $\Delta_H \leftarrow \frac{\|\mathbb{H}^{(\text{it})} - \mathbb{H}^{(\text{it}-1)}\|_F}{\|\mathbb{H}^{(\text{it}-1)}\|_F + 10^{-8}}$ 
13:  if  $\Delta_H < \text{tol}$  then
14:    break
15:  end if
16: end for
17: return  $\{\mu_p^{(\text{it})}, \Sigma_p^{(\text{it})}, \mathbb{H}^{(\text{it})}, \Sigma_\epsilon^{(\text{it})}\}$ 
```

5.5 Choice of the tuning parameters

In this section we discuss the choice of the penalty parameters λ_{sm} , λ_{tr} , and the number of vertices m .

Penalty parameters

The parameter λ_{sm} regulates the strength of the smoothing penalization. If the data are already smooth (e.g. in our simulations in Section 6) or if they have to keep their natural irregularity (e.g. in the case study in

Section 7), λ_{sm} can be set equal to 0. In general, $\lambda_{sm} > 0$ can be selected on the basis of the reconstruction error of the mean density process, i.e.:

$$\text{MSE}_{CV}(\lambda_{sm}) = \frac{1}{|V|} \sum_{i \in V} \|\mathbf{c}_i - \mathbb{H}\mathbf{p}_i\|_2^2$$

where $V \subset \{1, \dots, n\}$ is a held-out validation set disjoint from the training set, and $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^k$. For real case studies, whose statistical analyses are commonly supervised by experts on the scientific subject of the data, one can also choose λ_{sm} based on the interpretability of the results.

Unlike λ_{sm} , which enforces smoothness at the expense of fit to the data, the penalization of $\text{tr}(\Sigma_p)^{-1}$ can be interpreted as an effective increase, by a factor λ_{tr} , in the variance of each coefficient of $\psi(\mathbf{p})$ (see Equation (16)). Empirically, the algorithm typically exhibits convergence in the $\|\cdot\|_2$ norm of the coefficients $\mathbb{H}$ to the same result, across a wide range of values of λ_{tr} , indicating a limited sensitivity of the solution to this regularization parameter when the distribution of proportions saturates near the vertices – i.e. when 3P holds.

Number of vertices m

The number of mixture components m is treated as a model hyperparameter; its selection is validated empirically in C for the simulation settings considered in Section 6. A classical approach consists in setting m equal to the number of principal components selected via the elbow rule, or via a threshold on the cumulative sum of eigenvalues. Alternatively, one can fit the model for $m \in \{2, \dots, m_{\max}\}$ and select m via reconstruction error on a hold-out validation set (analogous to MSE_{CV} in Section 5.3), an information criterion, or domain knowledge. The sensitivity of the results to m should always be assessed by comparing solutions for adjacent values.

6 A simulation study: probability density unmixing

This section investigates the behavior of the proposed method under two relevant *stress* scenarios, with increasing degree of difficulty. Each simulated dataset consists of $n = 500$ mixtures of m density components (vertices) defined on the unit interval. Each component approximates a Beta distribution with varying shape parameters. Beta distributions are a well-known and flexible parametric family defined on the unit interval, making them a natural choice for simulating heterogeneous density shapes on $[0, 1]$. However, as shown in Section A, exact Beta distributions are linearly dependent in the Bayes Hilbert space, which would violate the assumptions of Section 2. To circumvent this issue, each component is defined as a sample density obtained from a histogram generated from the corresponding Beta distribution, yielding vertices that are a.s. linearly independent. The densities are then transformed using the centered log-ratio (clr) transformation, which maps them from the Bayes space to an unconstrained Hilbert space. The resulting clr functions are represented through cubic compositional B-spline basis expansions with a second-derivative roughness penalty, as in [21], yielding a finite-dimensional coefficient representation.

Mixture observations are generated by simulating compositional weights in isometric log-ratio (ilr) coordinates from the multivariate Gaussian distribution

$$\mathcal{N}_{m-1}(\mu_p, \Sigma_p),$$

and subsequently transforming them back to the simplex. Functional mixtures are then obtained as linear combinations of the vertices in clr space. To mimic realistic data conditions, additive multivariate Gaussian noise,

$$\varepsilon \sim \mathcal{N}_k(\mathbf{0}, \Sigma_\varepsilon),$$

is introduced in the coefficient space.

This framework generates synthetic density-valued functional observations with a known underlying structure, enabling a systematic evaluation of estimation accuracy, robustness to noise, and recovery of both mixture components and mixing proportions.

The most challenging scenarios are expected to arise when: (i) the residual variance increases, i.e., $\text{tr}(\Sigma_\varepsilon)$ becomes larger (Study A); (ii) the mean of the mixing distribution μ_p moves further away from the uniform composition $[0]_{S^m}$ towards a corner of the simplex (Study B).

Datasets are simulated under the following configuration. The number of mixture components is set to $m = 3$. The noise covariance matrix is specified as $\Sigma_\varepsilon = \sigma^2 I_{k \times k}$, where $\sigma^2 = 0.01 \cdot \text{tr}(\widehat{\Sigma}_c)$, with $\widehat{\Sigma}_c$ the sample covariance of the clr-coefficient vectors $\{\mathbf{c}_i\}_{i=1}^n$, unless otherwise stated (see Study A). The mean of the mixture coefficients is set to $\mu_p = \mathbf{0}_{B^2}$ in Study A, while it varies in Study B, in such a way to study its effect on the estimation.

The estimation phase is carried out according to Algorithm 1. To ensure comparability across scenarios, the following hyperparameters are kept constant in every simulation. The model distribution is initialized as follows: the mean vector is set to

$$\mu_p^{(0)} = \mathbf{0},$$

the covariance matrix of the mixture parameters to

$$\Sigma_p^{(0)} = I_{(m-1) \times (m-1)},$$

and the noise covariance matrix is initialized as

$$\Sigma_\varepsilon^{(0)} = I_{k \times k}.$$

The matrix $\mathbb{H}^{(0)}$ is initialized such that its columns correspond to the first m simplicial principal components of $\mathbf{c}_1, \dots, \mathbf{c}_n$ with m set to 3. The regularization parameters are fixed at

$$(\lambda_{sm}, \lambda_{tr}) = (0, 1),$$

and the learning rate is set to $\alpha_r = 20\%$, meaning that each update is computed using 100 randomly sampled observations.

We claim that, apart from $m = 3$, the selected hyperparameters are intentionally not optimised for each individual realization of the synthetic datasets. This choice is motivated by the need to ensure comparability across the different scenarios and, at the same time, to avoid relying on a fully automatic tuning procedure, that would be computationally expensive in an extensive simulation study. The selection of $m = 3$ should not be regarded as an artificial advantage derived from the simulation design. Although $m = 3$ corresponds to the true underlying dimensionality used to generate the data, this choice is also strongly supported empirically (see C).

6.1 Study A: amplifying the dispersion from the convex hull

In Study A, we investigate how the dispersion of the residuals affects the performance of our maximum-likelihood unmixing strategy. As $\text{tr}(\Sigma_\varepsilon)$ increases, the proportion of variance explained by the mixture correspondingly decreases. This makes the identification of the mixture more challenging, as the mixture structure becomes increasingly obscured by the overall variability of the data.

Figure 2 (top row) illustrates a representative realization of Study A: the discrepancy between the true and estimated vertices is negligible when σ^2 equals 0.01, corresponding to observations that are almost noiseless mixtures. However, the mismatch becomes more severe when σ^2 reaches 0.5, that is, when the variability of the residual is comparable to half the variability of the vertices themselves. In this regime, the noise makes the estimated vertices also noisy. In situations like these the solutions can be:

- Avoiding subsampling at each iteration, setting $\alpha_r = 1$ – see Algorithm 1.
- Increase the smoothing penalization parameter λ_{sm} .

Figure 3 (left) reports the L^2 error distribution across 50 independent repetitions. The representative realization in Figure 2 (top row) is selected as the one closest to the median error and should be interpreted as a typical outcome. Two features emerge: the median error remains roughly stable across values of σ^2 , suggesting that the algorithm consistently identifies the correct vertex directions even under substantial noise; the interquartile range widens markedly as σ^2 increases, reflecting growing variability across repetitions. Two conditions are compared: in both cases, BH unmixing is applied, but the data are generated either from a Bayes Hilbert mixture (light blue boxplots) or from a linear mixture (violet boxplots). The latter case corresponds to a misspecified model, where the unmixing assumption does not match the data-generating process. As expected under misspecification, the linear mixture case yields systematically higher errors, confirming that the BH unmixing estimator is sensitive to the assumed mixing geometry.

6.2 Study B: moving the mean proportion towards one vertex

In Study B, we examine how the proximity to one particular vertex, and consequently the distance of μ_p from the uniform proportion $[0]_{S^m}$, influences the performance of our maximum-likelihood unmixing strategy. As μ_p 's norm increases, at least one of the vertices becomes underrepresented in the dataset of the mixtures. This aspect can make the recognition of that vertex difficult. In particular, as the proportions relative to a vertex are distributed near 0, the influence of that vertex becomes almost null. In that specific case, the dataset will

be approximately a mixture of $(m - 1)$ vertices. For this simulation, we consider the following μ_p – recalling that $m = 3$:

$$(0, 0), (0, 1/2), (0, 1), (0, 2)$$

Figure 11 illustrates the four choices of μ_p , represented as points in $S^3 \subset \mathbb{R}^3$. As $\|\mu_p\|$ increases, the mean proportion moves away from the uniform composition towards one vertex. It is important to note that for all these simulations, the starting value for μ_p in the iterative procedure is always equal to $[0]_{S^m}$. As one can easily observe in the example in Figure 2 (bottom row), qualitatively, the estimated vertices progressively get worse from $[0]_{S^m}$ to $\psi^{-1}((0, 1)) \approx (0.18, 0.18, 0.64) \in S^3$. When discrepancy between the proportion of one vertex is too high with respect to the ones of the other vertices, e.g. in our example for $\psi^{-1}((0, 2)) \approx (0.07, 0.07, 0.86)$, the two less represented vertices collapse into one unique vertex, becoming indistinguishable.

Figure 3 (right) summarises estimation accuracy across 50 independent repetitions; the representative realization in Figure 2 (bottom row) corresponds to the median error. Performance remains stable for $\mu_p \in \{(0, 0), (0, 1/2), (0, 1)\}$ and deteriorates sharply only at $\mu_p = (0, 2)$, where one vertex carries proportion approximately 0.86 and the remaining two become nearly indistinguishable. This threshold behaviour suggests robustness of the estimation procedure to moderate imbalances in the mixing distribution, with breakdown occurring only when one vertex is so dominant that the others are effectively unobserved. As in Study A, BH unmixing is applied to both conditions: data generated from a BH mixture (light blue) and data generated from a linear mixture (violet). The BH mixture case consistently yields lower errors throughout, confirming that the advantage of the correct model specification results also in this case – regardless of $\|\mu_p\|$.

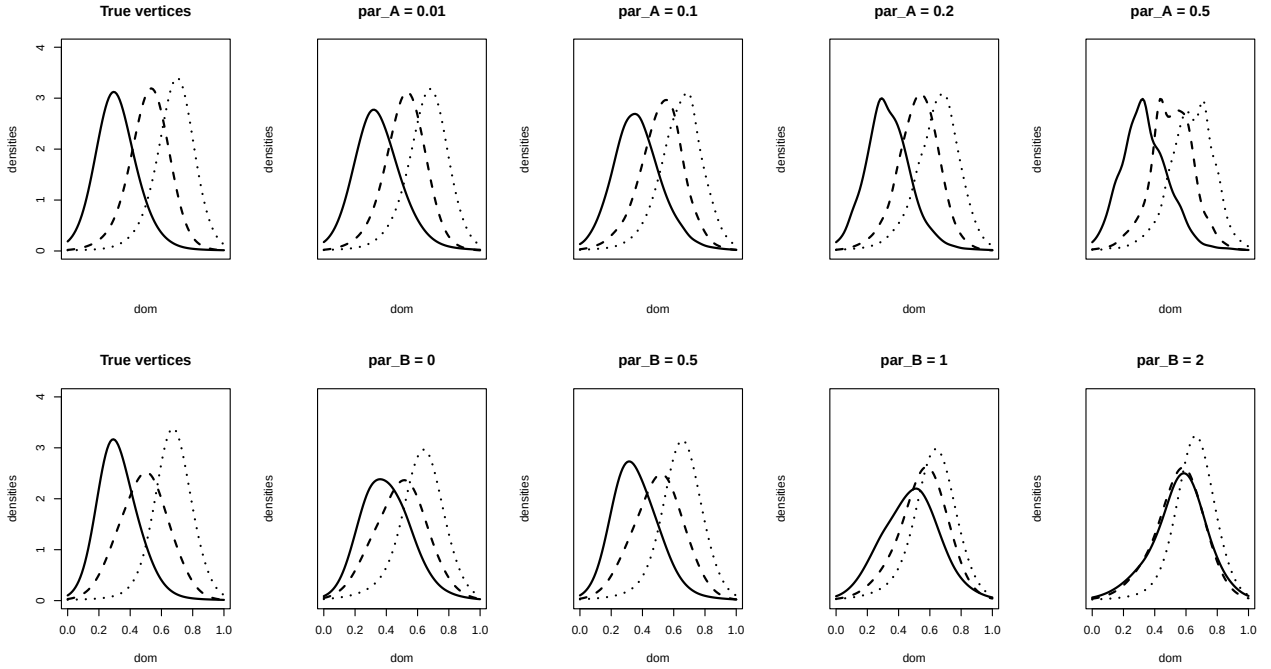


Figure 2: Representative realizations of Study A (top row) and Study B (bottom row): true vertices (leftmost panel) and estimated vertices for increasing values of $\text{par_a} = \sigma^2$ (Study A) and $\text{par_b} = \|\mu_p\|$ (Study B). Each realization is selected as the one closest to the median L^2 error over the 50 repetitions.

