Unmixing compositional data with Bayesian techniques

R. TOLOSANA-DELGADO

Maritime Engineering Laboratory - Universitat Politècnica de Catalunya, Spain raimon.tolosana@upc.edu

Abstract

A general problem in compositional data analysis is the unmixing of a composition into a series of pure endmembers. In its most complex version, one does neither know the composition of these endmembers, nor their relative contribution to each observed composition. The problem is particularly cumbersome if the number of endmembers is larger than the number of observed components. This contribution proposes a possible solution of this under-determined problem.

The proposed method starts assuming that the endmember composition is known. Then, a geometric characterization of the problem allows to find the set of possible endmember proportions compatible with the observed composition. Within this set any solution may be valid, but some are more likely than other. To use this idea and choose the "most likely" solution in each case, the problem can be tackled with Bayesian Markov-Chain Monte-Carlo techniques. Finally, once we are familiar with MCMC, it is quite straightforward to allow the endmember compositions to randomly vary, and use the same MCMC to estimate the endmember composition most compatible with the studied data.

1 Definitions

An end-member problem considers a set of D-variate observations (the N rows of $\mathbf{X}$) as generated by a *composition* (a convex linear *mixture*) of P "pure" members,

$$\mathbf{X} = \mathbf{Z} \cdot \mathbf{T}, \tag{1}$$

where

- each endmember composition in the original D-dimensional simplex is identified as a row of $\mathbf{T}$,
- and their relative contributions to the observations are given in the rows of $\mathbf{Z}$.

The so-called *bilinear problem* aims at obtaining estimates of both $\mathbf{Z}$ and $\mathbf{T}$. In general, most applications of end-member modelling assume $P << D$ (e.g. Weltje, 1997; Aitchison and Bacon-Shone, 1999; Billheimer, 2001). On the contrary, this contribution tackles this problem in the uncommon case that $P > D$, motivated by the need to recast a geochemical sediment composition into a "probable" mineralogical composition. The complete developments may be found in the work of Tolosana-Delgado et al. (2011).

Through this paper the classical compositional concepts of centered log-ratio transformation (clr, Aitchison, 1986), isometric log-ratio transformation (ilr, Egozcue et al., 2003), perturbation ($\oplus$, Aitchison, 1986), powering ($\odot$, Aitchison, 2002), as well as the additive logistic normal, a.k.a. normal distribution on the simplex ($\mathcal{N}_{\mathcal{S}}^{P}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, Mateu-Figueras et al., 2003) will be used, as well as the inverse perturbation $\ominus$. Classical addition $+$ and product $\cdot$ will keep their meanings with respect to the conventional real geometry. If not explicitly stated, vectors will be considered always row-vectors and denoted by boldface lowercase characters, while matrices will be denoted by boldface uppercase characters. The superindex ^{t} will denote matrix transposition.

2 Geometric characterization of the set of solutions

In a first step towards a solution, we will assume end-members with fixed characteristics. In terms of our application, that means assuming a known and fixed stoichiometry for all present minerals (i.e., set $\mathbf{T}$), and finding all possible mineral assemblages that would be compatible with a given geochemical composition. The rows of $\mathbf{T}$ and $\mathbf{X}$ belong to a D-part simplex $\mathcal{S}^D$, while the rows of $\mathbf{Y}$ belong

to a P-part simplex $\mathcal{S}^P$. Matrix $\mathbf{T}$ is not square, hence not invertible. We will thus make use of the generalized Moore-Penrose inversion, based on its singular value decomposition and that of its symmetrization,

$$\begin{aligned} \mathbf{T} &= \mathbf{U}_T \cdot \mathbf{D} \cdot \mathbf{V}^t = \mathbf{U}_T \cdot \mathrm{diag}(d_1, \ldots, d_D) \cdot \mathbf{V}^t, \\ \mathbf{T} \cdot \mathbf{T}^t &= \mathbf{U} \cdot \mathbf{D}^2 \cdot \mathbf{U}^t = \underbrace{[\mathbf{U}_T, \mathbf{W}]}_{\mathbf{U}} \cdot \underbrace{\mathrm{diag}(d_1^2, \ldots, d_D^2, 0, \ldots, 0)}_{\mathbf{D}^2} \cdot \mathbf{U}^t, \end{aligned}$$

where:

