Tracking Proterozoic–Triassic sediment routing to western Laurentia via bivariate non-negative matrix factorization of detrital provenance data

Joel E. Saylor1* and Kurt E. Sundell2

1 Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia, Vancouver, BC, Canada
2 Department of Geosciences, University of Arizona, Tucson, AZ, USA

Abstract: When reconstructing sediment provenance, the challenge posed by multiple univariate (i.e., age-only) detrital geochronology data sets with similar one-dimensional distributions can be mitigated by incorporation of a second variable such as Hf isotopic data. However, there is no commonly applied method of sample intercomparison, much less forward or inverse modeling of these bivariate data sets. In this paper we explore application of non-negative matrix factorization (NMF) to bivariate data sets. Factorization of univariate mixed (a.k.a., sink or daughter) data sets has been demonstrated to successfully recover both the one-dimensional endmember (a.k.a., source or parent) distributions and their mixture weightings. We show that NMF can successfully recover both the two-dimensional distributions and mixing weights in synthetic data sets. Application of the method to 24 published Neoproterozoic–Triassic samples from western Laurentia yields six two-dimensional endmember distributions that are a close match to empirical sediment sources on the southern, eastern, and northern margin of Laurentia. The results are broadly consistent with previous interpretations and confirm that the method can characterize unknown sediment sources based on data from analyzed sink samples. Correlating between factorized endmembers and empirical sources indicates that the Transcontinental Arch was not a barrier to east–west sediment transport until the late Cambrian. Quantitative comparison also shows that the closest known match for a factorized endmember with highly evolved Permo-Triassic zircons is from northern South America and suggests northward transport of this detritus along the margin of Laurentia. The inability to resolve sediment sources using one-dimensional age distributions has confounded interpretive approaches that rely on the presence or absence of age modes in specific age ranges. These limitations are partly addressed using statistical measures of similarity between sink and source samples (e.g. Leary et al. 2020). Nevertheless, uncertainty persists even with statistical comparison, as indicated by variable attribution of late Neoproterozoic–early Paleozoic zircons to peri-Gondwana terranes accreted to eastern Laurentia or those from South or Central America (Gleason et al. 2007; Kissock et al. 2018; Liu and Stockli 2019). The inability to resolve sediment sources using one-dimensional (age) distributions necessitates an alternative approach. This has been successfully addressed with multivariate approaches (e.g. Gehrels and Pecha 2014; Linde et al. 2016, 2017; Thomas et al. 2020; Waite et al. 2020). Distinct sources may have unique thermal histories, Hf isotopic compositions at the time of zircon crystallization (εHf), oxygen isotope compositions, rare-earth element concentrations, or Th/U ratios (e.g. Belousova et al. 2002; Hoskin and Schaltegger 2003; Rahl et al. 2003; Fu et al. 2008;
Mueller et al. 2008; Flowerdew et al. 2009; Appleby et al. 2010; Cecil et al. 2011; Anfinson et al. 2012). Bivariate petrochronological approaches use this secondary variable in addition to grain ages to refine attribution of detrital grains or groups of grains to specific sources (e.g. Gehrels and Pecha 2014). For example, in the case of the late Paleozoic samples discussed above, systematic variations in the Hf isotopic compositions of zircons have been used to discriminate between sources in the Franklinian Basin, which have positive (‘juvenile’) $\varepsilon_{\text{HfT}}$ values, and sources along the Iapetus suture in eastern Laurentia, which have more negative (‘evolved’) $\varepsilon_{\text{HfT}}$ values.

Qualitative interpretations of bivariate data sets limit both the objectivity of those interpretations and also the ability to model the two-dimensional distributions (Sundell and Saylor 2021). For example, cross-plots of zircon U–Pb ages and $\varepsilon_{\text{HfT}}$ values are often compared to fields representing source areas in order to assess the overlap between samples and sources (e.g. Gehrels and Pecha 2014 and many others). Although providing insight and potentially discriminatory power, this approach does not leverage the modal proportions of age/$\varepsilon_{\text{HfT}}$ distributions to determine whether multiple sources may have contributed to a particular sample. Comparison of measured age/$\varepsilon_{\text{HfT}}$ pairs to source fields also provides no basis for forward or inverse modeling of sediment mixing.

In this paper we develop a non-negative matrix factorization (NMF) approach to reconstructing unknown two-dimensional endmember (a.k.a. source or parent) distributions based on known two-dimensional distributions from mixed (a.k.a. sink or daughter) samples. We first demonstrate that the algorithm is able to reconstruct both the age and $\varepsilon_{\text{HfT}}$ distributions, and the mixing weights based on factorization of sinks randomly mixed from known endmembers. We next apply the algorithm to a suite of 24 Neoproterozoic–Triassic samples from central–western Laurentia.

The algorithm identifies endmembers that match empirical sediment sources and suggests sediment transport pathways that are largely consistent with previous interpretations. It also refines identification of the source of Paleoproterozoic and Permo-Triassic zircons.

Methods

Non-negative matrix factorization

Here we extend the NMF approach, which Sharman and Johnstone (2017) and Saylor et al. (2019) applied to univariate data sets and one-dimensional distributions, to bivariate data sets and two-dimensional distributions. The difference between these approaches is similar to that between signal processing applications of NMF to waveforms (e.g. Raj et al. 2010) and image processing applications of NMF to hyperspectral images (e.g. Rajabi and Ghassemian 2014). In both instances, given a matrix $V$ (composed of $m$ features from $n$ samples) NMF seeks to develop two matrices $W$ (m-by-$k$) and $H$ (k-by-$n$) such that

$$V = WH + E$$  \hspace{1cm} (1)

where $E$ is the final residual. In the NMF algorithm implemented here $E$ is calculated as the Frobenius norm of $V-WH$ and is calculated as

$$E = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} [V_{ij} - (WH)_{ij}]^2}.  \hspace{1cm} (2)$$

Data are input to the algorithm as two-dimensional kernel density estimates (KDEs) with fixed bandwidths (Fig. 1) (Andersen 2014; Roberts and Spencer 2015; Spencer et al. 2019). Discrete data
points (i.e. age and $\varepsilon_{\text{Hf}}$) are transformed to two-dimensional distributions following the approach outlined by Sundell et al. (2019) and Sundell and Saylor (2021). Briefly, the transformation proceeds by first performing a discrete cosine transform of each discrete data set (Ahmed et al. 1974). The resulting two-dimensional matrix of points is multiplied by a Gaussian function based on the defined kernel bandwidths. The resulting matrix is then inverted using the discrete cosine transform. Finally, the matrix is normalized so that the resulting three-dimensional volume integrates to 1. The result is a two-dimensional matrix of $z$-axis values that can be visualized as a three-dimensional volume where the density is displayed as height in the $z$-axis direction (Fig. 1a) or as a two-dimensional plot of clustering of ages and $\varepsilon_{\text{Hf}}$ (‘density’, color-coded as intensity (Fig. 1b)). For all analyses presented here the bandwidth for age is 20 Myr and that for $\varepsilon_{\text{Hf}}$ values is four $\varepsilon_{\text{Hf}}$ units.