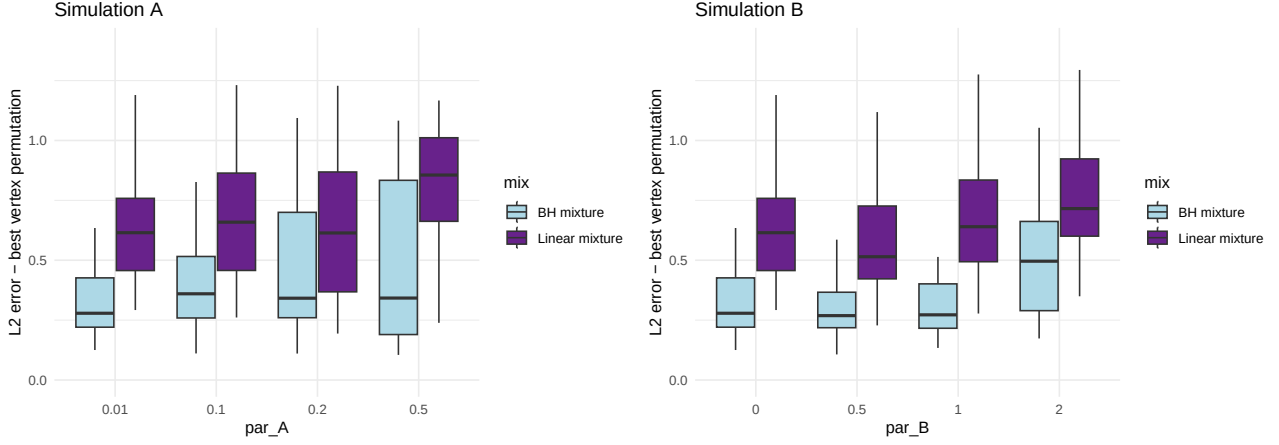


Figure 3: Distribution of the L^2 estimation error – defined as $\frac{1}{m} \min_{\pi} \sum_j \|h_j - \hat{h}_{\pi(j)}\|_{B^2}$, minimised over all vertex permutations π – across 50 independent repetitions, as a function of σ^2 (Study A, left) and $\|\mu_p\|$ (Study B, right). Light blue: data generated as Bayes Hilbert mixtures; violet: data generated as linear mixtures.

7 A case study: AVIRIS Indian Pines hyperspectral unmixing

The AVIRIS Indian Pines dataset is a widely used benchmark in hyperspectral image analysis [41], providing a challenging setting for the evaluation of spectral analyses: from pixel classification [28] to unsupervised unmixing [26]. Acquired by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) in 1992 over agricultural areas in northwestern Indiana, the scene consists of 145×145 pixels with reflectance measurements across 224 spectral bands spanning the wavelength range from approximately $0.5\mu\text{m}$ to $2.5\mu\text{m}$ [3]. After removing bands affected by atmospheric absorption and noise, the dataset exhibits high spectral redundancy, subtle class variability, and significant spectral mixing due to limited spatial resolution and heterogeneous land cover. Moreover, the Indian Pines site is predominantly characterised by the coexistence of vegetation, water, and soil, with only a limited presence of artificial structures such as streets and buildings – see Figure 16. As a consequence, each pixel is likely to simultaneously represent multiple land-cover types. The combined effect of low spatial resolution – resulting in spatial blurring – and the intrinsic heterogeneity of the scene makes Indian Pines particularly suitable for testing advanced unmixing models that go beyond linear and pointwise representations of spectra. In this setting, Bayes Hilbert mixtures are a natural modelling choice: rather than describing each pixel as a discrete combination of pure endmembers, they yield a single coherent distribution that integrates the spectral characteristics of all contributing land-cover types, consistently with the fuzzy nature of mixed pixels.

In this work, we interpret each hyperspectral pixel not as a finite-dimensional vector but as a functional compositional object, viewed as a distribution over the continuous wavelength domain. This perspective naturally leads to an embedding of hyperspectral data into the Bayes Hilbert space, despite it being originally designed for analysing probability density functions. This approach enables us to focus on the *shape* of the spectra, rather than on their amplitude. Indeed, it is widely recognised in the literature that amplitude has no physical meaning in hyperspectral analysis.

Within this functional-compositional framework, spectral unmixing is reinterpreted as a density unmixing problem. We model each observed normalised hyperspectrum as a BH mixture – as mixing is thought to occur within pixels, i.e. fuzzy-like, as discussed in Section 3 – of latent *endmember* densities (vertices) defined over the wavelength domain. Figure 4 on the left shows the dataset of normalised hyperspectra. Applying the simplicial FPCA [12], we select $m = 4$. We set the starting values for $\{h_j\}_{j=1}^m$ as the first m principal components, and for Σ_{ϵ} as $\rho_{m+1} I_{k \times k}$, where ρ_{m+1} is the $(m+1)$ -th eigenvalue, i.e. the variance explained by the first discarded principal component. The first 4 principal components, obtained without imposing smoothness so as to preserve the roughness typical of hyperspectral data, and the variance explained are reported in Figure 14 in G.

The choice of $m = 4$ is guided by Figure 14 (center), i.e. the plot of variance explained varying the number of retained principal components. This plot displays an elbow in $m = 4$ – retaining 96% of the total variance.

Selecting $m = 4$, we discuss the result of our MCEM unmixing, employed to estimate the tuple of vertices $\{h_1, h_2, h_3, h_4\}$, the mean ilr-proportion μ_p and covariance Σ_p , and residual coefficients' covariance Σ_{ϵ} . In Figure 4 we show the data (left) compared to the estimated vertices (right). The vertices seem reasonable convex generators of the data. In order to interpret the results, we plot, for each pixel of the Indian Pines site, the *maximum a posteriori* estimate of each component of the proportion p_j , for every $j \in \{1, \dots, m\}$, obtaining m maps – see Figure 5.

Comparing the results obtained with historical imagery from 1993's Google Earth, with the ten years later

map – Figure 16 – and keeping as a reference the experts’ manually labeled landcover map – Figure 18 – we can state that the proposed spectral unmixing strategy is capable of accurately distinguishing the following land-cover classes: h_1 – cultivated lands of type 1, h_2 – cultivated lands of type 2, h_3 – impervious surfaces, including streets and building roofs, h_4 – trees, and By considering the sum of the proportions of h_1 and h_2 , we can easily recognise the profile of the cultivated fields in 1993. The difference between h_1 and h_2 is unknown a priori. However, we can observe that h_1 has a stronger decay in the band $[0.55\mu m, 1\mu m]$, classified as the *Visible to Near Infrared* (NVIR). In this band, water reacting with chlorophyll – the green pigment, present in all green plants and in cyanobacteria, responsible for the absorption of light to provide energy for photosynthesis – typically present a higher reflectance [24]. Indeed, this band is observed by scientists in order to assess if a plant is healthy – high reflectance – or not. Thus, we can suggest that h_1 represents fields that are less healthy than the ones with a high proportion of h_2 . Analogously, the difference could also be due to a higher presence of water in h_2 .

By comparison, we report in G (Figure 15) the result of Vertex Component Analysis (VCA, [25]) on AVIRIS Indian Pines, obtained on the normalized hyperspectra, trained on 25% of the pixels – to reduce redundancy induced by spatial correlation –, fixing $m = 4$ to make the result comparable to ours. VCA is a statistical method for *linear* unmixing that requires a pure-pixel assumption. It extracts the endmembers $(\nu_j)_{j=1}^m$ among the statistical units, and computes the pixels’ abundances vectors $\mathbf{a} = (a_j)_{j=1}^m$, which, in contrast to proportions, are only constrained to be positive, not to sum to 1. Consequently, two endmembers can have a high abundance on the same pixel. Employing as a reference Figures 17 and 16 shows that ν_3 highlights the presence of trees, similarly to h_4 , ν_2 the grass-pastures (see Figure 18), while ν_1, ν_4 , though spatially coherent, are of difficult interpretability both in terms of manually labeled map and of Google Earth’s photographs (Figure 16).

The interpretable findings and the coherence of our results highlight the potential of functional-compositional representations for hyperspectral unmixing, particularly in scenarios where spectral variability and spatial blurring challenge traditional linear models. Indeed, as illustrated in Section 3, the Bayes Hilbert mixture naturally models spatially blurred observations, where distinct endmembers are smoothly integrated within a single pixel.

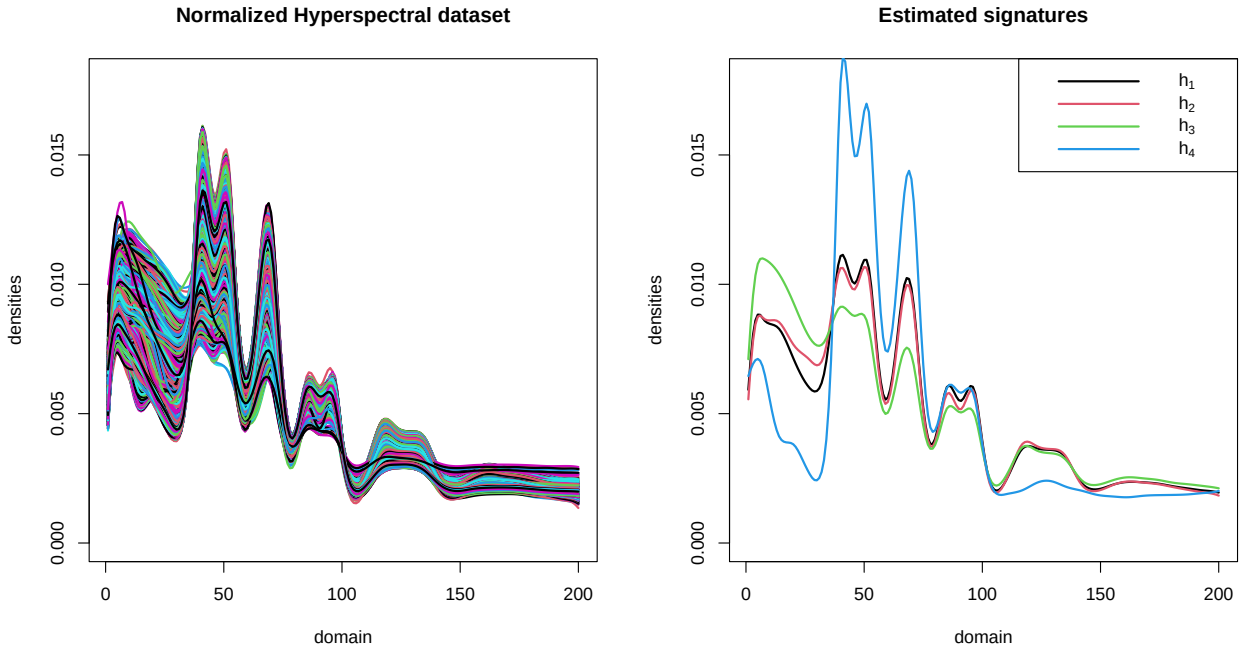


Figure 4: On the left the normalised hyperspectral mixtures, to be unmixed. On the right the $m = 4$ estimated vertices.