- the singular vectors of the symmetrized $\mathbf{T} \cdot \mathbf{T}^t$ are given by the columns of $\mathbf{U}$, and identify a basis of $\mathbb{R}^P$; from them, the first D are stored in matrix $\mathbf{U}_T$ (and coincide with the left singular vectors of $\mathbf{T}$), and the following in a matrix $\mathbf{W}$, by columns; the column-vectors of $\mathbf{W}$ will be called *null singular vectors* of $\mathbf{T}$;
- the singular values are provided as the diagonal elements of matrix $\mathbf{D}$, with D non-zero diagonal values; $\mathbf{T} \cdot \mathbf{T}^t$ has the singular values of $\mathbf{T}$ squared, but the symmetrized matrix has $(P - D)$ extra zero values;
- the right singular vectors of $\mathbf{T}$ are given in the columns of $\mathbf{V}$, and they identify a basis of $\mathbb{R}^D$.

With these elements, the Moore-Penrose inverse $\mathbf{T}^-$ of a matrix $\mathbf{T}$ is

$$\mathbf{T}^- = \mathbf{V} \cdot \mathbf{D}^{-1} \cdot \mathbf{U}_T^t \tag{2}$$

Then, for each row $\mathbf{x}_i$ of $\mathbf{X}$, it may be shown that:

1. the set S of possible solutions is embedded on a hyperplane $\mathbb{H}$ in the P-dimensional real space (not in the Aitchison geometry of $\mathcal{S}^P$), with parametric equation

$$\mathbf{z}_i(\boldsymbol{\lambda}) = \mathbf{x}_i \cdot \mathbf{T}^- + \boldsymbol{\lambda} \cdot \mathbf{W}^t; \tag{3}$$

2. the orientation of $\mathbb{H}$ depends only on $\mathbf{W}$, thus on the singular vectors of $\mathbf{T} \cdot \mathbf{T}^t$, while its actual position may be found with $\mathbf{T}^-$, in Eq. (2);
3. the set of possible solutions S is the intersection $S = \mathbb{H} \cap \mathcal{S}^P$, as embedded subsets of $\mathbb{R}^P$. That is natural, because the set of valid solutions must belong to both the hyperplane of candidate solutions $\mathbb{H}$ and the simplex $\mathcal{S}^P$. This implies that S is a convex subset of $\mathbb{H}$ with the geometry of $\mathbb{R}^P$. Moreover, given that Eq. (3) establishes a linear equivalence between $\mathbf{z}_i \in \mathbb{H}$ and $\boldsymbol{\lambda} \in \mathbb{R}^{P-D}$, the set of $\boldsymbol{\lambda}$ leading to valid compositions $\mathbf{z}_i$ is also a convex hull in $\mathbb{R}^{P-D}$, which we denote by S_λ to distinguish it from $S \subset \mathcal{S}^P$.

Example: *take a "geochemical composition" of $D = 2$ parts called A and B, i.e. $\mathbf{x} = [x_A, x_B] \in \mathcal{S}^D$, and a "mineral composition" formed by $P = 3$ parts called a, b and c, i.e. $\mathbf{x} = [x_a, x_b, x_c] \in \mathcal{S}^P$. Assume that the stoichiometry of the minerals is*

$$\mathbf{T} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} \equiv \begin{pmatrix} 1/3 & 2/3 \\ 1/2 & 1/2 \\ 0 & 1 \end{pmatrix}$$

From the first expression of $\mathbf{T}$, it is obvious that $a = b + c$, or with regard to the second expression, $3a = 2b + c$. Thus, if we had a mineral composition $\mathbf{z}_0$, the set of mineral compositions

$$\mathbf{z} = \mathbf{z}_0 + \lambda^* \cdot [-1, 1, 1]^* = \mathbf{z}_0 + \lambda \cdot \underbrace{[-3, 2, 1]/\sqrt{14}}_{\mathbf{w}} \tag{4}$$

would give exactly the same geochemical composition for any value of λ. Note that the asterisk marks the equivalence with regard to non-closed stoichiometric compositions, but we will use the second one

because it uses closed normalized vectors. We need further the SVD of $\mathbf{T}$*. If we consider now the geochemical composition* $\mathbf{x} = (0.1, 0.9)$*, the hyperplane* $\mathbb{H}$ *of candidate solutions would be the line of Eq. (4) passing through the point* $\mathbf{z}_0 = (0.2571, 0.0286, 0.7143)$*. This is within* $\mathcal{S}^P$*, but we may still move along the null singular vector* $\mathbf{w}$ *to find other valid solutions. Given the values of* $\mathbf{w}$*, we may have a 0 component either at* x_a *or at* x_b*, using respectively* $\lambda = 0.3207$ *or* $\lambda = -0.0535$*. Thus, the set of valid solutions correspond to* $\lambda \in (-0.0535, 0.3207)$*. If we now consider* $\mathbf{x} = (0.9, 0.1)$*, we may see that there is no intersection of its space of solutions and* $\mathcal{S}^P$*, thus that geochemical composition would not be compatible with the stoichiometry chosen. A mineral with a larger proportion of A with respect to B would be necessary.*