**Rank estimation**

We tested two methods to identify the optimum rank for factorization. These are the segmented linear regression breakpoint-analysis used by Saylor et al. (2019) and a bi-cross-validation method developed by Owen and Perry (2009). We also considered Akaike’s information criterion (Akaike 1974), Bayesian information criterion (Schwarz 1978), and minimum length description (Squires et al. 2017). However, we rejected all of these because in the analysis of Fu et al. (2019) bi-cross-validation more accurately estimated the true rank at low ranks (i.e. ranks <25) than any of the alternatives, and we anticipate that factorization of detrital data sets should yield fewer than 25 potential sources. As an example of the expected number of endmembers, Laskowski et al. (2013) grouped samples spanning the North American Cordilleran retroarc into seven characteristic groups based on their detrital zircon U–Pb age spectra. Similarly, May et al. (2013) grouped Late Mississippian–Cretaceous samples from the Big Horn Basin into four tectonostratigraphic assemblages, also based on their U–Pb age spectra.

Breakpoint analysis proceeds as described by Saylor et al. (2019) and is summarized here. We first plot the final residual v. the rank and apply a segmented linear regression with one break. We calculate the sum of squared residuals (SSR; Draper and Smith 1981) and compare it to the sum of squared error for the full rank solution.

**Fig. 2.** Comparison of known sources and factorized endmembers for the eight-source data set. (a) Eight source kernel density estimates that were used as sources and mixed to form 40 random sink samples. (b) Factorization of the 40 samples to eight endmembers yields distributions that closely match the original five sources as indicated by cross-correlation values approaching 1 and D values approaching 0. (c) Factorization of the 40 sink samples to seven endmembers yields distribution that match the original five more poorly. KDEs are based on a 20 Myr bandwidth for age, and a 4 $\varepsilon$-unit bandwidth for $\varepsilon_{\text{Hf}}$ and are plotted as the contour of the median (50th percentile). Abbreviations: CC: cross-correlation coefficient, K-S: Kolmogorov–Smirnov D-value.
for each segment as

$$SSR_1 = \sum_{i=2}^{n} \frac{(x_i - f(x_i))^2}{B}$$

(3)

and

$$SSR_2 = \sum_{i=2}^{n} \frac{(x_i - g(x_i))^2}{C}.$$  

(4)

In these equations, $x$ is the breakpoint, which is calculated for all integers between 2 and $n$, where $n$ is the maximum rank being considered. $R$ is the final residual for each rank ($k$), and $f(x)$ and $g(x)$ are the expected final residual calculated by linear regression over the lower-rank or higher-rank segments, respectively. $A = 2$ and $x_2$ is the breakpoint, which is calculated for all integers between 2 and $n$, where $n$ is the maximum rank being considered. $R$ is the final residual for each rank ($k$), and $f(x)$ and $g(x)$ are the expected final residual calculated by linear regression over the lower-rank or higher-rank segments, respectively. For $x_2 = 2$ and $x_2 = n$, $f(x_2)$ and $g(x_2)$ are equal to the final residual, yielding $SSR_1$ and $SSR_2 = 0$, respectively. The optimal rank is then the $x$ that minimizes the sum $SSR_1 + SSR_2$.  

Bi-cross-validation proceeds following the algorithm developed by Owen and Perry (2009). We divided the input matrix $V$ from Equation (1) (with dimensions $i \times j$) into four quadrants as follows:

$$V = \begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

(5)

where $A \in \mathbb{R}^{i \times k}$, $B \in \mathbb{R}^{j \times (j-k)}$, $C \in \mathbb{R}^{j \times (i-k)}$, and $D \in \mathbb{R}^{j \times j}$. We then conducted NMF on all quadrants ($A$, $B$, $C$, $D$) for all ranks from $k = 2$ to $k = n$ and used the factorized matrix to calculate the residual of the matrix on its diagonal. Hence, $D = \tilde{W}_D H_D$, is the NMF of quadrant $D$ for rank $k$ and is used to calculate the residual for $A$ in Equations (6–8) below. We calculated the residual difference ($R^2$) between the factorized matrices and their known counterparts as the Frobenius norm of

$$A - \tilde{W}_D H_D,$$

(6)

where

$$\tilde{W}_D = B H_D^{-1} C,$$

(7)

$$\tilde{H}_D = \tilde{W}_D D C.$$  

(8)

In this scenario the optimum rank is the $k$ that minimizes the residual $R$. We also considered the rank that minimizes the sum of residuals $R_{tot} = R_A + R_B + R_C + R_D$.  

**Synthetic data sets**

We first tested the method on three numerically generated bivariate data sets. Inputs to the NMF algorithm were comprised of 24–40 samples of paired Pb and εHe data mixed from three, five, or eight known sources. Mixing was done by first assigning a random weight to each of the sources using the randomization algorithm of Sundell and Saylor (2017). Age and εHe was determined from each of the sources in proportion to their randomly assigned weight. Factorization of the data sets with three and five sources proceeded similarly to that with eight sources (Supplemental Tables S1–S6). Therefore, for brevity we discuss the data sets with eight sources below, while the others are fully discussed in the Supplemental text.  