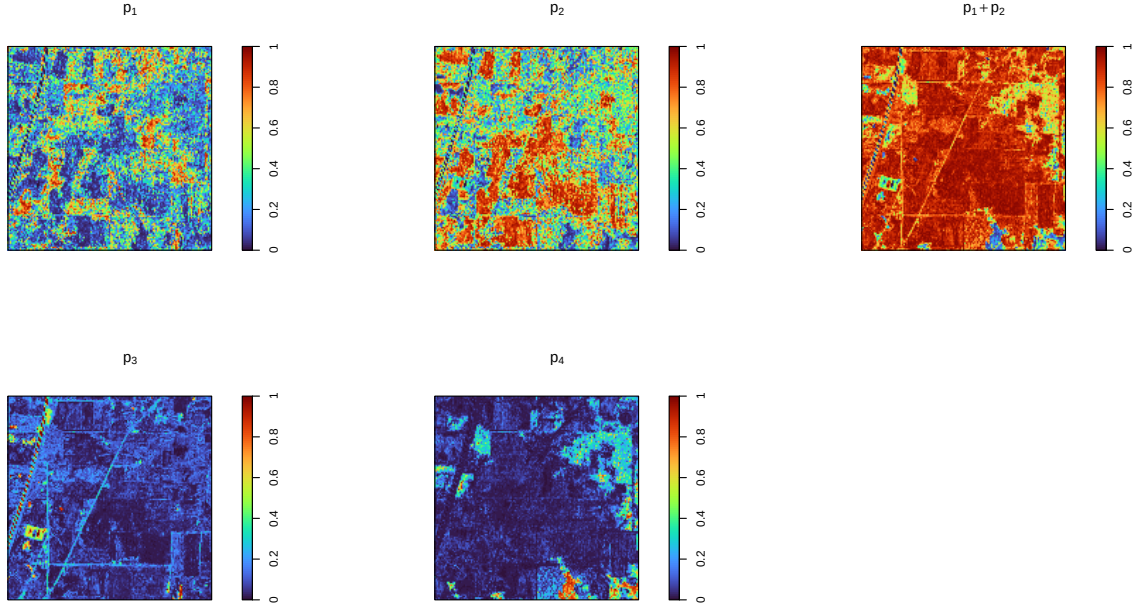


Figure 5: Unmixing in Bayes Hilbert space. The six panels show the estimated proportions $p_1, \dots, p_4 \in (0, 1)$ as well as $p_1 + p_2$.

Conclusion

In this work, we introduced a framework for the analysis and unmixing of random density mixtures in the Bayes Hilbert space $B^2(I)$. We studied the identifiability of a statistically space-efficient representation of a random mixture, defined as the convex combination of vertices in $B^2(I)$, with vertex proportions following a distribution that maximizes the covered region of the simplex. General identifiability results for mixtures in Hilbert spaces were established and subsequently applied to the Bayes Hilbert space setting.

Based on these results, we proposed a penalized maximum likelihood approach for the unmixing of Bayes Hilbert mixtures aimed at recovering the statistically space-efficient representation. The resulting optimization problem can be implemented through a coordinate-wise maximization algorithm.

The proposed methodology was illustrated through a hyperspectral data application, where observations can be naturally embedded in the Bayes Hilbert space and analyzed in terms of distributional shape rather than amplitude. A complementary simulation study based on noisy mixtures of Beta distributions further demonstrated the interpretability and practical performance of the method, particularly in the presence of fuzzy-type mixtures.

The performance of the approach may deteriorate when observations lie far from the center of the convex hull of the vertices, are very noisy or when the mixture type is misspecified, for instance when the underlying mixture is linear. An interesting direction for future research is to relax the assumption of strictly Bayes-Hilbert or linear mixtures and investigate which α -mixture within the α -mixture family [7, 38] provides the most appropriate representation of a generic random mixture of densities. Concerning the model for the proportions, the ilr framework may become restrictive under the pure-pixel assumption, since logarithmic transformations are undefined for zero components; in such cases, the discrete α -transformations [7, 38] may provide a more suitable alternative, as they admit pure proportions and remain well-defined on the boundary of the simplex. In our model, we assume that the proportion vector $\mathbf{p}$ follows an ilr-normal distribution, that is, its isometric log-ratio (ilr) transformation is multivariate normally distributed. This choice yields a unimodal distribution for $\mathbf{p}$. When the observed realisations of G exhibit multimodality – suggesting the presence of distinct subpopulations – a more flexible model for the distribution of $\mathbf{p}$ may be considered [8].

Finally, one could consider modelling correlation –temporal, spatial, or spatio-temporal, depending on the application – in the proportions. Notably, in the hyperspectral case study of Section 7, spatially contiguous proportion maps are obtained even without explicitly accounting for spatial dependence (see Figure 5).

We believe that the Bayes Hilbert space setting provides a principled and interpretable foundation for fuzzy-like mixture modelling of distributional data, and that the extensions outlined above represent promising directions for broadening its scope and applicability.

Acknowledgments

GP and AM acknowledge the support provided by the European Commission under the “HORIZON-CL4-2021-DIGITALEMERGING-01 project BioProS - Biointelligent Production Sensor to Measure Viral Activity” (grant agreement no. 101070120), 2022-2026”. GP and AM acknowledge the initiative “Dipartimento di Eccellenza 2023–2027”, MUR, Italy, Dipartimento di Matematica, Politecnico di Milano. This research has received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project number 513634041. GP acknowledges Maarten Jung, Manuel Pfeuffer, Johannes M. Feeser and Dr. Georg Keilbar for the fruitful scientific exchanges on functional and density data analysis, and Alfredo Gimenez Zapiola for sharing his expertise on hyperspectral unmixing.

References

- [1] M. Arngren, M. N. Schmidt, and J. Larsen. Unmixing of hyperspectral images using bayesian nonnegative matrix factorization with volume prior. *Journal of Signal Processing Systems*, 65(3):479–496, 2011. doi:10.1007/s11265-010-0533-2.
- [2] M. Asadi, N. Ebrahimi, and E. S. Soofi. The alpha-mixture of survival functions. *Journal of Applied Probability*, 56(4):1151–1167, 2019. doi:10.1017/jpr.2019.72.
- [3] M. F. Baumgardner, L. L. Biehl, and D. A. Landgrebe. 220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3, Sep 2015. URL <https://purrr.purdue.edu/publications/1947/1>.
- [4] J. Beck, B. M. Dia, L. F. R. Espath, Q. Long, and R. Tempone. Fast bayesian experimental design: Laplace-based importance sampling for the expected information gain. *Computer Methods in Applied Mechanics and Engineering*, 334:523–553, 2018. ISSN 0045-7825. doi:10.1016/j.cma.2018.01.053.
- [5] P. Billingsley. *Probability and Measure*. Wiley, New York, NY, USA, 3rd edition, 1995. ISBN 978-0471007106.
- [6] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [7] L. Clarotto, D. Allard, and A. Menafoglio. A new class of α -transformations for the spatial analysis of compositional data. *Spatial Statistics*, 47:100570, 2022. doi:10.1016/j.spasta.2021.100570.
- [8] M. Comas-Cufí, J. A. Martín-Fernández, and G. Mateu-Figueras. Log-ratio methods in mixture models for compositional data sets. *SORT – Statistics and Operations Research Transactions*, 40(2):349–374, 2016. doi:10.2436/20.8080.02.47.
- [9] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003. doi:10.1023/A:1023818214614.
- [10] S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer New York, New York, NY, 1 edition, 2006.
- [11] K. Hron. Multivariate densities in bayes spaces: The novel concept of marginals and its implications. In G. Aneiros, E. G. Bongiorno, A. Goia, and M. Hušková, editors, *New Trends in Functional Statistics and Related Fields*, pages 241–248, Cham, 2025. Springer Nature Switzerland. doi:10.1007/978-3-031-92383-8_30.
- [12] K. Hron, A. Menafoglio, M. Templ, K. Hrušová, and P. Filzmoser. Simplicial principal component analysis for density functions in bayes spaces. *Computational Statistics & Data Analysis*, 94:330–350, 2016. doi:10.1016/j.csda.2015.07.007.
- [13] K. Hron, J. Machalová, and A. Menafoglio. Bivariate densities in bayes spaces: orthogonal decomposition and spline representation. *Stat Papers*, 64(5):1629–1667, 2023. doi:10.1007/s00362-022-01359-z.
- [14] P. Jašková, J. Palarea-Albaladejo, A. Gába, D. Dumuid, Ž. Pedišić, J. Pelclová, and K. Hron. Compositional functional regression and isotemporal substitution analysis: Methods and application in time-use epidemiology. *Statistical Methods in Medical Research*, 32(10):2064–2080, 2023. doi:10.1177/09622802231192949.
- [15] Q. Jiang, Y. Dong, J. Peng, M. Yan, and Y. Sun. Maximum likelihood estimation based nonnegative matrix factorization for hyperspectral unmixing. *Remote Sensing*, 13(13):2637, 2021. doi:10.3390/rs13132637.

- [16] Z. Ju and H. Liu. Fuzzy gaussian mixture models. *Pattern Recognition*, 45(3):1146–1158, 2012. ISSN 0031-3203. doi:10.1016/j.patcog.2011.08.028.
- [17] L. Jun and J. M. Bioucas-Dias. Minimum volume simplex analysis: A fast algorithm to unmix hyperspectral data. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 250–253, Boston, MA, USA, 2008. doi:10.1109/IGARSS.2008.4779330.
- [18] G. Keilbar and S. Greven. Counterfactual density effects and the German East–West income gap, 2026. URL <https://arxiv.org/abs/2603.28470>.
- [19] J. Machalová, R. Talská, K. Hron, and A. Gába. Compositional splines for representation of density functions. *Comput Stat*, 36(2):1031–1064, 2021. doi:10.1007/s00180-020-01042-7.
- [20] E. M. Maier, A. Fottner, S. Greven, and A. Stöcker. Additive density regression. 2025. URL <https://arxiv.org/abs/2510.14502>.
- [21] E. M. Maier, A. Stöcker, B. Fitzenberger, and S. Greven. Additive density-on-scalar regression in bayes hilbert spaces with an application to gender economics. *Ann. Appl. Stat.*, 19(1):680–700, 2025. doi:10.1214/24-AOAS1979.
- [22] S. Mazdeyasna, M. S. Arefin, A. Fales, S. J. Leavesley, T. J. Pfefer, and Q. Wang. Evaluating normalization methods for robust spectral performance assessments of hyperspectral imaging cameras. *Biosensors*, 15(1):20, 2025. doi:10.3390/bios15010020.
- [23] A. Menafoglio, P. Secchi, and A. Guadagnini. A class-kriging predictor for functional compositions with application to particle-size curves in heterogeneous aquifers. *Math Geosci*, 48(4):463–485, 2016. doi:10.1007/s11004-015-9625-7.
- [24] NASA. URL http://science.nasa.gov/ems/08_nearinfraredwaves. National Aeronautics and Space Administration, Science Mission Directorate. (2010). Reflected Near-Infrared Waves. Retrieved February 10, 2026, from NASA Science website:.
- [25] J. M. P. Nascimento and J. M. B. Dias. Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(4):898–910, 2005. doi:10.1109/TGRS.2005.844293.
- [26] J. M. P. Nascimento and J. M. B. Dias. Does independent component analysis play a role in unmixing hyperspectral data? *IEEE Transactions on Geoscience and Remote Sensing*, 43(1):175–187, 2005. doi:10.1109/TGRS.2004.839806.
- [27] R. M. Neal. Mcmc using hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 113–162. CRC Press, 2011. doi:10.1201/b10905-6.
- [28] B. L. N. Phaneendra Kumar, R. Vaddi, P. Manoharan, et al. A new band selection framework for hyperspectral remote sensing image classification. *Scientific Reports*, 14:31836, 2024. doi:10.1038/s41598-024-83118-8.
- [29] R. Rajabi and H. Ghassemian. Hyperspectral data unmixing using gnmf method and sparseness constraint. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1450–1453. IEEE, 2013. doi:10.1109/IGARSS.2013.6723058.
- [30] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, 2005.
- [31] O. Shojaee and R. Momeni. The α -mixture of cumulative distribution functions: Properties, applications to parallel system and stochastic comparisons. *Journal of the Indian Society for Probability and Statistics*, 24(2):599–621, 2023. doi:10.1007/s41096-023-00169-2.
- [32] O. Shojaee, M. Asadi, and M. Finkelstein. On some properties of α -mixtures. *Metrika*, 84(8):1213–1240, 2021. doi:10.1007/s00184-021-00818-1.
- [33] O. Shojaee, M. Asadi, and M. Finkelstein. Stochastic properties of generalized finite α -mixtures. *Probability in the Engineering and Informational Sciences*, 36(4):1055–1079, 2022. doi:10.1017/S0269964821000243.
- [34] D. A. Simovici and C. Djeraba. *Mathematical Tools for Data Mining*. Springer London, 1 edition, 2008.
- [35] L. Steyer and S. Greven. Principal component analysis in bayes spaces for sparsely sampled density functions, 2023. URL <https://arxiv.org/abs/2309.11352>.

- [36] R. Talská, A. Menafoglio, J. Machalová, K. Hron, and E. Fišerová. Compositional regression with functional response. *Computational Statistics & Data Analysis*, 123:66–85, 2018. doi:10.1016/j.csda.2018.01.018.
