

Robust blind unmixing: A geometric approach to overcoming basis variation

Dumitru Mirauta¹, Vladimir V. Gusev^{*1}, Michael W. Gaultois², Matthew J. Rosseinsky², and Yannis Goulermas^{†1}

¹School of Computer Science and Informatics, University of Liverpool, Ashton Street, Liverpool, L69 3BX, Merseyside, United Kingdom

²Materials Innovation Factory, University of Liverpool, Oxford Street, Liverpool, L7 3NY, Merseyside, United Kingdom

Abstract

Signal separation problems are common in science. A prominent example of this occurs during the use of diffraction or spectroscopy to identify the individual components of a mixture by measuring it. In the simplest case, the measured signal is a linear combination of basis patterns corresponding to the constituent parts. The unmixing problem is to infer all or some of these basis patterns and abundances of components from measurements of distinct mixtures. One of the core challenges of this task is the variation of the basis from mixture to mixture due to noise and the exact physics of the measurement process. This is usually addressed with tailored model-based and parametric methods that are then limited in use to specific application domains by the nature of the assumptions made. We propose a novel geometric approach to unmixing problems which views the generation of data during measurement through a metric space lens, thereby shifting the focus from parametrised models to a general relationship between basis transformations and the corresponding geometry. We take advantage of the optimal transport distances to capture commonly occurring basis variations, and use minimisation of in-class variance of candidate solutions to drive the optimisation. We pay special attention to the one-dimensional case due to its practical importance and availability of efficient distance and transport map routines. The effectiveness of our approach is demonstrated on a range of unmixing tasks using random Gaussian mixture models, simulated powder X-ray diffraction, and laboratory hyperspectral imaging datasets.

Keywords: Unmixing, Basis variation, Optimal transport, Wasserstein distance, Earth movers distance, Error-in-variables

1 Introduction

The objects and processes of investigation in many scientific disciplines require advanced characterisation, often with multiple instrumental techniques to examine different aspects. In many cases, measurements generate a nonnegative response across some domain (e.g., across a range of angles, frequencies, energies, etc.). Examples of such responses include Powder X-ray Diffraction (PXRD) patterns, Hyperspectral Imaging (HI) pixels, and Raman spectra. For techniques producing real valued but periodic traces (e.g. infra-red spectroscopy, ion cyclotron resonance mass spectrometry, or even of audio waves) frequency distributions may be considered. These response profiles can be useful as a pseudoidentifier of the underlying object, where a similar profile is likely to refer to a similar—though not necessarily the same—object. A response

^{*}Corresponding author: vladimir.gusev@liverpool.ac.uk.

[†]This author has passed away.

may be characterised in itself, by a “pattern” of highs and lows, large and small features in different locations. As such, in this work, these response profiles will generally be referred to as patterns. For identification purposes, the magnitude of features relative to one another is more instructive than the absolute magnitude of any individual feature. As such, we will treat patterns as normalised distributions.

Crucially, the investigation of what may be considered a single underlying object, may produce a range of patterns. For instance, the PXRD pattern of a “pure phase” is composed of a series of peaks; these peaks can individually change position or shape, due to small variation in crystal structure induced as external parameters—such as temperature or pressure—change. Astronomical HI pixels may exhibit Doppler shift as relative speed to an object changes, or Earth HI may exhibit transformations due to atmospheric scattering, subsurface scattering, and illumination. In audio, the same structure (e.g. chord or overtone series of a voice) can change pitch and will have an associated (log) frequency shift. As such, multiple distinct points in pattern space can represent the same basis pattern, we call this basis variation.

Further, it is not always feasible to probe an object of interest alone. For instance, a single PXRD sample may contain responses from many “pure phases”. In such cases, the available response profiles will be a combination of responses from multiple objects. In order for the pattern(s) of interest to be isolated, it/they must first be separated. The set of all distinct patterns that can play a role in mixing will be referred to as the basis. In linear mixing, there is no basis variation, and mixtures are produced through linear combination of the basis.

Though linear combination is a common feature in many types of mixed data, across disciplines, so is basis variation. Here, it is variations of the basis patterns that are combined. With this, a set of patterns corresponding to the same set of objects, combined in the same proportions, may produce two different mixtures. As such, in this work a mixing model is defined by providing both a basis variation mechanism and a combination mechanism. This flexible umbrella model covers data encountered in materials discovery Ermon et al. (2015); Bai et al. (2018a); CHERNI et al. (2020), Remote sensing Borsoi et al. (2021); Heylen et al. (2014); Ma et al. (2014); Aharoni et al. (2022), Biology Hosseini-Asl et al. (2015); Yan et al. (2019); Glatz et al. (2011), Astrophysics Boulais, A. et al. (2021); Rameau, J. et al. (2021); Nakhostin et al. (2016); Liu et al. (2018), Food testing Holland et al. (1998), Music transcription Smaragdis (2004); Flamary et al. (2016) and other forms of audio separation (e.g. cocktail party problem). Image segmentation in computer vision Ho et al. (2017), multiple curve clustering/alignment Gaffney and Smyth (2004) and distribution template estimation Boissard et al. (2015) are also related.

While mixing produces a set of combined patterns from a basis, blind unmixing aims to recover a basis solely from a given set of combined patterns. While there is of course information loss in forming a single mixture, we may be able to patch together a good estimate of a basis from many mixed examples. Here domain-wise (cross-position) variation works in our favor, as different parts of a basis may be visible or obstructed in different mixtures depending on how the bases have transformed. Blind unmixing can of course be seen as an Inverse Problem (IP) Tarantola (2005); Vito et al. (2005), where the basis is a model parameter which essentially fixes a mixing model, which otherwise depends on per mixture variables. Otherwise, nonblind unmixing seeks to perform separation of the constituents in each mixture, according to a given basis. With sufficiently complex basis variation, there may be many ways of generating a given mixed example from a given basis, making even a nonblind “inversion” rather difficult.

Consider the importance of tackling the aforementioned basis variation in materials discovery; as sufficiently flexible unmixing methods are not available, much expert time can be directed towards manual analysis, instead of being able to focus on interpretation of higher level data and further experiment planning. The impact from the reduction of such a bottleneck may be felt far, as new materials enable new technologies and techniques for many other fields, including towards the improvement of the very hardware that could be helping accelerate the materials

discovery process. A mixing method that is not too specific to any single kind of characterisation technique is particularly beneficial here, as many kinds of characterisation techniques are in use.

We classify previous solutions to this problem as either (parametric) model fitting based, or geometric. Fitting based methods of course require the specification of a model to be fitted, however, it may be difficult to capture the multitude of transformation mechanisms required across the application domains, in a single, reasonably constrained model. A geometric method instead views mixing and its inverse problem geometrically. For instance, the linear unmixing case translates to a problem of recognising vertices in a point cloud that is expected to span a simplex. As many families of parametric mixing model can produce geometries with the same quality, they could all be covered by one geometric approach targeting this shared quality.

The more general form of the algorithm we suggest operates in an arbitrarily chosen metric space $(\mathcal{P}, d_{\mathcal{P}})$. If the spectra of interest exhibit transformations that are Lipschitz in this space, then this space has a key property. Namely, that up to some transformation parameter limit (which depends on the distinctiveness of the true basis), we show that the basis variations form clean, separable clusters (see Section A). This expected regularity can be exploited when searching for a way to pull apart our mixtures such that the resultant components correspond to plausible basis variants. The set of patterns (or representations thereof) $\mathcal{P}$ is a choice as well, for instance 1D peak patterns might be expressed through a space of vectors, of Gaussian Mixture Model (GMM) parameters, of abstract measures, etc. This in turn may affect the kind of distance that can be defined between patterns.

After this our focus shifts to the choice of Wasserstein space, and generating a practical algorithm that can execute in time scales feasible for interactive laboratory exploration. Wasserstein space can represent distribution-like data that is continuous or discrete, with weights on uniform or non-uniform grids, when grids are shared between discrete distributions or when not. Wasserstein distances can compare any combination of the above, and deal with transformations that would not be continuous if we were required to continue respecting a shared uniform grid for instance, such as when shifting data provided on a uniform grid. Wasserstein distances are Lipschitz to shifting, broadening and more Engquist et al. (2016). Overall, our approach to unmixing is novel in:

1. Allowing for non-Euclidean perturbation rather than explicitly fitting a fixed basis variation model.
2. Focusing on how a given mixture can be broken apart rather than on how a similar mixture might be put together.
3. Using the geometry of mixing to inform plausible decompositions, rather than a regressive cost.
4. Allowing for sparse and dynamic descriptors through signatures, rather than fixed grids, which can more accurately represent Wasserstein centroids and enable efficient separation.

This paper is structured as follows. In Section 2.1 we present our unmbrella model for mixing problems, establishing what is shared among them, and what can arbitrarily change. Section 2.2 gives an overview of Wasserstein distances and their place within Optimal Transport (OT) distances. In Section 2.3 we discuss existing fitting based and geometric approaches to unmixing. Section 3 contains our main contributions, our general approach, followed by the Wasserstein specialisation, and a practical implementation for 1D. A fast implementation hinges on efficient calculation of both Wasserstein distances, and crucially, transport plans, as such Section 4 discusses efficient computation of these. In Section 5 we compare the effectiveness of our method vs Non-negative Matrix Factorisation (NMF) and Overlapping NMF (oNMF), when dealing with data with basis variation in three contexts; GMM, PXRD and Hyperspectral data. Finally in Section 6 we will conclude and discuss how the approach could be further extended.

2 Preliminaries

Lower indices will correspond to vector indices while upper indices may sometimes enumerate non-vector quantities, such as discrete measures. By $X \subseteq \mathcal{X}$ we will implicitly mean measurable subsets.

List of Symbols

$a_i^j, \mathbf{A}, \mathbf{a}^j, \mathbf{a}_i$	Abundances: matrix entries, where we take a full slice over omitted indices.
$b^i, \mathbf{B}$	Bases: single pattern and tuple of respectively.
$\tilde{b}^{ij}, \tilde{\mathbf{B}}^j$	Basis variations: corresponding to combined pattern j , single and tuple of respectively.
$c^j, \mathbf{C}$	Combined/mixed patterns: single and tuple of respectively.
$\mathcal{U}(\mu, \nu)$	Couplings: a set containing all $\pi \in P(\mathcal{X} \times \mathcal{Y})$ that have μ and ν as marginals.
δ_x	Dirac: a member of $P(\mathcal{X})$ with weight only at $x \in \mathcal{X}$, where $\delta_x(X) = 1$ if $x \in X \subset \mathcal{X}$ and 0 otherwise.
$*_S$	Discrete cross-correlation: where $\mathbf{u} *_S \mathbf{v} = \sum_{s=1}^S v_s \mathbf{u}^{\downarrow s-1}$.
$\mathbf{D} \in \mathbb{R}^{N \times N}$	Distance matrix: usually between the supports of a given pair of finite measures.
$d_{\mathcal{X}}$	Ground space distance: distance between points in $\mathcal{X}$.
$\mathcal{X}, \mathcal{Y}$	Ground spaces: underlying spaces over which measures may be defined on.
K, M, N	Index limits: size of the basis, combined patterns set and measure support (maximum) respectively.
i, j	Indices: usually corresponding to the i^{th} basis pattern b^i and j^{th} combined pattern c^j respectively.
$[N]$	Integer range: i.e. the set $\{1, 2, \dots, N-1, N\} \subset \mathbb{N}$.
$\mathbb{R}_+^d$	Non-negative cone: i.e. $\{\mathbf{x} \in \mathbb{R}^d : x_i \geq 0\}$.
$\kappa(\mathbf{a}; \mathbf{B})$	Pattern combination function: mechanism for combining multiple patterns into a resultant mixed pattern.
$\mathcal{P}$	Pattern space: data space in which both pure and mixed examples live.
$d_{\mathcal{P}}$	Pattern space distance: distance between points in $\mathcal{P}$.
$\ \mathbf{x}\ _p^p$	P-norm: (to the p^{th} power) or $\sum_{i=1}^d x_i ^p$ where $\mathbf{x} \in \mathbb{R}^d$.
$P(\mathcal{X})$	Probability measures: i.e. those $m \in \mathcal{M}_+(\mathcal{X})$ with $m(\mathcal{X}) = 1$.
$T_{\#}\mu$	Pushforward: defined as $(T_{\#}\mu)(Y) = \mu(T^{-1}(Y))$ (for $Y \subset \mathcal{Y}$) for measure $\mu \in \mathcal{M}(\mathcal{X})$ and map $T : \mathcal{X} \rightarrow \mathcal{Y}$.
$\mathcal{M}_+(\mathcal{X})$	Radon cone: those $m \in \mathcal{M}(\mathcal{X})$ with $m(X) \geq 0$ for all $X \subset \mathcal{X}$.
$\mathcal{M}(\mathcal{X})$	Radon space: loosely, all m such that $m(X) \leq \infty$ for all $X \subset \mathcal{X}$.
$\mathbf{u}^{\downarrow s}$	Shift-down operator: with indices $(\mathbf{u}^{\downarrow s})_i = u_{i-s}$.
X_{μ}	Signature positions: set of positions of the Diracs that compose a finite measure μ .
W_{μ}	Signature weights: set of weights corresponding to each of the Diracs that compose a finite measure μ .
S^{d-1}	Simplex: i.e. $\{\mathbf{x} \in \mathbb{R}_+^d : \ \mathbf{x}\ _1 = 1\}$.
$(\mathbf{v}^1 \dots \mathbf{v}^N) \in V^N$	Tensor coordinates: expressed as concatenated slices of lower order tensor coordinates.
$\pi \in P(\mathcal{X} \times \mathcal{Y})$	Transport plan: for one given pair of measures.
$\mathbf{P} \in \mathbb{R}^{N \times N}$	Transport plan weights: for a discrete plan.
$(u^1, \dots, u^N) \in U^N$	Tuple: tuple or ordered set with elements $u \in U$.
$W_p^{d_{\mathcal{X}}}$	Wasserstein p distance: over measures with support in $(\mathcal{X}, d_{\mathcal{X}})$.

W_p

Wasserstein p distance: shorthand taking $d_{\mathbb{R}^N}(x, y) = \|x - y\|_p$.

Acronyms

AADPA	Average Absolute Difference in Percentage Abundances
AFD	Agile Factor Decomposition
BCCW	Basis-Component Centered Wasserstein
ED	Euclidean Distance
EMD	Earth Movers Distance
GC	Ground Cost
GD	Ground Distance
GMM	Gaussian Mixture Model
GS	Ground Space
HI	Hyperspectral Imaging
IBCW	Inter Basis Centered Wasserstein
ICCW	Inter Components Centered Wasserstein
IP	Inverse Problem
MLE	Maximum Likelihood Estimate
NMF	Non-negative Matrix Factorisation
NWCA	North-west Corner Algorithm
NWCR	North-west Corner Rule
oNMF	Overlapping NMF
OT	Optimal Transport
PCA	Principal Component Analysis
PXRD	Powder X-ray Diffraction
RMSE	Root Mean Square Error

2.1 Formal definition of mixing and unmixing

An unmixing problem seeks to invert the outputs of a mixing process for the corresponding inputs, as such, let us first define mixing. Let $\mathcal{P}$ be the space of allowed (pure or mixed) patterns, such as $\mathcal{P} = \mathbb{R}^N$ as in many cases, or the space of distributions over the real line $\mathcal{P} = P(\mathbb{R})$ as we will later consider.

The **basis** $\mathbf{B} = (b^1, \dots, b^K) \in \mathcal{P}^K$ is a tuple of such patterns, representing prototypes to be drawn from in the mixing process.

A **combination function** $\kappa : \mathbb{R}^K \rightarrow \mathcal{P}$ will describe the mechanism which prototypes may be combined through, such that each mixture is a function of its mixing proportions (here we take the basis to be fixed). The simplest and most common example is linear superposition $\kappa^{\text{lin}}(\mathbf{a}; \mathbf{B}) = \sum_{i=1}^K a_i b^i$. In the case of $\mathcal{P}$ being a space of distributions, mixing fractions will need to belong to a simplex $S^{d-1} = \{\mathbf{x} \in \mathbb{R}_+^d : \|\mathbf{x}\|_1 = 1\}$, for valid combinations, and so $\kappa : S^{K-1} \rightarrow \mathcal{P}$. Other examples include bilinear mixing $\kappa(\mathbf{A}; \mathbf{B}) = \sum_{m,n} a_{mn} b^m \odot b^n$ in some hyperspectral unmixing methods (where $\mathcal{P} = \mathbb{R}^N$ and $\odot$ is the Hadamard product) Heylen et al. (2014) or weighted Wasserstein centroids as in Schmitz et al. (2018).

While in the linear case a mixture might be constructed as $c^j = \kappa^{\text{lin}}(\mathbf{a}^j; \tilde{\mathbf{B}}^j)$, in the presence of **independent variation of each basis pattern**, we could interpret each mixture to be constructed according to its own, subtly different basis $c^j = \kappa^{\text{lin}}(\mathbf{a}^j; \tilde{\mathbf{B}}^j)$, with $\tilde{\mathbf{B}}^j = (\tilde{b}^{1j}, \dots, \tilde{b}^{Kj})$. The **basis variation model**, if parameterised, can be captured by a function $\phi : \Lambda \rightarrow \mathcal{P}$ in the variation parameters $\boldsymbol{\lambda} \in \Lambda$. For example, letting $\mathcal{P}$ be continuous functions over the real line, we could have shifting and broadening $\tilde{b} = \phi(\boldsymbol{\lambda}; b) = b(\lambda_2 \cdot (x - \lambda_1))$. For brevity, let this also act

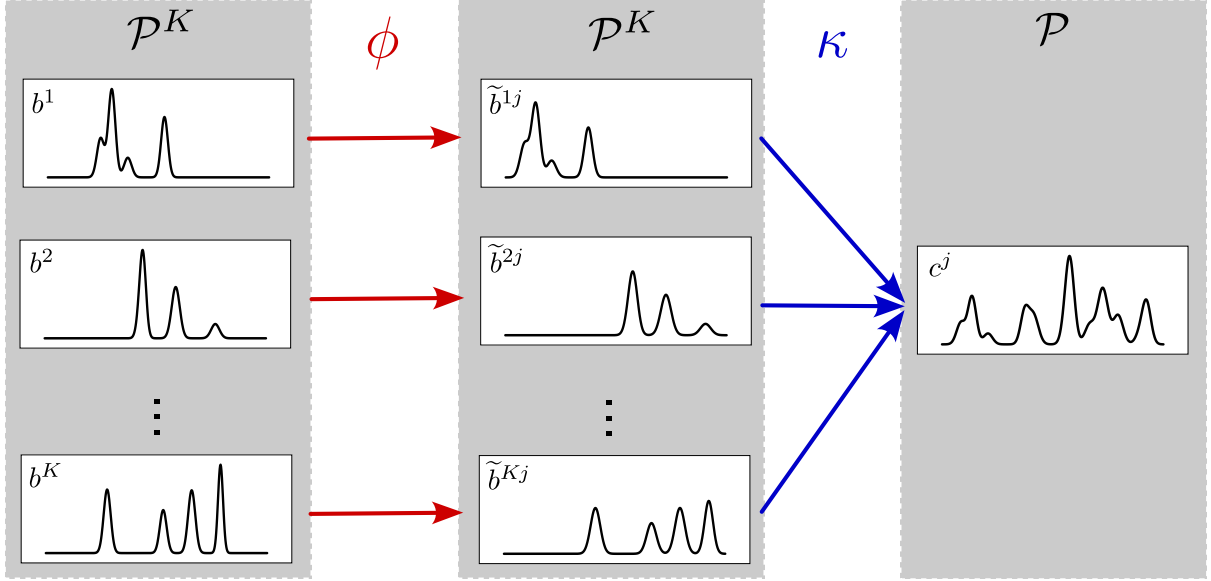


Figure 1: A mixing model operating in the space of pattern $\mathcal{P}$, incorporating independent variation of basis patterns prior to combination. Left: Basis patterns b^i . Middle: The varied basis copies $\tilde{b}^{ij}$ that correspond to the mixture c^j . Right: The resulting mixture c^j . Transitions from left to middle and from middle to right are provided through the variation ϕ and combination κ functions respectively (illustrated in red and blue respectively).

elementwise over a tuple of patterns, such that we may write $\tilde{\mathbf{B}}^j = \phi(\mathbf{\Lambda}^j; \mathbf{B})$ where $\mathbf{\Lambda}^j$ collects the individual parameters of each basis for mixture j .