Eight sources were derived from a global database of Pb and εHe data from igneous zircons compiled by Puetz and Condie (2019) (Fig. 2a and Supplemental Tables S7 and S8). We divide the data set into eight continental land masses following the Paleozoic suture boundaries identified by Domeier and Torsvik (2014). These include Africa ($n = 4882$), Antarctica ($n = 2084$), Asia ($n = 14881$), Australia ($n = 4654$), Baltic ($n = 2115$), North America ($n = 4264$), South America ($n = 3948$), and a suite of peri-Gondwanan terranes ($n = 10289$). We mixed these eight sources into 40 sinks where all sink samples have 2084 age/εHe pairs.

---

**Quantitative comparison**

We applied two statistical measures to compare the reconstructed sink samples to known sink samples and factorized sources to known sources. The first is the cross-correlation coefficient as applied in provenance studies (Saylor et al. 2012, 2013; Saylor and Sundell 2016; Sundell and Saylor 2021). This is applied to pairs of two-dimensional KDEs by comparing all elements of a square matrix (i.e. $m \times m$) which represents KDE1 with the corresponding elements of KDE2. For computational efficiency each matrix is vectorized ($\text{KDE}_{m \times n} \rightarrow \text{KDE}_{1 \times m^2}$), where $\text{KDE}$ has dimensions $m^2 \times 1$, and the cross-correlation coefficient is then calculated using the

![Fig. 3.](http://jgs.lyellcollection.org/)

Fig. 3. (a) Breakpoint analysis based on factorization of 40 sink samples from eight sources to ranks between 2–27 yields a minimum sum of squared residuals at rank eight, closely followed by rank seven. (b) Bi-cross-validation of based on factorization of four quadrants (i.e. A, B, C, D in equation (5)) from the 40 sink samples for ranks 2–27 yields minima in residual at ranks ranging from 5 to 26. (c) The sum of residuals of each of the four quadrants yields a minimum at a rank of 16. Data for this figure are presented in Supplemental Table S9.)
standard equation

\[ R^2 = \frac{\left( \sum_{i=1}^{n} (V_i - \bar{V})^2 \right)^2}{\left( \sum_{i=1}^{n} (V_i - \bar{V})^2 + \sum_{i=1}^{n} (\bar{V}_i - \bar{V})^2 \right)} . \]  

The second statistical measure is the Kolmogorov–Smirnov distance (D value). Unfortunately, unlike the univariate case, the cumulative distribution function (CDF) is not uniquely defined in more than one dimension. We therefore follow Peacock (1983) and Press and Teukolsky (1988) in calculating the D value for each quadrant surrounding each point in two two-dimensional CDFs. The final D value is then the maximum D value calculated from each quadrant (see Sundell and Saylor 2021, for further discussion).

We also use multi-dimensional scaling (MDS) to visualize the relationship between the six factorized endmembers and the 29 potential empirical sources (Vermeesch 2013). MDS attempts to present the dissimilarity between samples as distance in N-dimensional space. Samples are represented as a point, typically in two- or three-dimensional Cartesian space, with greater distances between two points indicating greater dissimilarity between the two samples. The transformation is accomplished by iterative rearrangement of the data in N-dimensional space to minimize the misfit (‘stress’) between the calculated distances and the disparities. In a low-stress MDS plot, the distances between points linearly correlate to the dissimilarities between samples. MDS was conducted using the D2mds software package and optimized the metric squared stress based on the coefficient of non-determination (i.e. complement of the cross-correlation coefficient, 1–\( R^2 \)) (Saylor et al. 2017).

Squared stress is calculated as

\[ \left( \frac{\sum_{j} (f(x_j) - d_j)^2}{\sum_{j} d_j^2} \right)^{0.5} , \]

where \( d_j \) is the distance and \( f(x_j) \) is the disparity between the \( i \)th and \( j \)th element. Disparity is calculated as a linear (1:1) transformation of the input dissimilarities.

Results

Rank estimation

For data sets randomly mixed from known sources, breakpoint analysis consistently yields optimum ranks within one of the known number of sources (Fig. 3a, Supplemental Figures S2A and S5A). In contrast bi-cross-validation yields different optimum ranks based on the residual for each of the four quadrants (i.e. A, B, C, D) in Equation (5) examined, as well as for the total residual (Fig. 3b and c, Supplemental Figures S2B–C and S5B–C). For example, breakpoint analysis of the 40 sink samples mixed from eight sources indicates that the optimum rank for factorization is eight, although rank seven is only slightly worse (Fig. 3a, Supplemental Table S9). In contrast, bi-cross-validation yields a range of optimum ranks between 5 and 26 (Fig. 3b) for the individual quadrants. The total residual indicates that the optimum rank is 16 (Fig. 3c).

Bivariate distributions and mixture weights

Since breakpoint analysis indicates that a rank of either seven or eight yields optimum factors (Fig. 3a), we factorized the 40 sink samples mixed from eight sources to both seven and eight endmembers (Fig. 2b and c). The cross-correlation coefficient and D value indicate close similarity between the known sources to both seven and eight endmembers factorized from the mixed sink samples. However, there is a slightly better match between the endmembers factorized for rank eight than for rank seven (compare Fig. 2b and c). The mean cross correlation coefficient between factorized endmembers for rank eight and the known sources is 0.99. In comparison the mean cross-correlation coefficient between factorized endmembers for rank seven and the known sources is 0.96, or 0.93 if the best match to Input 5 is included (Fig. 2c).

There is close correspondence between the mixture weighting used to mix the eight sources into 40 sink samples and the weightings based on factorization of the 40 sinks (Fig. 4). Cross-plotting mixture weights from the factorization to rank eight and known mixture weights yields a linear fit with a slope of 0.96 and a coefficient of determination of 0.98 (Fig. 4a). Cross-plotting mixture weights from the factorization to rank seven and known mixture weights yields slightly lower coefficient of determination (0.92), but a linear fit with a slope (1.01) which is closer to 1 than the slope for rank eight (Fig. 4b).