- [37] R. Talská, A. Menafoglio, K. Hron, J. J. Egozcue, and J. Palarea-Albaladejo. Weighting the domain of probability densities in functional data analysis. *Stat*, 9(1):e283, 2020. doi:10.1002/sta4.283.
- [38] M. Tsagris, S. Preston, and A. T. A. Wood. Improved classification for compositional data using the α -transformation. *Journal of Classification*, 33(2):243–261, 2016. doi:10.1007/s00357-016-9207-5.
- [39] K. G. van den Boogaart, J. J. Egozcue, and V. Pawlowsky-Glahn. Bayes hilbert spaces. *Australian & New Zealand Journal of Statistics*, 56(2):171–194, 2014. doi:10.1111/anzs.12074.
- [40] M. Vasile, L. Walker, A. Campbell, et al. Space object identification and classification from hyperspectral material analysis. *Scientific Reports*, 14:1570, 2024. doi:10.1038/s41598-024-51659-7.
- [41] A. G. Zapiola, A. Boselli, A. Menafoglio, and S. Vantini. Hyper-spectral unmixing algorithms for remote compositional surface mapping: a review of the state of the art. *Applied Computing and Geosciences*, 28: 100297, 2025. doi:10.1016/j.acags.2025.100297.

A BH mixtures in exponential families

As an illustrative example, we briefly discuss the special case of distributions belonging to the *exponential families*. We already know that the set of the exponential family of distributions is an affine subspace of $B^2(I)$ [39]. Here, we want prove that the Bayes Hilbert mixture operation preserves the *distribution* of the exponential family: if all elements in the convex hull of $\{h_j\} \subset B^2$ belong to the same distribution, then the Bayes Hilbert mixture also belongs to that distribution, yielding a density characterized by updated parameters.

Proposition 3. *Let $\mathcal{F}$ be the set of all densities obtained by truncating a fixed exponential family to a compact support $I' \subset I$. Then $\mathcal{F}$ is a convex subset of $B^2(I)$.*

Proof: Convex combinations in $B^2(I)$ are defined through the perturbation and powering operations $\oplus$ and $\odot$. Let $f_1, f_2 \in \mathcal{F}$ have the same compact support $I' \subset I$, and let $p \in [0, 1]$. Assume

$$\begin{cases} f_1(x) = h(x) \exp(\eta(\theta_1) \cdot T(x) - A(\theta_1)) \mathbb{1}_{I'}(x), \\ f_2(x) = h(x) \exp(\eta(\theta_2) \cdot T(x) - A(\theta_2)) \mathbb{1}_{I'}(x). \end{cases}$$

where $\mathbb{1}_I(x)$ denotes the indicator function of the set I at x . For $x \in I'$, the convex combination in $B^2(I)$ is

$$(p \odot f_1) \oplus ((1-p) \odot f_2).$$

By the algebraic properties of $B^2(I)$, this is equivalent to

$$\begin{aligned} & [p \odot (h(x) \cdot \exp(\eta(\theta_1) \cdot T(x) - A(\theta_1)))] \oplus \\ & \oplus [(1-p) \odot h(x) \cdot \exp(\eta(\theta_2) \cdot T(x) - A(\theta_2))] =_{B^2} \\ & =_{B^2} h(x)^{(p+(1-p))} \exp((p\eta(\theta_1) + (1-p)\eta(\theta_2))T(x) + \\ & \quad - (pA(\theta_1) + (1-p)A(\theta_2))) =_{B^2} \\ & =_{B^2} h(x) \exp(\eta(\theta_{12})T(x) - A(\theta_{12})) =_{B^2} \\ & =_{B^2} h(x) \exp(\eta(\theta_{12})T(x)) \end{aligned} \tag{17}$$

The existence of θ_{12} follow from the continuity of the natural parameter map η (by the intermediate value theorem). Hence, the convex combination belongs to the same exponential family truncated to I' , and therefore lies in $\mathcal{F}$. □

Hence, differently from the probabilistic mixture, which would generally exhibit multimodality, the BH mixture produces a distribution of the same exponential family but with modified parameters.

B Propositions for identifiability of the minimal representation

This appendix shows and discuss the details regarding the results, presented in Section 4, on identifiability of the representation of a random mixture G through a set of vertices $\{h_j\}_{j=1}^m$ and the distribution of the random proportion. Here, we assume that G is a *pure* random mixture, meaning that none of the possible realisations of G are allowed to fall outside the convex hull of a set of m vertices, where m is fixed – known or estimated. All the results in this Section are valid for any Hilbert space, although here presented in $B^2(I)$.

As a first step, in Section B.1 we prove that, if G is a random mixture, any set of vertices $\{h_j\}_{j=1}^m$ can be associated, according to G , with a unique random proportion $\mathbf{p} : \Omega \rightarrow S^m$. We will denote by $\mathfrak{D}(S^m)$ the set of all probability distributions on S^m , and write $\mathbf{p} \sim D \in \mathfrak{D}(S^m)$.

B.1 Partial identifiability

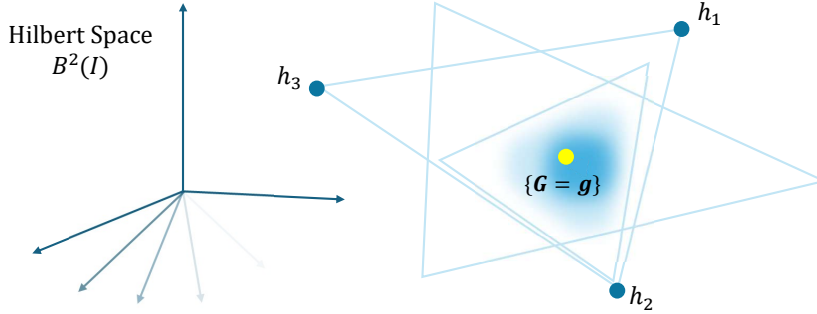


Figure 6: Schematic representation of a mixture G lying perfectly in $\mathcal{M}(h_1, h_2, h_3)$. The points h_1, h_2, h_3 are represented as points in the Bayes Hilbert space, and the triangle with vertices h_1, h_2, h_3 illustrates a convex hull. The blue cloud represents the distribution of G , with darker blue indicating higher probability mass, while the yellow point corresponds to a realization of G . Both the blue region and the realization of G lie within the depicted convex hull with probability 1. The other triangles, with unmarked vertices, represent alternative convex hulls in which G also lies with probability 1, but corresponding to a different distribution of the proportions. In this sense, the representation of G as a random mixture is not unique.

Let us assume that G is a random mixture of $\{h_j\}_{j=1}^m$, according to Definition 3.1. Then, by construction, for any $\omega \in \Omega$ we may find $\mathbf{p}(\omega)$ such that

$$G(\omega) = \bigoplus_{j=1}^m p_j(\omega) \odot h_j.$$

Hence, the random mixture G identifies at least one random variable $\mathbf{p} : \Omega \rightarrow S^m$. Proposition 4 states that the distribution of $\mathbf{p}$, given the distribution of G and $\{h_j\}_{j=1}^m$, is unique. The proof is given in the Appendix.

Proposition 4 (Identifiability in distribution of $\mathbf{p}|\{h_j\}_j$). *Let G be a random mixture of $\{h_j\}_{j=1}^m$, with h_j non-random elements in $B^2(I)$, $j = 1, \dots, m$, i.e. for $\omega \in \Omega$:*

$$G(\omega) = \bigoplus_{j=1}^m p_j(\omega) \odot h_j$$

for some $\mathbf{p} \sim D \in \mathfrak{D}(S^m)$. If it also holds:

$$G \stackrel{d}{=} \bigoplus_{j=1}^m \bar{p}_j \odot h_j$$

for some random vector $\bar{\mathbf{p}}$, then $\mathbf{p} \stackrel{d}{=} \bar{\mathbf{p}}$.

Proof of Proposition 4: By contradiction, let us assume that $\mathbf{p} \neq^d \bar{\mathbf{p}}$. Then, $\exists A \in S^m : \mathbb{P}(\mathbf{p} \in A) < \mathbb{P}(\bar{\mathbf{p}} \in A)$. Let us define $A_G := \left\{ g = \bigoplus_{j=1}^m p_j \odot h_j \mid \mathbf{p} \in A \right\}$ and $\forall \omega \in \Omega$, $\bar{G}(\omega) := \bigoplus_{j=1}^m \bar{p}_j(\omega) \odot h_j$. Then, by construction:

$$\mathbb{P}(G \in A_G) = \mathbb{P}(\mathbf{p} \in A) < \mathbb{P}(\bar{\mathbf{p}} \in A) = \mathbb{P}(\bar{G} \in A_G)$$

And this contradicts the assumption that $G = {}^d \bar{G}$. □

B.2 The vertices set

Not every m -tuple of vertices in $B^2(I)$ generates a valid convex hull for G (see Figure 7). To represent the distribution of G as a random mixture, we therefore restrict attention to those m -tuples whose elements are linearly independent and for which G lies in their convex hull with probability 1 – that is, $\text{supp}(G) \subseteq \mathcal{M}(\{h_j\}_{j=1}^m)$. The collection of such m -tuples defines the subset $\mathcal{H}_G^m \subset B^2(I) \times \dots \times B^2(I) = B^{2,m}(I)$ – Definition B.1. Furthermore, we show that $\mathcal{H}_G^m$ is closed in $B^{2,m}(I)$.

Definition B.1. We call $\mathcal{H}_G^m$ the subset of $B^{2,m}(I)$ defined as:

$$\mathcal{H}_G^m := \{(h_1, \dots, h_m) \in B^{2,m}(I) : \dim(\text{span}_{B^2(I)}(h_1, \dots, h_m)) = m, \\ \mathbb{P}(G \in \mathcal{M}(\{h_j\}_{j=1}^m)) = 1\}$$

where $\text{span}_{B^2(I)}$ is the span in the Bayes Hilbert space sense.

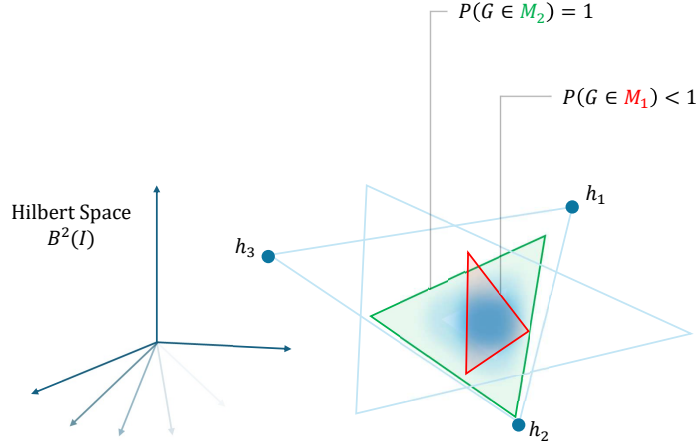


Figure 7: All triangles represent generic convex hulls. The green convex hull (M_2) is an example of element of $\mathcal{H}_G^m$, as $\text{supp}(G)$ is contained in it. Contrarily, the red one (M_1) does not entirely contain $\text{supp}(G)$, consequently it does not belong to $\mathcal{H}_G^m$.

Now, let us provide $B^{2,m}(I)$ with the norm $\|h\|_{B^{2,m}(I)} = \sum_{j=1}^m \|h\|_{B^2(I)}$. Hilbert spaces are by construction complete metric spaces, with respect to the norm induced by their own scalar product. Moreover, the cartesian product of m Bayes Hilbert spaces, $B^{2,m}(I)$, provided with a scalar product equal to the sum of the marginal scalar products $\|\cdot\|_{B^2(I)}$, is itself a Hilbert space. Thus, $(B^{2,m}(I), \|\cdot\|_{B^{2,m}(I)})$ is closed in itself. As a consequence of the previous proposition, we obtain the result in Proposition 5. The role of this result is being clarified in Section B.4, Theorem 1.

Proposition 5. *The set $B^{2,m}(I)$ is closed in $(B^{2,m}(I), \|\cdot\|_{B^{2,m}(I)})$*

Proof of Proposition 5: Consider a sequence of m -tuples of vertices $\{h_j^{(\iota)}\}_{j=1}^m$, $\iota \in \mathbb{N}$, converging to $\{h_j^*\}_{j=1}^m$. From the closure of $\mathcal{H}^m$, we know that $\{h_j^*\}_{j=1}^m \in \mathcal{H}^m$. However, we have to prove that, if $\mathcal{M}^* := \mathcal{M}(\{h_j^*\}_{j=1}^m)$, $\mathbb{P}(G \in \mathcal{M}^*) = 1$. We proceed as follows.

