Noname manuscript No.
(will be inserted by the editor)

A Bayesian nonparametric approach to unmixing detrital geochronologic data

John Tipton · Glenn Sharman · Samuel Johnstone

Received: date / Accepted: date

Abstract Sedimentary deposits constitute the primary record of changing environmental conditions that have acted on Earth's surface over geologic time. Clastic material is eroded from source locations (parents) in sediment routing systems and deposited at sink locations (children). Both parents and children have characteristics that vary across many different dimensions, including grain size, chemical composition, and the geochronologic age of constituent detrital minerals. During transport, sediment from different parents is mixed together to form a child, which in turn may serve as the parent for other sediment further down system or later in time when buried sediment is exhumed. The distribution of detrital mineral ages observed in parent and child sediments allows for investigation of the proportion of each parent in the child sediment which reflects the properties of the sediment routing system. To model the proportion of dates in a child sample that comes from each of the parent distributions, we use a Bayesian mixture of Dirichlet processes. This model allows for estimation of the mixing proportions with associated uncertainty while making minimal assumptions. We also present an extension to the model whereby we reconstruct unobserved parent distributions from multiple observed child distributions using mixtures of Dirichlet processes. The model accounts for uncertainty in both the number of mineral formation events that compose each parent distribution and the mixing proportions of each parent distribution that composes a child distribution. To demonstrate the model, we perform analyses using simulated data where the true age distribution is

John R. Tipton
Department of Mathematical Sciences, University of Arkanasas, Fayetteville, AR, USA
E-mail: jrtipton@uark.edu

Glenn R. Sharman
Department of Geosciences, University of Arkanasas, Fayetteville, AR, USA

Samuel A. Johnstone
U.S. Geological Survey, Geoscience and Environmental Change Science Center, Denver, USA

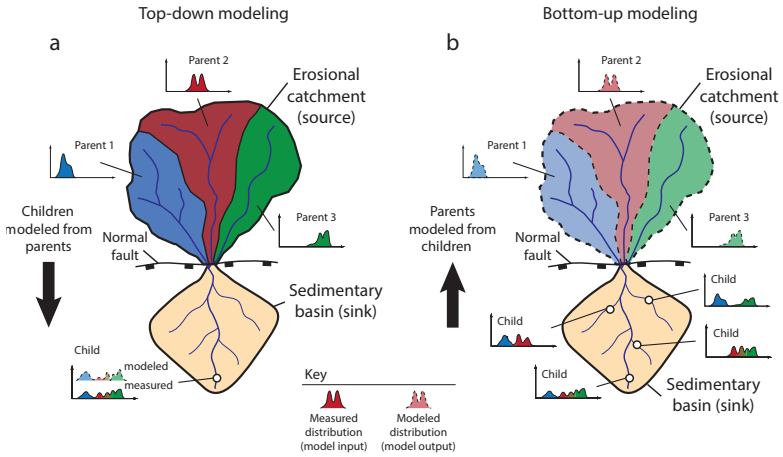


Fig. 1: Schematic depiction of a hypothetical sediment routing system with an erosional source region characterized by three parents (black, red, and green) and an associated sedimentary basin (yellow). The histograms represent the age distributions of detrital minerals from the respective parents and children with lines mapping these distributions to their respective locations in the landscape. Subfigure (a) shows the top-down modeling framework (sensu Sharman and Johnstone (2017)) where one or more children are modeled as a mixture of two or more parents. Subfigure (b) shows the bottom-up modeling framework where multiple children are used to reconstruct unobserved end-member sources (parents).

known as well as using a real world case study from the central California, USA coast.

1 Introduction

To understand the origins of modern and ancient landscapes one must understand how erosional processes and associated sedimentary basins evolve through time (Romans et al., 2016). As clastic material is generated by weathering and erosion, it is subsequently transported downstream, mixed, and ultimately deposited into a depositional sink. Statistical modeling of sediment mixing allows for inference about processes that generated the modern landscape. The ability to decipher the relative proportions of sources that eroded to produce sediment informs understanding of the underlying geologic processes controlling the evolution of the Earth's surface (Stock et al., 2006; Sharman et al., 2019; Mason et al., 2017; Kimbrough et al., 2015).

One of the most common ways to characterize the provenance of sediment is detrital geochronology – dating the time at which the individual miner-

als that make up a sedimentary rock formed or cooled. These mineralization events typically reflect the timing of igneous rock forming events or metamorphic alteration of source rocks (Gehrels, 2014). In other cases mineral ages reflect the history of mineral cooling (e.g., 'thermochronology', (Reiners and Brandon, 2006)). Detrital geochronologic ages are most commonly determined from measurements of radiogenic isotopes contained within individual mineral crystals. The decay of uranium (U) to lead (Pb) within zircon, a relatively robust mineral, makes this approach ideally suited for tracking sedimentary mixing (Amidon et al., 2005b; Sundell and Saylor, 2017; Sharman and Johnstone, 2017).