Empirical data set

As a second test of the algorithm as applied to an empirical data set, we factorized 24 Neoproterozoic–Triassic U–Pb and eHf data sets from basins spanning central and western North America (Figs 5 and 6) (Wooden et al. 2013; Gehrels and Pecha 2014; Linde et al. 2016, 2017; Thomas et al. 2020). We compared the results of this factorization to 29 potential sources in order to assess whether factorization of empirical samples results in recovery of plausible...
Bivariate distributions (Figs 5 and 6 and Supplemental Table S7) (Andersen et al. 2002, 2007, 2011; Bickford et al. 2008; Mueller et al. 2008; Rohr et al. 2008, 2010; Flowerdew et al. 2009; Appleby et al. 2010; Bickford et al. 2010; Rehnström 2010; Stewart et al. 2010; Weber et al. 2010; Brander et al. 2011; Anfinson et al. 2012; Augland et al. 2012; Malone 2012; Cochrane 2013; Doe et al. 2013; LaFlamme et al. 2013; Willner et al. 2013; Wooden et al. 2013; Holland et al. 2015; Pollock et al. 2015; Henderson et al. 2016; Thomas et al. 2017; Zirakparvar et al. 2017; Cecil et al. 2019). See Supplemental Table S10 for samples, sample groupings, and sources.

Breakpoint analysis indicates that the optimum factorization rank for the 24 Neoproterozoic–Triassic U–Pb and εHfT data sets is six (Fig. 7). Factorization to rank six yields the distributions shown in Figure 8a. Endmember 1 is characterized by U–Pb ages between 200–300 Ma and εHfT values ranging from 10 to −30 coupled with Mesozoic–Paleoproterozoic U–Pb ages with εHfT values >0 (Fig. 8a). Endmember 2 has similar Mesozoic–Paleoproterozoic U–Pb age distributions and associated positive εHfT values but lacks the late Paleozoic–Mesozoic U–Pb ages (Fig. 8a). Endmember 3 is characterized by a positively skewed mode between 1.0–1.2 Ga and εHfT values between 20 and −10 (Fig. 8a).

Fig. 5. Zircon U–Pb age and εHfT data sets from Neoproterozoic–Triassic samples from basins spanning central–western Laurentia. Factorization of these samples forms the basis for the paleogeographic reconstruction in Figure 10. White contour lines are plotted at the median (50th percentile). Roman numerals are keyed to locations in Figure 6. The data in this figure are presented in Supplemental Table S10.
Endmember 4 has a single primary U–Pb/εHf mode between 1.75 and 1.8 Ga, which is characterized by a wide range of εHf values from 20 to −30, and a secondary mode at c. 2.6 Ga that has εHf values of 0 ± 10 (Fig. 8a). Endmember 5 has a primary early Paleozoic U–Pb age mode that has a range of εHf values from 20 to −30 and a secondary late Neoproterozoic mode with similar εHf values (Fig. 8a). Endmember 6 is dominated by a positively skewed U–Pb age mode at c. 1.8 Ga that displays a range of εHf values from 10 to −30. It has a secondary U–Pb age mode at c. 2.7 Ga with εHf values of 0 ± 10 (Fig. 8a).

Discussion

Synthetic data sets

Analysis of the synthetic data sets indicates that breakpoint analysis provides a more consistent optimum factorization rank than bi-cross-validation. The optimal rank yielded by bi-cross-validation ranged from c. 0.75 to more than three times the known number of sources (Fig. 3b, Supplemental Figures S2B and S5B). The optimal rank estimated by breakpoint analysis also more closely matches the known number of sources than the rank estimated by bi-cross-validation. The final optimal rank based on the sum of residuals from bi-cross-validation of the four quadrants was consistently double the known number of sources (Fig. 3c, Supplemental Figures S2C and S5C). In contrast, breakpoint analysis yielded consistent optimal ranks, and the optimal ranks were within one of the known number of sources (Fig. 3a, Supplemental Figures S2A and S5A).

Factorization to the optimum rank indicated by breakpoint analysis yields distributions and mixing weights that closely match the known distributions and mixing weights. For example, factorizing the eight-source data set to seven or eight endmembers yields high mean cross-correlation coefficients (0.96–0.99). Similarly, for the eight-source data set the $R^2$ between known and factorized mixing weights is slightly higher for the eight-endmember factorization (0.98) than for the seven-endmember factorization (0.92) (Fig. 4).

These results indicate that NMF of bivariate data sets provides insight into the distributions and mixtures present in the data. We conclude that NMF of bivariate data sets is a reliable method to reconstruct both the distribution of characteristics of sediment sources and also the weighting of those sources in sink populations.

Empirical data set

Loci of potential sources

Neoproterozoic–Triassic strata from the western Laurentian margin, Ancestral Rocky Mountain basins, and mid-continent basins contain zircons with Archean–Triassic U–Pb ages (e.g. Gehrels et al. 2011; Kissock et al. 2018; Leary et al. 2020, and many others). Significant age ranges have multiple sources in basement provinces in Laurentia, as laid out below. In addition to direct derivation from these crustal domains, zircons could be recycled through sedimentary strata derived from the initial zircon source terrane. However,
Fig. 8. Comparison between sources factorized from the Laurentian Neoproterozoic–Triassic data set and their closest and second-closest empirical correlatives. (a) Factorized endmembers. (b) Closest matches sorted by row and shown adjacent to the factorized endmembers that they match. (c) Second-closest matches sorted by row and shown in the same row as the factorized endmembers that they match. CC denotes the cross-correlation coefficient between the empirical source and the factorized endmember on the same row. Data for this figure are presented in Supplemental Table S10.

Recycling is likely to mix sources in distinct proportions as described below in the section titled ‘Sediment routing in Paleozoic Laurentia’. The difference in proportions is exploited by the NMF algorithm to factorize out both the age/εHf distributions and the mixing proportions. On the other hand, if an age/εHf distribution is dispersed homogeneously throughout the sink data set, the NMF algorithm will be unable to factorize that source into contributing components (see the discussion in Saylor et al. 2019; Smith et al. 2020).

Mesoarchean–early Paleoproterozoic (>2.4 Ga) zircons may be derived from cratonic provinces and intervening Proterozoic orogens and arcs of northern Laurentia (Percival et al. 2004). This includes the Superior (3.8–2.6 Ga; Hoffman et al. 1989; Percival et al. 2004), Slave (3.3–2.5 Ga; Bleeker et al. 1999; Ootes et al. 2005), Medicine Hat (3.3–2.5 Ga; Villeneuve et al. 1993; Gifford et al. 2020), Rae (3.4–2.3 Ga), and Wyoming provinces (3.5–2.5 Ga; Ross 2002; Chamberlain et al. 2003).

