Geochemistry, Geophysics, Geosystems

TECHNICAL REPORTS: METHODS

10.1002/2015GC006070

New methods for unmixing sediment grain size data

Greig A. Paterson[1] and David Heslop[2]

[1]Key Laboratory of the Earth and Planetary Physics, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China, [2]Research School of Earth Sciences, Australian National University, Canberra, Australian Capital Territory, Australia

Key Points:
- We introduce fast, efficient, and precise algorithms for unmixing grain size data
- A new parametric EMA algorithm for identifying subpopulations is introduced
- These routines are available in AnalySize, which is new software for unmixing

Supporting Information:
- Supporting Information S1

Correspondence to:
G. A. Paterson,
greig.paterson@mail.iggcas.ac.cn

Citation:
Paterson, G. A., and D. Heslop (2015), New methods for unmixing sediment grain size data, *Geochem. Geophys. Geosyst.*, *16*, 4494–4506, doi:10.1002/2015GC006070.

Received 25 AUG 2015
Accepted 18 NOV 2015
Accepted article online 23 NOV 2015
Published online 19 DEC 2015

Abstract Grain size distribution (GSD) data are widely used in Earth sciences and although large data sets are regularly generated, detailed numerical analyses are not routine. Unmixing GSDs into components can help understand sediment provenance and depositional regimes/processes. End-member analysis (EMA), which fits one set of end-members to a given data set, is a powerful way to unmix GSDs into geologically meaningful parts. EMA estimates end-members based on covariability within a data set and can be considered as a nonparametric approach. Available EMA algorithms, however, either produce suboptimal solutions or are time consuming. We introduce unmixing algorithms inspired by hyperspectral image analysis that can be applied to GSD data and which provide an improvement over current techniques. Nonparametric EMA is often unable to identify unimodal grain size subpopulations that correspond to single sediment sources. An alternative approach is single-specimen unmixing (SSU), which unmixes individual GSDs into unimodal parametric distributions (e.g., lognormal). We demonstrate that the inherent nonuniqueness of SSU solutions renders this approach unviable for estimating underlying mixing processes. To overcome this, we develop a new algorithm to perform parametric EMA, whereby an entire data set can be unmixed into unimodal parametric end-members (e.g., Weibull distributions). This makes it easier to identify individual grain size subpopulations in highly mixed data sets. To aid investigators in applying these methods, all of the new algorithms are available in AnalySize, which is GUI software for processing and unmixing grain size data.

1. Introduction

Numerical unmixing of grain size distribution (GSD) data into constituent components, known as end-members, can yield valuable information on geological processes and paleoenvironments [e.g., *Prins et al.*, 2002; *Weltje and Prins*, 2003; *Vriend et al.*, 2011; *Meyer et al.*, 2013]. Attempts at unmixing GSD data have a long history in sedimentology (see *Weltje* [1997] for a more detailed review), but despite the potential of unmixing, numerous studies do not take advantage of this useful mathematical representation. This may, in part, be due to a lack of widely available, user-friendly software that allows the rapid processing of large data sets.

In this paper, we outline new methods for unmixing grain size data and introduce new software to aid researchers. In section 2, we briefly introduce the concept of unmixing grain size data. In section 3, we introduce a new approach for end-member analysis (EMA) inspired by hyperspectral image analysis and compare it to existing nonparametric EMA algorithms. We demonstrate that, by virtue of improved accuracy, this algorithm outperforms existing ones. Following on from *Weltje and Prins* [2007], we provide a demonstration in section 4 that widely used parametric approaches to unmix individual grain spectra are unlikely to yield physically meaningful models and are likely to lead to interpretations that are fundamentally different from the true mixing regime. However, because parametric approaches can represent unimodal grain size subpopulations, their use remains compelling. To overcome the shortcomings of single-specimen techniques, we introduce a new approach called parametric EMA in section 5, whereby a GSD data set can be characterized by single set of unimodal parametric end-members. Finally, in section 6, we introduce new software for analyzing grain size data obtained from laser diffraction particle grain size analyzers. The

AnalySize software provides our new algorithms in a user-friendly GUI alongside a suite of standard analysis statistics and plots.

2. Overview of Numerical Unmixing

Unmixing data into constituent end-members and their abundances is a key problem in many geoscience disciplines and has been extensively explored and discussed in hydrology [e.g., *Christophersen and Hooper*, 1992], sedimentology [e.g., *Weltje*, 1997], geochemistry [e.g., *Hannigan et al*., 2001], hyperspectral image analysis [e.g., *Bioucas-Dias et al*., 2012], and rock magnetism [*Heslop*, 2015]. Unmixing techniques assume that the observed data can be described as a linear mixture of the constituent end-members where the abundances of each end-member must be nonnegative and sum-to-one (or 100%) for each observation. Mathematically, the unmixing problem can be expressed in matrix notation as:

$$\mathbf{X}=\mathbf{AS}+\mathbf{E} \tag{1}$$

where $\mathbf{X}$ is the observed data (one specimen per row), $\mathbf{A}$ is the abundance matrix of the constitute end-members whose signatures are given by $\mathbf{S}$ (one end-member per row), and $\mathbf{E}$ represents sampling and measurement errors. Equation (1) is subject to abundance nonnegativity and sum-to-one constraints: $\forall_i : \mathbf{A}_i \geq 0$ and $\sum_i \mathbf{A}_i = 1$ (or 100%). These constraints correspond to the physical requirement that abundances cannot be negative and represent relative abundances that must sum to a constant. In the case of unmixing GSD data, it is also required that end-member signatures, which are normalized particle counts, are nonnegative and sum-to-one: $\forall_k : \mathbf{S}_k \geq 0$ and $\sum_k \mathbf{S}_k = 1$. Due to these constraints, there is no closed form solution to equation (1), which has to be solved numerically [*Renner*, 1993; *Heinz and Chang*, 2001].

In a geometric sense, *j* end-members represent the vertices of a *j-1* dimensional simplex that spans the data; three end-members map out a 2-D triangle, four end-members form a 3-D tetrahedron, and so on. In a noise-free case, the set of known end-members defines a simplex that bounds all of the specimens. In practice, however, the presence of noise and outlying data often mean that even when $\mathbf{S}$ is known, some data will lie outside the simplex. This means that when equation (1) is solved for $\mathbf{A}$ with the nonnegativity constraint removed, some of the observations in $\mathbf{X}$ will yield negative abundances. To quantify this, *Weltje* [1997] introduced the concept of convexity error, which is a measure of the proportion (P) of data points that lie outside of the simplex and their average squared distance from the simplex $\left(\bar{D}^2\right)$. Convexity error, C, is given by:

$$C=\log_{10}(P)+\log_{10}\left(\bar{D}^2\right) \tag{2}$$

As the simplex expands and fewer data fall outside or are closer the simplex edges, C decreases. For a simplex that fully encompasses all data, $C=-\infty$.

When the end-member signatures, $\mathbf{S}$, are known, the mixing abundances, $\mathbf{A}$, can be determined by numerically solving equation (1); this is known as linear or supervised unmixing. In general, however, both the end-member signatures and their abundances are unknown and must be estimated in what is known as a bilinear, or unsupervised unmixing problem. In context of unmixing GSD data, which generally deals with the unsupervised unmixing problem, two main approaches have been proposed. One approach involves using a set of empirically derived end-members to unmix all the specimens in a data set [*Weltje*, 1997]. Typically, the end-members for EMA are estimated from the data set itself, a process that we call "nonparametric EMA." An alternative approach is single-specimen unmixing (SSU), whereby end-members, described by parametric distributions (e.g., lognormal or Weibull), are fitted to individual grain size spectra [*Sun et al*., 2002].