The sequence of m -tuples of vertices induces equivalently a sequence of convex hulls

$$\mathcal{M}_\iota := \mathcal{M}(\{h_j^{(\iota)}\}_{j=1}^m) \subset \mathcal{H}_G^m \subset \mathcal{H}^m$$

converging to $\mathcal{M}^*$. If the sequence is increasing, namely $\mathcal{M}_\iota \subset \mathcal{M}_{\iota+1}$, then, straightforwardly $\mathbb{P}(G \in \mathcal{M}^*) \geq \mathbb{P}(G \in \mathcal{M}_\iota) = 1$. If it is decreasing, namely $\mathcal{M}_{\iota+1} \subset \mathcal{M}_\iota$, then, being $\mathbb{P}(G \in \mathcal{M}_\iota) = 1$ for any $\iota \in \mathbb{N}_0$, and following the properties of the probability as a measure,

$$1 = \lim_{\iota \rightarrow \infty} \mathbb{P}(G \in \mathcal{M}_\iota) = \mathbb{P}\left(\bigcap_{\iota=1}^{\infty} \{G \in \mathcal{M}_\iota\}\right) = \mathbb{P}(G \in \mathcal{M}^*)$$

This implies that $\{h_j^*\}_{j=1}^m \in \mathcal{H}_G^m$. Lastly, if the sequence is neither decreasing nor increasing, we can first extract the subsequence of convex hulls that contains $\mathcal{M}^*$. The elements of these sequence are ordered according to a new index $\iota_1, \iota_2, \dots$. Now, we proceed in this way:

- Assign the first element: $\mathcal{M}'_{\iota_1} = \mathcal{M}_{\iota_1}$.
- Update all the other elements of the sequence intersecting them with $\mathcal{M}'_{\iota_1}$.
- Discard the empty sets.
- Reassign the indexes $\iota_2, \iota_3, \dots$ respecting the previous order.

Then, repeat for $\iota_2, \iota_3, \dots$, obtaining a sequence of subsets of $\mathcal{H}$. Then, since:

$$\mathcal{M}^* \subset \bigcap_{v=1}^{\infty} \mathcal{M}'_{\iota_v} \subset \bigcap_{\iota=1}^{\infty} \mathcal{M}_{\iota} = \mathcal{M}^*$$

the intersection of these decreasing subsets converges to $\mathcal{M}^*$. Hence, as before, $\{h_j^*\}_{j=1}^m \in \mathcal{H}_G^m$. $\square$

B.3 A map from vertices to proportions

We next aim to identify the probability distribution D of the random proportion $\mathbf{p} \sim D$, knowing only G and the set of vertices $\{h_j\}_{j=1}^m$. Consequently, we are interested in measuring the distance between proportions in terms of their distributions.

For this reason, we equip $\mathfrak{D}(S^m)$ with the metric d_D defined by

$$d_D(\mathbf{p}, \mathbf{p}') = \lambda(\{\mathbf{q} \in S^m : F_{\mathbf{p}}(\mathbf{q}) \neq F_{\mathbf{p}'}(\mathbf{q})\}),$$

where λ denotes the Lebesgue measure on $\mathbb{R}^m \supset S^m$, and $F_{\mathbf{p}}$ is the joint cumulative distribution function associated with $\mathbf{p}$. We note a slight abuse of notation in writing $d_D(\mathbf{p}, \mathbf{p}')$, as this formally represents the distance between the distributions $D \sim \mathbf{p}$ and $D' \sim \mathbf{p}'$, although it is expressed directly in terms of the parameter vectors $\mathbf{p}$ and $\mathbf{p}'$. This metric explicitly quantifies the discrepancy between two random vectors in terms of the sets where their associated distribution functions differ, and is commonly referred to as a measure-theoretic Hamming distance [5].

Thanks to Proposition 4, we can define the function $\phi_G : \mathcal{H}_G^m \rightarrow \mathfrak{D}(S^m)$, a map that given the distribution of a random mixture G and a proper m -tuple of vertices, returns the distribution D of the relative random proportion $\mathbf{p}$, namely

$$\phi_G(\{h_j\}_{j=1}^m) = D \quad \text{iff} \quad G \stackrel{d}{=} \bigoplus_{j=1}^m p_j \odot h_j \quad \text{for } \mathbf{p} \sim D.$$

We can prove that ϕ_G is continuous.

Proposition 6 (ϕ_G is continuous). *Let G be a random mixture of densities and let ϕ_G be defined as above. Then, ϕ_G is continuous.*

Proof of Proposition 6: We want to prove that $(\|\epsilon\|_{\mathcal{H}^m} \rightarrow 0) \Rightarrow (d_D(\phi_G(\{h_j + \epsilon_j\}_j), \phi_G(\{h_j\}_j)) \rightarrow 0)$. Let $\mathbf{p} \sim \phi_G(\{h_j\}_j)$ and $\mathbf{p}^{(\epsilon)} \sim \phi_G(\{h_j + \epsilon_j\}_j)$. First, we claim some necessary and sufficient statements. By construction of ϕ_G , for any $\epsilon = (\epsilon_1, \dots, \epsilon_m) \in \mathcal{H}^m$ such that $\{h_j + \epsilon_j\}_{j=1}^m \in \mathcal{H}_G^m$,

$$G = \bigoplus_{j=1}^m p_j \odot h_j \stackrel{d}{=} \bigoplus_{j=1}^m p_j^{(\epsilon)} \odot (h_j + \epsilon_j) \stackrel{d}{=} \bigoplus_{j=1}^m p_j^{(\epsilon)} \odot h_j + p_j^{(\epsilon)} \odot \epsilon_j$$

meaning that both the original and the deviated pair of vertices and proportions still represent G in distribution. This is true if and only if:

$$\bigoplus_{j=1}^m p_j \odot h_j \ominus \bigoplus_{j=1}^m p_j^{(\epsilon)} \odot h_j \stackrel{d}{=} \bigoplus_{j=1}^m p_j^{(\epsilon)} \odot \epsilon_j$$

If and only if:

$$\bigoplus_{j=1}^m (p_j - p_j^{(\epsilon)}) \odot h_j \stackrel{d}{=} \bigoplus_{j=1}^m p_j^{(\epsilon)} \odot \epsilon_j$$

Let us now assume that $\|\epsilon\|_{\mathcal{H}^m} \rightarrow 0$. Then:

$$\forall \omega \in \Omega : \bigoplus_{j=1}^m p_j^{(\epsilon)}(\omega) \odot \epsilon_j \xrightarrow{\|\epsilon\|_{\mathcal{H}^m} \rightarrow 0} [0]_{B^2}$$

This implies that:

$$\lim_{\|\epsilon\|_{\mathcal{H}^m} \rightarrow 0} \bigoplus_{j=1}^m (p_j - p_j^{(\epsilon)}) \odot h_j \stackrel{d}{=} [0]_{B^2} \Rightarrow \forall j \lim_{\|\epsilon\|_{\mathcal{H}^m} \rightarrow 0} (p_j - p_j^{(\epsilon)}) \stackrel{d}{=} 0$$

Thus,

$$d_D(\mathbf{p}, \mathbf{p}^{(\epsilon)}) \rightarrow 0$$

□

Although in practical applications ϕ_G can often be regarded as invertible – since real data rarely exhibit exact symmetries – this is not the case in the present theoretical setting, where no assumption is made on the geometry of the distribution of G . In this section, we introduce the notion of *finite-symmetry*, a property ensuring that ϕ_G admits a finite preimage for every element of $\mathfrak{D}(S^m)$ and, consequently, that it is a closed map.

The closedness of ϕ_G will play a crucial role in Section B.4, in particular in Theorem 1.

Definition B.2 (Finite-symmetry). We say that a random mixture G is *finite-symmetric* if for every D the preimage $\phi_G^{-1}(D)$ has finite cardinality.

Notice that random mixtures G exhibiting infinitely many symmetries are degenerate and purely theoretical. An example is given by

$$G = h \oplus \epsilon,$$

where $\epsilon \sim \mathcal{N}([0]_{B^2}, \sigma^2 I(s, v))$, schematically represented in Figure 8. Here, $[0]_{B^2}$ denotes the neutral element of perturbation (which is the constant density under the Lebesgue reference measure), and I the identity operator. In this case, one may apply a rotation T to one representative convex hull by any angle $\theta \in [0, 2\pi)$ without altering the (spherical) distribution of the proportions D . In fact, G can be interpreted in this case as a noiseless mixture with *infinitely many* vertices – rather than a noisy unique vertex. Consequently, the preimage of D through ϕ_G is uncountable, being in one-to-one correspondence with the interval $[0, 2\pi)$.

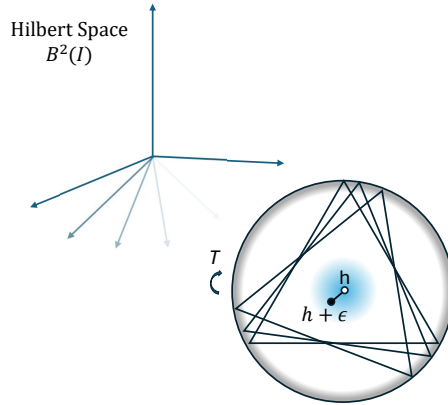


Figure 8: Schematic example of a non-finite-symmetric G . It can be seen as a density h with the addition of a noise, or alternatively as a noiseless mixture of infinitely many vertices around h .

Proposition 7. Let G be a *finite-symmetric* random mixture. Then ϕ_G is a closed map, that is, it maps closed sets into closed sets. In particular, $\text{Im}(\phi_G)$ is a closed subset of $\mathfrak{D}(S^m)$.

Proof of Proposition 7: Let C be a closed subset of the domain of ϕ_G , and let $\{\mathbf{p}_\iota\}_{\iota=1}^{+\infty} \subset \phi_G(C)$ be a convergent sequence with limit $\mathbf{p}$. For each ι there exists $x_\iota \in C$ such that $\phi_G(x_\iota) = \mathbf{p}_\iota$. Since $\phi_G^{-1}(\mathbf{p})$ is finite, there exists a subsequence (x_{ι_ρ}) converging to some $x \in C$. By continuity of ϕ_G , we obtain $\phi_G(x) = \mathbf{p}$, hence $\mathbf{p} \in \phi_G(C)$. Therefore $\phi_G(C)$ is closed. □

Henceforth, we restrict to finite-symmetric random mixtures G , which are the only ones relevant in real-case scenarios. Indeed, in real-case scenarios the distribution of G is never invariant under continuous families of transformations of the vertex set, as the underlying components carry distinct statistical signatures; exact continuous symmetries of the kind illustrated above are a purely theoretical artifact. In the next Section B.4, we introduce a strict partial order on the set of representations of G . More precisely, given two representations $(\{h_j\}_{j=1}^m, \mathbf{p})$ and $(\{\bar{h}_j\}_{j=1}^m, \bar{\mathbf{p}})$, we say that the first is smaller than the second if the associated trace of the covariance matrix of the corresponding proportions $\mathbf{p}$, denoted by $\text{tr}(\Sigma_p)$, is larger.

The quantity $\text{tr}(\Sigma_p)$ measures the total dispersion of $\mathbf{p}$. As discussed in Section 2, the trace of Σ_p provides a natural quantitative proxy for the geometric size of the representation. While related in concept, minimizing the trace encourages the distribution of $\mathbf{p}$ to spread out evenly within the convex hull, thereby utilising the available space efficiently. By contrast, directly minimising the volume of the convex hull tends to concentrate the vertices near the most extreme observations, which may distort the recovered statistical structure by overweighting atypical data points.

Figure 9 schematically illustrates the concept of a minimal element: it corresponds to the statistically space-efficient convex hull, i.e. to the representation in which the proportions exhibit the largest dispersion.

We aim to prove that a minimal element with respect to this partial order exists and is unique.

B.4 The minimal representation(s) of G

In this Section, we show that within the set of all the possible representations of the distribution of G there exists at least one representation that maximises the trace of the covariance of $\mathbf{p}$. Under a suitable assumption, this representation is unique.

Definition B.3. Let G be any random mixture, and let $(\{h_j\}_{j=1}^m, \mathbf{p})$, $(\{\bar{h}_j\}_{j=1}^m, \bar{\mathbf{p}})$ be such that:

$$G \stackrel{d}{=} \bigoplus_{j=1}^m p_j \odot h_j \stackrel{d}{=} \bigoplus_{j=1}^m \bar{p}_j \odot \bar{h}_j$$

We say that $(\{h_j\}_{j=1}^m, \mathbf{p}) <_G (\{\bar{h}_j\}_{j=1}^m, \bar{\mathbf{p}})$, if the trace of the covariance matrix Σ_p is higher than the one of $\Sigma_{\bar{p}}$, namely $\text{tr}(\Sigma_p) > \text{tr}(\Sigma_{\bar{p}})$.

Definition B.4 (Strict partial order, [34]). A relation $<$ on a set X is a strict partial order if it is:

- Irreflexive: $\forall x \in X, x \not< x$.