Early–middle Paleoproterozoic (c. 2.4–2.0 Ga) zircons are relatively rare in Laurentia, but have been reported in accreted terranes in northwestern Laurentia including the Tatson Orogen, the Buffalo Head terrane, the Chinchaga Domain (Ross et al. 1991; Théraulaz and Ross 1991; Villeneuve et al. 1993; Ross 2002), and sedimentary strata derived from them (Partin et al. 2014; Shiels et al. 2017, and references therein). Early–middle Paleoproterozoic crystallization ages have also been reported from limited intrusions in the Slave, Rae, and Hearns cratons (e.g. Hartlaub et al. 2007; Mumford 2013).

Paleoproterozoic–early Mesoproterozoic (2.0–1.6 Ga) zircon U–Pb ages may be from the Yavapai–Mazatzal terrane of southwestern Laurentia, basement of Greenland and Scandinavian (Hoffman 1988; Kalsbeek et al. 1993, 2008; Fonneland et al. 2004; Nutman et al. 2008; Gehrels et al. 2011; Slama et al. 2011; McClelland et al. 2016; Linde et al. 2017), and also orogens and arcs of the northern Laurentian Wopmay Orogen (1.8–1.9 Ga and 2.5–2.7 Ga) (Jackson et al. 2013). Similarly, zircons from the Mojave crustal block of southwestern Laurentia have crystallization ages of c. 1.8–1.4 Ga (Wooden et al. 1988; Hawkins et al. 1996; Barth et al. 2000). Depleted mantle Nd model ages and inherited zircon ages of 2.3–2.0 Ga in Mojave province granites indicate incorporation of older crustal material (Hawkins et al. 1996; Barth et al. 2000).

Late Mesoproterozoic (‘Grenville’) zircons have a wide variety of potential sources. These include Gondwanan terranes in the northern Appalachian Orogen, and exposed and subsurface sources in the Grenville Province of eastern and southern Laurentia (Percival et al. 2004; Fyffe et al. 2009; Macdonald et al. 2014; Karabinos et al. 2017; Thomas et al. 2017; Kissock et al. 2018). Alternatively, they have been attributed to the Ellesmerian Orogen and clastic wedge based on statistical similarity (Leary et al. 2020), Proterozoic basement of central Mexico (Weber et al. 2010), Gondwanan and peri-Gondwanan sources accreted to southern Laurentia (Liu and Stockli 2019), or Scandinavian Paleozoic strata and Proterozoic basement (Andersen et al. 2007, 2011, 2014; Brande et al. 2011; Augland et al. 2012; Kristoffersen et al. 2014).

Zircons with late Neoproterozoic (Cryogenian–Ediacaran)–early Paleozoic ages have been attributed to peri-Gondwanan terranes accreted to Laurentia during the assembly of Pangea (Kissock et al. 2018), subduction-related magmatism on the eastern and northeastern margin of Laurentia (Rehnström 2010), and Iapetan syn-rift rocks exposed in the northern Appalachian Orogen (Thomas 2014). Within this group, Cambrian-aged zircons have been attributed to sources ranging from the Amarillo–Wichita Uplift, to peri-Gondwanan terranes, or isolated Cambrian granites in New Mexico and Colorado (McMillan and McLenmore 2004; Martens et al. 2010; Thomas et al. 2016). Finally, Permo-Triassic magmatism occurred along the western margin of Laurentia and even into northern South America (Riggs et al. 2016; Cecil et al. 2019).
Comparison of factorized and empirical sources

Three-dimensional MDS enables a broad evaluation of the groups of samples that most closely approximate factorized endmembers (Fig. 9). Although the MDS plot yields coherent groupings we treat the results with caution because the final stress is relatively high (0.36). Numbers following source descriptions in this paragraph refer to the legend of Figure 9. Endmember 1 is closely correlated to zircons from the 1.4 Ga ‘anorogenic’ granite province (2), or Permo-Triassic plutons in Ecuador and Colombia (6) or the Mojave province of California (7). Endmember 2 is correlated to zircons from igneous and metamorphic basement rocks of the Mojave terrane (1), the 1.4 Ga ‘anorogenic’ granite province (2), Mesoproterozoic strata of the Belt Supergroup from Idaho and Wyoming (3), and igneous and metamorphic basement rocks of the Yavapai and Mazatzal terranes (4). Endmember 3 has multiple correlations in eastern and northern Laurentia (22, 24, 25, 27, 29), Scandinavia (23), and central Mexico (28). Endmembers 4 and 6 are most closely correlated to the >1.7 Ga age/εHfT distributions from the Severdup and Wendal Sea basins (8) or Paleoproterozoic metavolcanics and metagranites from Labrador (9). Endmember 5 correlates to Paleozoic strata from the Appalachian foreland basin (16, 20), Ordovician strata deposited on the Pearya Terrane (Ellesmere Island) prior to its accretion to northern Laurentia (15), Caledonia granites exposed in the UK (17) and eastern Greenland (18), and sedimentary strata along the Iapetus suture exposed in Nova Scotia (19).

Given the high stress associated with the MDS plot, we identify the closest matches to the six endmembers based on the cross-correlation coefficient in order to further hone our interpretation (Figs 8 and 10). Endmember 1 is most closely correlated to zircons from the 1.4 Ga ‘anorogenic’ granite province, and secondarily to Triassic plutons in Ecuador and Colombia. Endmember 2 is best matched by zircons from igneous and metamorphic basement rocks of the Mojave terrane, with the 1.4 Ga ‘anorogenic’ granite province providing a second-best match. The closest match to Endmember 3...
is the >900 Ma distribution of ages and εHf values from Paleozoic Appalachian foreland strata. Endmember 3 has a secondary match to Proterozoic basement and igneous clasts from glacial till from New York. Both Endmembers 4 and 6 are most closely correlated to the >1.7 Ga age/εHfT value distributions from the Sverdup and Wandel Sea basins and secondarily to Paleoproterozoic basement from Labrador. Finally, Endmember 5 is best matched either by the <900 Ma age/εHfT value distribution from Paleozoic strata of the Appalachian foreland or the entire distribution from the same strata.