3. Nonparametric EMA

Nonparametric EMA estimates the end-members from the data (i.e., unsupervised unmixing). Algorithms to perform this task [e.g., *Weltje*, 1997; *Dietze et al*., 2012] are briefly outlined below.

Weltje [1997] introduced a routine based on simplex expansion. The algorithm starts with an initialization of the end-members, which are assigned forms that meet the necessary nonnegativity and sum-to-one

constraints and which places them inside the mixing space defined by the distribution of GSD specimens. These initial end-members define a simplex, which does not encompass all of the data. This means that when the nonnegativity constraint is removed from **A**, equation (1) will yield one or more negative abundances for specimens located outside the initial simplex. The algorithm updates each end-member sequentially to expand the simplex and reduce the number of negative abundances, which reduces the convexity error. This process is repeated, progressively expanding the simplex until a desired level of convexity (e.g., $C \leq -6$) or a maximum number of iterations are reached.

Dietze et al. [2012] introduced an algorithm based on eigenvector rotation. For a given number of end-members, *j*, this algorithm rotates the first *j* eigenvectors of the data using the VARIMAX criterion. These vectors are then normalized to ensure that the constraints on equation (1) are satisfied and are taken to represent the end-member estimates. Using nonnegative least squares, the fractional abundances of the end-members are estimated. These abundances, however, are not constrained to sum-to-one and must be normalized to satisfy this constraint.

Both of these approaches apply a weight transformation, which is designed to reweight the residual misfits between the model reconstructions and the data and allow for sum-to-one normalization [*Manson and Imbrie*, 1964; *Miesch*, 1976]. Our implementations of these two algorithms scale the data by the maxima and minima of each grain size bin, which is recommended by *Weltje* [1997] and, for the data presented here, yield the most stable results for the *Dietze et al*. [2012] algorithm.

Here we introduce a new EMA approach, based on nonnegative matrix factorization (NMF) [*Lee and Seung*, 1999], which utilizes an existing algorithm from the hyperspectral image community [*Chen and Guillaume*, 2012]. NMF finds two nonnegative matrices, **W** and **H**, which satisfy $\mathbf{X} \approx \mathbf{WH}$. The similarities between the NMF formulation and equation (1) and the fact that both **W** and **H** satisfy the end-member and abundance nonnegativity constraints makes NMF an appealing approach to solving GSD unmixing problems. NMF solutions, however, are nonunique and multiple local minima exist. However, the nonuniqueness of NMF solutions can be reduced by appropriate initialization and the application of physically meaningful constraints [*Donoho and Stodden*, 2004; *Miao and Qi*, 2007; *Chen and Guillaume*, 2012].

The first step of our new approach is to make a robust estimate of a simplex that encompasses all of the data and which can act as the NMF initialization. To achieve this, we use the simplex identification via split augmented Lagrangian (SISAL) algorithm of *Bioucas-Dias* [2009]. SISAL is based on the principle of identifying the minimum volume simplex that bounds the data [*Craig*, 1994]. To ensure robustness against noise and outliers, however, the nonnegativity requirement, which forces the simplex to bound the observations, is replaced with a soft constraint that can be violated to a given degree. This acts much in the same fashion as the convexity error of *Weltje* [1997]. The main disadvantage of this algorithm is that it does not impose nonnegativity or sum-to-one constraints on the end-member signatures, which is physically unrealistic for grain size data. Nevertheless, the SISAL-derived end-members can be used to initialize an NMF-based routine.

The second step of the algorithm utilizes the hierarchical alternating least squares nonnegative matrix factorization (HALS-NMF) routine of *Chen and Guillaume* [2012]. This NMF algorithm uses the HALS update rules for optimization [*Cichocki et al*., 2008], which have been shown to outperform other routines in terms of efficiency and convergence [*Chen and Guillaume*, 2012, and references therein]. HALS-NMF minimizes the misfit between the observed data and the model reconstruction, but also allows for the inclusion of additional constraints. Of interest to GSD unmixing are the abundance sum-to-one constraint and an end-member minimum distance constraint. The minimum distance constraint encourages HALS-NMF to find a solution that minimizes the distance of each end-member to the centroid of the simplex, which encourages the simplex to shrink and bound the observations closely. It should be noted that the HALS-NMF algorithm does not allow for the explicit constraint that the end-member signatures must sum-to-one. However, as a consequence of the other constraints and optimization to fit the data, violations of this constraint are small. Nevertheless, after each iterative update of the nonnegative matrices, the end-member signatures are normalized to enforce end-member sum-to-one. Full details of the derivation of the constraints, the NMF update rules, and the algorithm are outlined in the supporting information.

The abundance sum-to-one and end-member minimum distance constraints are soft in the sense that they do not need to be fully satisfied. This has the advantage of making the algorithm robust against noise and outliers. The sum-to-one constraint is controlled by a single regularization term, α_1, which, through our testing