3 Selecting best solutions

In a second step, the goal is to select "best" solutions from this subset, in particular under some distributional assumptions. For instance, one may assume an additive logistic normal distribution on $\mathcal{S}^P$, i.e. $\mathbf{Z} \sim \mathcal{N}_{\mathcal{S}}^P(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This would allow to estimate the most likely mineral composition $\mathbf{z}_i$ linked to each geochemical composition $\mathbf{x}_i$, as well as the average mineral composition of the whole suite $\boldsymbol{\mu}$. The same framework can be used in the presence of co-variables, like in the illustration example. It is well-known that sediments of different grain size have very different mineral compositions. Thus it may be sensible to assume that the expected mineral composition $\boldsymbol{\mu} = \mathrm{E}[\mathbf{Z}] = \mathbf{a} \oplus t \odot \mathbf{b}$ is related linearly (with respect to the Aitchison geometry of $\mathcal{S}^P$) with the logarithm $t = -\log_2(\bar{d}[\mathrm{mm}])$ of the average grain diameter (in mm) as explanatory variable. One would then try to estimate the coefficients of the linear model. Provided that the observed geochemical compositions are all compatible with the given mineral stoichiometry (i.e. that there exists a $\mathbf{z}_i$ with positive components for each row $\mathbf{x}_i$), this framework is tractable with standard Markov Chain Monte Carlo techniques (MCMC), in particular with a Metropolis-Hastings (MH) algorithm for each $\boldsymbol{\lambda}_i$, together with standard Gibbs sampling for $\boldsymbol{\Sigma}$, and $\boldsymbol{\mu}$ (or $\mathbf{a}$ and $\mathbf{b}$).

MH algorithm is useful to simulate from a density function which is known up to a multiplicative constant, i.e. where we only have its likelihood function. Let the interest parameter θ have a likelihood $L(\theta)$. Assume we have a simulated value for θ^k at step k of the chain. MH simulates a new proposal θ^* from an arbitrary kernel distribution $K(\theta^*|\theta^k)$ depending on the preceding value in the chain (e.g. from a normal centered at θ^k), and then randomly chooses as new θ^{k+1} either θ^* or θ^k respectively with probability p or $(1-p)$, being

$$p = \max\left(1, \frac{K(\theta^*|\theta^k) \cdot L(\theta^*)}{K(\theta^k|\theta^*) \cdot L(\theta^k)}\right).$$

To efficiently explore the set of valid solutions S, kernel simulation must be done on $\boldsymbol{\lambda} \in \mathbb{R}^{P-D}$ (we drop here the row index $\boldsymbol{\lambda}_i$ for the sake of simplicity). For instance, we may use a normal distribution $\mathcal{N}^{P-D}(\boldsymbol{\lambda}^k, \boldsymbol{\Psi})$, with a cleverly chosen $\boldsymbol{\Psi}$. Discussing the role of this kernel variance is not the scope of this paper. If $\boldsymbol{\lambda}^*$ does not fall within the convex hull S_λ, then $\boldsymbol{\lambda}^*$ is rejected and we keep $\boldsymbol{\lambda}^k$. Otherwise, we may choose between $\boldsymbol{\lambda}^*$ and $\boldsymbol{\lambda}^k$ with a probability $p = \max\left(1, f(\mathbf{z}(\boldsymbol{\lambda}^*); \boldsymbol{\mu}, \boldsymbol{\Sigma}) / f(\mathbf{z}(\boldsymbol{\lambda}^k); \boldsymbol{\mu}, \boldsymbol{\Sigma})\right)$, with a non-closed additive logistic normal density

$$f(\mathbf{z}(\boldsymbol{\lambda}); \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\mathbf{z}(\boldsymbol{\lambda}) \cdot \mathbf{1}^t} \exp\left(-\frac{1}{2}\mathrm{ilr}(\mathbf{z}(\boldsymbol{\lambda}) \ominus \boldsymbol{\mu}) \cdot \boldsymbol{\Sigma}^- \cdot \mathrm{ilr}^t(\mathbf{z}(\boldsymbol{\lambda}) \ominus \boldsymbol{\mu})\right). \quad (5)$$

Running this chain sufficiently long we may obtain a sample from the distribution of $\boldsymbol{\lambda}_i$, for each row i. As usual in MCMC techniques, this sample is studied to characterize the variability of this parameter, or in our case, better to study the variability of $\mathbf{z}(\boldsymbol{\lambda}_i)$, computed using Eq. (3).