In full, we say that a mixture/combined pattern c^j arises according to the **mixing model** $\psi(\mathbf{a}^j, \mathbf{\Lambda}^j; \mathbf{B}) = \kappa(\mathbf{a}^j; \phi(\mathbf{\Lambda}^j; \mathbf{B}))$, generating the dataset $\mathbf{C} = \{c^j : j \in [M]\}$. The **blind unmixing problem** seeks to reverse this process, generating the basis $\mathbf{B}$ from the dataset $\mathbf{C}$ alone. In the framework of IPs Tarantola (2005); Vito et al. (2005), $\mathbf{B}$ could be seen to parameterise a family of mixing models $\psi_{\mathbf{B}} : S^{K-1} \times \Lambda^K \rightarrow \mathcal{P}$, in this context recovering the correct $\mathbf{B}$ is equivalent to recovering the correct mixing model.

2.2 Optimal transport primer

We give a brief overview of a few concepts in OT, as they form a basis for the intuition, analysis and the practical implementation of our approach. It will be useful to highlight the overlap between the Monge and Kantorovich models, discrete and continuous, as well as the Earth Movers Distance (EMD) and Wasserstein costs or their minimising maps. When equivalent, it can be instructive to switch between these different perspectives. For a deeper overview see Peyré and Cuturi (2019); Santambrogio (2015); Villani (2009).

The problem of OT is intuitively expressed in terms of the earth movers analogy. In this we have a mound of dirt, the shape of which is described by one distribution, and a hole described by another. We seek a set of movements from the former to the latter which would expend the least overall energy¹. While the “transport” in this analogy is literal, transport in the general problem should be understood more as association or matching².

More formally, the OT problem aims to find minimum cost associations between two distributions (defined by measures) $\mu \in P(\mathcal{X})$ and $\nu \in P(\mathcal{Y})$ over Ground Spaces (GSs) $\mathcal{X}, \mathcal{Y}$. Given a GS cost $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ of transferring mass from x to y , an OT cost can be formulated

¹The energy/work done in this analogy is proportional to both the “mass” of some piece of the mound and to the distance it needs to slide over.

²And can in fact be thought of as a generalised matching problem.

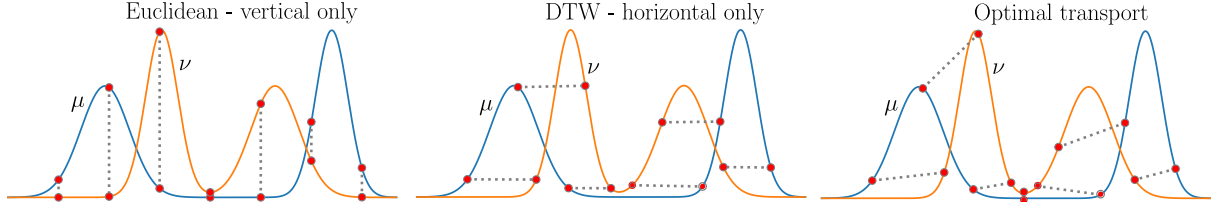


Figure 2: A distance between two distributions μ and ν can often be written as a sum of distances over parts of μ and ν that it takes to correspond. Here, DTW is dynamic time warping. Take μ to be an initial pattern (blue) and ν a transformed (orange) counterpart, with corresponding points (red markers) linked by dotted lines. In such a sum, change in one part is measured independently of change in another part, as if a point along μ had moved along its dotted line to become the corresponding point in ν .

as the average GS cost when following a particular transport map. Similar costs appear in information theory as measures of distortion Pollard (1982); Lloyd (1982). There are generally two types of maps that are considered, leading to two ways to write this cost.

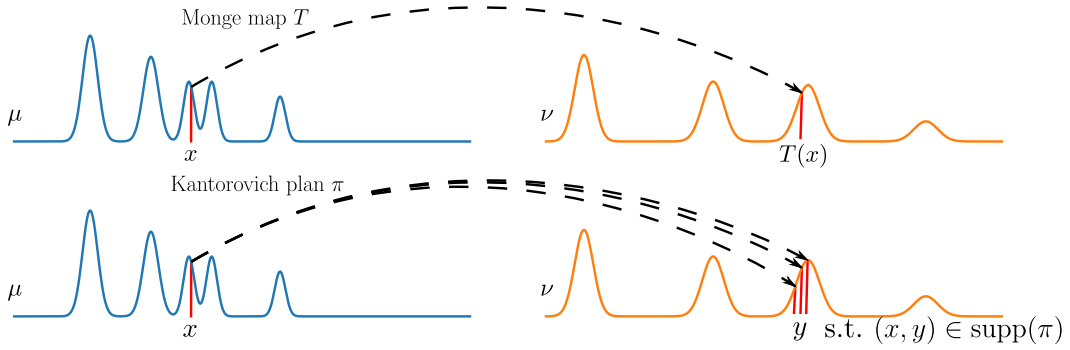


Figure 3: While a Monge map seeks a one to one correspondence, with a Kantorovich plan, mass can be broken up and sent to multiple locations. This splitting of mass can be necessary in the discrete setting, as a larger “stick” (Dirac) may need to be broken into parts and contribute towards multiple others for a mass preserving matching.

Definition 1 (Pushforward). The pushforward of $\mu \in \mathcal{M}(\mathcal{X})$ through $T : \mathcal{X} \rightarrow \mathcal{Y}$, results in a measure from $\mathcal{M}(\mathcal{Y})$ denoted $T_{\#}\mu$. This is defined as $(T_{\#}\mu)(Y) = \mu(T^{-1}(Y))$. That is to say, though T redistributes some mass such that it now is spread over Y rather than over X , the amounts of mass before and after this redistribution should agree.

Definition 2 (Monge cost). Provided a Monge map $T : \mathcal{X} \rightarrow \mathcal{Y}$ that satisfies mass conservation $T_{\#}\mu = \nu$, the average Ground Cost (GC) cost may be written

$$M^c(T) = \int_{\mathcal{X}} c(x, T(x)) d\mu(x). \quad (1)$$

Definition 3 (Space of couplings). Space of distributions π with μ, ν as marginals

$$\mathcal{U}(\mu, \nu) = \{\pi \in P(\mathcal{X} \times \mathcal{Y}) : \pi_{\mathcal{X}} = \mu, \pi_{\mathcal{Y}} = \nu\}, \quad (2)$$

where we denote the marginal $\pi_{\mathcal{X}}(X) = \int_{\mathcal{X} \times \mathcal{Y}} d\pi(x, y)$ for $X \subset \mathcal{X}$ and similarly $\pi_{\mathcal{Y}}(Y) = \int_{\mathcal{X} \times \mathcal{Y}} d\pi(x, y)$ for $Y \subset \mathcal{Y}$.

Definition 4 (Kantorovich cost). Provided a Kantorovich plan $\pi \in P(\mathcal{X} \times \mathcal{Y})$ that satisfies mass conservation $\pi \in \mathcal{U}(\mu, \nu)$, the average GC may be written

$$K^c(\pi) = \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y). \quad (3)$$

When $\mathcal{Y} = \mathcal{X}$ and $c(x, y)$ is a distance in $\mathcal{X}$, the problem can represent literal transport (movement across this common GS), as in an EMD. Here both the Monge and Kantorovich costs can be interpreted either as the average displacement during a rearrangement of μ into ν or equivalently the energy required by this rearrangement. In this case, an OT distance on $P(\mathcal{X})$ can be defined by taking the minimum cost mass preserving maps

$$d_M^c(\mu, \nu) = \min_{T: T_{\#}\mu = \nu} M^c(T) \quad \text{or} \quad d_K^c(\mu, \nu) = \min_{\pi \in \mathcal{U}(\mu, \nu)} K^c(\pi). \quad (4)$$

The Monge variant can be written in PDE form by considering the change of variables formula for distributions, where T acts as a reparameterisation. The Kantorovich variant is a linear functional (in π), where the dual problem— $\max_{\lambda, \gamma} \int \lambda(x) d\mu(x) + \int \gamma(y) d\nu(y)$ such that $c(x, y) \geq \lambda(x) + \gamma(y)$ —can be also be used.

If an optimal Monge map exists, it is equivalent to its Kantorovich counterpart in the sense that $\pi^* = (\text{id}, T^*)_{\#}\mu$, (see Fig. 4) where

$$\begin{aligned} K_c(\pi^*) &= \int c(x, y) d\pi^*(x, y) = \int c(x, y) d((\text{id}, T^*)_{\#}\mu)(x, y) \\ &= \int c(\text{id}(x), T^*(x)) d\mu(x) = M_c(T^*). \end{aligned} \quad (5)$$

Or loosely, the mass of π^* around the position $(x, T^*(x))$ is equal to that of μ around x , and 0 elsewhere, i.e. all of the mass of π^* is on the graph of the function T^* .

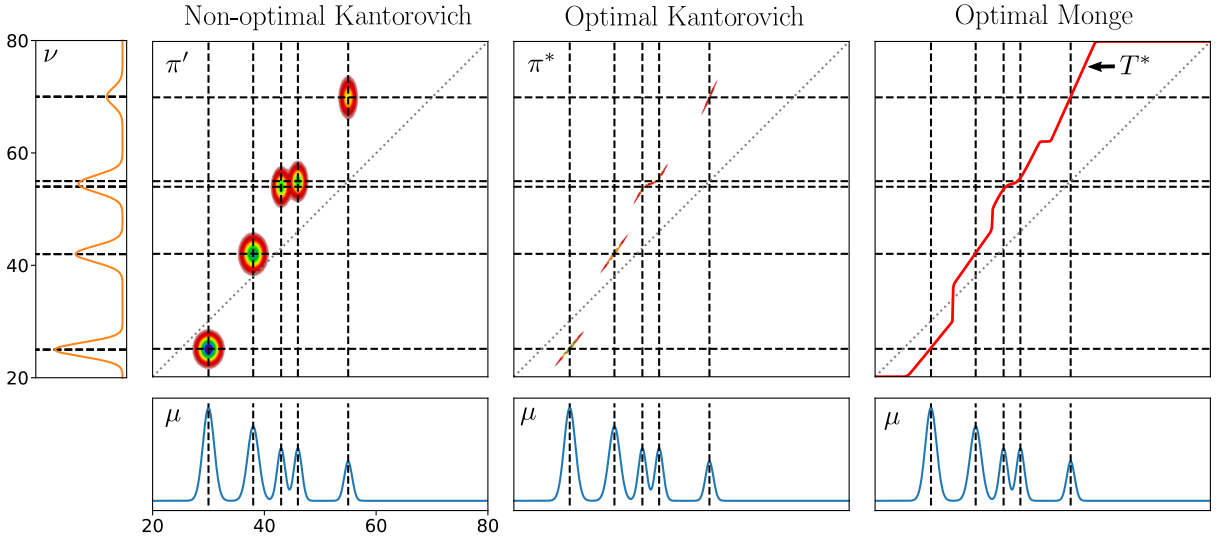


Figure 4: Overlap in Monge maps T and Kantorovich plans π . The density corresponding to bivariate distributions π are represented here through heatmaps, it is only noteworthy here where the density is zero (white) and where it is not. Nonzero density at (x, y) signifies transport from x to y , these (x, y) are in the “support” of the Kantorovich plan, $\text{supp}(\pi^*)$. While a Kantorovich plan allows for mass splitting if necessary, it is generally more economical to avoid splitting if possible. Therefore if possible $\text{supp}(\pi^*) \subseteq \text{graph}(T^*)$. Here, the pattern ν was generated by shifting and changing the breadth of the peaks in μ .

Valid distances can also be constructed in the format

$$W_p^c(\mu, \nu) = \sqrt[p]{d_M^{c^p}(\mu, \nu)} \quad \text{or} \quad W_p^c(\mu, \nu) = \sqrt[p]{d_K^{c^p}(\mu, \nu)}, \quad (6)$$

these are called Wasserstein distances. Let the omission of the superscript indicate the usual choice $c(x, y) = \|x - y\|_p$.

2.2.1 Discrete optimal transport

Let the measure $\mu \in P(\mathcal{X})$ have a finite support, so that $\mu = \sum_{i=1}^{N_\mu} w_\mu^i \delta_{x_\mu^i}$, where we may wish to refer solely to its positions $X_\mu = \{x_\mu^i \in \mathcal{X} : i \in [N_\mu]\}$ or its corresponding weights $W_\mu = \{w_\mu^i : i \in [N_\mu]\}$. The measure μ will be specified via its signature, a tuple of positions and weights (X_μ, W_μ) . The subscripts will be necessary later to distinguish the positions and weights of many different measures. Sometimes it may be more convenient to refer instead to the weight vector $\mathbf{w}_\mu \in \mathbb{R}_+^{N_\mu}$, which we will usually take to be normalised, i.e. $\mathbf{w}_\mu \in S^{N_\mu-1}$.

Let³ $\nu \in P(\mathcal{X})$ also be finite with signature (X_ν, W_ν) . The transport plan between these can only be nonzero on $X_\mu \times X_\nu$, taking the form $\pi = \sum_{i=1, j=1}^{N_\mu, N_\nu} p_{ij} \delta_{x_\mu^i, x_\nu^j}$, hence

$$K^c(\pi) = \sum_{i=1, j=1}^{N_\mu, N_\nu} p_{ij} \int_{\mathcal{X} \times \mathcal{X}} c(x, x') d\delta_{x_\mu^i, x_\nu^j}(x, x') = \sum_{i=1, j=1}^{N_\mu, N_\nu} c(x_\mu^i, x_\nu^j) p_{ij} \quad (7)$$

and so we arrive at the common form of EMD

$$\begin{aligned} d_K^c(\mu, \nu) &= \min_{\mathbf{P} \in \mathbb{R}_+^{N_\mu \times N_\nu}} \langle \mathbf{C}, \mathbf{P} \rangle \\ \text{s.t.} \quad &\mathbf{P}^T \mathbf{1} = \mathbf{w}_\mu \quad \text{and} \quad \mathbf{P} \mathbf{1} = \mathbf{w}_\nu \end{aligned} \quad (8)$$

where $c_{ij} = c(x_\mu^i, x_\nu^j)$.

Remark 1. Other applications of OT include: a distance between signed measures Engquist et al. (2016); a distance between combinations of elements in any set, as combinations may be treated as distribution, such as in Hargreaves et al. (2020); bootstrapping from intra as opposed to inter GS costs, with Gromov-Wasserstein distances; and more Peyré and Cuturi (2019); Flamary (2019). Use of the Wasserstein distance has been noted to produce a more stable optimisation in GANs Arjovsky et al. (2017). OT has been used in complex IPs, such as the inverse design of glass geometries that produces specific images with its caustics Schwartzburg et al. (2014).

2.3 Existing foundational techniques

There is a strong association between unmixing and factor based models, and indeed many solutions in which the **factor representation** is central. The factor representation of mixing is possible in **linear mixing**, where the basis is combined linearly, i.e. solely via κ^{lin} , with no basis variation and taking patterns to be expressed as vectors $\mathcal{P} \subseteq \mathbb{R}^N$. This representation usually corresponds to sampling on a shared uniform grid of positions. Here, multiple patterns (and similarly fractions) can be packed into the columns of a matrix. Noting that $\mathbf{c}^j = \kappa^{\text{lin}}(\mathbf{a}^j; \mathbf{B}) = \mathbf{B} \mathbf{a}^j$, the factor representation follows immediately $\mathbf{C} = (\mathbf{B} \mathbf{a}^1 \dots \mathbf{B} \mathbf{a}^M) = \mathbf{B} \mathbf{A}$, where $\mathbf{B} = (\mathbf{b}^1 \dots \mathbf{b}^k)$ and $\mathbf{A} = (\mathbf{a}^1 \dots \mathbf{a}^M)$. Since linear mixing can be expressed through matrix multiplication (in $\mathbb{R}^N$), linear unmixing, can here be seen as a matrix factorisation problem.

Matrix factorisation techniques such as Principal Component Analysis (PCA), Independent component analysis, and NMF can be therefore be seen to perform the required inversion, in the cases of orthogonal basis, independent basis, and nonnegative factor conditions being required respectively Mu et al. (2012); Agrawal et al. (2022); Lee and Seung (2001). Accordingly, NMF often forms a base for unmixing where nonnegative data is involved. Of course, while the model being fit may be linear, the fitting itself generally is a nonconvex optimisation.

The model utilised by oNMF Cichocki et al. (2009), diverges from the ordinary two factor model in NMF, by allowing bases to be varied via discrete convolution⁴ $\phi(\boldsymbol{\lambda}; \mathbf{b}) = \boldsymbol{\lambda} *_S \mathbf{b} = \sum_{s=1}^S \lambda_s \mathbf{b}^{\downarrow s-1}$. By $\mathbf{u}^{\downarrow s}$ we mean the shift-down operation where $(\mathbf{u}^{\downarrow s})_i = u_{i-s}$ if $i - s > 1$ else 0. Correspondingly, $\mathbf{U}^{\downarrow s}$ represents downward shifting each column vector. This allows for addition

³Here we chose to treat $\mathcal{X} = \mathcal{Y}$ for brevity, but this trivially applies to measures over separate GSs.

⁴As defined here, it is actually a cross correlation, though the kernel $\boldsymbol{\lambda}$ can be learned reversed.

of many overlapping shifted copies of each basis in each mixture, hence the name⁵. Similarly to NMF, a solution is sought by minimising a matrix dissimilarity between the mixtures $\mathbf{C}$ and their reconstruction according to the corresponding mixing model. By observing that the reconstructed mixture can be expressed as

$$\begin{aligned}\tilde{\mathbf{c}}^j &= \sum_{i=1}^K a_i^j \boldsymbol{\lambda}^{ij} *_S \mathbf{b}^i = \sum_{i=1}^K a_i^j \sum_{s=1}^S \lambda_s^{ij} (\mathbf{b}^i)^{\downarrow s-1} = \sum_{i=1}^K \sum_{s=1}^S \bar{\lambda}_{ij}^s (\mathbf{b}^i)^{\downarrow s-1} \\ &= \sum_{s=1}^S (\mathbf{B})^{\downarrow s-1} \bar{\mathbf{A}}_{*j}^s\end{aligned}\tag{9}$$

with matrices $(\bar{\mathbf{A}}^s)_{ij} = a_i^j \lambda_s^{ij}$, the usual matrix product sum form of the reconstructed mixtures $\tilde{\mathbf{C}} = \sum_{s=1}^S \mathbf{B}^{\downarrow s-1} \bar{\mathbf{A}}^s$ is obtained. This sum of products can be written as a single product of concatenated factors, hence this model can be seen through the lens of the linear mixing $\tilde{\mathbf{C}} = \bar{\mathbf{B}} \bar{\mathbf{A}}$, where $\bar{\mathbf{B}} = (\mathbf{B}^{\downarrow 0} \dots \mathbf{B}^{\downarrow S-1})$ and similarly $\bar{\mathbf{A}} = (\bar{\mathbf{A}}^0 \dots \bar{\mathbf{A}}^{S-1})$. Here the basis may be considered over-complete, holding multiple alternative representations of the same basis pattern. In mixing, only a single representation is desired for each basis pattern and as such sparse fractions are desired. An instance of this is Agile Factor Decomposition (AFD) Bai et al. (2018b), where varied bases are generated by a shift. Though additional constraints would be required so that different elements of the dictionary may correspond to different representations of the same basis, this is not dissimilar to the sparse dictionary learning problem. Accounts of methods more directly applying a sparse dictionary learning approach can be found in Borsoi et al. (2021).