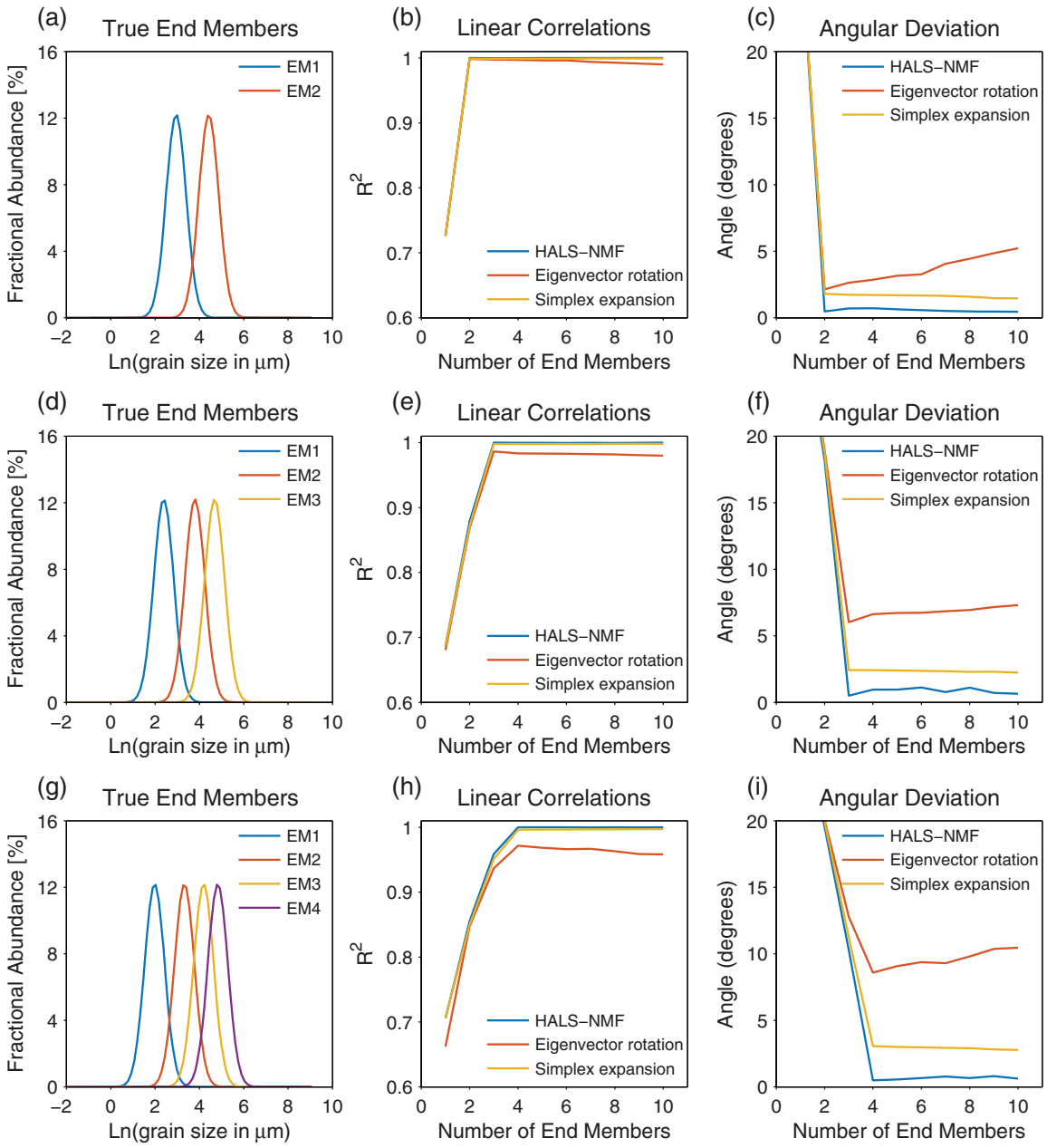

Figure 1. Synthetic lognormal end-member GSDs and goodness-of-fit statistics for unmixing of data sets of 200 specimens that are comprised of (a–c) two, (d–f) three, and (g–i) four end-members. In the goodness-of-fit plots, HALS-NMF refers to our new algorithm following *Chen and Guillaume* [2012], eigenvector rotation is the algorithm of *Dietze et al.* [2012], and simplex expansion is from *Weltje* [1997].

and development with real and synthetic data, we set to 5. The minimum distance regularization term, β_2, is first set to zero (i.e., no constraint). The HALS-NMF solution is determined and the convexity error calculated. If the error is extremely low (i.e., <-7), which would indicate overexpansion of the simplex and fitting of noise and/or outliers, we progressively increase β_2 until C is within an acceptable range ($-7 < C \leq -6$).

3.1. Algorithm Comparison

To compare the performances of the different algorithms, we generate three synthetic data sets comprised of two to four lognormal end-members (Figure 1). The abundances are randomly selected in the range [0,1] but constrained to sum-to-one. This represents a poorly mixed data set, which contains observations that

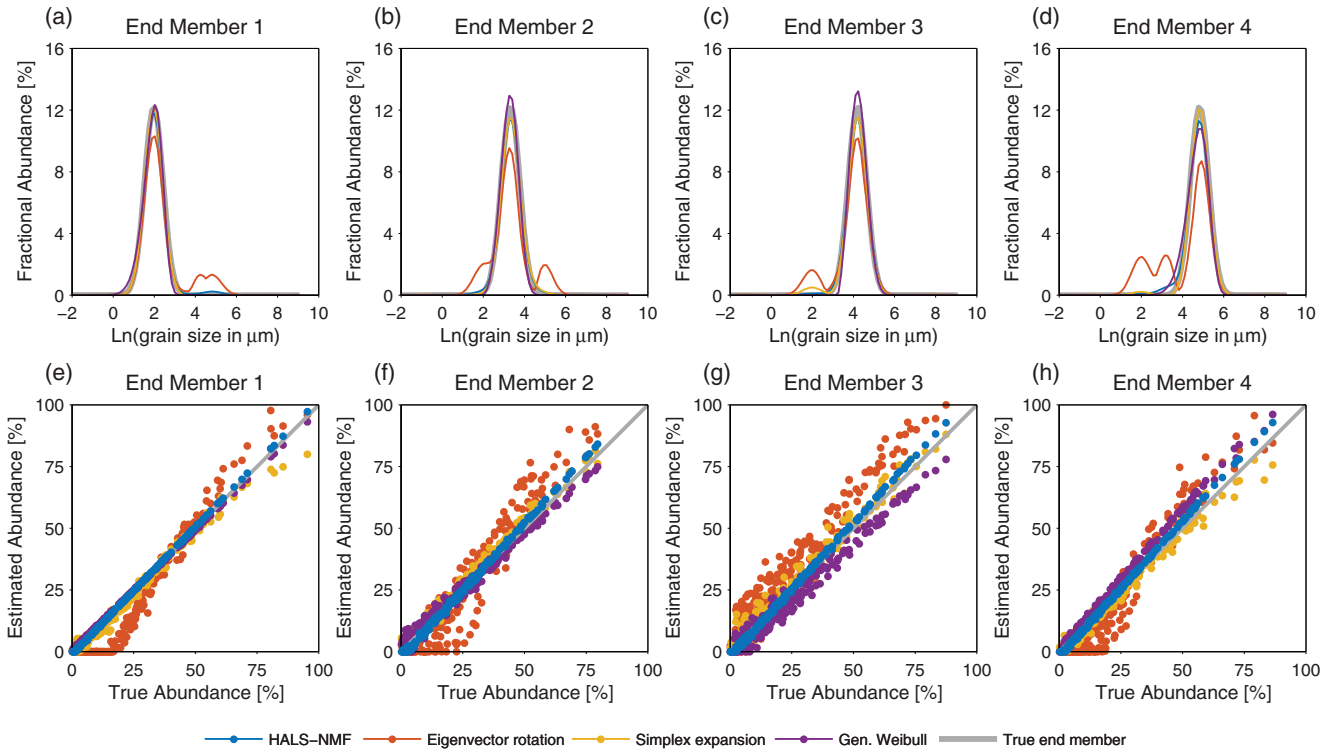

Figure 2. Fitting results from different EMA methods when applied to the four end-member data set in Figures 1g–1i. (a–d) Comparison between the fitted and true end-members. (e–h) Comparison between the estimated and true abundances.

lie close to the true end-members. It should therefore be possible for these algorithms to extract end-members and abundances close to the true values. The data set consists of 200 specimens each with 100 grain size bins. A small amount of noise, which is visually comparable to real data, was added to each specimen and the data renormalized to ensure nonnegativity and sum-to-one.

We compare performance of the algorithms by fitting 1–10 end-members to each of the data sets. We compare the coefficient of determination (R^2) and angular differences between the original data and the reconstruction produced by each algorithm (Figure 1). All algorithms identify peaks or plateaus in the goodness-of-fit statistics that correspond to the correct number of end-members. The eigenvector rotation routine of *Dietze et al.* [2012], however, does not plateau with increasing numbers of end-members. As the number of end-members increases, all goodness-of-fit statistics deteriorate. This is due to how this algorithm enforces the constraints on equation (1). First, a solution that minimizes the misfit to the data is found, this solution is then normalized to enforce the constraints. This normalization takes an optimal unconstrained solution and transforms it into a suboptimal constrained solution, which yields a relatively poor fit to the data. Both simplex expansion and HALS-NMF, however, incorporate the constraints as part of their iterative optimizations and produce optimal constrained solutions. For most cases of the simplex expansion algorithm, however, the maximum number of iterations is reached (set to a maximum of 2000), which indicates that an incomplete solution has been found. On our test system (MacBook Pro with a 2.8 GHz Intel Core i7), generating all of the fits in Figure 1, simplex expansion took >35 min, but HALS-NMF took only 30 s. Although *Weltje* [1997] suggests constructing the goodness-of-fit parameters based on an initial eigenvector reconstruction, which is a considerable faster process, both the eigenvector rotation and HALS-NMF algorithms yield final solutions. It is therefore fairer to compare the performances of the final solutions for all algorithms. We note that, increasing the number of iterations for the simplex expansion algorithm often leads to a better fit than shown in Figure 1.