**Sediment routing in Paleozoic Laurentia**

Factorization of 24 Neoproterozoic–Triassic sample groups from central–western Laurentia yields endmembers that match known potential sources of sediment on the southern, eastern and northern margins of Laurentia. This broadly supports previous interpretations of late Paleozoic sediment transport (Gehrels et al. 2011; Gehrels and Pecha 2014; Chapman and Laskowski 2019; Leary et al. 2020; Thomas et al. 2019) but refines previous source attributions and hence paleogeographic reconstructions. Zircon proportions in basin strata are a function of the volume of rock eroded, but also the lithology and zircon fertility of that rock, weathering, sediment flux, sorting during transport, sediment transport mechanism, and sampling and processing procedures (Moecher and Samson 2006; Key et al. 2012; Slama and Kosler 2012; Malusà et al. 2013, 2016; Lawton et al. 2015; Ibáñez-Mejía et al. 2018; Malkowski et al. 2019; Pettit et al. 2019). As a result, we treat proportions in the discussion below with caution, relying primarily on whether samples indicate a single dominant source or multiple sources. In the discussion below all coordinates are given in a modern reference frame.

**Proterozoic–Early Cambrian.** Factorization of 24 Neoproterozoic–Triassic sample groups from central–western Laurentia yields endmembers that match known potential sources of sediment on the southern, eastern and northern margins of Laurentia. This broadly supports previous interpretations of late Paleozoic sediment transport (Gehrels et al. 2011; Gehrels and Pecha 2014; Chapman and Laskowski 2019; Leary et al. 2020; Thomas et al. 2019) but refines previous source attributions and hence paleogeographic reconstructions. Zircon proportions in basin strata are a function of the volume of rock eroded, but also the lithology and zircon fertility of that rock, weathering, sediment flux, sorting during transport, sediment transport mechanism, and sampling and processing procedures (Moecher and Samson 2006; Key et al. 2012; Slama and Kosler 2012; Malusà et al. 2013, 2016; Lawton et al. 2015; Ibáñez-Mejía et al. 2018; Malkowski et al. 2019; Pettit et al. 2019). As a result, we treat proportions in the discussion below with caution, relying primarily on whether samples indicate a single dominant source or multiple sources. In the discussion below all coordinates are given in a modern reference frame.

**Late Cambrian–Ordovician.** There is a marked change in sediment source for samples from southwestern Laurentia (including Nevada–Utah and Sonora) during this interval. Factorized sources for the majority of the samples from western sources that shed detritus to the Mesozoic Sverdup and Wandel Sea basins (Rohr et al. 2008, 2010) and that match Proterozoic basement sources in Labrador and Newfoundland (LaFlamme et al. 2013). We interpret this as reflecting a sediment dispersal system that traversed northern Laurentia in a roughly east–west direction (Fig. 11a and b). With this dataset we cannot rule out the possibility of deposition in and recycling from intermediate depocenters. Likewise, factorization of samples from California, Nevada–Utah and Sonora are consistent with derivation from the same sources that later shed sediment to the Appalachian foreland basin (Fig. 8a, Endmember 3). We interpret this as reflecting an east–west transcontinental sediment routing system that delivered sediment from the Grenville basement of eastern Laurentia to western and southwestern Laurentia (Fig. 11a and b). These samples also indicate derivation from local sources in the Mojave province (Fig. 8a, Endmember 2).

The east–west transcontinental sediment dispersal indicated by this combination of sources and sinks is at odds with the conclusion of Gehrels and Pecha (2014) that the Transcontinental Arch formed a topographic barrier to westward sediment dispersal across the mid-latitudes of Laurentia. However, it provides support for a similar scenario of transcontinental sediment transport proposed by Linde et al. (2017). Furthermore, these conclusions are consistent across six (of seven) Mesoproterozoic–early Cambrian data sets from southwestern Laurentia (Fig. 11a and b), suggesting that the model results and application to bivariate data are robust. We conclude that this is an instance where quantitative modeling and comparison of bivariate data sets is able to refine and quantitatively assess sediment provenance and so provide objective discrimination between competing hypotheses.
Fig. 11. Paleogeographic maps of Laurentia from Neoproterozoic through Triassic time showing the location of sediment sinks (stars), sources (circles) and inferred sediment transport pathways (black lines and arrows). Circles are scaled based on the mixing weight estimated by non-negative matrix factorization. All maps are oriented in their modern spatial frame with north at the top of the figures. (a), (b) In the Mesozoic–Neoproterozoic and Neoproterozoic–Lower Cambrian timeslices factorization yields sources consistent with those in eastern Laurentia, suggesting transcontinental sediment transport. Samples in southwestern Laurentia also indicate derivation from local sources. (c) In the late Cambrian–Ordovician timeslice all sink sample models are dominated by...
Laurentia are consistent with the derivation from sources that shed sediment into the Mesozoic Sverdup and Wandel Sea basins (Rohr et al. 2008, 2010), and basement sources in Labrador and Newfoundland (LaFlamme et al. 2013) (Fig. 8a, Endmembers 4 and 6). This implies large-scale southward or southsouthwestward sediment transport and is likely a response to the onset of the Taconic phase of the Caledonian Orogeny (Fig. 11c) (McKerrow et al. 2000; van Staal et al. 2009). We follow previous authors in concluding that sediment was dispersed via a combination of westward fluvial and southward long-shore current transport (Chapman and Laskowski 2019; Leary et al. 2020). The exception to these transport pathways is the Viminii Formation (Nevada) which is largely attributed to sources in the Mojave province (Fig. 8a, Endmember 2), and suggests a northward sediment transport pathway. Consistent with emergence of the Transcontinental Arch as a topographic feature between the lower and upper Cambrian (Linde et al. 2017), NMF identifies no significant contribution of detritus attributable to eastern Laurentia in western Laurentian samples at this time (Fig. 11c).

Devonian. Detrital zircon data point to continued westward shedding of sediment from both northern and southwestern Laurentia. Samples from northern British Columbia continue to be dominated by a similar source as in the upper Cambrian–Ordovician (Fig. 8a, Endmembers 4 and 6). While this data set is compatible with either direct derivation from northern Laurentia or recycling Cambrian–Ordovician strata, the lack of mixing suggests direct derivation (i.e. little or no recycling). In comparison, samples from Nevada–Utah are mixtures of multiple sources, pointing to recycling of underlying strata. The simple distributions of samples from Sonora suggests direct derivation from local basement sources (Fig. 8a, Endmember 2). Thus, in southwestern Laurentia the mixture of sources attributed to northern Laurentia with local, Mojave block basement suggests that both the local basement and overlying strata were eroded and deposited in adjacent basins (Fig. 11d).