The IPs of NMF and oNMF are usually tackled through **reconstruction error minimisation**, through costs such as $\sum_j D(\psi(\mathbf{a}^j, \boldsymbol{\Lambda}^j; \mathbf{B}), c^j)$ where D is a distance, dissimilarity or divergence, for example, $\sum_j \|\mathbf{B}\mathbf{a}^j - c^j\|_2^2$. Recall that while all three metric axioms are satisfied by a distance, dissimilarities and divergences may violate the triangle inequality and divergences may also violate symmetry. A reconstructed data matrix $\tilde{\mathbf{C}}$ can also be defined, where the j^{th} column is given by the reconstructed mixture $\psi(\mathbf{a}^j, \boldsymbol{\Lambda}^j; \mathbf{B})$ and with the cost $D(\tilde{\mathbf{C}}, \mathbf{C})$ being minimised instead, for some matrix divergence, for example $\|\tilde{\mathbf{C}} - \mathbf{C}\|_2^2$, or column-wise EMD such as in EMD-NMF⁶ Sandler and Lindenbaum (2011). The Wasserstein barycenter fitting approach Schmitz et al. (2018), also forms such a cost.

A cost of course needs to act as a notion of proximity to a desirable solution, however, there are two reasons why a cost of the format described above may not do this as well as possible.

A good fit of the mixture may not require a good basis. If the bases may be arbitrarily contorted, as is the case in using an OT model for basis transformation, a perfect fit may be achieved with any choice of basis, and so does not serve to guide us to a plausible basis.

Avoidable local minima. As well as local minima corresponding to alternative solutions, a typical reconstruction cost can also be stationary at misaligned copies of the same solution. This is due to the use of pointwise distances and the many locally optimal comparisons that they can produce. Suppose $\mathbf{A}$ and $\mathbf{B}$ are fixed. Even if there exists a unique globally optimal $\boldsymbol{\Lambda}_*^j$ with $\psi(\mathbf{a}^j, \boldsymbol{\Lambda}_*^j; \mathbf{B}) = c^j$, there can still be other $\boldsymbol{\Lambda}^j$ that locally minimise $F^j(\boldsymbol{\Lambda}^j) = D(\psi(\mathbf{a}^j, \boldsymbol{\Lambda}^j; \mathbf{B}), c^j)$. For instance, consider registering a single uniformly shifted basis in a pointwise distance (see Fig. 5 - bottom left). Any combination of local minima of F^j is a local minimum of $\sum_{j=1}^M F^j(\boldsymbol{\Lambda}^j)$, hence if each F^j has L local minima, the sum has L^M local minima. If we have 100 mixed examples, and the alignment of each is akin to Fig. 5, then each feasible decomposition has roughly an additional 3.23×10^{84} misregistered counterparts, entirely due to the choice of distance used in construction of a fitting cost.

⁵Sometimes this may also be referred to by the name of convolutional NMF (cNMF), which is a closely related algorithm Smaragdis (2004), though relates to a different composition, of 2D objects.

⁶Note, while EMD-NMF performs a more robust fit of the linear mixing model, it still only fits a linear/two factor model.

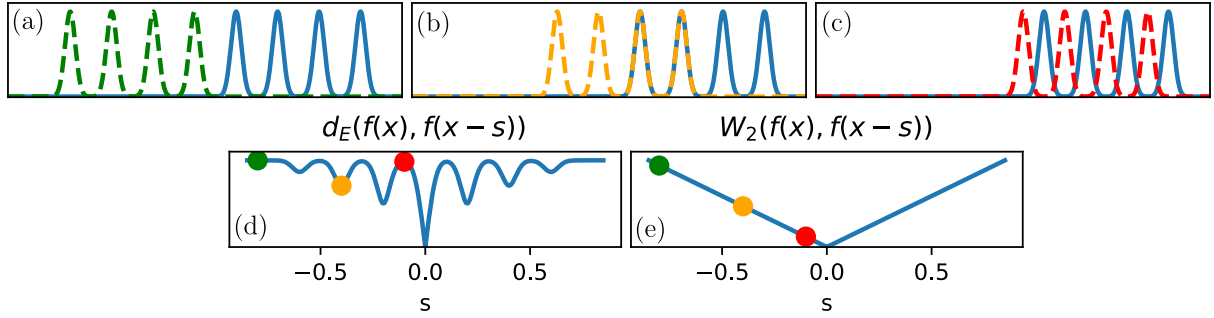


Figure 5: Distances after offsetting an original density f (a-c, solid blue), producing shifted counterparts (a-c, dashed) for various shifts s , the markers in panels (d) and (e) are colored according to corresponding values of s . Panels (d) and (e) track a pointwise distance (Euclidean) and an OT distance (Wasserstein) respectively, between the original and a shifted copy, as a function of shift. Note, since a pointwise distance only compares features that align vertically (see Fig. 2), we obtain local minima whenever non-corresponding peaks happen to align (such as in the yellow case). Note also the plateau in (d) when no features align (green case), where the gradient is near 0, and provides little direction in optimisation (even after perturbation).

2.3.1 Geometric approaches

Not only can the image of a mixing model ψ , and hence the point cloud of generated mixtures, form a specific structure within pattern space $\mathcal{P}$, but inputs such as the basis may be reflected through this structure as geometric features. Geometric approaches recognise this and look for such geometric features. We give these two classifications: those which consider the spatial structure arising from a particular form of the mixing model ψ , which strongly informs which geometric features should be sought, and those which seek some analogue to features present in the linear case, without necessarily considering whether the presence of such analogues will be preserved under the change in context.

In the linear case, taking all convex combinations of b^i simply forms a convex hull⁷ of these patterns, this contains all possible mixtures in the normalised case, with the bases at the vertices of the hull (Fig. 6 - left). Supposing a dataset provides a dense enough sampling of the possible mixtures, then to find the basis one needs only identify vertices. Simplex structure forms the foundations for simplex methods for linear mixing Ma et al. (2014). A modified structure may be obtained if one first allows bases to be perturbed prior to combination (Fig. 6 - middle) Stein (2003); Du et al. (2014); Zhou et al. (2018). While this modification considers variation of basis patterns, the distributions considered only assign significant likelihood to basis variation that does not drastically increase Euclidean distance (which can be large for non-pointwise variation).

Some approaches seek a simplex embedded within a manifold, for instance through search of vertices via counting graph geodesics Heylen et al. (2014). This offers an analogue for simplex fitting in the non-linear case, however, as basis variation is not treated independently in this model, it would not be able to inspect how each basis has changed individually, in the context of a given mixture. It is also not clear what kinds of mixing might produce such structure. Manifold analogues of PCA have also been explored Fletcher et al. (2004), however a manifold analogue of NMF might be more appropriate for non-negative data.

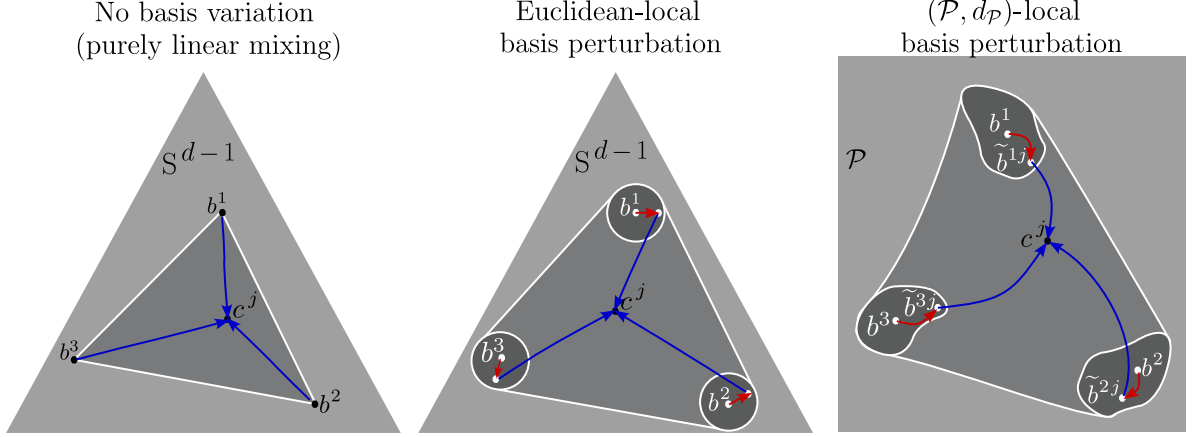


Figure 6: Geometric views of mixing under three kinds of basis variation (see Fig. 1) and linear combination, with the action of the variation ϕ and combination κ functions denoted with red and blue arrows respectively. Dark grey represents possible variations $\tilde{b}$ of basis pattern b , lighter grey - possible combinations, and lightest - the whole space of patterns; the simplex $S^{d-1} = \{\mathbf{x} \in \mathbb{R}_+^d : \|\mathbf{x}\|_1 = 1\}$ or an arbitrary pattern space $\mathcal{P}$.

3 Latent basis variant clustering

Consider the space of mixing inputs (fractions and basis variations) that combine to produce a given arbitrary pattern $p \in \mathcal{P}$ exactly

$$\mathcal{I}(p) = \{(\mathbf{a}, \tilde{\mathbf{B}}) \in S^{K-1} \times \mathcal{P}^K : \kappa(\mathbf{a}; \tilde{\mathbf{B}}) = p\}. \quad (10)$$

A decomposition that perfectly reconstructs all of our given mixtures can be obtained by simply selecting one element from each $\mathcal{I}(c^j)$. Such a selection of $\tilde{b}^{ij}$ may have no particular spatial structure (or connection to a shared basis), however under mild conditions (see Section A), we expect clustering of basis variants. There may be many solutions in $\prod_j \mathcal{I}(c^j)$ which appear to be clustered, a simple way to find one would be to optimise a clustering objective such as minimising in class variation. This has the desirable side effect of favouring bases requiring less modification on average to explain the data, which we might say provide a more natural explanation. To measure variation, we will employ the Fréchet variance.

The weighted Fréchet variance of a cluster of patterns $\tilde{\mathbf{P}} \subset \mathcal{P}$ is defined as follows

$$\text{FV}(\tilde{\mathbf{P}}, \mathbf{w}) = \min_{p \in \mathcal{P}} \sum_{j=1}^{|\tilde{\mathbf{P}}|} w_j d_{\mathcal{P}}(p, \tilde{p}^j)^2. \quad (11)$$

Much like ordinary variance, it measures average squared deviation from the mean (defined by the minimiser), with the modification that deviation is given by $d_{\mathcal{P}}$ rather than the Euclidean distance.

Our problem is therefore

$$\min_{(\mathbf{a}^j, \tilde{b}^{*j}) \in \mathcal{I}(c^j) \forall j} \sum_{i=1}^K \text{FV}(\tilde{b}^{i*}, a_i^*) = \arg \min_{\substack{(\mathbf{a}^j, \tilde{\mathbf{B}}^j) \in \mathcal{I}(c^j) \forall j \\ b^i \in \mathcal{P} \forall i}} \sum_{i=1}^K \sum_{j=1}^M a_i^j d_{\mathcal{P}}(b^i, \tilde{b}^{ij})^2 \quad (12)$$

where we weight according to abundances, as we wish components that play a lesser role in a given mixture, to have less overall influence. This is of course much too large a problem

⁷This is a simplex given a number of bases not more than one above the dimensionality of the patterns.

in practice, however, it can be decoupled into many smaller subproblems through alternating optimisation.

Averaging - given $\tilde{b}^{**}$, update cluster centroids

$$\forall i \in [K] \quad b^i = \arg \min_{b \in \mathcal{P}} \sum_{j=1}^M a_i^j d_{\mathcal{P}}(b, \tilde{b}^{ij})^2. \quad (13)$$

Separation - given b^* , if possible, select feasible mixing inputs that are less spread out from the cluster centers

$$\forall j \in [M] \quad \mathbf{a}^j, \tilde{\mathbf{B}}^j = \arg \min_{(\mathbf{a}, \tilde{\mathbf{B}}) \in \mathcal{I}(c^j)} \sum_{i=1}^K a_i d_{\mathcal{P}}(b^i, \tilde{b}^i)^2. \quad (14)$$

Clearly Eq. (12) is non-increasing under these alternating minimisations, as each updates a subsets of the overall sum in a non-increasing fashion. Note, either the cost is strictly decreasing, in which case we have convergence as the cost is bound below by 0, or neither averaging or separation have updated any terms at some iteration, in which case no further updates can be made, also converging.

For data such as that produced by PXRD; perturbation mechanisms such as shifting and broadening are of particular importance. While pointwise distances do not behave well with respect to these (see Fig. 5), Wasserstein distances are convex (and therefore locally Lipschitz) to these and more Engquist et al. (2016), and so we can expect clustering of basis variations (see Section A). Accordingly, from here we consider the set of all distributions $\mathcal{P} = P(\mathcal{X})$ on a geodesic GS $(\mathcal{X}, d_{\mathcal{X}})$. The distance on $\mathcal{P}$ will be the Wasserstein distance $d_{\mathcal{P}} = W_2^{d_{\mathcal{X}}}$.

3.1 Averaging

Let (S, d_S) be some metric space. For convenience, let the weighted centroid of two points $\mu, \nu \in S$ be expressed by

$$B_{d_S}(\mu, \nu, \lambda) = \arg \min_{\beta \in S} \left[(1 - \lambda) d_S(\beta, \mu)^2 + \lambda d_S(\beta, \nu)^2 \right], \quad (15)$$

this can in fact be obtained by walking a fraction of the way along the geodesic connecting the two points $g_{\mu}^{\nu} : [0, 1] \rightarrow S$, or $B_{d_S}(\mu, \nu, \lambda) = g_{\mu}^{\nu}(\lambda)$ (see Section C). In the Euclidean Distance (ED) case, where $d_S(\mu, \nu) = \|\mu - \nu\|_2$, the centroids lie on the line $(1 - \lambda)\mu + \lambda\nu$. Note, this connection between centroids and geodesics comes in handy both in the ground space and in the overarching Wasserstein space.

Definition 5 (The general McCann interpolation). Let $T^{\lambda}(x) = B_{d_{\mathcal{X}}}(x, T^*(x), \lambda)$ where T^* is the optimal transport map from μ to ν , then the general McCann interpolant is the pushforward $T_{\#}^{\lambda}\mu$ (for $\lambda \in [0, 1]$).

A Monge map can tell us that mass at x should go to $T(x)$, but does not specify how this should happen. In the McCann, or displacement interpolation, we continuously (and at constant speed) move mass along the GS geodesic connecting x and $T(x)$, obtaining a measure somewhere between μ and ν at each instant (see Fig. 7). Since the general McCann interpolant follows the Wasserstein space geodesic (see Section D), just as the centroid does (see Section C), one can use the McCann interpolant to calculate centroids in Wasserstein space and vice versa

$$B_{W_2^{d_{\mathcal{X}}}}(\mu, \nu, \lambda) = T_{\#}^{\lambda}\mu. \quad (16)$$

When a Monge map does not exist, an approximation—in the sense $W_2(\bar{T}_{\#}\mu, \nu) \approx 0$ —can be generated via the barycentric projection, $\bar{T}(x) = \int_{\mathcal{X}} y d\pi^*(x, y)$ in the ED GC case. Thanks to the displacement interpolation, there is no need to solve a coupled transport problem between

Algorithm 1 Pairwise barycenter

```
function  $\bar{B}_{W_2^{d_X}}(\mu, \nu, \lambda)$  ▷ Approximation due to projection better for lower  $\lambda$ .  
   $\mathbf{P} \leftarrow \text{DERIVEOTP}_{d_X}(\mu, \nu)$  ▷ Choose deriveOTP as appropriate, we consider the  
  North-west Corner Algorithm (NWCA) later  
  for  $i$  from 1 to  $N$  do  
     $\mathbf{x}_{\nu'}^i \leftarrow B_{d_X}(X_\nu, \mathbf{p}_{i*})$  ▷ Barycentric projection,  $\sum_{j=1}^M p_{ij} \mathbf{x}_\nu^j / w_\mu^i$  for ED  
     $\mathbf{x}_\beta^i \leftarrow B_{d_X}(\mathbf{x}_\mu^i, \mathbf{x}_{\nu'}^i, \lambda)$  ▷ Displacement interpolation,  $\lambda \mathbf{x}_\mu^i + (1 - \lambda) \mathbf{x}_{\nu'}^i$  for ED  
  end for  
  return  $(X_\beta, W_\mu)$   
end function
```

in $\pi^1 \in \mathcal{U}(\mu, \beta(\lambda))$ and $\pi^2 \in \mathcal{U}(\beta(\lambda), \nu)$ for each weight λ , knowing just $\pi \in \mathcal{U}(\mu, \nu)$ gives us $\beta(\lambda)$ for all $\lambda \in [0, 1]$.

When averaging more than two patterns, we approximate barycenters hierarchically, as in Boissard et al. (2015), see Algorithm 2.