Using model fits with the correct number of end-members (two end-members for the $j = 2$ case, three for $j = 3$ and so on; Figure 2), we also compare the estimated end-member spectra and abundance matrices to

Table 1. Run Times and Goodness-Of-Fit Statistics for Fitting the Correct Number of End-Members to the Data Sets in Figure 1 Using the Described EMA Methods

	HALS-NMF	Eigenvector Rotation	Simplex Expansion	Parametric EMA (Gen. Weibull)
$j = 2$				
Run time (s)	1.5	0.1	6.8	1.3
Data set R^2	1.000	0.998	0.999	0.980
Data set misfit (°)	0.5	2.1	1.8	7.3
Convexity error	−7.0	−8.9	−5.6	−11.0
EM misfit (°)	1.0	0.6	1.8	7.5
Abundance misfit (°)	0.8	2.4	2.0	1.0
$j = 3$				
Run time (s)	1.4	0.1	1.7	10.5
Data set R^2	1.000	0.986	0.998	0.990
Data set misfit (°)	0.5	6.0	2.4	5.0
Convexity error	−6.3	−2.9	−6.0	−4.2
EM misfit (°)	1.5	9.6	1.9	7.8
Abundance misfit (°)	1.5	9.4	3.0	6.2
$j = 4$				
Run time (s)	1.0	0.1	38.9	58.5
Data set R^2	1.000	0.971	0.996	0.997
Data set misfit (°)	0.5	8.6	3.1	2.7
Convexity error	−6.0	−2.1	−5.9	−5.0
EM misfit (°)	2.4	15.7	2.3	6.3
Abundance misfit (°)	2.4	14.6	5.3	6.6

their true values (all statistics are given in Table 1). For $j = 2$, 3, and 4, the estimated spectra from HALS-NMF have mean angular deviations of 1.0°, 1.5°, and 2.4°, respectively, from the true signatures. For the *Weltje* [1997] routine, the angles are 1.8°, 1.9°, and 2.3°, respectively, and for the *Dietze et al*. [2012] algorithm, these are 0.6°, 9.6°, and 15.7°, respectively. Similarly for abundances, HALS-NMF yields mean angular deviations of 0.8°, 1.5°, and 2.4°, respectively, between the estimates and known abundances; the *Weltje* [1997] algorithm yields 2.0°, 3.0°, and 5.3°, respectively; the *Dietze et al*. [2012] algorithm yields angles of 2.4°, 9.4°, and 14.6°, respectively. It should be noted that for the $j = 2$ and 4, the simplex expansion did not fully converge to the convexity error criterion, but both solutions have $C < -5.6$. The convexity errors for all fitting results presented in Figure 1 are given in the supporting information.

Although fastest (Table 1), the eigenvector rotation method of *Dietze et al*. [2012] is the poorest performing of all algorithms particularly as the number of end-members increases. The simplex expansion algorithm [*Weltje*, 1997] yields comparable results to the HALS-NMF algorithm but can be relatively slow (1.2–40 times slower than HALS-NMF; Table 1). Overall, however, HALS-NMF not only fits the data sets better than both these other methods but is also more capable of identifying the true end-members and their abundances.

In the above examples, the data sets are poorly mixed and therefore nonparametric EMA can extract end-members that closely resemble the true end-members. However, when the data are highly mixed (i.e., no single specimen is near a pure end-member), nonparametric EMA yields multimodal distributions. These are sometimes referred to as dynamic populations (DP) and represent size fractions that covary throughout a data set. High mixing levels, however, are common in natural settings [e.g., *Prins et al*., 2002; *Weltje and Prins*, 2003; *Vriend et al*., 2011; *Dietze et al*., 2012].

We consider the situation where a suite of specimens is composed of two sources, Source 1 and Source 2, represented by the lognormal distributions shown in Figure 3a. In this synthetic data set of 99 specimens, both sources are always present with each source never having a relative contribution less than ∼13% (Figure 3b). The resultant end-members from nonparametric EMA (using our HALS-NMF algorithm) fit the data well ($R^2 = 1.000$, angular misfit = 0.3°) and are shown in Figure 3c. Because both end-members are always present, nonparametric EMA identifies two bimodal end-members; end-member 1 is dominated by Source 1, but with a secondary peak, which corresponds to Source 2 and vice versa for end-member 2 (Figure 3c). The fitted end-members have an angular misfit of 13.8° from the true end-members. Comparing the estimated abundances of nonparametric EMA end-member 1 with the true abundance, it can be seen that nonparametric EMA systematically misestimates the abundance of the sources (Figure 3d). This corresponds to a total angular misfit of 8.2° between the estimated and true abundances.

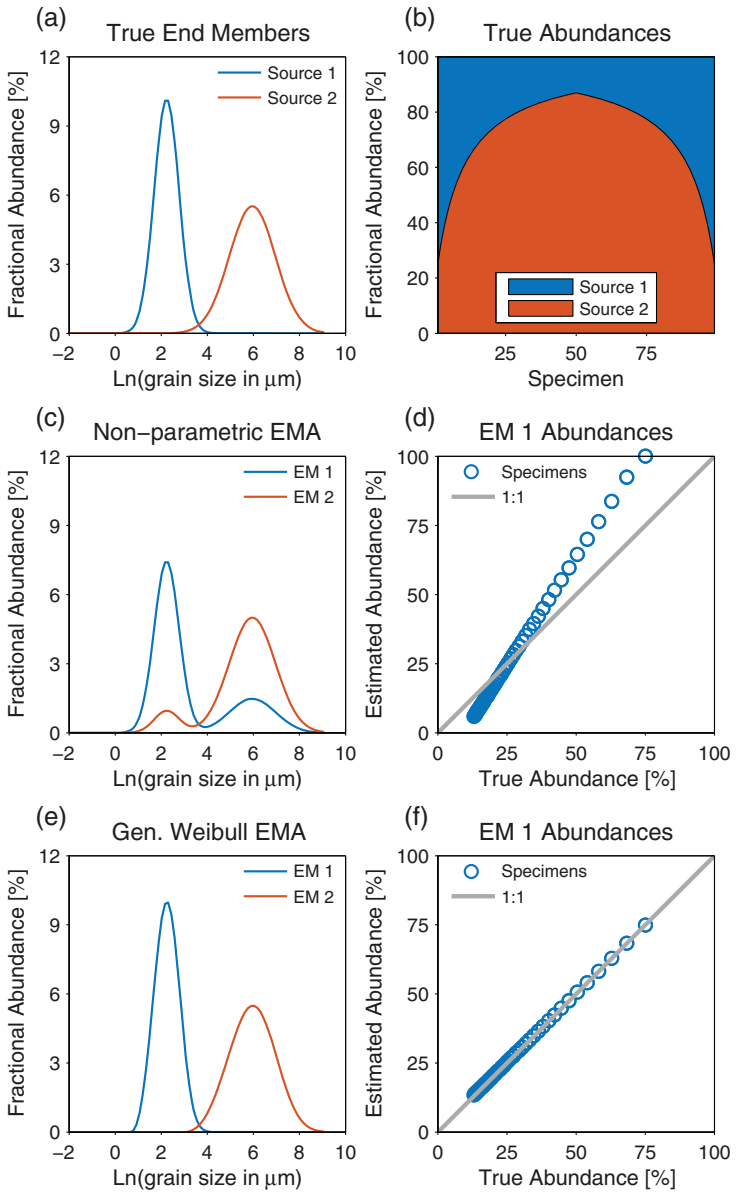

Figure 3. (a and b) Two lognormal source end-members and their abundance variations in a synthetic data set of 99 specimens. This is a moderately mixed data set where the minimum abundance of each source never falls below ~13%. (c and d) Nonparametric EMA unmixing results. (e and f) General Weibull parametric EMA unmixing results.