4 Estimating the endmember composition

The final step is to allow the end-member properties (the mineral stoichiometry) to vary in a controlled, interpretable fashion. As we are already in a Bayesian framework, this is naturally encoded using a

set of prior distributions for the observable composition of each endmember (i.e. the geochemical stoichiometry of each mineral). Again, MCMC techniques are necessary to solve the problem. That delivers the same set of results as before (most likely mineral composition of each observation $\mathbf{z}(\boldsymbol{\lambda}_i)$, together with the compositional variability $\boldsymbol{\Sigma}$, and the mean mineral composition $\boldsymbol{\mu}$ or equivalent coefficients $\mathbf{a}$ and $\mathbf{b}$ of a linear model) plus a posterior assessment of the probable stoichiometric composition $\mathbf{T}$ of the end-members most compatible with the observed data set. Two key issues appear here: an adequate, low-dimension parametrization of $\mathbf{T}$; and an unequivocal characterization of the convex hull of solutions S_λ.

Each row of $\mathbf{T}$ gives the chemical composition of a mineral. From a practical point of view, some minerals (e.g. quartz, apatite), have a fixed composition, whereas other vary in a limited fashion between some extreme members (e.g., garnets, amphiboles, biotites, epidotes). The composition of those of the second kind needs to be parametrized cleverly. Take the j-th row $\mathbf{t}_j$ of $\mathbf{T}$ to represent one of these minerals having N_j different extreme members. Consider the compositions of these end-members as the rows of a matrix $\mathbf{M}_j$ of dimension (N_j, D). Following Eq. (1), one may write $\mathbf{t}_j = \boldsymbol{\omega}_j \cdot \mathbf{M}_j$, where $\boldsymbol{\omega}_j$ may be seen as a random composition distributed on a simplex $\mathcal{S}^{N_j}$, different for each j, with coordinates $\boldsymbol{\tau}_j = \mathrm{ilr}(\boldsymbol{\omega}_j)$. Denoting by $\boldsymbol{\tau} = (\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_K)$ the concatenation of all coordinates of the K simplexes linked to each variable-composition mineral, we may see $\mathbf{T} = \mathbf{T}(\boldsymbol{\tau})$. As will be seen in the illustration example, this idea strongly reduces the dimension needed to represent the stoichiometry, yet without loss of flexibility, generality and geological sense.

Consider now a stoichiometry $\mathbf{T}(\boldsymbol{\tau}) = \mathbf{T}$ and a slight perturbation of it $\mathbf{T}(\boldsymbol{\tau} + \boldsymbol{\epsilon}) := \mathbf{T}_\epsilon$. Because of the numerics of SVD, there is no reason why their respective null singular vectors $\mathbf{W}$ and $\mathbf{W}_\epsilon$ should be similar. Their column-vectors represent two different orthonormal bases of $\mathbb{H}$, but there is no need that they are the same. Even in the case of having one single null vector, we may find $\mathbf{w}_\epsilon \simeq \mathbf{w}$, or $\mathbf{w}_\epsilon \simeq -\mathbf{w}$. It is thus necessary to force the column vectors of $\mathbf{W}$ to be oriented in a given reference way. In the case of having one single null vector, the trick consists on forcing a chosen coefficient to be always positive. When having more vectors, the generalization of this concept can be obtained by forcing a suitable minor of dimension $P - D$ from $\mathbf{W}$ to have a lower triangular form with positive diagonal elements. This can be done in the following way. Assume that we have identified a minor $\mathbf{A}$ of $\mathbf{W}$ which has almost surely full rank for any value of $\boldsymbol{\tau}$. Then we can compute the LU decomposition of $\mathbf{A} = \mathbf{L} \cdot \mathbf{U}$, and re-define our null vectors as $\mathbf{W}' = \mathbf{W} \cdot \mathbf{U}^{-1}$. By construction the minor $\mathbf{A}'$ of $\mathbf{W}'$ will be $\mathbf{A} = \mathbf{L} \cdot \mathbf{U} \cdot \mathbf{U}^{-1}\mathbf{L} =$, of the desired form.