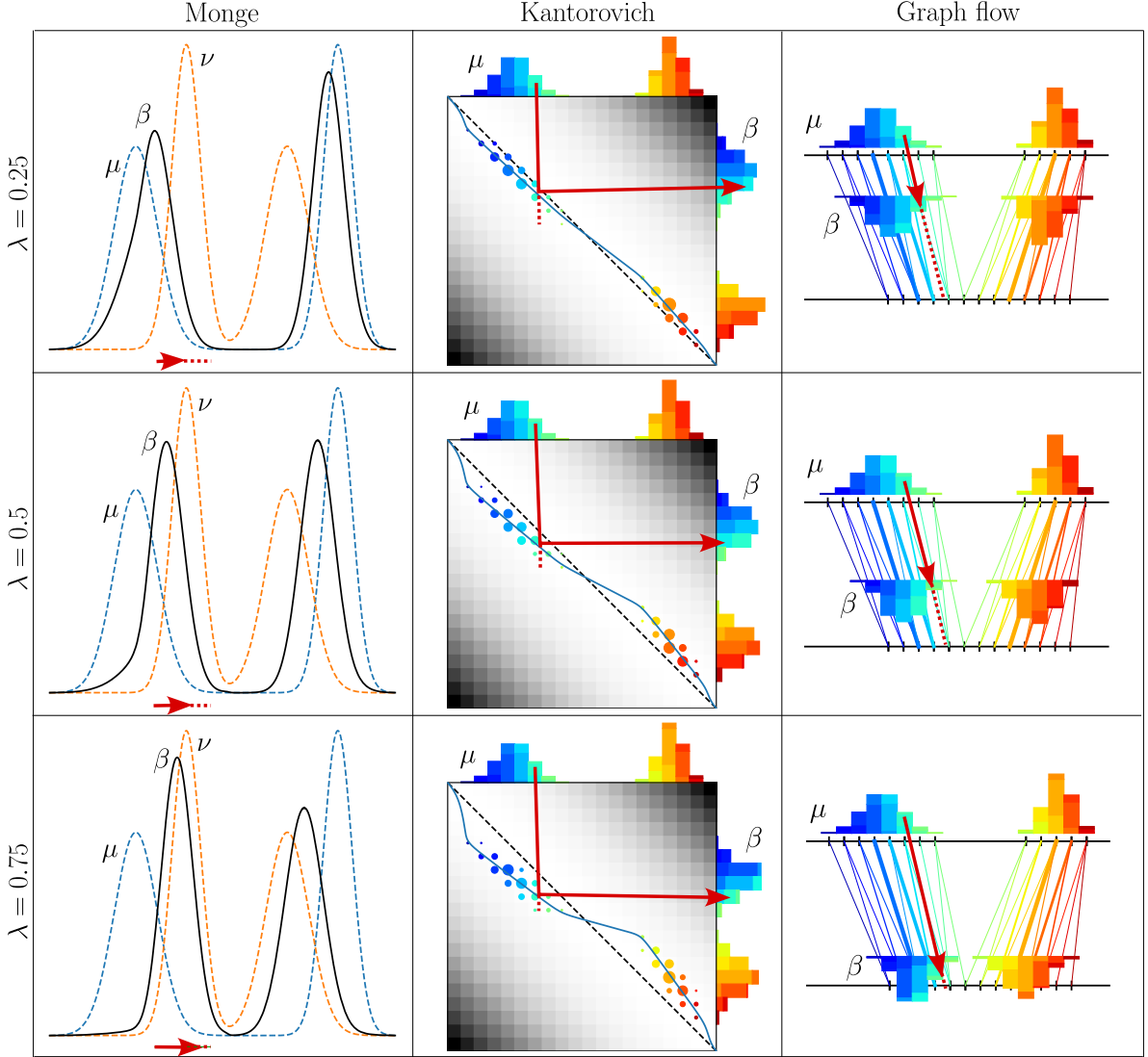


Figure 7: The displacement interpolation (OT barycenter construction) from three perspectives (one per column). The barycenter captures aspects of multiple patterns by striving to be ‘in-between’ them, attempting to be similar to both. With an appropriate distance, this can combine information in a meaningful way. The barycenter β will be more μ -like or ν -like according to the weighting coefficient λ . Ordinarily a transport map associates some mass at position x in μ to mass at position y in ν (see Fig. 3). With this association, we can consider continuously moving this mass from its location within μ to its location within ν , along the shortest path between x and y . With $\lambda = 0$, all mass elements are at their source position and constitute μ . With $\lambda = 1$, all mass elements are at their target position and constitute ν . Mass elements in intermediate positions, constitute an intermediate pattern β . In latter columns, a segment of one colour in μ corresponds to that of the same colour in β and in ν . The blue curve in the Kantorovich column illustrates the Monge map from μ to β .

Algorithm 2 Hierarchical Wasserstein centroid approximation

```

function HWCA( $W = (w^1, \dots, w^{N_M})$ ,  $M = (\mu^1, \dots, \mu^{N_M})$ )
     $\triangleright$  For the sake of brevity, let  $N_M = |M|$  be a power of two.
     $\triangleright$  Take  $W, M$  to be sorted such that weights are decreasing.

    while  $|M| > 1$  do
         $M' \leftarrow (), W' \leftarrow ()$ 
        for  $k$  from 1 to  $\frac{|M|}{2}$  do  $\triangleright$  Can be run in parallel
             $w'_k \leftarrow w_{2k-1} + w_{2k}$ 
             $\mu'_k \leftarrow \bar{B}_{W_2^{d_X}}(\mu^{2k-1}, \mu^{2k}, w_{2k-1}/w'_k)$ 
        end for
         $M \leftarrow M', W \leftarrow W'$   $\triangleright$  Note, these can be edited in place
    end while
    return  $\mu^1$ 
end function

```

3.2 Separation

Recall the notation of Section 2.2.1. Discretising and writing in terms of abundance scaled versions of the transport maps $\bar{\mathbf{P}}^i = a_i \mathbf{P}^i$ between b^i and $\tilde{b}^i$, the discrete W_2 separation subproblem (see Eq. (14)) becomes

$$\begin{aligned}
 & \min_{\bar{\mathbf{P}}^1, \dots, \bar{\mathbf{P}}^K \in \mathbb{R}_+^{N \times N}} \sum_{i=1}^K \langle \mathbf{D}^i, \bar{\mathbf{P}}^i \rangle \\
 \text{s.t.} \quad & (\bar{\mathbf{P}}^i)^T \mathbf{1} = \|\bar{\mathbf{P}}^i\|_1 \mathbf{w}_{b^i} \quad \forall i \in [K] \quad \text{and} \quad \sum_{i=1}^K \bar{\mathbf{P}}^i \mathbf{1} = \mathbf{w}_c
 \end{aligned} \tag{17}$$

where the entries of $\mathbf{D}^i$ are given by $d_{kl}^i = d_X(\mathbf{x}_{b^i}^k, \mathbf{x}_c^l)^2$ and all measures have been taken to have the same size support for brevity⁸. The combination constraint

$$\sum_{i=1}^K a_i \tilde{b}^i = c \implies \sum_{i=1}^K a_i \sum_{k=1}^N w_{b^i}^k \delta_{\mathbf{x}_{b^i}^k} = \sum_{k=1}^N w_c^k \delta_{\mathbf{x}_c^k} \tag{18}$$

gives that the varied basis patterns cannot have weight where the mixture does not or $\text{supp}(\tilde{b}^i) \subset \text{supp}(c)$, hence we can set $\mathbf{x}_{b^i}^k = \mathbf{x}_c^k$. And so, during the separation subproblem, minimisation with respect to weights alone is sufficient. The weight vector $\mathbf{w}_{b^i}^\sim$ is obtained from the marginal $\bar{\mathbf{P}}^i \mathbf{1} / \|\bar{\mathbf{P}}^i\|_1$.

Concatenating

$$\begin{aligned}
 \tilde{\mathbf{D}} &= (\mathbf{D}^1 \mathbf{D}^2 \dots \mathbf{D}^K) \\
 \tilde{\mathbf{P}} &= (\bar{\mathbf{P}}^1 \bar{\mathbf{P}}^2 \dots \bar{\mathbf{P}}^K) \\
 \mathbf{w}_c^T &= (a_1 \mathbf{w}_{b^1}^T \dots a_K \mathbf{w}_{b^K}^T),
 \end{aligned} \tag{19}$$

we can now express the separation as

$$\begin{aligned}
 & \min_{\mathbf{a} \in \mathbb{S}^{K-1}} \langle \tilde{\mathbf{D}}, \tilde{\mathbf{P}} \rangle \\
 \text{s.t.} \quad & \tilde{\mathbf{P}}^T \mathbf{1} = \mathbf{w}_c^\sim \quad \text{and} \quad \tilde{\mathbf{P}} \mathbf{1} = \mathbf{w}_c
 \end{aligned} \tag{20}$$

⁸Uneven measures can be trivially padded with zero weight Diracs.

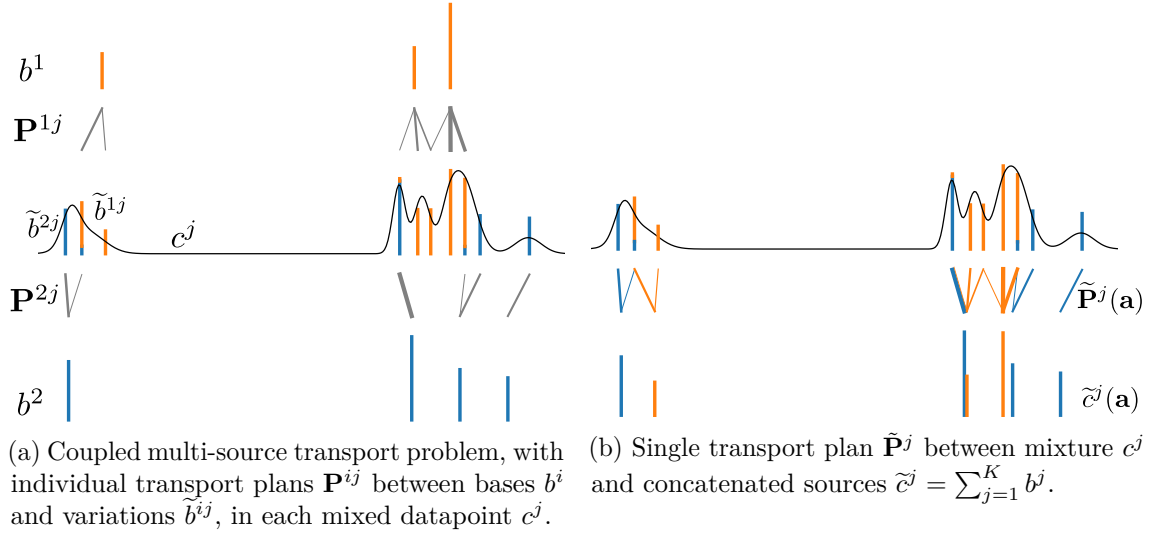


Figure 8: The segmentation problem as it emerges from the variation cost in the Wasserstein context (a), minimising with respect to multiple coupled plans, and an equivalent, single transport plan/distance evaluation simplification (b). Note, in the former case we inherently discover the abundances $\mathbf{a}$ when optimising for unnormalised b^j , but in the latter case we must evaluate the distance/plan for each choice of $\mathbf{a}$. However, since we can construct exact plans for the ordinary 1D optimal transport problem easily, now we only have to optimise for $\mathbf{a}$.

where $\tilde{c}(\mathbf{a}) = \kappa^{\text{lin}}(\mathbf{a}; \mathbf{B})$, as supposed to $c = \kappa^{\text{lin}}(\mathbf{a}; \Phi \mathbf{B})$, so we can in fact write

$$\min_{\mathbf{a} \in \mathbb{S}^{K-1}} W_2^{d_X}(\tilde{c}(\mathbf{a}), c)^2 \quad (21)$$

which is a convex optimisation problem that benefits from being able to efficiently evaluate W_2^2 and its optimal transport map.

For the initial retrieval of abundances, this optimisation is indeed equivalent to $W_2^{d_X}$ regression in the linear model, such as in Sandler and Lindenbaum (2011), however, we do not stop there (see Algorithm 3), as we are not after a linear representation, we are after the transformed basis patterns inferred from the unconstrained marginals. If the initial distance computation cannot expose the concatenated transport plan, we must use another method to recover it afterwards. We then separate this into individual plans (see Fig. 8). Note the advantage in non-overlapping supports here, where some positions can be associated with a particular basis pattern and therefore transport to/from such positions in the combined map can be identified as transport to/from the corresponding pattern.

Algorithm 3 OT separation (unmixing with known basis)

```

function SEPARATE( $\mathbf{B}, c$ )
   $\mathbf{a} \leftarrow \arg \min_{\bar{\mathbf{a}}} W_2^{d_X}(\tilde{c}(\bar{\mathbf{a}}), c)^2$ 
   $(\bar{\mathbf{P}}^1 \bar{\mathbf{P}}^2 \dots \bar{\mathbf{P}}^K) \leftarrow \text{DERIVEOTP}_{d_X}(\tilde{c}(\mathbf{a}), c)$ 
  for  $i \in \{1, \dots, K\}$  do
     $\mathbf{w}_{\tilde{b}^i}^T \leftarrow \bar{\mathbf{P}}^i \mathbf{1} / \|\bar{\mathbf{P}}^i\|_1$ 
  end for
end function

```

3.3 Proposed unmixing algorithm

Given a method for efficiently computing transport plans (see Section 4.1 for 1D) and provided that centroids can be computed (or well approximated) in the given GS (in a feasible timescale), such as via a hierarchical approach, we arrive at the general unmixing algorithm Algorithm 4. This algorithm is highly parallelisable⁹, as due to the degree of decoupling in the problem, there are many independent subproblems and for large datasets, GPUs can be leveraged to solve a larger subset of these concurrently. We provide a concrete implementation for the $(P(\mathbb{R}), W_2)$ case.

Algorithm 4 Warping robust unmixing for discrete measures

```

function W2UNMIX( $C, N^{\text{iter}}, K$ )
  Generate random initial  $\tilde{B}$  (random weights over each  $x \in \text{supp}(c)$ )
  for  $k \in \{1, \dots, N^{\text{iter}}\}$  do
    for  $i \in \{1, \dots, K\}$  do ▷ Averaging step
       $b^i \leftarrow \text{HWCA}(\mathbf{a}_i, (\tilde{b}^{i1}, \dots, \tilde{b}^{iM}))$ 
    end for
    for  $c \in C$  do ▷ Can be performed independently per mixture
      SEPARATE( $B, c$ )
    end for
  end for
  return  $\tilde{B}, B$ 
end function

```

3.4 Variations of the unmixing problem

Depending on what information is originally known, the problem of interest may in fact be more constrained. For instance our separation step provides a solution to the problem of separating components when the basis is known. For partially blind unmixing, the known patterns are included in the initial basis instead of fully random data and can be kept fixed by skipping the corresponding basis update. However, if we wish to detect novel patterns, as one often may in materials discovery, these will not have comparably weighted clusters in our geometry, as these patterns are effectively outliers and are drowned out by the more common bases. As such, we cannot simply use the partially blind approach in this case. A dedicated anomaly detection approach is necessary.

Problem	Amount of basis known	Primary goal
Separation (non-blind)	All	For each mixture, determine parts likely arising from each basis
Partially blind unmixing	Some	Find remaining basis patterns
Fully blind unmixing	None	Find basis B
+ Anomaly detection	-	Identify datapoints containing trace uncommon patterns

Table 1: Table of variations of the unmixing problem. The anomaly detection relaxation can be optionally made from any of the first three variants, it is treated separately from partially blind unmixing since a sought additional component may be too statistically insignificant to estimate (though may be indirectly detected).

The presence of sufficiently novel patterns can still be detected through our usual formulation,

⁹The averaging step is broken in to a series of pairwise reduction passes, these can be run in parallel.

via an uncharacteristically large segmentation cost for a particular mixture. This can be seen as attempting to fit a mixture of $K + 1$ distinct basis patterns with only a K subset, will involve additional deformation of these K basis patterns. By constraining the deformation of these basis patterns, to less than what would minimally be required for fully fitting with a K subset, we can attempt to estimate the unknown through the residual, after a best partial fit. Again, residuals need only be enabled for those mixtures where anomalies are reasonably suspected.

This changes the separation problem to the following linear program

$$\begin{aligned}
& \min_{\substack{\bar{\pi}^i \in M_+(\mathbb{R} \times \mathbb{R}) \\ \bar{\eta} \in M_+(\mathbb{R})}} \sum_{i=1}^K \langle d, \bar{\pi}^i \rangle + \beta \|\bar{\eta}\|_1 \\
& \text{s.t.} \quad (\text{proj}_x)_\# \bar{\pi}^i = \|\bar{\pi}^i\|_1 b^i \quad \forall i \in [K], \quad (\text{proj}_y)_\# \sum_{i=1}^K \bar{\pi}^i + \bar{\eta} = c \\
& \quad \text{and} \quad \langle d, \bar{\pi}^i \rangle \leq \overline{\Delta^2}
\end{aligned} \tag{22}$$

where the $\|\bar{\eta}\|_1 = \bar{\eta}(\mathbb{R}) = \varepsilon$ penalty discourages unnecessarily attributing mass to η , and $\overline{\Delta^2}$ is a specified maximum average squared displacement, set to a value less than the minimum of the unconstrained segmentation cost. In the discrete linear program, we can alternatively (or additionally) enforce a maximum displacement constraint, this can be achieved efficiently by disregarding variables $\bar{p}_{kl}^i$ that correspond to $d_{kl}^i > \Delta^2$, reducing the amount of variables that need to be solved for.

4 Fast Wasserstein

Here we will cover two ways in which to exploit structure in the OT problem, for a fast exact and an approximate distance calculation.

4.1 The north west corner algorithm

Definition 6 (Monotone increasing). A transport plan π is monotone increasing if

$$\langle x' - x, y' - y \rangle \geq 0 \quad \forall (x, y), (x', y') \in \text{supp}(\pi). \tag{23}$$

In the 1D case, the monotonicity of transport plans fully constrains them, such that no minimisation is required. Further, for GCs that admit monotone increasing solutions, such as the ED, this unique feasible map can be constructed through the NWCA (in $\mathcal{O}(N)$ time). Though this fact is known in OT literature Bai et al. (2023), it does not appear to be always exploited downstream.

Lemma 1. *Though a Monge map is not sufficient to convey optimal transport between a pair of general measures (such that $\text{supp}(\pi^*) \subseteq \text{graph}(T^*)$), a parameterised trace $(x, y) = (T_x(z), T_y(z))$ —with monotone T_x, T_y —is in 1D with the ED GC. Similarly, in the discrete case, a sequence (x_μ^i, x_ν^j) with monotone x_μ^i and x_ν^j is sufficient.*

Proof. In the discrete case we have the following. Take the points in the support of the optimal (Kantorovich) map in any initial order, forming the sequence $z^i = (x^i, y^i) \in \text{supp}(\pi^*)$. Suppose there are no z^i with the same x value, then we can order these such that $\{x^i\}$ are strictly increasing. Since the optimal plan is monotone increasing π^* , $(x^{i+1} - x^i)(y^{i+1} - y^i) \geq 0$ must hold, where $x^{i+1} - x^i > 0$ by ordering and so $y^{i+1} - y^i \geq 0$, hence $\{y^i\}$ are non-decreasing.

More generally, we may order such that $\{x^i\}$ are non-decreasing, containing both strictly increasing and constant subsequences. Over strictly increasing subsequences of x , we have non-decreasing y as before. By uniqueness of z^i (points in the support), y cannot also be equal

for constant x subsequences, so over these, z^i can be ordered such that y^i are strictly increasing. Finally, note that for z in a preceding strictly increasing sequence, and z' the next subsequence, $x' - x > 0$, hence all y in the preceding subsequence are lower than those in the next.

Hence the support of any optimal transport map (for discrete, euclidean gd, 1D) can be expressed through two non-decreasing subsequences over the GS. Similarly, in the limit we require monotone T_x, T_y in the continuous case. $\square$

Corollary 1. *If x_μ^i and x_ν^j are ordered, then it follows from the monotonicity of the support trace that*

$$p_{ij}p_{i'j'} > 0 \quad \text{with} \quad i \leq i' \quad \Rightarrow \quad j \leq j'. \quad (24)$$

Definition 7 (Slacks).

$$s_\mu^i = w_\mu^i - \sum_{j=1}^M p_{ij} \quad \text{and} \quad s_\nu^j = w_\nu^j - \sum_{i=1}^N p_{ij}, \quad (25)$$

where $s_\mu^i = 0$ signals the i^{th} row constraint being met.

Definition 8 (North-west corner rule (NWCR)). Set the north-west corner of the transport plan to $p_{11} = \min(w_\mu^1, w_\nu^1)$, so that at least one of $s_\mu^1 = 0$ or $s_\nu^1 = 0$ is satisfied.