In some case, it is reasonable to expect physical processes to produce multimodal end-members (e.g., glacial grinding) [*Weltje and Prins*, 2003], and the geological context of the end-members needs to be carefully considered. However, in situations where specific information about individual subpopulations (SPs) is sought, or two physically unrelated processes yield covarying end-members, the above-described type of model bias is undesirable, leading some workers to pursue methods of identifying SPs.

4. Single-Specimen Unmixing

Single-specimen unmixing (SSU) is the most commonly used approach to estimate unimodal grain size subpopulations. The basic approach is to unmix individual GSDs into parametric end-members, such as

Table 2. Parametric Distributions Supported in AnalySize

Type	Parameters to Fit	Comment
Lognormal	2	Parameters control the location and scale of the distribution. Lognormal distributions are fitted in linear space
Weibull	2	Parameters control the shape and scale of the distribution. Weibull distributions are fitted in bin number space, where each size bin is numbered consecutively
Gen. Weibull	3	General Weibull distribution. Adds an additional location parameter to the Weibull distribution. General Weibull distributions are fitted in bin number space, where each size bin is numbered consecutively
SGG	3	Skewed Generalized Gaussian (SGG) distribution [*Egli*, 2003]. The SGG is a four parameter distribution, but a maximum entropy approximation is used to reduce the number of free parameters to fit to three [*Egli*, 2004]. Parameters control the location, scale, and skewness/kurtosis of the distribution. SGG distributions are fitted in log space

lognormal or Weibull distributions using a least squares approach [*Sun et al*., 2002]. SSU is widely used to analyze and unmix grain size spectra, despite having well-documented issues [*Weltje and Prins*, 2007]. *Weltje and Prins* [2007] outline many of the challenges of unmixing a single grain size spectrum. Specifically, they draw attention to: (1) the difficulty of deciding how many end-members are sufficient to represent a given specimen and (2) the lack of geological context to constrain the general behavior of the end-members (i.e., using a single specimen as opposed to the entire data set). Nevertheless, SSU is still used widely [e.g., *Li et al*., 2014; *Wang et al*., 2015].

The issues raised by *Weltje and Prins* [2007] demonstrate that SSU solutions are inherently nonunique. This nonuniqueness can be further illustrated by considering two sets of three general Weibull [*Weibull*, 1951] (see section 5 and Table 2) end-members (Figures 4a and 4b), which are similar, but clearly distinct. When combined with different mixing proportions, however, each set of end-members can yield noise-free total grain size spectra that appear nearly identical, but are not exactly the same (Figures 4c and 4d). Despite the underlying mixing process being different (different end-members and abundances differing by factors up to 2), the angular difference between the two resultant spectra is only 0.0132° and the average difference between specimen A and B across all bins is ~0.0003%. If we assume that the cumulative difference between the total spectra of specimens A and B up to the fiftieth percentile can be used to represent the error on the median value, then we can use it to approximately estimate a coefficient of variation of the median of ~0.01%. This is an approximate estimate of the maximum level of noise that would allow the identification of these grain size spectra as two distinct spectra. Above this noise level, distinction would be extremely difficult. Modern particle size analyzers are compliant with the ISO-13320 standard, which stipulates that the coefficient of variation of the median should be less than 3% for median grain sizes above 10 μm. Many manufacturers, however, often quote values better than 1%. Nevertheless, the noise level necessary to discern between different unmixing regimes using SSU is currently beyond the capabilities of particle size analyzers.

We note that, although the true statistical nature of noise influencing GSD data is poorly quantified, our noise-free estimate of the coefficient of variation of the median is likely an overestimate of the level of noise required to clearly distinguish these two spectra. Furthermore, in this example, we know the exact form and number of end-members to fit, which is information that is not readily available in real situations. In cases where neither are known or where the number of end-members is misestimated, equally similar fits may be achieved that are poor representations of the true mixing regime. Therefore, SSU is not a feasible approach for unmixing GSD data, and such models are unlikely to be representative of the true mixing system.

5. Parametric EMA

The popularity of SSU demonstrates a clear desire to explore GSD data in terms of subpopulations (SPs). Therefore, an approach that can identify SPs that are consistent with the variability of an entire data set is needed.

To unmix GSD data into SPs, we have developed a form of parametric EMA that allows end-members described by parametric distributions to be fitted to the entire data set. Starting from an appropriate parametric distribution (discussed below and in Table 2) and an assumption of the number of end-members to fit, we make an initial guess of the distribution parameters. This initial estimate of distribution parameters is

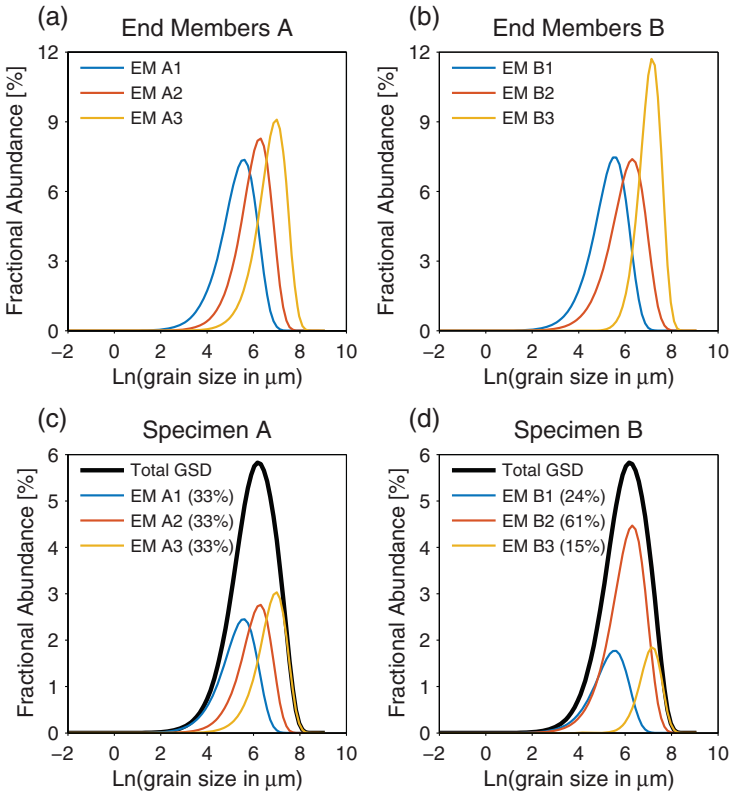

Figure 4. (a and b) Two sets of general Weibull end-members that when combined in differing abundances can yield identical (c and d) grain size distributions.

a key step in parametric fitting. As can be seen in Figure 4, an observed grain size spectrum can appear unimodal, but disguises the presence of multiple end-members. To overcome this, the algorithm first performs nonparametric EMA and identifies the maxima of the nonparametric end-members to estimate the parametric end-members (this includes identifying multiple maxima in multimodal end-members). A fast optimization search is performed to unmix the nonparametric end-members and the result of this forms the initial guess for the parametric end-members. The initial guess is then fitted to the observed data set to estimate the abundances of the initial end-members. This step uses the simplex projection unmixing algorithm (SPU) [*Heylen et al.*, 2011; *Heylen and Scheunders*, 2012], which is a rapid alternative to constrained least squares [*Heinz and Chang*, 2001]. The distribution parameters are then optimized to find a solution that minimizes the Frobenius norm between the observed data and the model. SPU is used at each step to determine the appropriate end-member abundances. Since this form of parametric fitting is applied to the entire data set, it accounts for variance across the mixing space that is not considered by SSU. As is the case for nonparametric EMA, a collection of models with differing numbers of end-members can be produced to estimate the optimal model complexity (cf. Figures 1b and 1c).