Carboniferous–Permian. The Carboniferous–Permian saw orogenesis on all margins of Laurentia as well as intraplate deformation in the Ancestral Rocky Mountains (e.g. Leary et al. 2017). Sediment provenance reflects this complexity and our conclusions are correspondingly tentative given our small sample size. Samples from southwestern Laurentia continue to reflect a mixture of multiple sources including northeastern Laurentia, Mojave block, and 1.4–1.5 Ga magmatic sources (Fig. 8a, Endmembers 1, 2, 4 and 6). We follow previous interpretations and attribute this signal to extensive recycling of strata on the western margin of Laurentia in response to Ancestral Rocky Mountain orogenesis and recycling of Roberts Mountain Allochthon or Antler foreland basin strata (Gehrels and Dickinson 2000; Gehrels and Pecha 2014; Linde et al. 2017).

Nevertheless, Mississippian strata from Arizona (Gehrels et al. 2011), Permian strata from Sonora (Gehrels and Pecha 2014), and Carboniferous strata from mid-continental basins (Thomas et al. 2020) record an influx of zircons with ages of 1.0–1.2 Ga and juvenile–evolved εHfT values. NMF suggests that these are derived from sources that delivered sediment to the Appalachian foreland basin, and mid-continental basins (Fig. 8a, Endmember 3) (Thomas et al. 2017, 2020). With regard to basement lithologies, this source is most similar to basement and clasts from glacial tillites in New York (Zirakparvar et al. 2017). Samples from Laurentian mid-continental basins and British Columbia have age/εHfT distributions that are similar to the <900 Ma portion of the age/εHt distribution from Appalachian foreland basin strata (Thomas et al. 2017) (Fig. 8a, Endmember 5). With regard to basement sources, this source is most similar to Caledonian granites in eastern Greenland (Rehnström 2010).

The presence of zircons with ages of 1.0–1.2 Ga and juvenile–
evolved εHfT values (Fig. 8a, Endmember 3) in Arizona, Sonora, and the mid-continental basins is consistent with the hypothesis that the sediment routing systems had buried and/or bypassed the Transcontinental Arch (Fig. 11e). Such an alternative interpretation is local derivation: on the western margin of Laurentia via recycling from underlying pre-Devonian strata, and in the mid-continental basins directly from Grenville crust. Following previous research, and for the reasons below, we favor the hypothesis that at least some sediment was transported from eastern to western Laurentia and represents re-integration of a transcontinental sediment transport system in southwestern Laurentia (Gehrels et al. 2011; Chapman and Laskowski 2019; Leary et al. 2020).

The Carboniferous–Permian interval saw the appearance of a source with a major Neoproterozoic–early Paleozoic age mode and juvenile–
evolved εHfT values (Fig. 8a, Endmember 5) in mid-continental basins and British Columbia. This source is closely matched by the <900 Ma portion of the distribution from Appalachian foreland basin strata and also Caledonian granites from eastern Greenland. At the same time, a portion of the distribution from Sonora is attributed to the same source that delivered sediment to the Appalachian foreland basin and mid-continental basins (Fig. 8a, Endmember 3) (Thomas et al. 2017, 2020), and is similar to basement and clasts from glacial tillites in New York (Zirakparvar et al. 2017). These sources are not present in pre-Devonian strata of either Sonora or British Columbia (Fig. 11e), suggesting that they were introduced for the first time in the Carboniferous–Permian via transcontinental transport. As has been highlighted by previous research, this interval represents an overtopping and/or bypass of the Transcontinental Arch and re-initiation of major sediment delivery from central eastern to western Laurentia (Chapman and Laskowski 2019).

Triassic. Triassic strata record a reorganization of sediment dispersal from southwestward to northwestward (Lawton 1994; Dickinson and Gehrels 2006). Mixture weights of Triassic samples in Sonora, Nevada–Utah and British Columbia indicate contributions from Endmember 1 (Fig. 8a). This source is mostly matched by 1.4–1.5 Ga basement from southern and central Laurentia (Fig. 10). Its second closest match is to highly evolved Permo-Triassic sources from Colombia and Ecuador (Cochrane 2013). There is a greater proportion of this source in Sonoran samples than in samples from Nevada–Utah, and it constitutes only a minor percentage of the zircons in samples from southern British Columbia. The northward dilution of this source is consistent with it being located in southwestern Laurentia or northwestern Gondwana (i.e. modern Colombia and Ecuador) and progressive dilution during northward transport (Fig. 11f).

Quantitative comparison between Endmember 1 and potential empirical sources identifies Permo-Triassic sources in northwestern Gondwana, rather than the more juvenile Permo-Triassic granitoids

Quantitative comparison between Endmember 1 and potential empirical sources identifies Permo-Triassic sources in northwestern Gondwana, rather than the more juvenile Permo-Triassic granitoids.
or associated volcanism of the Mojave region (Riggs et al. 2016; Cecil et al. 2019), as the closest known match. This leaves two interpretive possibilities. First, at a minimum it requires relocating the source for Triassic strata away from the early Cordilleran arc of southern California and northern Mojave region to somewhere else. Based on the reasoning above, we suggest that potential sources should be sought to the southeast. Second, if the northwestern Gondwanan provenance is confirmed by future analyses, this conclusion requires defining a non-marine sediment distribution system that could transport sediment from Gondwana to Laurentia. In either case, this represents an instance where bivariate modeling and quantitative comparison can significantly revise sediment source identifications and therefore sediment dispersal pathways.

Summary. Several revisions of previous interpretations can be made based on NMF of the bivariate data sets. This includes attribution of Proterozoic–lower Cambrian detritus to east–west sediment dispersal, supporting a model in which the Transcontinental Arch did not emerge as a barrier to east–west transport until the late Cambrian (cf. Gehrels and Pecha 2014; Linde et al. 2017). Similarly, derivation of upper Cambrian–Ordovician detritus from western Laurentia as indicated by the bivariate data set confirms previous interpretations of SSW-directed sediment dispersal. Finally, attribution of Permo-Triassic zircons to a source in Gondwana revises previous source attributions.