Definition 9 (North-west corner algorithm (NWCA)). Having satisfied at least one row or column constraint, we move on to determine the remainder of the transport plan, which is structured

$$\mathbf{P}(\mu, \nu) = \begin{pmatrix} p_{11} & \mathbf{0} \\ \mathbf{P}(\mu^1, \nu^1) \end{pmatrix} \quad \text{or} \quad \mathbf{P}(\mu, \nu) = \begin{pmatrix} p_{11} & \mathbf{P}(\mu^1, \nu^1) \\ \mathbf{0} \end{pmatrix} \quad (26)$$

where $\mu^1 = \mu - p_{11}\delta_{x_\mu^1}$ and $\nu^1 = \nu - p_{11}\delta_{x_\nu^1}$. To determine $\mathbf{P}(\mu^1, \nu^1)$, and therefore $\mathbf{P}(\mu, \nu)$, we iterate this rule until $\mu^K(\mathcal{X}) = \nu^K(\mathcal{X}) = 0$. This yields a feasible transport plan (as all slacks are set to 0) in at most $N + M - 1$ steps¹⁰ (see 3.4.2 in Peyré and Cuturi (2019)).

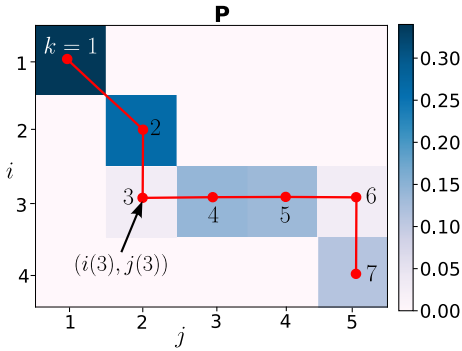
¹⁰Too see why, consider that each step satisfies at least one additional constraint (than the previous). There are a total of $N + M$ constraints, and the final step must satisfy both the last row and column constraints. Alternatively, the longest monotonic trace through the matrix $\mathbf{P}(\mu, \nu)$ visits $N + M - 1$ entries.

Algorithm 5 North-west corner algorithm

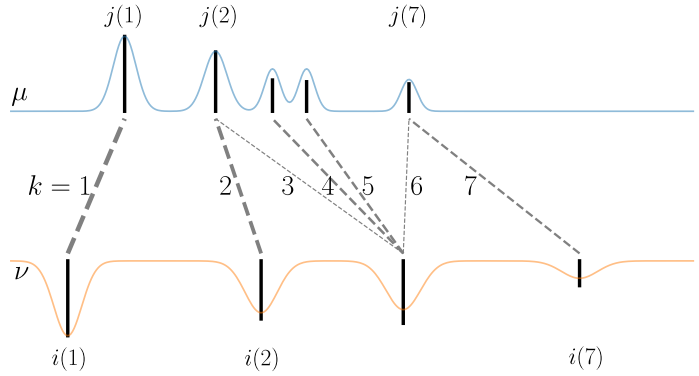
```

function NWCA( $\mathbf{w}_\mu, \mathbf{w}_\nu$ )
   $i_1 \leftarrow 1, j_1 \leftarrow 1, s_\mu \leftarrow w_\mu^1, s_\nu \leftarrow w_\nu^1, \mathbf{P} \leftarrow \mathbf{0}_{N \times M}, \text{KC} = 0$ 
  for  $k = 1$  to  $k = N + M - 1$  do
     $p_{i_k j_k} \leftarrow \min(s_\mu, s_\nu)$ 
     $s_\mu \leftarrow s_\mu - p_{i_k j_k}$ 
     $s_\nu \leftarrow s_\nu - p_{i_k j_k}$ 
     $\text{KC} = \text{KC} + c(x_\mu^{i_k}, x_\nu^{j_k}) p_{i_k j_k}$ 
    if  $s_\mu = 0$  then
       $i_{k+1} \leftarrow i_k + 1, j_{k+1} \leftarrow j_k$ 
      if  $i_{k+1} = N$  then
        break
      end if
       $s_\mu \leftarrow w_\mu^{i_{k+1}}$ 
    end if
    if  $s_\nu = 0$  then
       $i_{k+1} \leftarrow i_k, j_{k+1} \leftarrow j_k + 1$ 
      if  $j_{k+1} = M$  then
        break
      end if
       $s_\nu \leftarrow w_\nu^{j_{k+1}}$ 
    end if
  end for
end function

```



(a) Optimal Kantorovich plan.



(b) Edges corresponding to $p_{ij}^* > 0$ do not cross.

Figure 9: The North-west corner algorithm (Algorithm 5) and the monotonic trace produced, corresponding to monotonic parameterisations $i(k)$ and $j(k)$, or in the continuous case $x(t)$ and $y(t)$. As the name suggests, we start in the north-west corner of the transport plan, and at each step k , compare how much mass is left at each marginal index i and j , transfer the most mutually available, and increment the saturated index/indices (i , j , or both). Note, the horizontal stepping of the Kantorovich plan in the middle panels of Fig. 7, how a Monge function cannot do this, and how a function pair $(x(t), y(t))$ can (see Lemma 1), as one can linger at the same position for some time, e.g. $i(k) = 3$ for $k \in [3, 6]$.

4.1.1 Optimality for 1D sorted measures

For $\mu \in P(\mathbb{R})$, its signature may be sorted so that

$$x_\mu^1 < x_\mu^2 \dots < x_\mu^{N-1} < x_\mu^N \quad (27)$$

noting that if $x_\mu^i = x_\mu^{i'}$, we can simplify $w_\mu^i \delta_{x_\mu^i} + w_\mu^{i'} \delta_{x_\mu^{i'}} = (w_\mu^i + w_\mu^{i'}) \delta_{x_\mu^i}$.

Lemma 2. *Taking 1D μ and ν to be sorted without loss of generality, the NWCR is a necessary condition for a monotone plan.*

Proof. Supposing this is not true, if neither the mass constraint at x_μ^1 or x_ν^1 is satisfied by p_{11} alone then

$$s_\mu^1 > 0 \Rightarrow \exists p_{i1} > 0 \quad \text{and} \quad s_\nu^1 > 0 \Rightarrow \exists p_{1j'} > 0 \quad (28)$$

with $i, j' > 1$, but by monotonicity (recall Corollary 1), since $p_{1j'} p_{i1} > 0$ (where $1 < i$) we must have $j' \leq 1$, which contradicts $1 < j'$. $\square$

Corollary 2. *If $\mathbf{P}(\mu, \nu)$ is to be monotone, $\mathbf{P}(\mu^k, \nu^k)$ must be monotone for all k , hence the application of the NWCA is necessary for $\mathbf{P}(\mu, \nu)$ to be monotone.*

Theorem 1. *In 1D, application of the NWCA on sorted signatures produces an optimal map, provided the GC satisfies $c(x, y) + c(x', y') \leq c(x, y') + c(x', y)$ for all $\langle x' - x, y' - y \rangle \geq 0$.*

Proof. As by Proposition 2.5 of Rieger (2006), there is only one monotone increasing transport plan with marginals μ, ν in 1D and as we require a monotonic increasing plan for optimality, we have the unique optimal solution. $\square$

4.2 Optimal distance approximation quantisers

Though the computation of a complex distance calculation might scale badly in the size of the original patterns compared, we may be able to economically approximate this via a much smaller but nearby quantised counterparts. In other words, there may exist much smaller but similar patterns that are much faster to compare. For instance the ‘signatures’ used in Rubner et al. (2000), generated via k-means clustering from some original distribution (which in practice will also be discrete but larger).

It is well known that k-means clustering in a distribution is equivalent to looking for its W_2 optimal k point quantiser Lloyd (1982); Pollard (1982); Canas and Rosasco (2012) (see Section B for a quick justification). For the OT case this can be thought of as approximating the cost of delivering some goods between two sets of consumers, by the cost of the transport between major warehouses/hubs, where so long as the hubs are appropriately placed, this will represent the bulk of the overall transport (scale separation Shirdhonkar and Jacobs (2008); Bronstein et al. (2021)).

Let the closest quantised pattern of at most size N be

$$\begin{aligned} \bar{\mu}_N &= \arg \min_{m \in \mathcal{P}} d_{\mathcal{P}}(\mu, m) \\ \text{size}(m) &\leq N, \end{aligned} \quad (29)$$

where for a space of measures we can take $\text{size}(m) = \text{supp}(m)$. So long as $\lim_{N \rightarrow \infty} d_{\mathcal{P}}(\bar{\mu}_N, \mu) = 0$, then $\exists N'$ for which

$$|d_{\mathcal{P}}(\mu, \nu) - d_{\mathcal{P}}(\bar{\mu}_{N'}, \bar{\nu}_{N'})| \leq d_{\mathcal{P}}(\bar{\mu}_{N'}, \mu) + d_{\mathcal{P}}(\bar{\nu}_{N'}, \nu) < 2\tau, \quad (30)$$

from the triangle inequality. Additionally, with monotonically improving proxies $d_{\mathcal{P}}(\bar{\mu}_{N+1}, \mu) < d_{\mathcal{P}}(\bar{\mu}_N, \mu)$, the distance approximations also improve monotonically. Appendix Section B.1 shows that these conditions do indeed hold for W_2 .

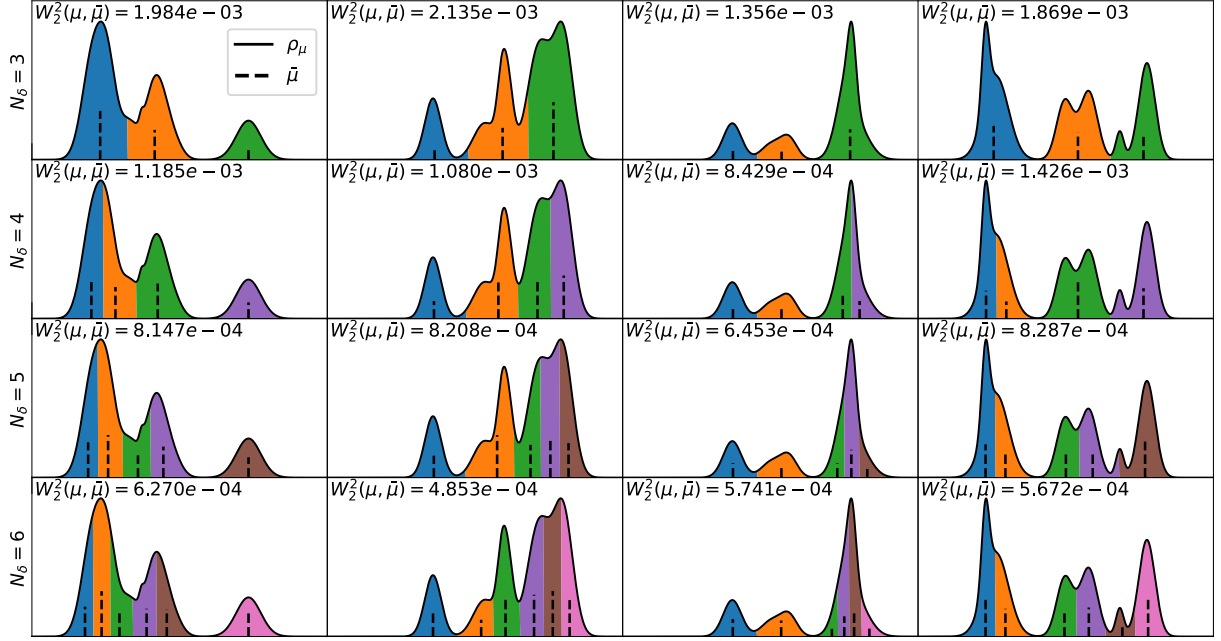


Figure 10: Quantisation of various continuous measures μ with density ρ_μ (one per column), into a discrete measure $\bar{\mu}$ of various numbers of Diracs N_δ (one per row). The colored region surrounding a dashed stick denotes the continuous area that is condensed into that stick, with taller sticks (larger weight) having absorbed larger areas. With lower quantisation cost (higher N_δ), we of course have better proxies for estimating the Wasserstein (p=2) distance between different μ s.

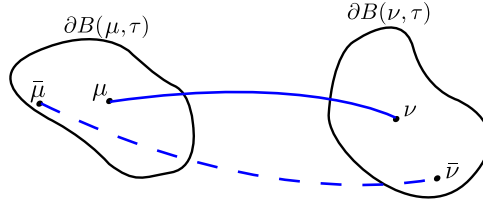


Figure 11: Approximation of $d(\mu, \nu)$ through $d(\bar{\mu}, \bar{\nu})$ where $\bar{\mu}, \bar{\nu}$ are within balls of radius τ (solid black boundary), where $d(\bar{\mu}, \bar{\nu}) \rightarrow d(\mu, \nu)$ as $\tau \rightarrow 0$ by the triangle inequality. For a geodesic distance, the distance $d(\mu, \nu)$ is the length of the geodesic connecting μ and ν (solid blue line, while dashed connects $\bar{\mu}$ and $\bar{\nu}$).

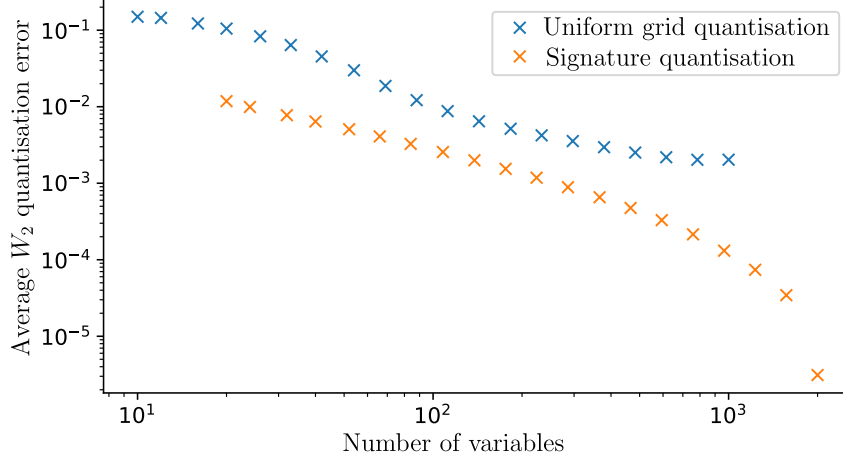


Figure 12: Average error (evaluated over 1000 random GMM pattern pairs) vs variables required for two quantisation choices. GMMs are generated as in Section 5.1, their density is initially sampled on a uniform grid of 1024 points. Signatures require twice the amount of variables for each grid size as we are specifying positions as well as weights (plot is adjusted accordingly). The signature quantisation is achieved through hierarchical agglomerative centroid-linkage clustering.

When using this approximation along with our unmixing framework, a parallel may be noted with multilevel clustering Ho et al. (2017), as we are quantising/clustering both between and within our patterns. Though of course in our case the varied basis distribution must also be learned and our discrete measures have weighted atoms.

In practice, hierarchical agglomerative centroid-linkage provides good proxies, while allowing a quantiser to be built iteratively. Intuitively, similarity to k-means can be understood from noting that repeatedly merging Diracs that cause the smallest change in W_2 is likely to merge measures from the same Voronoi partition/cluster, when all points from a cluster are merged, we obtain a hierarchical approximation of the W_2 average of that cluster (see Section B). The merging continues until a criterion is met, such as target size or a target distance error (resulting in signatures of various sizes per input measure). Fig. 12 suggests that one can represent a distribution more accurately even with half as many “sticks” (Diracs) when compared to a uniform grid downsampling, such that there is better value in controlling positions as well as weights (compared to just a greater number of weights). In the context of the anomaly detection linear program, marginals of half the size results in transport plans of a quarter the size, which may result in solve time speed up factors of 16 – 64 based on $\mathcal{O}(n^2) - \mathcal{O}(n^3)$ scaling for solving a naive linear program (while using more accurate marginals).

5 Experimental evaluation

To demonstrate the flexibility of our method, we performed three experiments, each representing a different mixing context. In Section 5.1, we explore blind unmixing performance across 1200 randomly generated large datasets, featuring various levels of basis variation. The independent variation of each peak represents a more extreme version of what might otherwise occur in PXRD patterns. In Section 5.2, we explore non-blind unmixing within a dataset corresponding to a more realistic, physically informed, simulation of PXRD patterns. This features a large basis, with some basis patterns being hard to distinguish after variation. In Section 5.3, we explore blind unmixing across three smaller laboratory hyperspectral imaging datasets corresponding to different physical mixing scenarios, each with different resulting mixing models.

We compare against two other general¹¹ unmixing methods. These are NMF and oNMF (with

¹¹General in that they do not make any assumptions corresponding to a particular discipline or kind of spectra.

Kullback-Leibler and Frobenius divergence based costs respectively). During blind unmixing, each method is tested for a constant $N_{\text{iter}} = 60$. In our visualisations, our derived signatures will be represented via kernel density estimation¹², using a Gaussian kernel of specified width.

As it is standard to provide Root Mean Square Error (RMSE) between ground truth and inferred quantities to measure performance, we will do so as well, wherever possible, however, as we are working with signatures, in order to measure these we must first approximate densities and evaluate them on a common grid. In such a case $\mathbf{B}$ refers to this further approximation, while $\mathbf{B}$ is a tuple of measures as usual.

$$\text{RMSE}(\mathbf{B}, \mathbf{B}') = \sqrt{\frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M (b_{ij} - b'_{ij})^2} \quad (31)$$

To facilitate a more direct comparison, we will define a dissimilarity based a shift invariant modification of W_2 , which we will refer to as “centered Wasserstein”

$$\begin{aligned} \bar{W}_2(\mu, \nu) &= W_2(T_{\#}^{-\mu_b} \mu, T_{\#}^{-\nu_b} \nu) \\ T^{\Delta x}(x) &= x + \Delta x \quad \text{with} \quad \mu_b = \int x d\mu(x). \end{aligned} \quad (32)$$

When ground truth is available we will use the following metrics to gauge performance, ground truth quantities will be decorated with a “t” subscript.

1. Inter Basis Centered Wasserstein (IBCW)

$$\text{IBCW}(\mathbf{B}, \mathbf{B}') = \frac{1}{K} \sum_{i=1}^K \bar{W}_2(b^i, b'^i) \quad (33)$$

We can similarly define inter component and basis-component variants, though these seemed to follow similar trends in our experiments.

2. Average absolute difference in percentage abundances (AADPA).

$$\frac{100}{KM} \sum_{i=1}^K \sum_{j=1}^M |a_i^j - (a_t)_i^j| \quad (34)$$

Though the RMSE between abundances is perfectly appropriate, this is offered as a more interpretable alternative.

As a standard oNMF implementation can only shift basis patterns to the right, the learned basis patterns will be offset, we center these in our plots and for RMSE calculations. Also note that with discrete convolutions we can only broaden and not thin patterns, so the basis patterns learned correspond to thinnest, rather than average variations.