This new approach allows continuous distributions to be used to unmix grain size data sets and we have primarily focused on convenient forms, which have been previously suggested and fit the basic requirement of nonnegative grain sizes (Table 2). In addition to commonly used distributions (e.g., lognormal and Weibull), the algorithm supports two distributions that are characterized by three parameters, which offer greater flexibility in terms of shape. The general Weibull distribution [*Weibull*, 1951] includes an additional location parameter, which controls the limit of the left-hand tail of the distribution, thus controlling skewness. The skewed generalized Gaussian (SGG) distribution [*Egli*, 2003] is a modification of a Gaussian distribution to include skewness and kurtosis parameters. SGGs are fitted to the grain size data in log space, and we use the maximum entropy approximation of *Egli* [2004], which relates skewness and kurtosis, to reduce the number of free parameters to three.

Application of parametric EMA to the mixed example given in Figure 3a is shown in Figure 3e. Here we unmix the synthetic data using two general Weibull distributions, which correlation and angular deviation analysis confirm is the most appropriate number of end-members. Even though the true form of the end-members is lognormal rather than general Weibull, parametric EMA identifies unimodal subpopulations (Figure 3e) that fit closely with the expected signatures (Figure 3a; total angular misfit of 1.4°), and yields accurate abundance estimates (Figure 3f; total angular misfit of 0.2°).

Using general Weibulls, we also apply parametric EMA to the data sets in Figure 1 to unmix the appropriate number of end-members. The unmixing results are summarized in Figure 2 and Table 1. Overall, fitting general Weibulls unmixes the data sets well, but fails to fully capture the details of the tails of the end-members (e.g., the left-hand tails of end-members 3 and 4, Figures 2c and 2d). This is a limitation of the form of general Weibull end-members and fitting lognormal or SGG end-members can overcome this. The general Weibull abundances lie close to the expected trends and are generally less scattered than those for the simplex expansion and eigenvector rotation algorithms, but more scattered than for the HALS-NMF algorithm. When unmixing three or more end-members, fitting general Weibulls provide a better fit to the data and yield a model closer to the true unmixing regime than can be obtained by the eigenvector rotation algorithm. Parametric EMA, however, is the slowest method and takes ~1 min for four end-members, in comparison to ~1 s for the HALS-NMF algorithm. Using nonparametric EMA to estimate the best range for the number parametric end-members to test is recommended to minimize computation time.

We further apply parametric EMA (using SGG end-members) to a data set consisting of the three highly overlapping general Weibull end-members shown in Figure 4a. These data sets consist of 200 spectra with random abundances that meet the sum-to-one constraint, but where the minimum abundance is varied between 0 and 30% to simulate poorly to highly mixed cases, respectively. All 200 grain size spectra have only a single peak and could not be successfully unmixed by SSU techniques.

In Figure 5, the misfit between the known and estimated end-members and abundances is plotted against increasing degree of mixing. At low degrees of mixing, nonparametric EMA (HALS-NMF) tends to estimate the true end-members and abundances more accurately. This is because, at low levels of mixing, near-pure end-members are present in the data and can be identified by the algorithm. For parametric EMA, however, differences in the exact form of the end-members hinder the ability to fit the true end-members accurately (Figure 5). As the level of mixing increases, pure end-members are no longer sampled in the data set and nonparametric EMA cannot clearly identify the true mixing regime. With increasing mixing, parametric EMA fits the true mixing regime, with a slight, but unstable improvement as mixing increases. By a minimum abundance of 10–15%, parametric EMA fits the true mixing regime better than nonparametric EMA, despite the difference in the form of the estimated end-members compared to the true members. The overall fit to the data sets from both methods remains consistently low (Figure 5c). This highlights the usefulness of parametric unmixing in highly mixed data sets.

The end-members and abundances for a minimum abundance of 15% are shown in Figure 6. The estimated end-members from nonparametric EMA exhibit contamination with each other. This is most evident in the tail toward larger grain sizes for end-member 1 and the tail toward smaller grain sizes for end-member 3 (Figures 6a and 6c, respectively). This end-member impurity results in a large difference between the expected and true abundances (up to a 31% difference). SGG end-members approximate the true end-members well, but overestimate the peaks of end-members 2 and 3. This causes a systematic underestimation of the abundances of end-members 2 and 3 and, hence, an overestimate of the abundance of end-member 1. Nevertheless, the estimated abundances are close to the expected values (Figure 5b) and follow a similar trend (maximum abundance difference of 15%).

5.1. Interpreting Parametric Distributions

Although the use of parametric distributions to model GSDs has no basis in theory, it is supported by the observation of *Weltje and Prins* [2007] that often nonparametric EMA yields unimodal end-members that can be reasonably approximated by parametric distributions (see also examples in *Prins et al*. [2002], *Weltje and Prins* [2003], *Vriend et al*. [2011], and *Vandenberghe* [2013]). In such cases, it is likely that the variability in the respective data sets is sufficient that a near-"pure" end-member has been sampled.

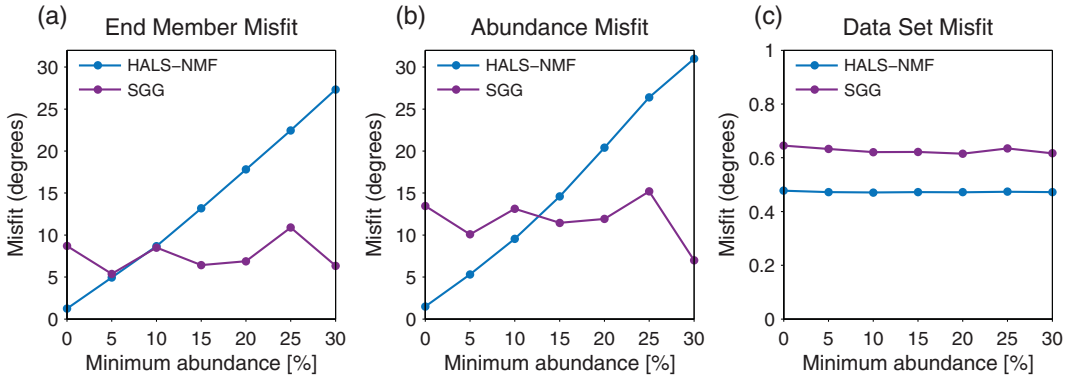

Figure 5. The misfit between the true and estimated (a) end-members, (b) abundances, and (c) the whole data set as a function of minimum end-member abundance (i.e., the level of end-member mixing).

Weltje and Prins [2007] also raise concerns about the interpretation and physical significance of the distribution parameters. Fundamentally, many distribution parameters are dimensionless numbers to control shape and scale; this is particularly true for the Weibull family. In this sense, the distribution parameters carry no physical meaning, and interpretation as such may lead to flawed conclusions. Interpretation of parametric end-members should, therefore, be based on physical descriptive statistics. This could be geometric or logarithmic graphic measures [*Folk and Ward*, 1957], or grain size percentiles, which provide a means to quantify parametric end-members in a physically intuitive way. It should be noted that these methods can be inappropriate for multimodal distributions, such as the observed grain size spectra or some nonparametric end-