- Asymmetric: $\forall x, y \in X$ if $x < y$ then $y \not< x$.
- Transitive: $\forall x, y, z \in X$ if $(x < y) \wedge (y < z)$ then $x < z$.

Definition B.5 (Minimality [34]). An element $x \in (X, <)$, where $<$ is a (generic) order relation is called minimal if $\nexists y \in X : y < x$.

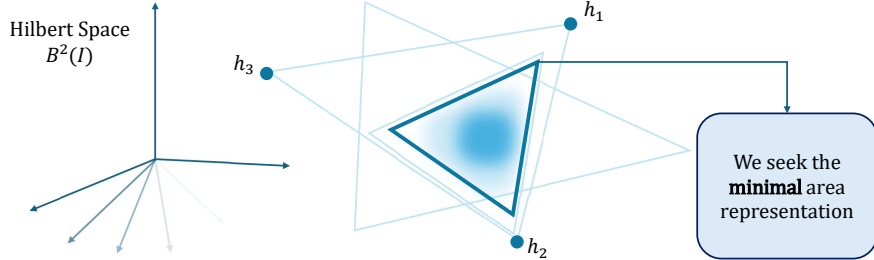


Figure 9: Schematic illustration of the concept of a minimal element: it corresponds to the convex hull in which the associated proportions exhibit the largest dispersion.

Let $\sim_\pi$ denote the equivalence relation on m -tuples of vertices defined as follows: two m -tuples $(h_1, \dots, h_m)$ and $(h'_1, \dots, h'_m)$ are equivalent, $(h_1, \dots, h_m) \sim_\pi (h'_1, \dots, h'_m)$, if there exists a permutation π of $\{1, \dots, m\}$ such that

$$(h_1, \dots, h_m) = (h'_{\pi(1)}, \dots, h'_{\pi(m)}).$$

Let $\mathcal{H}_G^m \times \mathfrak{D}(S^m) / \sim_\pi$ be the equivalence class with respect to $\sim_\pi$.

Proposition 8. The relation $<_G$ is a **strict partial order** relation on $\mathcal{C}_G \subset \mathcal{H}_G^m \times \mathfrak{D}(S^m) / \sim_\pi$, where $\mathfrak{D}(S^m)$ is the space of all possible distributions on S^m , $\mathcal{H}_G^m \times \mathfrak{D}(S^m) / \sim_\pi$ is the quotient space w.r.t. the equivalence relation π and

$$\begin{aligned} \mathcal{C}_G &= \left\{ (\{h_j\}_{j=1}^m, \mathbf{p}) \in \mathcal{H}_G^m \times \mathfrak{D}(S^m) : G \stackrel{d}{=} \bigoplus_{j=1}^m p_j \odot h_j \right\} / \sim_\pi = \\ &= \left\{ (\{h_j\}_{j=1}^m, \mathbf{p}) \in \mathcal{H}_G^m \times \mathfrak{D}(S^m) : \mathbf{p} = \phi_G(\{h_j\}_{j=1}^m) \right\} / \sim_\pi . \end{aligned}$$

Proof of Proposition 8: We only have to check all the properties of strict partial orders:

1. Irreflexive: trivial
2. Asymmetric: $(\{h_j\}_{j=1}^m, \mathbf{p}) <_G (\{\bar{h}_j\}_{j=1}^m, \bar{\mathbf{p}})$ implies that $\text{tr}(\Sigma_p) > \text{tr}(\Sigma_{\bar{p}})$, hence it cannot be true that $(\{h_j\}_{j=1}^m, \mathbf{p}) >_G (\{\bar{h}_j\}_{j=1}^m, \bar{\mathbf{p}})$.
3. Transitive: let $(\{h_j\}_{j=1}^m, \mathbf{p}) <_G (\{\bar{h}_j\}_{j=1}^m, \bar{\mathbf{p}})$ and $(\{\bar{h}_j\}_{j=1}^m, \bar{\mathbf{p}}) <_G (\{\hat{h}_j\}_{j=1}^m, \hat{\mathbf{p}})$. Then $\Sigma_p > \Sigma_{\bar{p}} > \Sigma_{\hat{p}}$. This implies that $(\{h_j\}_{j=1}^m, \mathbf{p}) <_G (\{\hat{h}_j\}_{j=1}^m, \hat{\mathbf{p}})$.

□

Proposition 9 (Existence of a minimal element). *If G is such that*

$$\sup_{(\{h_j\}_j, \mathbf{p}) \in \mathcal{C}_G} (\text{tr}(\Sigma_p)) < +\infty$$

then $(\mathcal{C}_G, <_G)$ has at least one minimal element.

Proof of Proposition 9: Let us consider a generic sequence $(\{h_j^{(\iota)}\}_j, \mathbf{p}^{(\iota)})_\iota \subset \mathcal{C}_G$ such that

$$\text{tr}(\Sigma_p^{(\iota)}) \rightarrow^{\iota \rightarrow \infty} \sup_{(\{h_j\}_j, \mathbf{p}) \in \mathcal{C}_G} (\text{tr}(\Sigma_p)) =: M < +\infty$$

From Proposition 7, we know that $\text{Im}(\phi_G) \subset \mathfrak{D}(S^m)$ is closed, hence also $\mathcal{H}_G^m \times \text{Im}(\phi_G)$ is closed. Then, $\exists(\{h_j^*\}_j, \mathbf{p}^*) \in \mathcal{H}_G^m \times \text{Im}(\phi_G) : (\{h_j^*\}_j, \mathbf{p}^*) = \lim_{\iota \rightarrow \infty} (\{h_j^{(\iota)}\}_j, \mathbf{p}^{(\iota)})$ and by construction $\text{tr}(\Sigma_p^*) = M$. Hence, there exist at least a minimal element, that is $(\{h_j^*\}_j, \mathbf{p}^*) \in \mathcal{H}_G^m \times \text{Im}(\phi_G)$. Consequently, there exists at least a minimal element in $\mathcal{C}_G = (\{h_j^*\}_j, \mathbf{p}^*) \in \mathcal{H}_G^m \times \text{Im}(\phi_G)/\pi$.

□

So far, we proved the existence of a minimal representation of G , according to the order $<_G$. To ensure uniqueness of the minimal representation, it is crucial to define and assume a property of G , which we call the *Probabilistic Pure Pixel* (3P) assumption. Intuitively, the 3P assumption requires that each vertex can be observed with positive probability, up to arbitrary precision; in other words, no vertex is probabilistically hidden. For a more formal definition see 4.1

For readers familiar with Hyperspectral Unmixing, the 3P assumption differs from the classical pure pixel assumption: 3P only requires that the probability of observing each noiseless mixture is theoretically positive, whereas the pure pixel assumption assumes it occurs with probability one.

When 3P assumption holds, we can conclude with Theorem 1, that states the uniqueness of the desired minimal representation of G .

Proof of Theorem 1: Let us assume that the minimal element is not unique. Then, we can intersect two convex hulls, obtaining a new set Θ . Let $\{h_1^*, \dots, h_m^*\}$ be one of these two convex hulls. If $h_j^* \in \Theta$ for every j , then for convexity, $\mathcal{M}(h_1^*, \dots, h_m^*) \subset \Theta$ and this is against the assumption of minimality. Then, at least one of the m vertices does not belong to Θ . Consequently, since Θ is the intersection of two closed subsets of $\mathcal{H}$, it is closed itself. Moreover, its complementary Θ^c is open. Therefore, $\exists \epsilon > 0$ such that $\mathbb{P}(G \in D_\epsilon(h_j) \subset \Theta) = 0$, then 3P does not hold. Due to the generality of the choice of $\{h_1^*, \dots, h_m^*\}$, 3P does not hold. This proves that 3P is a sufficient condition for uniqueness.

□

C Proofs of Section 5

Proof of Equation 12: First, we can rewrite Equation 11, fixing $\Sigma_\epsilon, \mu_p, \Sigma_p$ as follow:

$$\begin{aligned} \hat{l}(\mathbb{H}) = & -\frac{1}{2} \sum_{b=1}^B \sum_{i=1}^n w_{b,i} \cdot (\mathbf{f}_i - \mathbb{H} \cdot \mathbf{p}_{b,i})' \Sigma_\epsilon^{-1} (\mathbf{f}_i - \mathbb{H} \cdot \mathbf{p}_{b,i}) + \\ & -\lambda_{sm} \text{tr}(\mathbb{H}' \cdot D_2 \cdot \mathbb{H}) + \text{const} \end{aligned} \quad (18)$$

where $\mathbf{p}_{b,i}$ is the b -th sampled proportion for the i -th unit, $\mathbf{p}_{b,i}$ sampled according to the fixed μ_p , and Σ_p . Notice that $\forall b, i : \dim(\mathbf{p}_{b,i}) = m \times 1$. Lastly, $D_2 = \left[\langle \zeta_r'', \zeta_s'' \rangle \right]_{r,s}$, where ζ_r is the r -th element of the

k -dimensional basis of $\mathcal{H}$.

The gradient of $\hat{l}$ with respect to the matrix $\mathbb{H}$ has the same dimensions as $\mathbb{H}$, being the codomain of $\hat{l}$ one-dimensional, and it is equal to

$$\begin{aligned}\nabla_{\mathbb{H}}\hat{l} &= \sum_{b=1}^B \sum_{i=1}^n w_{b,i} \cdot \Sigma_{\epsilon}^{-1}(\mathbf{f}_i - \mathbb{H} \cdot \mathbf{p}_{b,i}) \mathbf{p}_{b,i}' - 2\lambda_{sm}(D_2 \cdot \mathbb{H}) \\ &= \Sigma_{\epsilon}^{-1} \left(\sum_{b=1}^B \sum_{i=1}^n w_{b,i} (\mathbf{f}_i - \mathbb{H} \cdot \mathbf{p}_{b,i}) \mathbf{p}_{b,i}' \right) - 2\lambda_{sm}(D_2 \cdot \mathbb{H})\end{aligned}$$

By imposing the gradient to be equal to the $k \times m$ dimensional null matrix, we get the following expression:

$$\begin{aligned}\Sigma_{\epsilon}^{-1} \left(\sum_{b=1}^B \sum_{i=1}^n w_{b,i} \cdot \mathbf{f}_i \mathbf{p}_{b,i}' \right) &= \\ &= \Sigma_{\epsilon}^{-1} \mathbb{H} \left(\sum_{b=1}^B \sum_{i=1}^n w_{b,i} \cdot (\mathbf{p}_{b,i} \mathbf{p}_{b,i}') \right) + 2\lambda_{sm}(D_2 \cdot \mathbb{H})\end{aligned}$$

That implies

$$\begin{aligned}\left(\sum_{b=1}^B \sum_{i=1}^n w_{b,i} \cdot \mathbf{f}_i \mathbf{p}_{b,i}' \right) &= \\ &= \mathbb{H} \left(\sum_{b=1}^B \sum_{i=1}^n w_{b,i} \cdot (\mathbf{p}_{b,i} \mathbf{p}_{b,i}') \right) + (2\lambda_{sm} \cdot \Sigma_{\epsilon} \cdot D_2) \cdot \mathbb{H}\end{aligned}$$

Hence, the problem is equivalent to finding $\mathbb{H}$ such that:

$$A_l \mathbb{H} + \mathbb{H} A_r = C$$

and the solution is:

$$\text{vec}(\mathbb{H}^*) = (A_r' \otimes I_{k \times k} + I_{m \times m} \otimes A_l)^{-1} \text{vec}(C)$$

where $\otimes$ denotes the Kroenker product, $I_{k \times k}$ and $I_{m \times m}$ are, respectively, the identity matrix of dimension $k \times k$ and $m \times m$, and $\text{vec}(\cdot)$ is the operation that vectorizes the matrices by column. □

Notice that if $\lambda_{sm} = 0$, the maximum of $\hat{l}$ – see Equation 11 – with respect to $\mathbb{H}$ is equal to:

$$\mathbb{H} = \left(\sum_{b=1}^B \sum_{i=1}^n w_{b,i} \cdot \mathbf{f}_i \mathbf{p}_{b,i}' \right) \left(\sum_{b=1}^B \sum_{i=1}^n w_{b,i} \cdot \mathbf{p}_{b,i} \mathbf{p}_{b,i}' \right)^{-1} = C A_r^{-1}$$

The matrix C can be interpreted as a weighted mean of the observations $\{\mathbf{f}_i\}$, where the weights are given by their posterior proportions. Let us examine the inverse of the second term, namely A_r^{-1} .

- The diagonal entries of A_r represent average squared posterior proportions associated with each vertex.
- The off-diagonal entries, say in position (j, j') , quantify the joint contribution of vertices j and j' , and can be interpreted as a *probability* of their coexistence within the data. Such coexistence indices tend to increase when many units i exhibit a uniform coexistence in two or more vertices. This is the case when the number of vertices m is overestimated.