Using these two ideas, we fully identify the stoichiometry and its null singular vectors. Now we can add to our Gibbs and MH samplers a new MH loop that updates $\boldsymbol{\tau}$, either in one single step or each $\boldsymbol{\tau}_j$ separately. In any case, fixed $\boldsymbol{\lambda}$ and having a previous simulation $\boldsymbol{\tau}^k$ and a candidate simulation $\boldsymbol{\tau}^*$, we may compute their stoichiometry $\mathbf{T}^k$ and $\mathbf{T}^*$, their null vectors and build their mineral compositions $\mathbf{Z}^k = (\mathbf{z}_1^k, \ldots, \mathbf{z}_N^k)$ and $\mathbf{Z}^* = (\mathbf{z}_1^*, \ldots, \mathbf{z}_N^*)$ respectively from Eq. (3). Then, fixed $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ (or $\mathbf{a}$ and $\mathbf{b}$ instead of the mean), we may compute the likelihood of $\mathbf{Z}^k$ and $\mathbf{Z}^*$ by means of Eq. (5) used N times.

Example: *Take the set of oxides SiO_2, Al_2O_3, Fe_2O_3, MgO, MnO, CaO and Na_2O, and the minerals quartz, plagioclase and amphiboles. We consider that these minerals may be ideally modelled as combinations of extreme members described in Table 1. With this hierarchical definition, quartz is not varying and does not need to be parametrized. Plagioclase can be parametrized with a scalar τ_{plg} which describes the log-ratio of proportions of albite/anortite constituents. Finally, amphiboles are parametrized with a 2-dimensional ilr vector $\boldsymbol{\tau}_{amph}$, describing the proportions of the three extreme members considered. Therefore, a (3,7)-dimensional matrix $\mathbf{T}$ has been meaningfully parametrized with a 3-coefficient vector $\boldsymbol{\tau}$, which space of definition is the whole $\mathbb{R}^3$. These minerals, together with those described in Tolosana-Delgado et al. (2011), may be used in the reconstruction of the mineral compositions of glacial sediments from amphibolitic rocks, described and analysed in the contribution by von Eynatten and Tolosana-Delgado (2011).*

Table 1: Some mineral examples, to complement those appearing in Tolosana-Delgado et al. (2011) for the mineral reconstruction of amphibolitic rocks.

mineral	member	SiO_2	Al_2O_3	Fe_2O_3	MgO	MnO	CaO	Na_2O
quartz	quartz	1	0	0	0	0	0	0
plagioclase	albite	3	0.5	0	0	0	0	0.5
	anortite	2	1	0.5	0	0	1	0
amphibole	tremolite-actinolite	8	0	1.25	2.5	0	2	0
	pargasite	6	1.5	1	2	0	2	0.5
	tschermakite	8	2	0.75	1.5	0	2	0

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Aitchison, J. (2002). Simplicial inference. In M. A. G. Viana and D. S. P. Richards (Eds.), *Algebraic Methods in Statistics and Probability*, Volume 287 of *Contemporary Mathematics Series*, pp. 1–22. American Mathematical Society, Providence, Rhode Island (USA), 340 p.

Aitchison, J. and J. Bacon-Shone (1999). Convex linear combination of compositions. *Biometrika 86*(2), 351–364.

Billheimer, D., 2001. Compositional receptor modeling. Environmetrics 12(5), 451–467.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*(3), 279–300.

von Eynatten, H. and R. Tolosana-Delgado (2011). Geochemistry versus grain-size relations of sediments in the light of comminution, chemical alteration, and contrasting source rocks. In Proceedings of CoDaWork'11 (this volume)

Mateu-Figueras, G., V. Pawlowsky-Glahn, and C. Barceló-Vidal (2003). Distributions on the simplex. In S. Thió-Henestrosa and J. A. Martín-Fernández (Eds.), *Compositional Data Analysis Workshop – CoDaWork'03, Proceedings*. Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/.

Tolosana-Delgado, R., von Eynatten, H., Karius, V. (2011). Constructing modal mineralogy from geochemical composition: a geometric-Bayesian approach *Computers & Geosciences* (in press). online, doi:10.1016/j.cageo.2010.08.005.

Weltje, J. G. (1997). End-member modeling of compositional data: numerical-statistical algorithms for solving the explicit mixing problem. *Mathematical Geology 29*, 503–549.