5.1 Random GMM datasets

Suppose our patterns are 1D GMM’s. We express our random measures with respect to the Gaussian density $g(x; \mu, \sigma)$ for mean μ and standard deviation σ

$$\rho_{b^i}(x) = \sum_{k=1}^{N_G} I_k^i g(x; \mu_k^i, \sigma_k^i), \quad (35)$$

¹²If our signatures are taken to approximate continuous objects, this process approximates them by spreading out mass that was previously concentrated during discretisation.

where a continuous measure ν can be related to its density via $\nu(S) = \int_S \rho_\nu(x) dx$ for $S \subset \mathbb{R}$. This distribution is itself randomly drawn by randomly selecting

$$I_k^i \sim |\mathcal{N}(1, 0.25)|, \quad \mu_k^i \sim U(0.2, 0.8) \quad \text{and} \quad \sigma_k^i \sim 0.02 \cdot 2^{\mathcal{N}(0, 0.5)}. \quad (36)$$

Peak intensity is denoted by I (where we must normalise $\sum_{k=1}^{N_G} I_k^i = 1$), location by μ and breadth by σ . Variation is then induced by subsequently perturbing the parameters of each peak k away from those of the basis patterns, independently, per each warped basis j , per each mixture i

$$\rho_{b^{ij}}(x) = \sum_{k=0}^{N_G} I_k^i g(x; \mu_k^{ij}, \sigma_k^{ij}) \quad (37)$$

where

$$\mu_k^{ij} = \mu_k^i + \Delta\mu_k^{ij}, \quad \sigma_k^{ij} = \sigma_k^i \cdot \gamma_k^{ij}. \quad (38)$$

with

$$\Delta\mu_k^{ij} \sim \mathcal{N}(0, \sigma_{\Delta\mu}) \quad \text{and} \quad \gamma_k^{ij} \sim 2^{\mathcal{N}(0, \sigma_p)}. \quad (39)$$

The exponential is used as we would like a broadening such as $\sigma = \sigma \times 2$ to be roughly as likely as thinning by the same factor $\sigma = \sigma/2$. Mixed examples can then be created by combining $c^j = \sum_{i=1}^K h_i^j \tilde{b}^{ij}$ where $\mathbf{h}^j$ is drawn from a uniform distribution on S^{K-1} . We perform our assessment across three different basis variation strengths summarised in Table 2.

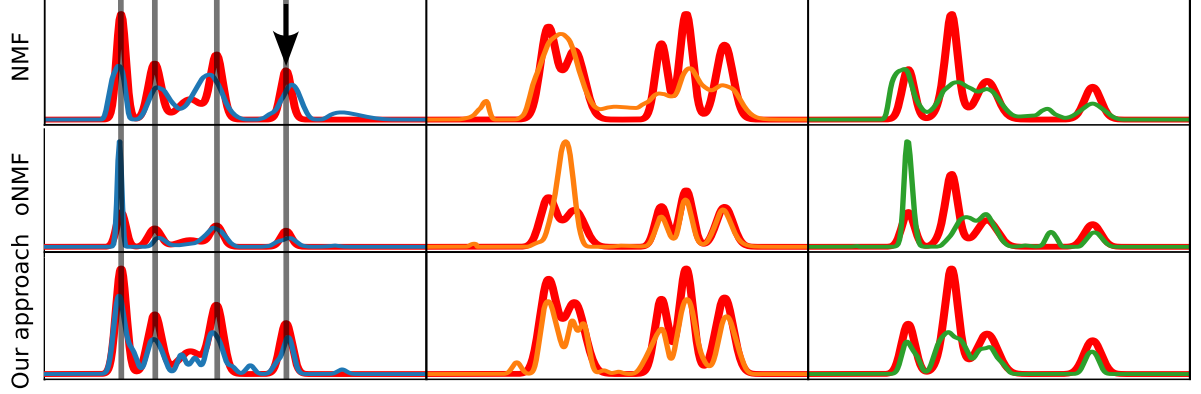
Preset	$\sigma_{\Delta\mu}$	σ_p	95% of $\Delta\mu_k^{ij}$ are in	95% of σ_k^{ij} are in
Low	0.015	0.15	$[-0.03, +0.03]$	$[0.81\sigma_k^i, 1.23\sigma_k^i]$
Medium	0.025	0.3	$[-0.05, +0.05]$	$[0.66\sigma_k^i, 1.52\sigma_k^i]$
High	0.035	0.5	$[-0.07, +0.07]$	$[0.50\sigma_k^i, 2.00\sigma_k^i]$

Table 2: Basis variation strength presets used, where two variations of the same peak can be found as far as 14% of the domain width apart.

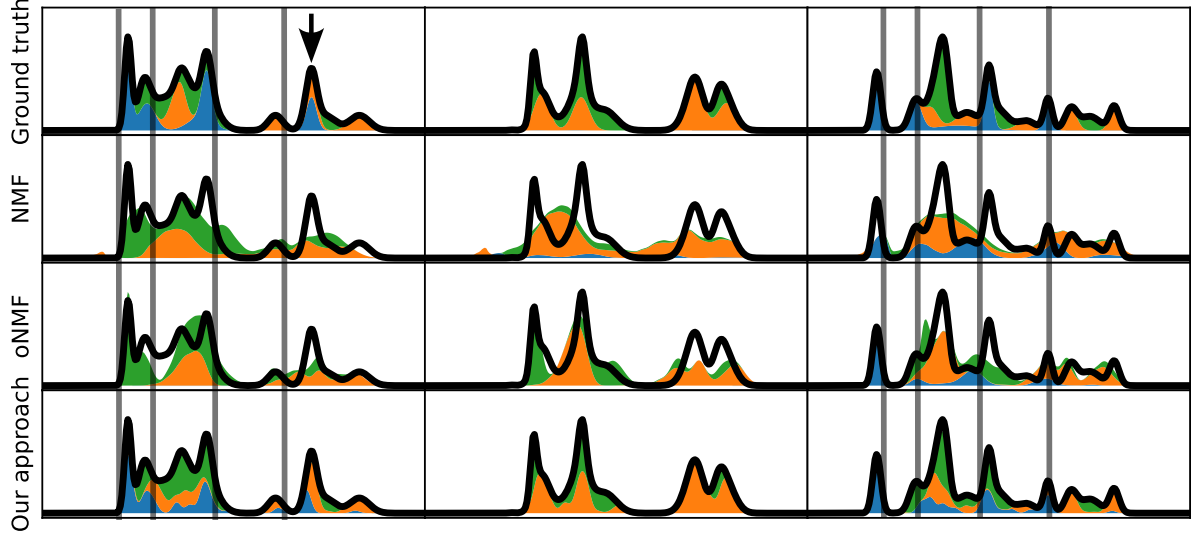
Variability	Approach	Metric			
		AADPA($\mathbf{A}, \mathbf{A}_t$)	RMSE($\mathbf{A}, \mathbf{A}_t$) $\times 100$	RMSE($\mathbf{B}, \mathbf{B}_t$) $\times 1000$	IBCW($\mathbf{B}, \mathbf{B}_t$) $\times 100$
Low	NMF	5.783 $\pm$ 2.469	7.909 $\pm$ 3.421	1.742 $\pm$ 0.569	1.294 $\pm$ 0.602
	oNMF	8.300 $\pm$ 2.921	10.882 $\pm$ 3.722	4.611 $\pm$ 1.454	3.559 $\pm$ 1.507
	Our approach	4.274 $\pm$ 1.973	5.544 $\pm$ 2.757	1.668 $\pm$ 0.485	1.756 $\pm$ 0.894
Medium	NMF	10.247 $\pm$ 3.800	14.142 $\pm$ 5.120	3.130 $\pm$ 0.926	2.262 $\pm$ 1.056
	oNMF	11.955 $\pm$ 3.134	15.539 $\pm$ 3.846	7.138 $\pm$ 1.710	5.220 $\pm$ 1.771
	Our approach	6.019 $\pm$ 2.064	8.154 $\pm$ 2.932	2.574 $\pm$ 0.587	1.837 $\pm$ 0.853
High	NMF	14.167 $\pm$ 4.023	19.276 $\pm$ 5.180	4.271 $\pm$ 1.010	3.299 $\pm$ 1.243
	oNMF	14.413 $\pm$ 2.928	18.743 $\pm$ 3.550	8.723 $\pm$ 1.636	5.769 $\pm$ 1.589
	Our approach	8.371 $\pm$ 2.493	11.509 $\pm$ 3.451	3.625 $\pm$ 0.732	2.154 $\pm$ 0.800

Table 3: Mean decomposition metrics over 400 random GMM datasets per variability preset. These were generated with $K = 3$ and $M = 800$, where the density of each c^j is initially sampled uniformly with $N = 256$ then discretised with $N_\delta = 64$. The best of eight decompositions with different initialisation (as measured by final cost), is taken for each dataset/method. The ground truth basis ($\mathbf{B}_t$) is also similarly sampled (producing $\mathbf{B}_t$), as is the kde density of the inferred basis signature for basis RMSEs. The oNMF is performed with a maximum shift of 25 gridpoints and the inferred basis patterns are centered by shifting for 13 gridpoints before measuring RMSE($\mathbf{B}, \mathbf{B}_t$).

We can see from Table 3 that sometimes the oNMF model can perform worse than NMF in extracting the information we are after, this is despite the fact that our oNMF solutions are



(a) The basis attained with each decomposition method tested, one method per row. Each column corresponds to one basis pattern, which also has an associated color (e.g. b^1 in blue), while the ground truth basis patterns are overlaid in red.



(b) Mixed datapoints c^j (black line), one per column, as a sum of transformed bases $\tilde{b}^{ij}$ (area with coloring corresponding to b^i), according to ground truth, and the tested methods, one per row.

Figure 13: Visual comparison of decompositions of one randomly generated GMM dataset (high variation preset), according to each tested method. Signatures are visualised by smoothing with a Gaussian kernel with $\sigma_{\text{KDE}} = 0.004$. Note that the mixture in the first column of (b), has a component of the first basis (blue) in which some peaks are found further to the left and others are found further to the right, when compared to the left column of (a). Also, note that the mixture in the third column of (b), has a component with a different set of non-uniform shifts. Therefore, for a decomposition method to be able to explain both of these mixtures with respect to the same basis, it must have the ability to perturb each peak independently.

initialised with NMF solutions. Even if the overall fit can improve, the accuracy of inferred constituents can simultaneously suffer (see Fig. 13). In fact, the more flexible a composition model, the more trivially a profile can be fit, regardless of basis. For instance, if one were to use OT as a basis variation model $\phi(T; b) = T_{\#}b$, one could arbitrarily warp any given basis to perfectly fit any mixture, not giving us any indication as to whether one basis is more appropriate over another. This supports our original reasoning for defining a model driven by a non-reconstruction cost.

5.2 PXRD unmixing

As an indicator of suitability to PXRD patterns, we can generate a benchmarking dataset with the Synthetic instance generator Bras et al. (2014), which provides a more physically informed model of such data (compared to GMMs). Similarly to the previous, this too features non-uniform shifting and broadening. The default preset of this generator (the AlliFe system) features a relatively large basis (6 patterns, corresponding to pure phases). As we found that none of the methods we compare here could *consistently* blindly discover all 6 of the basis patterns, we assess separation/abundance estimation in the non-blind (see Table 1) regime here, the performance in this task is summarised in Table 4 and illustrated in Fig. 14.

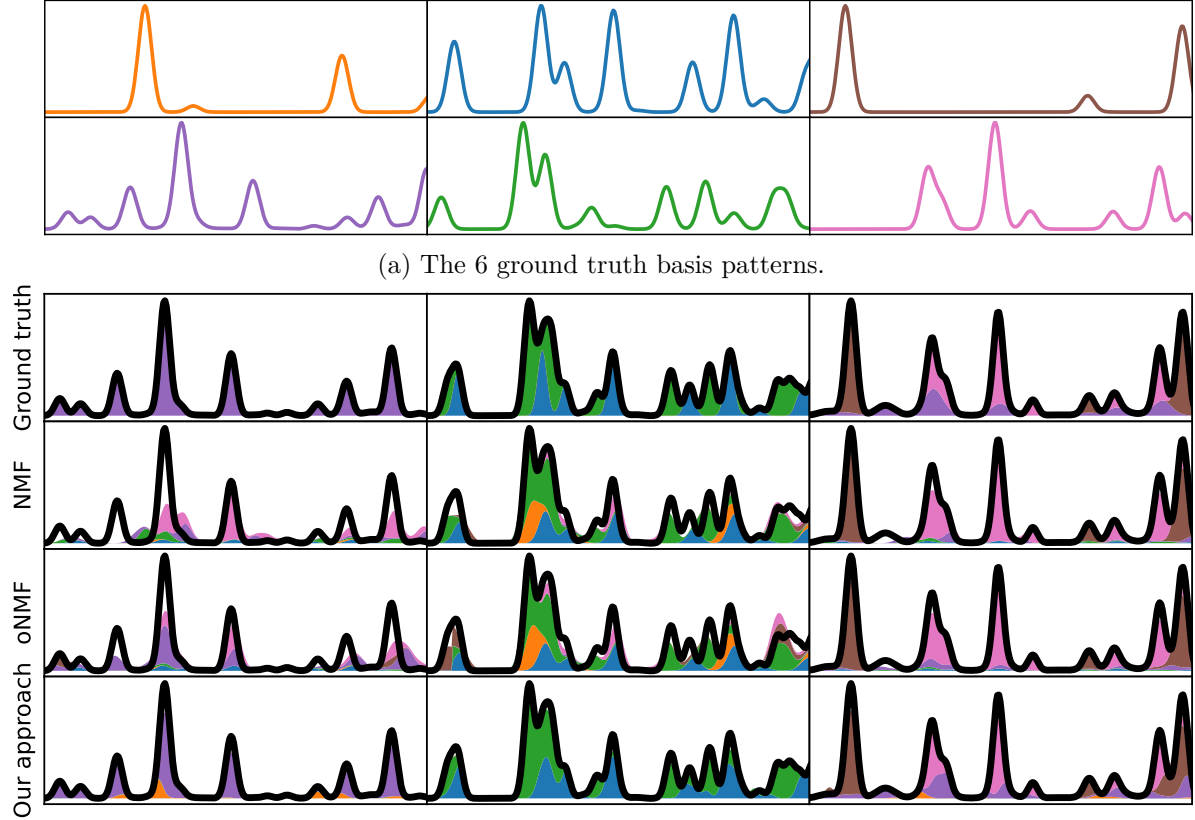


Figure 14: Visual comparison of basis separation (non-blind unmixing) according to the tested methods (one per row), of a Synthetic instance generator dataset with $K = 6$, $m = M = 1275$, $N_\delta = 64$. To visualise signatures, $\sigma_{\text{KDE}} = 0.188$ has been used, x domain (2θ) has been truncated to $[16, 45]$ in these plots.

Table 4 shows significantly lower error in abundance prediction for our approach. Fig. 14 again demonstrates the merit of a non-reconstruction based cost, in that, though the sums of

	Metric	
Approach	AADPA($\mathbf{A}, \mathbf{A}_t$)	RMSE($\mathbf{A}, \mathbf{A}_t$)
NMF	10.467	0.203
oNMF	6.825	0.118
Our approach	3.115	0.055

Table 4: Abundance derivation metrics for the compared approaches (in the separation-only/fixed-basis regime), over the synthetic instance generator dataset.

the components come close to matching the profile of the mixture for NMF and oNMF in the middle column of Fig. 14b, they are not able to correctly relate features of the mixture with features from the basis patterns, and therefore wrongly partition the area under a mixture, and derive less accurate varied bases and abundances. In the right column of Fig. 14b, where the purple pattern cannot be adequately varied by NMF and oNMF, they over assign mass to the similar, pink pattern instead.

5.3 Hyperspectral imaging

This dataset Zhao et al. (2019) contains 3 scenes containing hyperspectral images of colored powders prepared in a laboratory in known amounts to showcase some different mixing scenarios in HI, these are as follows.

Scene 1 (Linear), 3 distinct powders are arranged in a grid of squares smaller than that of the hyperspectral instruments pixels, the subpixel pattern is repeated periodically. A light source points directly at the powders, light generally travels from source, to a single powder (without significant scattering), to a hyperspectral pixel, though rays passing by each of the subpixel squares can make it to the same pixel, adding to the measurement independently.

Scene 2 (Intimate), 4 distinct powders are mixed uniformly, the same ray of light can scatter off of multiple differently colored grains before reaching the sensor, as such this contains second-order reflections (and possibly higher) involving every combination of basis patterns.

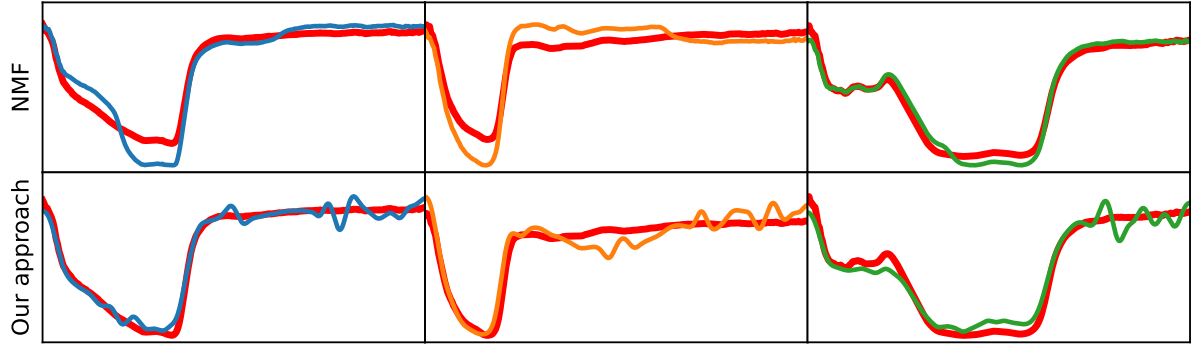
Scene 3 (Limited bilinear), 3 distinct powders are mixed as in the first case, though now the light source points at an adjacent plastic board, such that most rays will go from source, to board, to a subpixel square, to a pixel. This contains second-order reflections of the board with each of the first three basis patterns.

Scene	Approach	Metric			
		AADPA($\mathbf{A}, \mathbf{A}_t$)	RMSE($\mathbf{A}, \mathbf{A}_t$) $\times 10$	RMSE($\mathbf{B}, \mathbf{B}_t$) $\times 1000$	IBCW($\mathbf{B}, \mathbf{B}_t$)
1	NMF	4.561	0.549	2.072	45.269
	Our approach	4.437	0.595	1.947	41.683
2	NMF	11.697	1.486	1.830	35.212
	Our approach	9.787	1.384	1.648	40.198
3	NMF	12.788	1.602	2.458	49.157
	Our approach	3.065	0.459	1.985	40.403

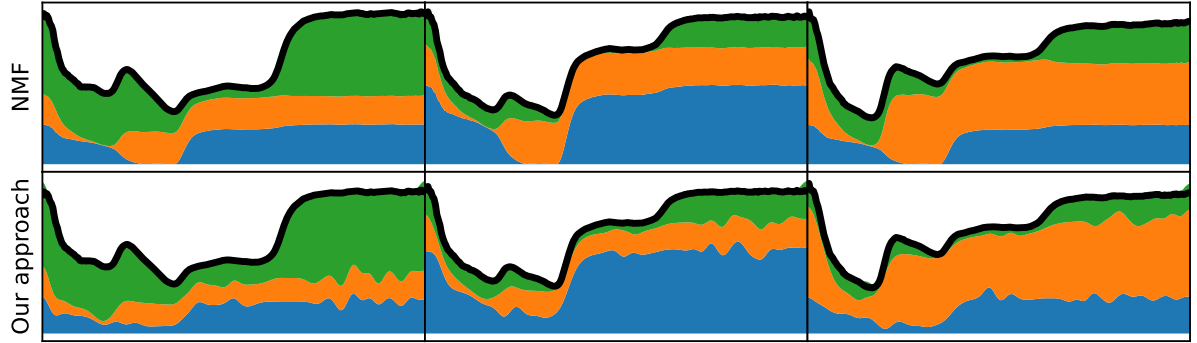
Table 5: Performance summary of NMF vs our method (in the fully blind regime), on a HI dataset. The best of 16 decompositions with different initial starts (as measured by final cost), is taken for each scene/method. oNMF has been omitted in this instance due to instability.