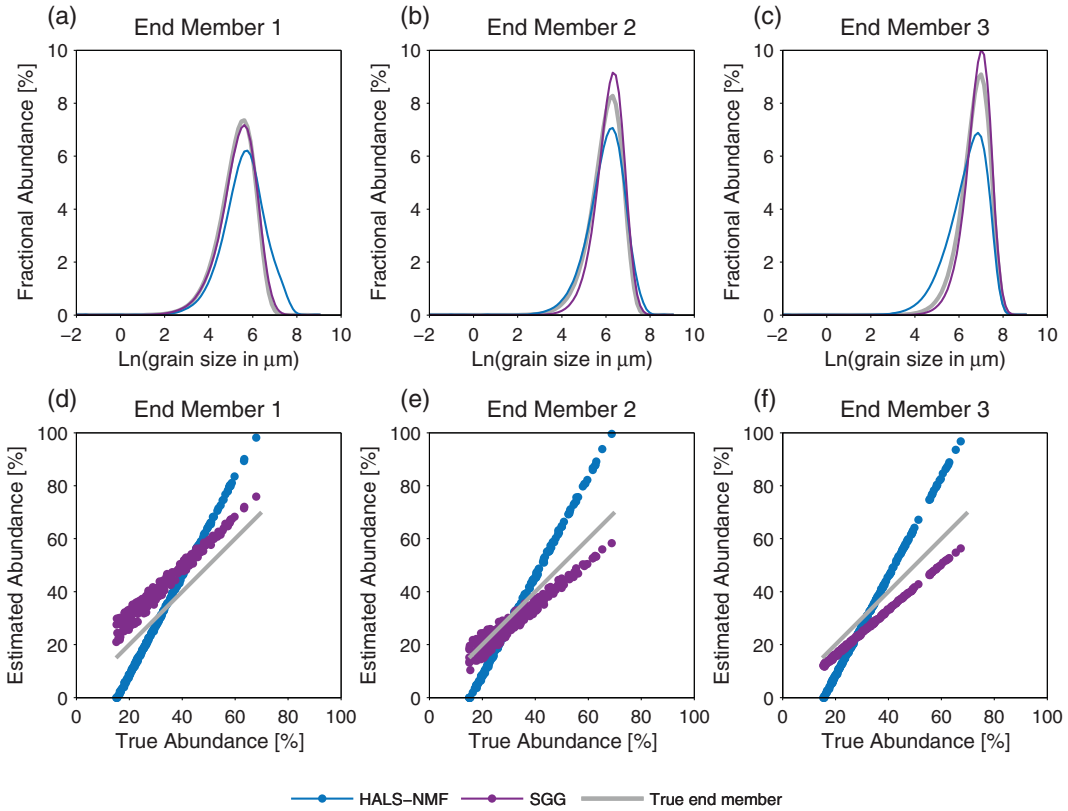

Figure 6. Fitting results from nonparametric and parametric EMA methods to a data set consisting of the end-members shown in Figure 4a when the minimum end-member abundance is 15%. (a–c) Comparison between the fitted and true end-members. (d–f) Comparison between the estimated and true abundances.

members. In this sense, parametric EMA provides end-members that are most suitable for this kind of analysis.

6. The AnalySize Software Package

To help users take advantage of the new algorithms introduced here, we have developed the AnalySize software package. AnalySize is a MATLAB GUI that provides a comprehensive set of tools for analyzing GSD data. AnalySize is capable of processing a wide range of data formats primarily, but not exclusively, obtained from laser diffraction particle size analyzers. Given that the exploration of large data sets can be time consuming, AnalySize is also capable of saving a fitting session to a standard MATLAB data file. Users can then load this data file at a later time and continue their analyses or transfer their results and data to another user.

AnalySize can be download from https://www.github.com/greigpaterson/AnalySize and can be run using MATLAB v8 or above (no additional Toolboxes required). AnalySize is compatible with both Windows and Mac systems and includes a detailed manual with the test data sets used in this paper.

All plots generated by AnalySize can be saved to encapsulated postscript files suitable for publication with no or minimal adjustment. Examples of these figures are given in the supporting information. AnalySize also provides tools for exploring and analyzing grain size data. This includes descriptive statistics and percentiles for the data and the fitted end-members. The available statistics include mean, standard deviation, skewness, and kurtosis determined by geometric and logarithmic methods of moments and geometric and logarithmic graphic analysis, as well as commonly used percentiles [*Krumbein and Pettijohn*, 1938; *Folk and Ward*, 1957; *Blott and Pye*, 2001]. Ternary plots from both Folk [*Folk*, 1954, 1974] and Shepard [*Shepard*, 1954; *Schlee*, 1973] fine and course classifications can also be generated. The clay, sand, silt, and gravel sizes follow those defined by Wentworth [1922] and the data can be exported in text format. AnalySize also supports the plotting of multiple GSDs and the CM plot of *Passega* [1964], which plots the first cumulative grain size percentile against the median grain size.

7. Summary

We have introduced and developed new algorithms for unmixing grain size distribution data. These approaches provide an improvement over existing algorithms and address major issues associated with identifying grain size subpopulations using single-specimen unmixing techniques. All of these new algorithms are available for use in AnalySize, which is a new GUI software package to analyze and unmix GSD data. In addition, AnalySize includes a suite of standard plot and statistical analyses.

These algorithms and AnalySize are designed to aid researchers to better understand their data. The software incorporates EMA techniques that yield different information. For example, nonparametric EMA identifies covarying components and hence reveals information about mixing dynamics or nonselective physical processes that inherently yield multimodal end-members. Parametric EMA provides more detail about subpopulations, which may be related to different sources or source process that can appropriately be unmixed into unimodal end-members. It should be kept in mind, however, that AnalySize is simply a set of mathematical tools and that geological information and physical plausibility should always guide users choice of analysis procedure and interpretation of the results. We warmly welcome users with suggestions, comments, or bug reports to contact us and help to improve AnalySize.

Acknowledgments
We thank Chunxia Zhang for helping initiate this work and for providing useful discussion and data to test and develop AnalySize. Liang Yi and Caicai Liu are also thanks for discussions and data. G.A.P. acknowledges funding from NSFC grant 41374072. D.H. was supported by the Australian Research Council (grant DP120103952).

References

Bioucas-Dias, J. M. (2009), A variable splitting augmented Lagrangian approach to linear spectral unmixing, in *First IEEE GRSS Workshop on Hyperspectral Image and Signal: Evolution in Remote Sensing Processing. WHISPERS 2009*, Grenoble, France, pp. 1–4, IEEE, doi:10.1109/WHISPERS.2009.5289072.

Bioucas-Dias, J. M., A. Plaza, N. Dobigeon, M. Parente, D. Qian, P. Gader, and J. Chanussot (2012), Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches, *IEEE J. Selec. Top. Appl. Earth Obs. Remote Sens.*, *5*, 354–379, doi:10.1109/JSTARS.2012.2194696.

Blott, S. J., and K. Pye (2001), GRADISTAT: A grain size distribution and statistics package for the analysis of unconsolidated sediments, *Earth Surf. Processes Landforms*, *26*, 1237–1248, doi:10.1002/esp.261.

Chen, W., and M. Guillaume (2012), HALS-based NMF with flexible constraints for hyperspectral unmixing, *EURASIP J. Adv. Signal Processes*, 2012, 54, doi:10.1186/1687-6180-2012-54.