Consequently, when A_r^{-1} is applied to C , it acts as a correction term that mitigates the artificial coexistence between different vertices, thereby improving their identifiability. This mechanism is analogous to the decorrelation of a dataset, which is obtained by applying the inverse of the upper triangular Cholesky factor of the covariance matrix.

Proof of Equation 14: First, we can rewrite Equation 11, fixing $\mathbb{H}, \boldsymbol{\mu}_p, \Sigma_p$ as follow:

$$\begin{aligned}\hat{l}(\Sigma_{\epsilon}) &= -\frac{1}{2} \sum_{b=1}^B \sum_{i=1}^n w_{b,i} \cdot (\mathbf{f}_i - \mathbb{H} \cdot \mathbf{p}_{b,i})' \Sigma_{\epsilon}^{-1} (\mathbf{f}_i - \mathbb{H} \cdot \mathbf{p}_{b,i}) + \\ &\quad - \log(\det(\Sigma_{\epsilon})) + \text{const}\end{aligned}\tag{19}$$

this maximization problem is equivalent to finding the maximum likelihood covariance matrix when the mean is fixed [6], whose solution is:

$$\begin{aligned}
\Sigma_\epsilon^* &= \\
&= \frac{1}{\sum_b \sum_i w_{b,i}} \sum_{b=1}^B \sum_{i=1}^n w_{b,i} ((\mathbf{f}_i - \mathbb{H}\mathbf{p}_{b,i}) - \mathbf{0})((\mathbf{f}_i - \mathbb{H}\mathbf{p}_{b,i}) - \mathbf{0})' = \\
&= \frac{1}{\sum_b \sum_i w_{b,i}} \sum_{b=1}^B \sum_{i=1}^n w_{b,i} (\mathbf{f}_i - \mathbb{H}\mathbf{p}_{b,i})(\mathbf{f}_i - \mathbb{H}\mathbf{p}_{b,i})'
\end{aligned} \tag{20}$$

□

Proof of Equation 15: this result is well known and employed in modern statistics, e.g. in [35].

□

Proof of Equation 16:

$$\begin{aligned}
\nabla_{\Sigma_p^{-1}} \hat{l} &= \\
&= \nabla_{\Sigma_p^{-1}} \left(\sum_b \sum_i w_{b,i} (\psi(p_{b,i}) - \boldsymbol{\mu}_p)' \Sigma_p^{-1} (\psi(p_{b,i}) - \boldsymbol{\mu}_p) + \right. \\
&\quad \left. - \log(\det(\Sigma_p)) - \lambda_{tr} \text{tr}(\Sigma_p^{-1}) \right) = \\
&= \nabla_{\Sigma_p^{-1}} \left(\sum_b \sum_i w_{b,i} (\psi(p_{b,i}) - \boldsymbol{\mu}_p)' \Sigma_p^{-1} (\psi(p_{b,i}) - \boldsymbol{\mu}_p) + \right. \\
&\quad \left. + (\sum_b \sum_i w_{b,i}) \log(\det(\Sigma_p^{-1})) - \lambda_{tr} \text{tr}(\Sigma_p^{-1}) \right) = \\
&= - \sum_b \sum_i w_{b,i} (\psi(p_{b,i}) - \boldsymbol{\mu}_p)(\psi(p_{b,i}) - \boldsymbol{\mu}_p)' + \\
&\quad + (\sum_b \sum_i w_{b,i}) \frac{\det(\Sigma_p^{-1}) \Sigma_p}{\det(\Sigma_p^{-1})} - \lambda_{tr} I_{(m-1) \times (m-1)}
\end{aligned}$$

Imposing the gradient equal to 0, we proved our statement.

□

D Justified choice of m in the simulations

The scree plots of all simulated datasets (see Figure 10) clearly display that the explained variance is essentially exhausted beyond this point. Therefore, $m = 3$ would be selected based on standard diagnostic criteria, independently of prior knowledge of the data-generating process. Finally, we compare the performance of the BH unmixing procedure under two scenarios: when the data are generated according to the BH mixture model, and when this assumption is violated, i.e., when the data are in fact linear mixtures.

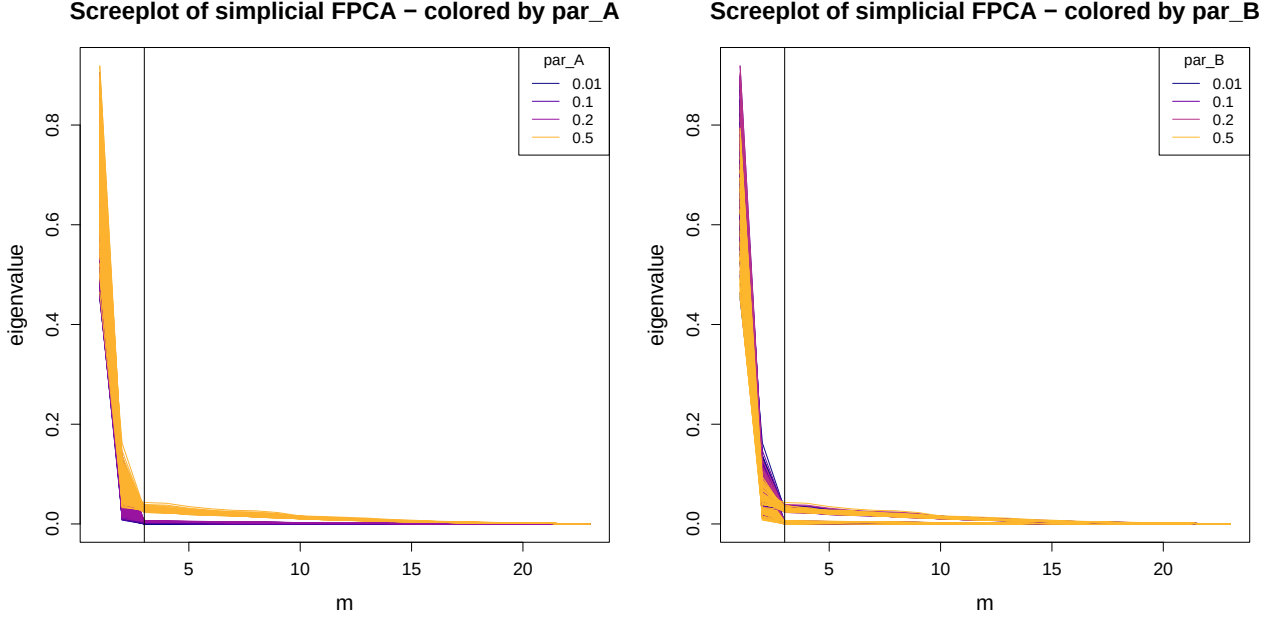


Figure 10: Scree plots of all the simulated datasets – both Study A and B. As one can notice, the variance exhaust in $m = 3$.

E Visualisation of the parameter in Study B

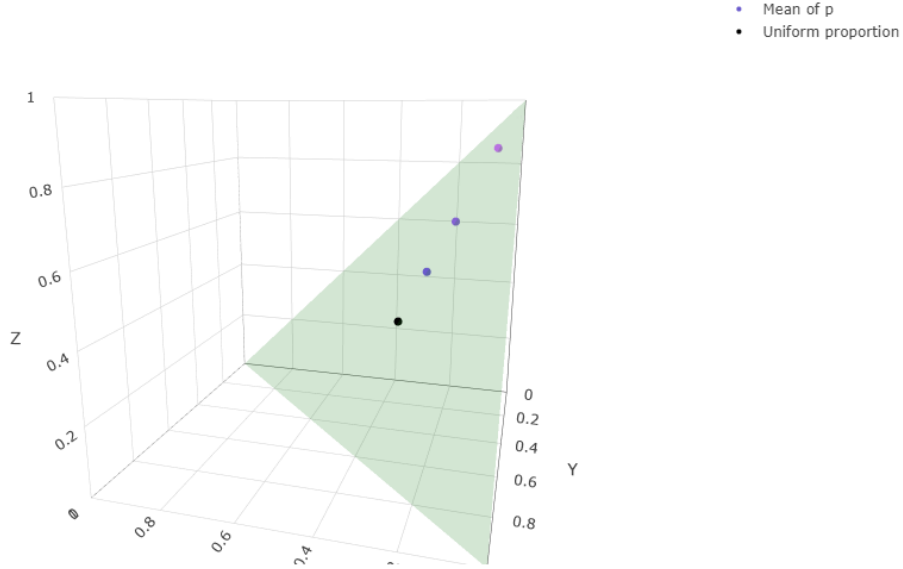


Figure 11: The options of μ_p in simulation study B, represented in $\mathbb{R}^3$, strictly belonging to $S^3 \subset \mathbb{R}^3$. The black point is $\psi^{-1}((0,0))$, the darkest violet is $\psi^{-1}((0,1/2))$ and the lightest is $\psi^{-1}((0,2))$.

F Effect of the number of vertices on vertex estimation error

In addition to the Study A (noise level, Section 6.1) and Study B (separation of mean proportions, Section 6.2), in this appendix we assess the effect of the number of vertices m on the estimation accuracy of the BH model. We compare three configurations: the baseline case $m = 3$ (shared by the Simulation A and B studies, with $\sigma^2 = 0.01$) and the cases $m = 5$ and $m = 8$ from Simulation C, keeping all other parameters fixed ($\sigma^2 = 0.01$, $\mu_p = 0$, $n = 500$, $k = 23$, 50 replicates per configuration). For each replicate we compute the vertex estimation

error as a weighted Frobenius norm between the estimated and true vertices, after optimal realignment via linear assignment (as in Section 6.1 and 6.2). As discussed below, the results reveal a trade-off between two competing effects of m on estimation accuracy, rather than a single monotonic relationship.

Figure 12 shows the distribution of the error for $m = 3, 5, 8$. Two distinct patterns emerge. First, the spread of the error across replicates decreases monotonically with m : the interquartile range and whisker extent are markedly larger for $m = 3$ than for $m = 5$ and $m = 8$. Second, the median error level does not follow the same monotonic trend: it is lowest for $m = 5$, higher for $m = 3$, and higher still for $m = 8$. The relationship between estimation error and the number of vertices is therefore non-monotonic, reflecting two competing effects that we discuss below.

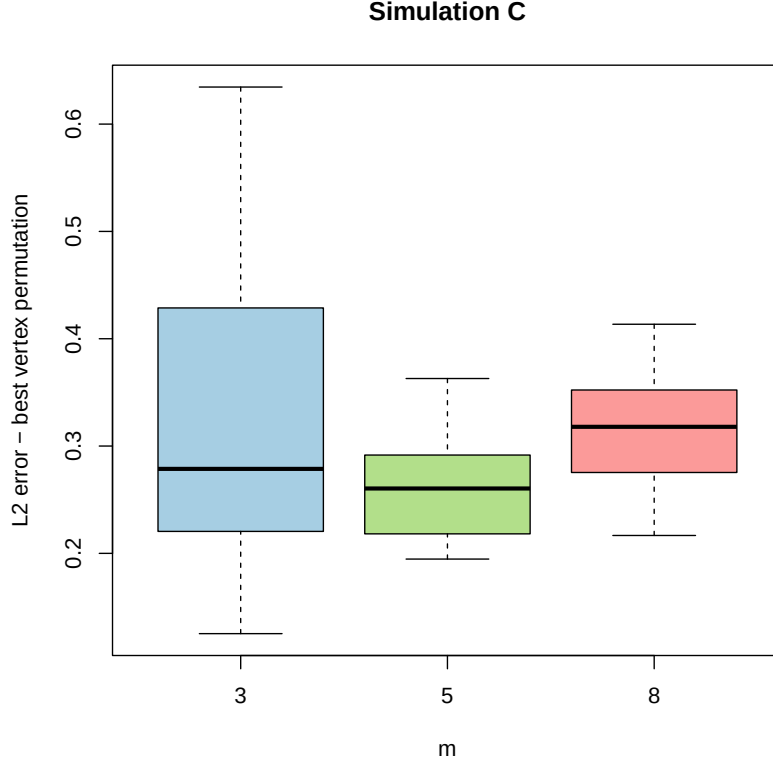


Figure 12: Vertex estimation error (weighted Frobenius norm, outliers not shown) for $m = 3, 5, 8$ vertices, BH model, 50 replicates per configuration ($\text{sd_perc} = 0.01$, $\mu_p = 0$, $n = 500$).