Each scene contains a number of images (10, 14 and 8 respectively), where each image corresponds to different mixing ratios. While each pixel in each image could contain unique basis variation, due to the previously mentioned redundancy, we treat the Euclidean average¹³

¹³The Euclidean average is used simply to avoid resampling for NMF, which requires the inputs to be on the same grid.



(a) The basis attained with NMF vs our method. Each column corresponds to one basis pattern, which also has an associated color (e.g. b^1 in blue), while the ground truth basis patterns are overlayed in red.



(b) Mixed datapoints c^j (black line), one per column, as a sum of transformed bases $\tilde{b}^{ij}$ (area with coloring corresponding to b^i), according to NMF in the top row and our approach in the bottom row.

Figure 15: Visual comparison of decompositions of scene 1. To visualise signatures, $\sigma_{\text{KDE}} = 7.53$ has been used.

pixel of each mixed image as our dataset. Each pixel is a vector with 256 entries representing intensity at each point in the linearly spaced spectral band from 400 to 1000nm, we generate signatures with $N_\delta = 128$ for our method. For each scene we perform fully blind unmixing.

Though a reconstruction based method might have otherwise used any additional degrees of freedom in order to try to provide a closer fit, injecting unnecessary variation (at the expense of ground truth recovery), scene 1 shows (see Table 5) that our method does not suffer from this issue (as we minimise basis variation). Thus we provide similar decompositions in purely linear cases, as the appropriately constrained model, while having flexibility should more complex cases show up.

In cases featuring bilinear contributions $\kappa_{\mathbf{A}}^{\text{bin}}\mathbf{B} = \sum_{p,q} a_{pq}^j b^p \odot b^q$, we can still make progress with a linear combination function, noting that our method has the flexibility to implicitly construct basis variations $\tilde{b}^{ij}$ that are themselves some combination of the first-order and the second-order reflections of b^i (with each of the other basis patterns). The more second-order contributions in c^j , the more $\tilde{b}^{ij}$ deviates from b^i . Our approach somewhat improves on a linear decomposition on scene 2, but improves greatly on a linear decomposition of scene 3 where the presence of second-order reflections is more limited, but still seemingly significant.

6 Conclusion

Many scientific disciplines work with “mixed data,” where each measurement results from a combination of underlying basis patterns that can themselves vary between samples (see Fig. 1). Prevailing unmixing methods largely derive from a specific choice of variation function, meaning that datasets that feature different forms of variation likely need to be processed with different specific methods. If no such method is available, significant adaptation is then required. These methods typically optimize for reconstruction accuracy, that is, how closely the model’s output matches the measured signal. Reconstruction accuracy alone does not provide sufficient guidance, as multiple combinations of inputs might achieve a perfect fit for some mixtures. Not all inputs are equally plausible, yet reconstruction does not distinguish between them. Lastly, pointwise distances are often used in constructing a cost, which can vary erratically with respect to non-pointwise basis variation (see Fig. 5).

Our work introduces a different paradigm for unmixing based on geometric approaches (see Section 2.3.1). One can imagine a set of seed inputs which transform into a latent point cloud after passing through the variation function ϕ , and into the final measurements after passing through the combination function κ . Geometric approaches typically draw conclusions from geometric features of the point cloud of measurements, which can have a particular structure as a result of the mixing model. We take this further by considering all latent point clouds that exactly produce the given measurements, and weigh these by considering the kinds of point cloud that the variation function ϕ can feasibly produce. We show that for any ϕ that is Lipschitz with respect to some distance $d_{\mathcal{P}}$, one can expect clustering in the metric space $(d_{\mathcal{P}}, \mathcal{P})$, where a smaller bound on maximum variation parameters (e.g. shifting/broadening) produces tighter clusters (see Section A). As such, we propose a (in-class) Fréchet variance based cost to search for a latent point cloud that correspond to the least basis variation possible.

To handle common non-pointwise basis variations like shifting and broadening, a corresponding non-pointwise distance metric is required. The Wasserstein distance is a natural choice when signals are treated as distributions. Our practical implementation leverages several key features of OT to ensure efficiency: quantization of distributions into sparse, representative “signatures”; a fast algorithm (NWCA) to compute exact 1D Wasserstein distances and transport maps in linear time; and displacement interpolation for the efficient calculation of Wasserstein centroids (averages). These innovations make our method computationally feasible for exploratory data analysis. As demonstrated through extensive experiments on random GMM, simulated PXRD, and laboratory HI data, our approach consistently outperforms other general-purpose methods

in a range of unmixing scenarios.

This work provides a strong foundation for a wide range of applications and opens several avenues for future research. These include clustering with a full Fisher discriminant by including between-class variance terms, exploring application to higher-dimensional data (this requires fast nD transport map calculation), bidirectional partial matching during anomaly detection, and, finally, exploring other ground distances, including learned distances, or even non-OT distances all together.

Declarations

These results were obtained as part of a PhD funded by Johnson Matthey. The authors acknowledge support from the Leverhulme Trust via the Leverhulme Research Centre for Functional Materials Design (RC-2015-036), UK Engineering and Physical Sciences Research Council (EPSRC) through grant number EP/V026887 as well as through AI for Chemistry: Alchemy Hub (EPSRC grant EP/Y028775/1 and EP/Y028759/1). V. V. Gusev is supported by the Leverhulme Research Fellowship.

Appendix A Clustering and separability of metric space perturbations

Theorem 2 gives us that, so long as $\phi(\lambda; b)$ is Lipchitz with respect to $d_{\mathcal{P}}$, it is possible to guarantee well defined basis variant clusters for any basis $B \in \mathcal{P}^K$ (according to the criteria of Lemma 3), up to some (basis dependent) bound on the parameter λ .

Lemma 3 (Voronoi separable). *With a maximum radius among clusters $\max_{ij} d_{\mathcal{P}}(b^i, \tilde{b}^{ij}) < R$ and adequate basis pattern spacing $\min_{i \neq j} d_{\mathcal{P}}(b^i, b^j) > 2R$ (or maximum similarity), the corresponding basis of some varied pattern can be easily identified, from the Voronoi diagram of the bases*

$$\tilde{b}^{jk} \in X^i \quad \Rightarrow \quad j = i$$

for Voronoi cell X^i (see Eq. (45)), and in fact

$$d_{\mathcal{P}}(b^i, \tilde{b}^{jk}) < R \quad \Rightarrow \quad j = i,$$

where each cluster is enclosed in a ball which is enclosed in its own Voronoi cell, $b^{i*} \subset B_R(b^i) \subset X^i$.

Proof. We require that no variations spill into a neighbouring Voronoi cell, i.e. no two clusters touch/share a common point

$$\min_{\substack{\forall i \neq k \\ \forall j, l}} d_{\mathcal{P}}(\tilde{b}^{ij}, \tilde{b}^{kl}) > 0. \quad (40)$$

From the triangle inequality

$$\begin{aligned} d_{\mathcal{P}}(b^i, b^k) &\leq d_{\mathcal{P}}(b^i, \tilde{b}^{ij}) + d_{\mathcal{P}}(\tilde{b}^{ij}, \tilde{b}^{kl}) + d_{\mathcal{P}}(\tilde{b}^{kl}, b^k) \\ &\Rightarrow d_{\mathcal{P}}(\tilde{b}^{ij}, \tilde{b}^{kl}) \geq d_{\mathcal{P}}(b^i, b^k) - 2R, \end{aligned} \quad (41)$$

holds for all i, j, k, l , hence holds also for the choice

$$\min_{\substack{\forall i \neq k \\ \forall j, l}} d_{\mathcal{P}}(\tilde{b}^{ij}, \tilde{b}^{kl}) \geq \min_{i \neq j} d_{\mathcal{P}}(b^i, b^j) - 2R > 0. \quad \square$$

Corollary 3. *Clearly if $d_{\mathcal{P}}(\phi(\lambda; b^i), b^i) \leq R \quad \forall b^i \in B$ then $\max_{ij} d_{\mathcal{P}}(b^i, \tilde{b}^{ij}) < R$.*

Theorem 2. Let ϕ be Lipschitz continuous s.t. $\exists L_b$ with $d_{\mathcal{P}}(\phi(\lambda; b), \phi(\lambda'; b)) \leq L_b d_{\Lambda}(\lambda, \lambda')$, and let $\phi(\lambda_0; b) = b$ at some λ_0 . Voronoi separability can be guaranteed through a parameter limit $d_{\Lambda}(\lambda, \lambda_0) \leq d_{\Lambda \max}$, so long as $d_{\Lambda \max} < \min_{i \neq j} d_{\mathcal{P}}(b^i, b^j) / 2 \max_i L_{b^i}$.

Proof. We have $d_{\mathcal{P}}(\phi(\lambda; b^i), b^i) \leq L_{b^i} d_{\Lambda}(\lambda, \lambda_0) \leq \max_i L_{b^i} d_{\Lambda}(\lambda, \lambda_0) \leq \max_i L_{b^i} d_{\Lambda \max}$, giving $R = \max_i L_{b^i} d_{\Lambda \max}$. $\square$

Remark 2. With a compact pattern space, $\min_{i \neq j} d_{\mathcal{P}}(b^i, b^j)$ cannot be greater than some $d_{\mathcal{P} \max}$, so it is necessary that $d_{\mathcal{P}}(\phi(\lambda; b), b) \leq R < d_{\mathcal{P} \max} / 2$. For instance, with Variations generated

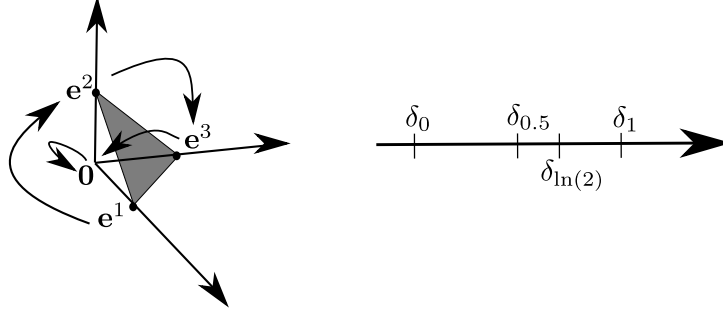


Figure 16: Left: Simplex S^2 , where weights on a static grid such as $\mathbf{x} = (0, 0.5, 1)$ can be represented. The repeated application of the shift-down operation (initially on $\mathbf{e}^1$) is illustrated. Right: The Dirac measures $\delta_x \in P(\mathbb{R})$ representing distributions with weights concentrated at a single point x , which we can continuously interpolate between.

by discrete shifting where $\phi(s; \mathbf{p}) = \mathbf{p}^{\downarrow s} \in \text{Hull}(\{\mathbf{e}^i : i \in [K]\} \cup \{\mathbf{0}\})$ (accounting for mass loss from shifting past the end of a vector) with $s \in \{0, 1, 2, \dots\}$ and $\mathbf{p} \in S^{N-1}$, we can have $d_{\mathcal{P}}(\mathbf{p}^{\downarrow 1}, \mathbf{p}) = d_{\mathcal{P} \max} \not\leq d_{\mathcal{P} \max} / 2$ (see Fig. 16) when $d_{\mathcal{P}}$ is the Euclidean distance, so basis variation as a result of shifting, is not Voronoi separable over $\mathcal{P} = \text{Hull}(\{\mathbf{e}^i : i \in [K]\} \cup \{\mathbf{0}\})$ with the Euclidean distance.

Appendix B Wasserstein quantisation and weighted k-means

Minimising

$$\begin{aligned} \bar{\mu}_N &= \arg \min_{m \in P(\mathcal{X})} W_2^{d_{\mathcal{X}}}(\mu, m)^2 \\ |\text{supp}(m)| &\leq N, \end{aligned} \quad (42)$$

is generally a semidiscrete OT problem¹⁴ for the discrete measure $m = \sum_{i=1}^N w_m^i \delta_{x_m^i}$, where transport maps between μ and m are necessarily of the form $T(x) = x_m^i$ for $x \in X^i$ where X^i are disjoint so that they partition $\mathcal{X}$. The weights are determined from mass conservation $m(\{x_m^i\}) = w_m^i = \mu(T^{-1}(\{x_m^i\})) = \mu(X^i)$, hence Eq. (42) can be stated as

$$\min_{\substack{x_m^i \in \mathcal{X} \\ X^i \subset \mathcal{X}}} \sum_{i=1}^N \int_{X^i} d_{\mathcal{X}}(x, x_m^i)^2 d\mu(x). \quad (43)$$

This can be solved in an alternating fashion, where given regions $\{X^i \forall i\}$, the corresponding optimal x_m^i are centroids¹⁵ of each region

$$x_m^i = \arg \min_{x' \in \mathcal{X}} \int_{X^i} d_{\mathcal{X}}(x, x')^2 d\mu(x) \quad (44)$$

¹⁴Where μ could be discrete as well, but only m is constrained to be.

¹⁵Where $x_m^i = \int_{X^i} x d\mu(x) / \int_{X^i} d\mu(x)$ when $d_{\mathcal{X}}$ is the Euclidean distance

and given $\text{supp}(m) = \{x_m^i \forall i\}$, the optimal regions are then given by the Voronoi tessellation

$$X^i = \{x \in \mathcal{X} : d_{\mathcal{X}}(x, x_m^i) < d_{\mathcal{X}}(x, x_m^j) \forall j \neq i\}. \quad (45)$$

To see this, suppose X^i violated this condition, then for some j , $\hat{X} = \{x \in X^i : d_{\mathcal{X}}(x, x_m^i) > d_{\mathcal{X}}(x, x_m^j)\}$ is non-empty, then one can reassign each $x \in \hat{X}$ according to $\bar{X}^i = X^i \setminus \hat{X}$ and $\bar{X}^j = X^j \cup \hat{X}$ and by doing so improve the cost

$$\begin{aligned} & \int_{\bar{X}^i} d_{\mathcal{X}}(x, x_m^i)^2 d\mu(x) + \int_{\hat{X}} d_{\mathcal{X}}(x, x_m^i)^2 d\mu(x) + \int_{X^j} d_{\mathcal{X}}(x, x_m^j)^2 d\mu(x) \\ & > \int_{\bar{X}^i} d_{\mathcal{X}}(x, x_m^i)^2 d\mu(x) + \int_{\hat{X}} d_{\mathcal{X}}(x, x_m^j)^2 d\mu(x) + \int_{X^j} d_{\mathcal{X}}(x, x_m^j)^2 d\mu(x), \end{aligned} \quad (46)$$

hence violating Eq. (45) is suboptimal. Regardless of the strict optimality of this step, the sequence of updates produced by alternating these steps are non-increasing and the cost is bound below by 0, hence this approach converges to a local minimiser of Eq. (42).

This alternating approach is a generalised, weighted version of the k-means algorithm which can be seen as a generalisation of Lloyds algorithm Lloyd (1982), for which $\mathcal{X} = \mathbb{R}$, $X^i = \{x : x_{\text{mid}}^{i-1} < x < x_{\text{mid}}^i\}$ and $x_{\text{mid}}^i = (x_m^i + x_m^{i+1})/2$ with $x_{\text{mid}}^0 = -\infty$ and $x_{\text{mid}}^N = \infty$.

B.1 Wasserstein quantiser monotonicity and convergence

Theorem 3 (Monotonicity). *Let $\mu \in P(\mathcal{X})$, its optimal W_2 quantisers of size N ($\bar{\mu}_N$) have monotonically decreasing error*

$$W_2^2(\mu, \bar{\mu}_N) > W_2^2(\mu, \bar{\mu}_{N+1}). \quad (47)$$

This follows from being able to construct a suboptimal quantiser of size $N+1$ that is better than the optimal one of size N .

Proof. Let $\mu \in P(\mathcal{X})$ and its optimal quantiser of size N be $\bar{\mu}_N$. Pick $x' \in X^N \setminus \{x^N\}$ and let A, B be the corresponding Voronoi sets of x^N, x' partitioning X^N , we may construct

$$\tilde{\mu}_{N+1} = \bar{\mu}_N - \mu(X^N)\delta_{x^N} + \mu(A)\delta_{x^N} + \mu(B)\delta_{x'}, \quad (48)$$

where we have let $x_{\bar{\mu}_N}^i = x^i$ for brevity. We have that

$$\begin{aligned} \int_{X^N} d_{\mathcal{X}}(x, x^N)^2 d\mu(x) &= \int_A d_{\mathcal{X}}(x, x^N)^2 d\mu(x) + \int_B d_{\mathcal{X}}(x, x^N)^2 d\mu(x) \\ &> \int_A d_{\mathcal{X}}(x, x^N)^2 d\mu(x) + \int_B d_{\mathcal{X}}(x, x')^2 d\mu(x), \end{aligned} \quad (49)$$

since by definition $B = \{x \in X^N : d(x, x') < d(x, x^N)\}$. Hence $W_2^2(\mu, \bar{\mu}_N) - W_2^2(\mu, \tilde{\mu}_{N+1}) > 0$ and so we have monotonicity

$$W_2^2(\mu, \bar{\mu}_N) > W_2^2(\mu, \tilde{\mu}_{N+1}) \geq W_2^2(\mu, \bar{\mu}_{N+1}). \quad \square$$

Theorem 4 (Convergence). *Let $\mu \in P(\mathcal{X})$, its optimal W_2 quantisers of size N ($\bar{\mu}_N$) weakly converge to it.*

$$W_2^2(\mu, \bar{\mu}_N) \rightarrow 0 \quad \text{as } N \rightarrow \infty \quad (50)$$

Proof. Consider a sequence of suboptimal quantisers $\tilde{\mu}_N$ that converge to μ , then optimal quantisers must also converge. Construct such a sequence by noting that at each N , one can pick the largest diameter Voronoi region, where $\text{diam}(X) = \max_{x, y \in X} d_{\mathcal{X}}(x, y)$, over which one

has a 1 Dirac distribution, and by monotonicity we can generate a 2 Dirac quantisation over this region with a better cost. Clearly the diameter of all regions shrinks to 0 in the limit.

$$\begin{aligned}
W_2^2(\mu, \bar{\mu}_N) &< W_2^2(\mu, \tilde{\mu}_N) = \sum_{i=1}^N \int_{X_{\bar{\mu}_N}^i} d_{\mathcal{X}}(x, x_{\bar{\mu}_N}^i)^2 d\mu(x) \\
&\leq \sum_{i=1}^N \mu(X_{\bar{\mu}_N}^i) \text{diam}(X_{\bar{\mu}_N}^i)^2 \\
&\leq \left[\sum_{i=1}^N \mu(X_{\bar{\mu}_N}^i) \right] \max_i \text{diam}(X_{\bar{\mu}_N}^i)^2 \rightarrow 0
\end{aligned} \tag{51}$$

where $\sum_{i=1}^N \mu(X_{\bar{\mu}_N}^i) = 1$. □

Appendix C Weighted centroids parameterise the geodesic

Lemma 4. *Provided that $\forall p^1, p^2 \in \mathcal{P}$ and $\forall d_{1b}, d_{2b} \geq 0$ with $d_{1b} + d_{2b} \geq d(p^1, p^2) = d_{12}$ $\exists p^b \in \mathcal{P}$ s.t. $d(p^1, p^b) = d_{1b}$ and $d(p^2, p^b) = d_{2b}$, then the centroids $p^b(\lambda) = B_d(p^1, p^2, \lambda)$ lie on the geodesic between p^1 and p^2 . Further, the curve given by $p^b(\lambda)$ is constant speed and is therefore the geodesic between p^1 and p^2 .*