Christophersen, N., and R. P. Hooper (1992), Multivariate analysis of stream water chemical data: The use of principal components analysis for the end-member mixing problem, *Water Resour. Res.*, *28*, 99–107, doi:10.1029/91WR02518.
Cichocki, A., A. H. Phan, and C. Caiafa (2008), Flexible HALS algorithms for sparse non-negative matrix/tensor factorization, in *IEEE Workshop on Machine Learning for Signal Processing, 2008. MLSP 2008*, pp. 73–78, IEEE, Cancun, Mexico, doi:10.1109/MLSP.2008.4685458.
Craig, M. D. (1994), Minimum-volume transforms for remotely sensed data, *IEEE Trans. Geosci. Remote Sens.*, *32*, 542–552, doi:10.1109/36.297973.
Dietze, E., K. Hartmann, B. Diekmann, J. Ijmker, F. Lehmkuhl, S. Opitz, G. Stauch, B. Wünnemann, and A. Borchers (2012), An end-member algorithm for deciphering modern detrital processes from lake sediments of Lake Donggi Cona, NE Tibetan Plateau, China, *Sediment. Geol.*, *243–244*, 169–180, doi:10.1016/j.sedgeo.2011.09.014.
Donoho, D., and V. Stodden (2004), When does non-negative matrix factorization give correct decomposition into parts?, in *Advances in Neural Information Processing Systems*, vol. 16, edited by S. Thrun, L. K. Saul, and B. Schölkopf, pp. 1141–1148, MIT Press, Cambridge, Mass.
Egli, R. (2003), Analysis of the field dependence of remanent magnetization curves, *J. Geophys. Res.*, *108*(B2), 2081, doi:10.1029/2002JB002023.
Egli, R. (2004), Characterization of individual rock magnetic components by analysis of remanence curves: 2. Fundamental properties of coercivity distributions, *Phys. Chem. Earth*, *29*, 851–867, doi:10.1016/j.pce.2004.04.001.
Folk, R. L. (1954), The distinction between grain size and mineral composition in sedimentary-rock nomenclature, *J. Geol.*, *62*, 344–359, doi:10.2307/30065016.
Folk, R. L. (1974), *Petrology of Sedimentary Rocks*, Hemphill Publ. Co., Austin, Tex.
Folk, R. L., and W. C. Ward (1957), Brazos River bar [Texas]; a study in the significance of grain size parameters, *J. Sediment. Petrol.*, *27*, 3–26.
Hannigan, R. E., A. R. Basu, and F. Teichmann (2001), Mantle reservoir geochemistry from statistical analysis of ICP-MS trace element data of equatorial mid-Atlantic MORB glasses, *Chem. Geol.*, *175*, 397–428, doi:10.1016/S0009-2541(00)00335-1.
Heinz, D. C., and C.-I. Chang (2001), Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery, *IEEE Trans. Geosci. Remote Sens.*, *39*, 529–545, doi:10.1109/36.911111.
Heslop, D. (2015), Numerical strategies for magnetic mineral unmixing, *Earth Sci. Rev.*, *150*, 256–284, doi:10.1016/j.earscirev.2015.07.007.
Heylen, R., and P. Scheunders (2012), A fast geometric algorithm for solving the inversion problem in spectral unmixing, in *4th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pp. 1–4, IEEE, Shanghai, China, doi:10.1109/WHISPERS.2012.6874221.
Heylen, R., D. Burazerovic, and P. Scheunders (2011), Fully constrained least squares spectral unmixing by simplex projection, *IEEE Trans. Geosci. Remote Sens.*, *49*, 4112–4122, doi:10.1109/TGRS.2011.2155070.
Krumbein, W. C., and F. J. Pettijohn (1938), *Manual of Sedimentary Petrography*, Appleton-Century-Crofts, N. Y.
Lee, D. D., and H. S. Seung (1999), Learning the parts of objects by non-negative matrix factorization, *Nature*, *401*, 788–791, doi:10.1038/44565.
Li, Z., D. Sun, F. Chen, F. Wang, Y. Zhang, F. Guo, X. Wang, and B. Li (2014), Chronology and paleoenvironmental records of a drill core in the central Tengger Desert of China, *Quat. Sci. Rev.*, *85*, 85–98, doi:10.1016/j.quascirev.2013.12.003.
Manson, V., and J. Imbrie (1964), *Fortran Program for Factor and Vector Analysis of Geologic Data Using an IBM 7090 or 7094/1401 Computer System, Kansas Geol. Surv. Spec. Distrib. Publ.*, vol. 13, 46 pp., Kansas Geol. Surv., Lawrence.
Meyer, I., G. R. Davies, C. Vogt, H. Kuhlmann, and J.-B. W. Stuut (2013), Changing rainfall patterns in NW Africa since the Younger Dryas, *Aeolian Res.*, *10*, 111–123, doi:10.1016/j.aeolia.2013.03.003.
Miao, L., and H. Qi (2007), Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization, *IEEE Trans. Geosci. Remote Sens.*, *45*, 765–777, doi:10.1109/TGRS.2006.888466.
Miesch, A. T. (1976), Q-mode factor analysis of geochemical and petrologic data matrices with constant row-sums, U.S. Geol. Surv. Prof. Pap., 32, 217–225.
Passega, R. (1964), Grain size representation by CM patterns as a geologic tool, *J. Sediment. Petrol.*, *34*, 830–847.
Prins, M. A., L. M. Bouwer, C. J. Beets, S. R. Troelstra, G. J. Weltje, R. W. Kruk, A. Kuijpers, and P. Z. Vroon (2002), Ocean circulation and iceberg discharge in the glacial North Atlantic: Inferences from unmixing of sediment size distributions, *Geology*, *30*, 555–558, doi:10.1130/0091-7613(2002)030<0555:OCAIDI>2.0.CO;2.
Renner, R. M. (1993), A constrained least-squares subroutine for adjusting negative estimated element concentrations to zero, *Comput. Geosci.*, *19*, 1351–1360, doi:10.1016/0098-3004(93)90034-3.
Schlee, J. S. (1973), Atlantic continental shelf and slope of the United States—Sediment texture of the northeastern part, U.S. Geol. Surv. Prof. Pap., 529-L, 64 pp.
Shepard, F. P. (1954), Nomenclature based on sand-silt-clay ratios, *J. Sediment. Petrol.*, *24*, 151–158, doi:10.1306/D4269774-2B26-11D7-8648000102C1865D.
Sun, D., J. Bloemendal, D. K. Rea, J. Vandenberghe, F. Jiang, Z. An, and R. Su (2002), Grain-size distribution function of polymodal sediments in hydraulic and aeolian environments, and numerical partitioning of the sedimentary components. *Sediment. Geol.*, *152*, 263–277, doi:10.1016/S0037-0738(02)00082-9.
Vandenberghe, J. (2013), Grain size of fine-grained windblown sediment: A powerful proxy for process identification, *Earth Sci. Rev.*, *121*, 18–30, doi:10.1016/j.earscirev.2013.03.001.
Vriend, M., M. A. Prins, J.-P. Buylaert, J. Vandenberghe, and H. Lu (2011), Contrasting dust supply patterns across the north-western Chinese Loess Plateau during the last glacial-interglacial cycle, *Quat. Int.*, *240*, 167–180, doi:10.1016/j.quaint.2010.11.009.
Wang, F., D. Sun, F. Chen, J. Bloemendal, F. Guo, Z. Li, Y. Zhang, B. Li, and X. Wang (2015), Formation and evolution of the Badain Jaran Desert, North China, as revealed by a drill core from the desert centre and by geological survey, *Palaeogeogr. Palaeoclimatol. Palaeoecol.*, *426*, 139–158, doi:10.1016/j.palaeo.2015.03.011.
Weibull, W. (1951), A statistical distribution function of wide applicability, *J. Appl. Mech.*, *18*, 293–297.
Weltje, G. J. (1997), End-member modeling of compositional data: Numerical-statistical algorithms for solving the explicit mixing problem, *Math. Geol.*, *29*, 503–549, doi:10.1007/BF02775085.
Weltje, G. J., and M. A. Prins (2003), Muddled or mixed? Inferring palaeoclimate from size distributions of deep-sea clastics, *Sediment. Geol.*, *162*, 39–62, doi:10.1016/S0037-0738(03)00235-5.
Weltje, G. J., and M. A. Prins (2007), Genetically meaningful decomposition of grain-size distributions, *Sediment. Geol.*, *202*, 409–424, doi:10.1016/j.sedgeo.2007.03.007.
Wentworth, C. K. (1922), A scale of grade and class terms for clastic sediments, *J. Geol.*, *30*, 377–392, doi:10.2307/30063207.