The results of this study set the stage for interpretation of provenance records in the Jurassic–Paleogene retroarc foreland basin. First, we demonstrate that NMF can identify sediment sources in transcontinental sediment transport systems and we can therefore anticipate success applying this method to foreland basin strata. Second, recognizing these sources and their distribution in Proterozoic–Triassic strata on the western Laurentian margin provides a framework in which to interpret Jurassic–Paleogene provenance data. Particularly, the ability to discriminate zircons that are attributable to Yavapai–Mazatzal basement or Belt Supergroup strata from those attributable to northern or northeastern Laurentian sources suggests that these differences can be discriminated as they are recycled into younger foreland basin strata.

Software availability
We have developed an open-source software package for one-dimensional and two-dimensional NMF of detrital provenance data sets: DZnmf2D. This package is developed as a graphical user interface that can be run in a MATLAB environment or as a stand-alone executable in Windows or MacOS. The software, example data sets, and a user manual are available in the supplemental material to this paper and in public repositories including GitHub.

Conclusion
In this paper we develop an application of NMF to bivariate detrital data sets. We show that the algorithm can recover both the bivariate endmember distribution and also the mixing weights by factorizing mixtures of known sources. The recovered endmember distributions are almost identical to the known sources and the mixing weights show close correlation to the known mixing weights. We conclude that NMF provides a robust method of determining the endmember distributions and mixing weights in mixed two-dimensional detrital distributions.

After evaluating possible approaches to determining the optimum factorization rank, we conclude that breakpoint analysis provides the most robust method of determining the optimum rank. Bi-cross-validation yielded inaccurate and imprecise estimates of the factorization rank and number of sources. In contrast breakpoint analysis consistently yielded an optimum rank within one of the known number of sources.

Application of the method to 24 Proterozoic–Triassic samples from western Laurentia yielded an optimum rank of six. Factorization to six endmembers yielded distributions that closely match known potential sources on the northern, eastern, and southern margins of Laurentia. The mixing weight of these endmembers is broadly consistent with previous interpretations of Paleozoic–Mesozoic sediment dispersal and suggests broad compatibility amongst previously published qualitative and quantitative interpretations of univariate detrital geochronology data, previously published qualitative interpretations of bivariate data, and the interpretations based on quantitative analysis of bivariate data presented here. Although broadly compatible, NMF of bivariate data sets revises some sediment transport reconstructions and clarifies early Paleozoic and Triassic sediment sources. Most notably, the revised provenance attributions support emergence of the Transcontinental Arch as a barrier to east–west transport in the late Cambrian and a Gondwanan source for Permo-Triassic zircons with evolved Hf isotopic compositions found in southern Laurentian strata.

Acknowledgements
We are extremely grateful to Dr. E. Szymanski and an anonymous reviewer for helpful and detailed comments and to Dr. F. Cheng for editorial handling.

Author contributions JES: conceptualization (lead), data curation (equal), formal analysis (lead), methodology (equal), software (equal), writing – original draft (lead); KES: conceptualization (supporting), data curation (equal), formal analysis (equal), methodology (equal), software (equal), writing – original draft (supporting)

Funding
This work was funded by the University of British Columbia, the Natural Sciences and Engineering Research Council of Canada (Discovery Grant Program), and National Science Foundation (EAR-1824557).

Data availability
All data generated or analysed during this study are included in this published article (and its supplementary information files).

Scientific editing by Feng Cheng

Correction notice: Equations 7 and 8 were incorrect and have been updated.

References
Ahmed, N., Natarajan, T. and Rao, K.R. 1974. Discrete cosine transform. IEEE Transactions on Computers, 100, 90–93, https://doi.org/10.1109/T-C.1974.223764
Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716–723, https://doi.org/10.1109/TAC.1974.1100705
Andersen, T. 2014. The detrital zircon record: supercontinents, parallel evolution —Or coincidence? Precambrian Research, 244, 279–287, https://doi.org/10.1016/j.precamres.2013.10.013
Andersen, T., Griffin, W.L. and Pearson, N.J. 2002. Crustal evolution in the SW Part of the Baltic Shield: the Hf isotope evidence. Journal of Petrology, 43, 1725–1747, https://doi.org/10.1093/petrology/43.9.1725
Andersen, T., Griffin, W.L. and Sylvester, A.G. 2007. Sveconorwegian crustal underplating in southwestern Fennoscandia: LAM-ICPMS U–Pb and Lu–Hf isotope evidence from granites and gneisses in Telemark, southern Norway. Lithos, 93, 273–287, https://doi.org/10.1016/j.lithos.2006.03.068
Andersen, T., Saeed, A., Gabrielsen, R.H. and Olausen, S. 2011. Provenance characteristics of the Brumunddal sandstone in the Oslo Rift derived from U–Pb, Lu-Hf and trace element analyses of detrital zircons by laser ablation ICPMS. Norwegian Journal of Geology/Norsk Geologisk Forening, 91
Andresen, A., Agerić-Dwarko, N.V., Kristoffersen, M. and Hanken, N.-M. 2014. A Timanian foreland basin setting for the late Neoproterozoic–Early Palaeozoic cover sequences (Dividal Group) of northeastern Balska. Geological Society, London, Special Publications, 390, 157–175, https://doi.org/10.1144/SP390.29
Anfinson, O.A., Leier, A.L., Gaschnig, R., Embry, A.F. and Dewing, K. 2012. U–Pb and Hf isotopic data from Franklinian Basin strata: insights into the nature of Crockerland and the timing of accretion, Canadian Arctic Islands. Canadian Journal of Earth Sciences, 49, 1316–1328, https://doi.org/10.1139/e2012-067
Appleby, S.K., Gillespie, M.R., Graham, C.M., Hinton, R.W., Oliver, G.J. and Kelly, N.M. 2010. Do S-type granites commonly sample infracrustal sources?

Applied Geochemistry, 2001, 16, 525–530, https://doi.org/10.1016/S0305-1978(00)00056-5
Appleby, S.K., Gillespie, M.R., Graham, C.M., Hinton, R.W., Oliver, G.J. and Kelly, N.M. 2010. Do S-type granites commonly sample infracrustal sources?