Proof. $B(p^1, p^2, \lambda)$ is attained when the barycentric cost takes the value

$$\begin{aligned}
&\min_{d_{1b}, d_{2b} \in \mathbb{R}_+} (1 - \lambda)d_{1b}^2 + \lambda d_{2b}^2 \\
&\text{s.t. } d_{1b} + d_{2b} \geq d_{12}
\end{aligned} \tag{52}$$

with Lagrangian

$$\mathcal{L}(d_{1b}, d_{2b}, \gamma) = \lambda d_{1b}^2 + (1 - \lambda)d_{2b}^2 + \gamma(d_{12} - d_{1b} - d_{2b}). \tag{53}$$

It is necessary for optimality that we have stationarity (Karush-Kuhn-Tucker conditions), $\partial_{d_{1b}} \mathcal{L} = \partial_{d_{2b}} \mathcal{L} = 0$, hence $\gamma = 2(1 - \lambda)d_{1b} = 2\lambda d_{2b}$. Outside the trivial cases $p^1 = p^2$ or $\lambda \in \{0, 1\}$, it follows that $\gamma > 0$. Since $\gamma \neq 0$, the bound is tight, i.e. $d_{1b} + d_{2b} = d_{12}$ indicating that $\forall \lambda \in (0, 1)$, $p^b(\lambda)$ sits on the geodesic connecting p^1, p^2 . Further, $d_{2b} = d_{1b}(1 - \lambda)/\lambda$, so that $d_{12} = [1 + (1 - \lambda)/\lambda] d_{1b} = d_{1b}/\lambda$. Since we know $p^b(\lambda)$ at worst reparameterises the geodesic, $d(p^1, p^b(\lambda)) = \lambda d(p^1, p^2)$ is enough to know it is constant speed. □

Appendix D General McCann interpolation

The following is related to the Monge OT minimisation

$$\min_{T^t} \left\{ \int C[(T^t(x))_{0 \leq t \leq 1}] d\mu(x) : T^0 = \text{Id} \quad \text{and} \quad T^1_{\#} \mu = \nu \right\}, \tag{54}$$

that is to say, rather than evaluating the final displacement of x to y , evaluate the whole trajectory taken between these. This is one dynamic formulation of the OT problem. At time t , the mass at x is instead at $T^t(x)$ and only at $t = 1$ does it arrive at y . These two problems are equivalent (with $T^{*1} = T^*$) provided that the static cost agrees with the cost of the best trajectory, i.e. that

$$c(x, y) = \min_{(T^t(x))_{0 \leq t \leq 1}} \left\{ C[(T^t(x))_{0 \leq t \leq 1}] : T^0(x) = x \quad \text{and} \quad T^1(x) = y \right\} \tag{55}$$

is met Villani (2003).

Definition 10 (Length of a trajectory). For the trajectory $\tau : [0, 1] \rightarrow \mathcal{X}$, the length is defined as

$$L_{d_{\mathcal{X}}}[\tau] = \sup \sum_{i=1}^N d_{\mathcal{X}}(\tau(t_{i-1}), \tau(t_i)), \quad (56)$$

with the supremum taken over all N and sequences $0 = t_0 \leq \dots \leq t_N = 1$.

Definition 11 (Length space). A metric space $(\mathcal{X}, d_{\mathcal{X}})$ with $d_{\mathcal{X}}(x, y) = \inf_{\tau} L[\tau]$ over all $\tau : [0, 1] \rightarrow \mathcal{X}$ with $\tau(0) = x$ and $\tau(1) = y$.

Definition 12 (Geodesic). A geodesic from $x \in \mathcal{X}$ to $y \in \mathcal{X}$, is a trajectory $\tau : [0, 1] \rightarrow \mathcal{X}$ with $\tau(0) = x$, $\tau(1) = y$ which has (1.) minimal length and (2.) constant speed. (1.) can be ensured by $L[(\tau)_{t^1 \leq t \leq t^2}] = d_{\mathcal{X}}(\tau(t^1), \tau(t^2))$ for all $0 \leq t^1 < t^2 \leq 1$, and (2.) by $L[(\tau)_{t^1 \leq t \leq t^2}] = (t^2 - t^1)\lambda$ for some λ , again for all $0 \leq t^1 < t^2 \leq 1$. By Definition 10, (1.) is given by (2.) when $\lambda = d_{\mathcal{X}}(x, y)$.

Definition 13 (Geodesic space). A metric space $(\mathcal{X}, d_{\mathcal{X}})$ in which $\forall x, y \in \mathcal{X}$, there exists a geodesic connecting x and y . Geodesic spaces are also length spaces.

Suppose $c = d_{\mathcal{X}}^p$ where the ground space $(\mathcal{X}, d_{\mathcal{X}})$ is a geodesic space, if we let $C[\tau] = L_{d_{\mathcal{X}}}[\tau]^p$ and without loss of correspondence to the Monge problem, let Eq. (54) also require that all considered paths are constant speed, then clearly the optimal trajectories $(T^{*t}(x))_{0 \leq t \leq 1}$ are ground space geodesics between x and $T^*(x)$. By Lemma 4, this optimal trajectory can then be expressed in terms of ground space barycenters $T^{*t}(x) = B_{d_{\mathcal{X}}}(x, T^*(x), t)$, where T^* is the optimal map in the static problem.

Lemma 5. *Let the OT problem between μ and ν admit a Monge solution and let T^{*t} be the (constant speed) dynamic OT solution. Let T^{12} be the (static) optimal transport map between $T_{\#}^{*t^1} \mu$ and $T_{\#}^{*t^2} \mu$. Provided that $T^{12}(T^{*t^1}(x)) = T^{*t^2}(x)$ for all $0 \leq t^1 < t^2 \leq 1$, then $(T_{\#}^{*t} \mu)_{0 \leq t \leq 1}$ is a Wasserstein space geodesic between μ and ν .*

Proof. Note that the shortest path between $T^{*t^1}(x)$ to $T^{*t^2}(x)$ is given by the geodesic segments $(T^{*t}(x))_{t^1 \leq t \leq t^2}$, then

$$\begin{aligned} W_p^{d_{\mathcal{X}}}(T_{\#}^{*t^1} \mu, T_{\#}^{*t^2} \mu)^p &= \int d_{\mathcal{X}}(x, T^{12}(x))^p d(T_{\#}^{*t^1} \mu)(x) \\ &= \int d_{\mathcal{X}}(T^{*t^1}(x), T^{*t^2}(x))^p d\mu(x) \\ &= \int L_{d_{\mathcal{X}}}[(T^{*t}(x))_{t^1 \leq t \leq t^2}]^p d\mu(x) \\ &= \int [(t^2 - t^1) d_{\mathcal{X}}(x, T^*(x))]^p d\mu(x) = (t^2 - t^1)^p W_p^{d_{\mathcal{X}}}(\mu, \nu)^p. \quad \square \end{aligned}$$

Where Lemma 5 applies, such as in 1D, where feasible monotonic plans are unique, Lemma 4 can be employed again, allowing one to calculate Wasserstein barycenters via $B(\mu, \nu, \lambda) = T_{\#}^{*\lambda} \mu$, this is the displacement interpolation. This is usually stated for an Euclidean ground space where $T^{*\lambda} = (1 - \lambda)\text{Id} + \lambda T^*$.

References

Jharna Agrawal, Manish Gupta, and Hitendra Garg. Blind source separation in perspective of ica algorithms: A review. In *2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, pages 78–85, 2022. doi: 10.1109/CISES54857.2022.9844373.

- Ran Aharoni, Asaf Zuck, David Peri, and Shai Kendler. Reflectance spectra analysis algorithms for the characterization of deposits and condensed traces on surfaces. In Prof. Ashis Kumar Samanta, editor, *Colorimetry*, chapter 10. IntechOpen, Rijeka, 2022. doi: 10.5772/intechopen.101301.
- Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, page 214–223, Sydney, NSW, Australia, 2017. JMLR.org.
- Junwen Bai, Sebastian Ament, Guillaume Perez, John Gregoire, and Carla Gomes. An efficient relaxed projection method for constrained non-negative matrix factorization with application to the phase-mapping problem in materials science. In Willem-Jan van Hoeve, editor, *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pages 52–62, Cham, 2018a. Springer International Publishing. ISBN 978-3-319-93031-2.
- Junwen Bai, Yexiang Xue, Johan Bjorck, Ronan Le Bras, Brendan Rappazzo, Richard Bernstein, Santosh K. Suram, R. Bruce van Dover, John M. Gregoire, and Carla P. Gomes. Phase-mapper: Accelerating materials discovery with ai. *AI Mag.*, 39(1):15–26, mar 2018b. ISSN 0738-4602. doi: 10.1609/aimag.v39i1.2785. URL <https://doi.org/10.1609/aimag.v39i1.2785>.
- Yikun Bai, Bernhard Schmitzer, Matthew Thorpe, and Soheil Kolouri. Sliced optimal partial transport. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13681–13690, 2023. doi: 10.1109/CVPR52729.2023.01315.
- Emmanuel Boissard, Thibaut Le Gouic, and Jean-Michel Loubes. Distribution’s template estimate with Wasserstein metrics. *Bernoulli*, 21(2):740 – 759, 2015. doi: 10.3150/13-BEJ585. URL <https://doi.org/10.3150/13-BEJ585>.
- Ricardo Augusto Borsoi, Tales Imbiriba, José Carlos Moreira Bermudez, Cédric Richard, Jocelyn Chanussot, Lucas Drumetz, Jean-Yves Tournet, Alina Zare, and Christian Jutten. Spectral variability in hyperspectral data unmixing: A comprehensive review. *IEEE Geoscience and Remote Sensing Magazine*, 9(4):223–270, 2021. doi: 10.1109/MGRS.2021.3071158.
- Boulais, A., Berné, O., Faury, G., and Deville, Y. Unmixing methods based on nonnegativity and weakly mixed pixels for astronomical hyperspectral datasets. *A&A*, 647:A105, 2021. doi: 10.1051/0004-6361/201936399. URL <https://doi.org/10.1051/0004-6361/201936399>.
- R. Bras, R. Bernstein, J.M. Gregoire, Santosh Suram, Carla Gomes, Bart Selman, and Robert van Dover. A computational challenge problem in materials discovery: Synthetic problem generator and real-world datasets. *Proceedings of the National Conference on Artificial Intelligence*, 1: 438–443, 01 2014.
- Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021.
- Guillermo Canas and Lorenzo Rosasco. Learning probability measures with respect to optimal transport metrics. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, Lake Tahoe, Nevada, USA, 2012. Curran Associates, Inc.
- Afef CHERNI, Sandrine ANTHOINE, and Caroline CHAUX. β -nmf and sparsity promoting regularizations for complex mixture unmixing - application to 2d hsqc nmr. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5795–5799, 2020. doi: 10.1109/ICASSP40776.2020.9053154.

- Andrzej Cichocki, Rafal Zdunek, Anh-Huy Phan, and Shun-ichi Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. 10 2009. ISBN 978-0-470-74666-0. doi: 10.1002/9780470747278.
- Xiaoxiao Du, Alina Zare, Paul Gader, and Dmitri Dranishnikov. Spatial and spectral unmixing using the beta compositional model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):1994–2003, 2014. doi: 10.1109/JSTARS.2014.2330347.
- Bjorn Engquist, Brittany Hamfeldt, and Yunan Yang. Optimal transport for seismic full waveform inversion. *Communications in Mathematical Sciences*, 14, 02 2016. doi: 10.4310/CMS.2016.v14.n8.a9.
- Stefano Ermon, Ronan Le Bras, Santosh Suram, John Gregoire, Carla Gomes, Bart Selman, and Robert van Dover. Pattern decomposition with complex combinatorial constraints: Application to materials discovery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015. doi: 10.1609/aaai.v29i1.9233. URL <https://ojs.aaai.org/index.php/AAAI/article/view/9233>.
- R. Flamary. *Optimal Transport for Machine Learning*. PhD thesis, Université Cote d’Azur, 2019.
- Rémi Flamary, Cédric Févotte, Nicolas Courty, and Valentin Emiya. Optimal spectral transportation with application to music transcription. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS’16, page 703–711, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.
- P.T. Fletcher, Conglin Lu, S.M. Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8):995–1005, 2004. doi: 10.1109/TMI.2004.831793.
- Scott Gaffney and Padhraic Smyth. Joint probabilistic curve clustering and alignment. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS’04, page 473–480, Cambridge, MA, USA, 2004. MIT Press.
- Jürgen Glatz, Nikolaos C. Deliolanis, Andreas Buehler, Daniel Razansky, and Vasilis Ntziachristos. Blind source unmixing in multi-spectral optoacoustic tomography. *Opt. Express*, 19(4):3175–3184, Feb 2011. doi: 10.1364/OE.19.003175. URL <http://opg.optica.org/oe/abstract.cfm?URI=oe-19-4-3175>.
- Cameron J. Hargreaves, Matthew S. Dyer, Michael W. Gaultois, Vitaliy A. Kurlin, and Matthew J. Rosseinsky. The earth mover’s distance as a metric for the space of inorganic compositions. *Chemistry of Materials*, 32(24):10610–10620, 2020. doi: 10.1021/acs.chemmater.0c03381.
- R. Heylen, M. Parente, and P. Gader. A review of nonlinear hyperspectral unmixing methods. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):1844–1868, 2014.
- Nhat Ho, Xuan Long Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. Multilevel clustering via wasserstein means. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, page 1501–1509, Sydney, NSW, Australia, 2017. JMLR.org.
- J K Holland, E K Kemsley, and R H Wilson. Use of fourier transform infrared spectroscopy and partial least squares regression for the detection of adulteration of strawberry purées. *Journal of the Science of Food and Agriculture*, 76(2):263–269, 1998. doi: [https://doi.org/10.1002/\(SICI\)1097-0010\(199802\)76:2<263::AID-JSFA943>3.0.CO;2-F](https://doi.org/10.1002/(SICI)1097-0010(199802)76:2<263::AID-JSFA943>3.0.CO;2-F).

- E. Hosseini-Asl, J. M. Zurada, and A. El-Baz. Automatic segmentation of pathological lung using incremental nonnegative matrix factorization. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 3111–3115, 2015.
- Daniel Lee and Hyunjune Seung. Algorithms for non-negative matrix factorization. *Adv. Neural Inform. Process. Syst.*, 13, 02 2001.
- Jun Liu, Bin Luo, Sylvain Douté, and Jocelyn Chanussot. Exploration of planetary hyperspectral images with unsupervised spectral unmixing: A case study of planet mars. *Remote Sensing*, 10(5), 2018. ISSN 2072-4292. doi: 10.3390/rs10050737. URL <https://www.mdpi.com/2072-4292/10/5/737>.
- S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982. doi: 10.1109/TIT.1982.1056489.
- W. Ma, J. M. Bioucas-Dias, T. Chan, N. Gillis, P. Gader, A. J. Plaza, A. Ambikapathi, and C. Chi. A signal processing perspective on hyperspectral unmixing: Insights from remote sensing. *IEEE Signal Processing Magazine*, 31(1):67–81, 2014.
- Tingting Mu, John Yannis Goulermas, Jun’ichi Tsujii, and Sophia Ananiadou. Proximity-based frameworks for generating embeddings from multi-output data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2216–2232, 2012. doi: 10.1109/TPAMI.2012.20.
- Sina Nakhostin, Harold Clenet, Thomas Corpetti, and Nicolas Courty. Joint anomaly detection and spectral unmixing for planetary hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):6879–6894, 2016. doi: 10.1109/TGRS.2016.2586188.
- Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11 (5-6):355–602, 2019. URL <https://arxiv.org/abs/1803.00567>.
- D. Pollard. Quantization and the method of k-means. *IEEE Transactions on Information Theory*, 28(2):199–205, 1982. doi: 10.1109/TIT.1982.1056481.
- Rameau, J., Chanussot, J., Carlotti, A., Bonnefoy, M., and Delorme, P. Spectral unmixing for exoplanet direct detection in hyperspectral data. *A&A*, 649:A143, 2021. doi: 10.1051/0004-6361/202140337. URL <https://doi.org/10.1051/0004-6361/202140337>.
- Marc Oliver Rieger. Monotonicity of transport plans, 2006. URL <http://cvgmt.sns.it/paper/948/>. cvgmt preprint.
- Yossi Rubner, Carlo Tomasi, and Leonidas Guibas. The earth mover’s distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121, 11 2000. doi: 10.1023/A:1026543900054.
- R. Sandler and M. Lindenbaum. Nonnegative matrix factorization with earth mover’s distance metric for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1590–1602, 2011.
- Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Springer International Publishing, Cham, 2015. doi: 10.1007/978-3-319-20828-2.
- Morgan A. Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngolè, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, jan 2018. doi: 10.1137/17m1140431.

- Yuliy Schwartzburg, Romain Testuz, Andrea Tagliasacchi, and Mark Pauly. High-contrast computational caustic design. *ACM Trans. Graph.*, 33(4), jul 2014. ISSN 0730-0301. doi: 10.1145/2601097.2601200. URL <https://doi.org/10.1145/2601097.2601200>.
- Sameer Shirdhonkar and David W. Jacobs. Approximate earth mover’s distance in linear time. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. doi: 10.1109/CVPR.2008.4587662.
- Paris Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In Carlos G. Puntonet and Alberto Prieto, editors, *Independent Component Analysis and Blind Signal Separation*, pages 494–499, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-30110-3.
- D. Stein. Application of the normal compositional model to the analysis of hyperspectral imagery. In *IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data, 2003*, pages 44–51, 2003. doi: 10.1109/WARSD.2003.1295171.
- Albert Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, Philadelphia, 2005. doi: 10.1137/1.9780898717921.
- Cédric Villani. Topics in optimal transportation. In *Graduate Studies in Mathematics*, 2003. URL <https://api.semanticscholar.org/CorpusID:118448577>.
- Cédric Villani. *Optimal Transport*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9.
- Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Umberto De Giovannini, and Francesca Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6(30):883–904, 2005. URL <http://jmlr.org/papers/v6/devito05a.html>.
- Jiexi Yan, Cheng Deng, Lei Luo, Xiaoqian Wang, Xiaohui Yao, Li Shen, and Heng Huang. Identifying imaging markers for predicting cognitive assessments using wasserstein distances based matrix regression. *Frontiers in Neuroscience*, 13:668, 2019. ISSN 1662-453X. doi: 10.3389/fnins.2019.00668. URL <https://www.frontiersin.org/article/10.3389/fnins.2019.00668>.
- Min Zhao, Jie Chen, and Zhe He. A laboratory-created dataset with ground truth for hyperspectral unmixing evaluation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2170–2183, 2019. doi: 10.1109/JSTARS.2019.2905099.
- Yuan Zhou, Anand Rangarajan, and Paul D. Gader. A gaussian mixture model representation of endmember variability in hyperspectral unmixing. *IEEE Transactions on Image Processing*, 27(5):2242–2256, 2018. doi: 10.1109/TIP.2018.2795744.