We will follow the convention that sediment sources are called *parents* and sink locations are called *children*. Because the statistical model is not defined mechanistically, the definition of the parent source is flexible enough to accommodate different scales. For example, a river could be considered the parent source sediment to a marine basin (as in the natural case study presented later in the manuscript). If rivers are considered the children, then the parent sources may be the upstream tributaries or distinctive bedrock domains within the upstream catchment. Using this language, the manuscript aims to address two questions. First, can we estimate the proportion of each parent age distribution in a child age distribution with associated uncertainty? Second, can we estimate the age distributions for unobserved parents given a set of child age distributions? These questions are answered using "top-down" and "bottom-up" approaches to sediment unmixing, respectively (see Fig. 1; (Sharman and Johnstone, 2017)). The top-down approach (Fig. 1a) models one or more child samples as a mixture of specified parent samples. The bottom-up approach (Fig. 1b) uses multiple child samples to model likely parents which are more generally referred to as end-members in mixture modeling efforts (Weltje, 1997; Paterson and Heslop, 2015).

Bayesian mixture modeling of geologic data, including detrital data, has the advantage of explicitly quantifying uncertainty (Ward et al., 2010; Cooper and Krueger, 2017; Blake et al., 2018). For geochronologic data, Bayesian methods have been used to estimate the age distribution when addressing single samples with probabilistic estimates (Jasra et al., 2006). We extend the work of Jasra et al. (2006) to multiple sample locations by modeling the geologic mixing of sediments derived from source areas containing minerals recording different crystallization events. A key feature of these data are that the age distributions are mutlimodal due to a range of mineral formation events for a given mineral. Thus, an implicit assumption is that each mode in the age distribution is the result of a distinct mineral formation event.

We demonstrate the utility of the Bayesian approach using both a synthetic dataset and a well-constrained, natural case study in central California, USA (Sickmann et al., 2016). In the simulated data, the model is shown empirically to recover the simulated age distributions. After better understanding model performance with simulated data, the model is applied to the case study dataset from central California. The top-down mixing approach is able to successfully reconstruct parent contributions in both the synthetic and

natural datasets. Although the bottom-up unmixing model is able to successfully reconstruct parents in the synthetic dataset, there is evidence of non-identifiability – where parents cannot be uniquely characterized from the children – when applied to the natural dataset. Although the focus of this work is on sediment age distributions, the framework presented here can also give guidance about other scientific questions that relate to mixing of non-parametric sum-to-one data in Earth sciences and other disciplines (e.g., unmixing sediment grain size distribution; Weltje (1997) and references within)

2 Model Overview

To define the statistical model, we follow the convention that letters represent data and Greek symbols represent parameters. A plaintext symbol (y/θ) represents a scalar, a bold lowercase symbol ($\mathbf{y}/\boldsymbol{\theta}$) represents a vector, and a bold uppercase symbol ($\mathbf{Z}/\boldsymbol{\Theta}$) is a matrix whose columns are vectors written as $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_p)$ or $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)'$ where the appropriate dimensions are implied. We use the notation $[y]$ to represent the probability distribution/mass function (pdf/pmf) and let $[y|\theta]$ represent the conditional pdf/pmf of the random variable y given θ.

Following Berliner (2003), the statistical model described below is divided into three components: the data model, the process model, and the prior model. In general, the data model defines probability distributions that describe the variability in the data due to the measurement process. The data model can be modified to account for non Gaussian measurement processes like counts, outliers, spatial/temporal correlations, etc (Tipton et al., 2017; Hefley et al., 2017). Process models describe the best scientific understanding of the process of interest. For example, process models have been used to describe the monthly response of trees to climate (Tipton et al., 2016), the relationship between climate and pollen in sediments (Tipton et al., 2019), and the movement of ice sheets in Antarctica (Chang et al., 2016; Guan et al., 2018). For the sediment mixing model, the two processes we attempt to capture with the statistical model are 1) the mineral formation events and 2) the erosional/weathering, transport,f and filtering of grain ages from parent sources to children sinks. The prior model describes the range of parameter values that are plausible and completes the formal mathematical definition of the model by guaranteeing that that posterior distribution is proper (i.e., integrates to 1).

3 Top-down mixing model

The model framework presented below, which is appropriate for situations where the parent and children sediment have been independently characterized, will answer the first research question: can one estimate the proportion of each parent that comprises the child sediment with appropriate estimates of uncertainty?

3.1 Top-down mixing data model

Let $\mathbf{y}$ be a n_y-vector of observed age measurements of a single child of interest and let $\mathbf{z}_b$ be a n_b-vector of observed date measurements for each of the $b = 1, \ldots, B$ parents. Because the observed ages include measurement uncertainty reported as a standard deviation, we explicitly account for this source of uncertainty in the data model. In the case of U-Pb dating of detrital zircon grains, dates are most commonly determined using laser ablation-inductively coupled plasma-mass spectrometry (Gehrels, 2012). Such date measurements typically have relative 2σ analytical precision of 1-4%, with relative uncertainty increasing for younger analyses (Puetz et al., 2018). For each detrital mineral, the analytical uncertainty (in standard deviations) for the child is reported as the n_y-vector $\boldsymbol{\sigma}_y$ and for each of the $b = 1, \ldots, B$ parents as a n_b-vector $\boldsymbol{\sigma}_{z_b}$. We assume the date measurement uncertainty follows a Gaussian distribution where the observed sediment grain date is