Effect 1: vertex separation reduces variability The first effect is explained by the vertex-generation mechanism in the simulator. For a given m , the j -th vertex ($j = 1, \dots, m$) is defined as a $\text{Beta}(j+1, m-j+2)$ density. Figure 13 shows these densities for $m = 3, 5, 8$. For $m = 3$, the Beta parameters remain close to the centre ($\text{Beta}(2, 4)$, $\text{Beta}(3, 3)$, $\text{Beta}(4, 2)$), producing wide, strongly overlapping vertices. As m increases, the parameters move progressively toward the tails of the distribution (e.g. $\text{Beta}(2, 9)$ and $\text{Beta}(9, 2)$ for $m = 8$), yielding narrower, better-separated vertices. Lower geometric separation between vertices makes the estimation problem less identifiable: small perturbations in the simulated data can drive the EM algorithm toward qualitatively different solutions across replicates. This effect alone would predict a monotonic decrease in both the level and the variability of the error as m increases.

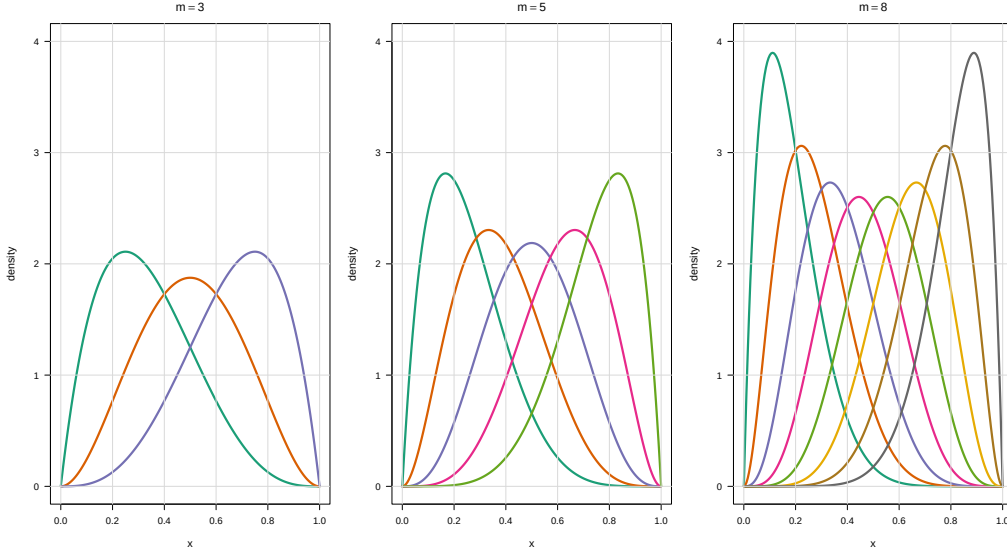


Figure 13: Densities that our vertices approximate, generated by the simulator under the $\text{Beta}(j+1, m-j+2)$, $j = 1, \dots, m$ parametrisation, for $m = 3, 5, 8$. For small m the vertices are wider and overlapping; as m increases they become narrower and better separated, reducing the ambiguity of the estimation problem and, consequently, the error variability shown in Figure 12.

Effect 2: more vertices means more parameters to estimate The second effect works in the opposite direction on the median error level. The vertex matrix H has dimension $k \times m$ with $k = 23$ basis functions, so the number of free parameters to estimate grows linearly with m (69, 115 and 184 parameters for $m = 3, 5, 8$ respectively), while the sample size is held fixed at $n = 500$ across all configurations. The ratio of vertex parameters to observations therefore increases from 0.14 at $m = 3$ to 0.23 at $m = 5$ and 0.37 at $m = 8$. With a fixed amount of information in the data, estimating more parameters leaves less information available per parameter, which tends to inflate the estimation error for each individual vertex. We verified that this increase is not an artefact of how the error is aggregated across columns of H : computing the error separately for each vertex (i.e. without summing over columns) yields the same pattern, with the per-vertex error markedly higher at $m = 8$ than at $m = 5$.

Net effect The two mechanisms act on different aspects of the error distribution and, in this experiment, partially offset each other. Increasing m from 3 to 5 is dominated by the separation effect: vertices become easier to tell apart, and both the level and the variability of the error decrease. Increasing m further from 5 to 8 is dominated by the parameter-count effect: vertices are even better separated, so the variability of the error continues to decrease, but the larger number of parameters relative to the fixed sample size increases the typical magnitude of the error. The configuration $m = 5$ in this study appears to sit close to the point at which the two effects balance, combining a comparatively low error level with low variability across replicates. We report the results as observed, without correcting for this trade-off, as it reflects a genuine property of the estimation problem under a fixed sample size rather than an instability of the proposed method.

G Additional figures for the AVIRIS Indian Pines case study

This appendix collects auxiliary figures for the AVIRIS Indian Pines case study of Section 7. Figure 14 reports the diagnostic output of the simplicial FPCA used to select $m = 5$, while Figure 15 displays the abundance maps obtained by the VCA competitor, included here for completeness.

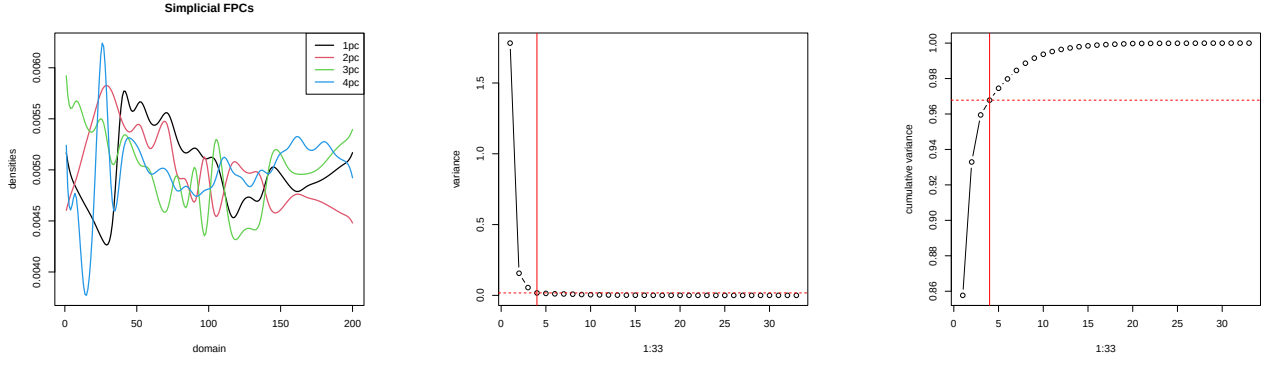


Figure 14: First 4 principal components (left). Variance explained by the first 4 components (center). Cumulative variance explained by the first 4 components (right).

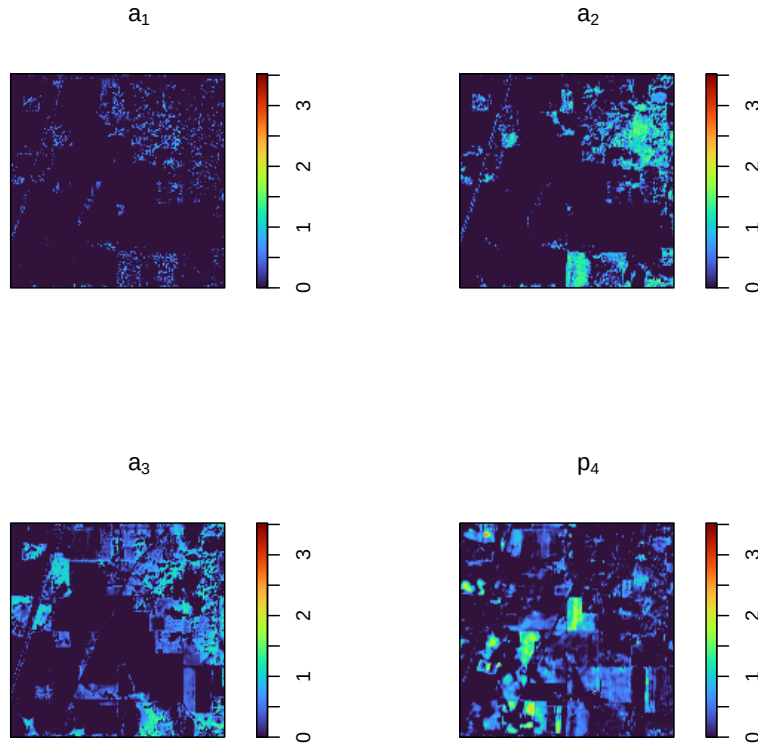


Figure 15: Vertex Component Analysis. In this map, we report the obtained abundances maps, where each abundance a_j is a positive number expressing the linear amplitude of the effect of the endmember ν_j .



Figure 16: Indian Pines site in 1993 (left), one year after AVIRIS recorded the hyperspectral data. In 1992, we have an almost identical image, with some parts missing. Indian Pines site in 2003 (right) – the first colored one –, 11 years after AVIRIS recorded the hyperspectral data. This image works as a reference to understand the image in 1993, which is black and white. Source: Google Earth



Figure 17: Experts' landcover map with 4 summarising categories. For the detailed subcategories see Figure 18 in G.

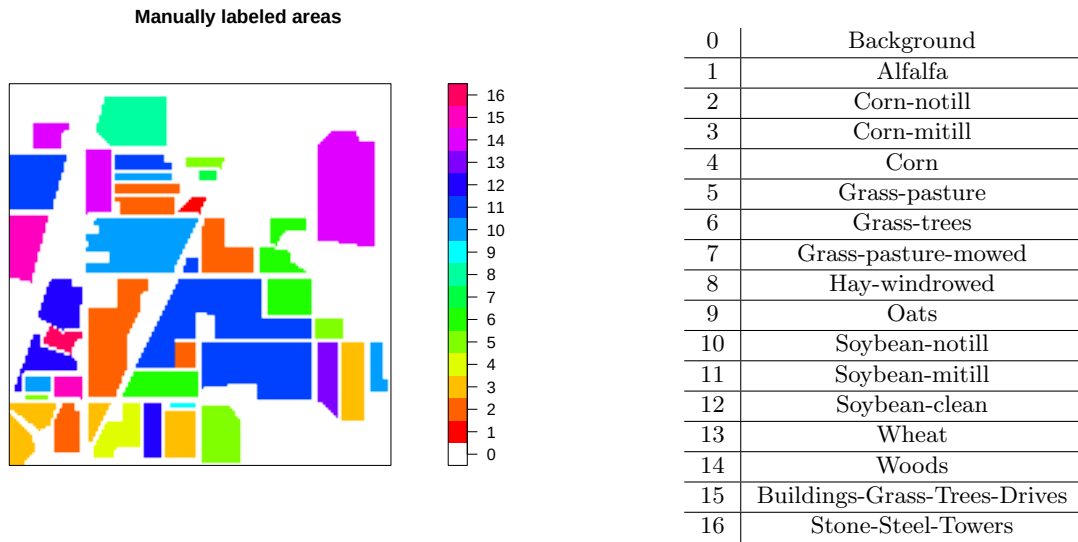


Figure 18: Experts' landcover map (left), with legend (right), with the detailed subcategories

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

- 67/2026** Ciardulli, S.; Fontana, N.; Vantini, S.; Ieva F.
Generalized propensity score weighting for functional causal inference framework
- 66/2026** Antonietti, P.F.; Bonizzoni, F.; Corti, M.; De March, N.; Di Noto, S.; Regazzoni, F.
A stability-preserving polytopal discontinuous Galerkin method for the Fisher-Kolmogorov model with applications to neurodegenerative diseases
- 65/2026** De Sanctis, M.F.; Arnone, E.; Ieva, F.; Sangalli, L.M.
Modeling group heterogeneity in spatio-temporal data via physics-informed regression
- 64/2026** Radisic, I.; Tirotta, R.; Zingaro, A.; Pagani, S.; Dede', L.
Physics-Informed Neural Networks for the High-Resolution Reconstruction of Flow Measurement Indicators in Fluid Dynamics
- 62/2026** Ciaramella, G.; Gong, W.; Kwok, F.; Tan, Z.
On the Uniform Convergence Analysis of the Schwarz Alternating Method for Optimal Control Problems
- 63/2026** Ciaramella, G.; Gong, W.; Kwok, F.; Tan, Z.
Uniform Convergence of the Schwarz Alternating Method for Optimal Control Problems
- 60/2026** Ciaramella, G.; Gander, M.J.; Van Crieckingen, S.; Vanzan, T.
Algebraic and Two-Level Parallel Substructured Schwarz Methods
- 58/2026** Mapelli, A.; Massi, M.C.; Cuccuru, G.; Di Angelantonio, E.; Ieva, F.
Prior-informed conditional Gaussian graphical models: an application to protein interaction network reconstruction
- 57/2026** Fontana, N.; Secchi, P.; Di Angelantonio, E.; Ieva, F.
Modeling time-varying genetic effects on binary disease risk via functional Mendelian Randomization
- 56/2026** Botta, P.; Vitullo, P.; Ventimiglia, T.; Linninger, A.; Zunino, P.
Physics-Informed Learning of Microvascular Flow Models using Graph Neural Networks