$$\begin{aligned} \mathbf{y}|\tilde{\mathbf{y}}, \boldsymbol{\sigma}_y^2 &\sim \mathrm{N}(\tilde{\mathbf{y}}, \mathrm{diag}(\boldsymbol{\sigma}_y^2)), \\ \mathbf{z}_b|\tilde{\mathbf{z}}_b, \boldsymbol{\sigma}_{z_b}^2 &\sim \mathrm{N}(\tilde{\mathbf{z}}_b, \mathrm{diag}(\boldsymbol{\sigma}_{z_b}^2)), \end{aligned} \tag{1}$$

where $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The notation $\mathrm{diag}(\boldsymbol{\sigma^2})$ represents a diagonal covariance matrix with i, ith element σ_i^2 and off diagonal elements all equal to 0. We break the variable naming convention and let $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{z}}_b$ be latent parameters that represent the true, unobserved age of the sediments where $\boldsymbol{y}$ ($\boldsymbol{z}_b$) will be close to $\tilde{\mathbf{y}}$ ($\tilde{\mathbf{z}}_b$) because the measurement uncertainty is small relative to the variability in the data (i.e., the average coefficient of variation of measured dates, defined as the dating standard deviation divided by the estimated date, is about 2-3%). The reason we account for age dating uncertainty is twofold. First, in a perfect world we could measure the mineral dates exactly; however, in practice our measurements introduce some uncertainty such that our data are approximations of a true, unknown age. Second, because of this, the model can give more weight to dates with less uncertainty and vice versa. In this way, it is possible within the model framework to combine data from different minerals that have different dating uncertainties in a principled manner. In addition, it is possible to account for more uncertainty in the data or to account for asymmetric measurement uncertainties using a Student's-t, log-normal, or other distribution instead of the normal distribution (Jasra et al., 2006). However, these extensions were not explored in the current work to minimize the number of parameters to estimate in this proof of concept model framework.

3.2 Top-down mixing process model

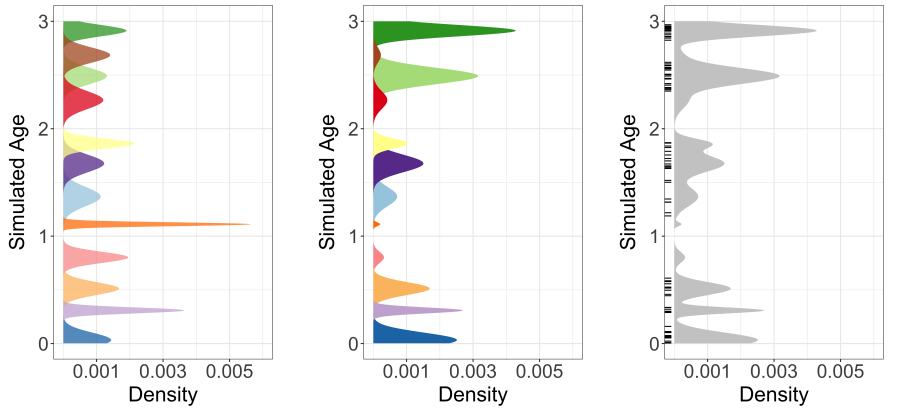
The process model addresses two scientific questions. First, what are the estimates of the true, unobserved detrital mineral age distributions at the parent

and child locations? Second, what proportions of those detrital minerals did each parent source contribute to the child? There are many different methods available to model geochronological age distributions from sample data, including kernel density estimation (Vermeesch, 2012), non-negative matrix factorization (Saylor et al., 2019), and Bayesian nonparametric models of mineral formation event mixing (Jasra et al., 2006; Tye et al., 2019). The following section details how our understanding of the geologic processes that generated the observed data inform the development of the statistical model.

Over geologic time, individual minerals may be repeatedly recycled into sedimentary rocks by erosion, transport, deposition, and exhumation. However, in many cases the dates recorded by individual minerals contained in these deposits are distinctive and unaffected by these recycling processes (e.g., excluding burial reheating of low temperature thermochronometers (Fosdick et al., 2015)). We assume that minerals created by the same geologic event share an age distribution that is relatively homogeneous with only small variability. Furthermore, we characterize episodes of rock and mineral formation as punctuated events (typically lasting 10^5 to 10^7 years (Chen and Moore, 1982; Irwin and Wooden, 1999; Wotzlaw et al., 2013)), which are nearly discrete events relative to the age of Earth (approximately 4.5×10^9 years). While minerals often show overgrowths of different ages, this provides a useful approximation. Under the conceptual model (Fig. 1), sediment at each parent is formed by the decomposition of rocks containing minerals created at different times where the potential mineral creation events are shared across parents. Although all parents share the same potential mineral formation events, this does not imply that each parent will actually contain minerals from these potential events. The parent sediments are then mixed producing sediment that has the potential to be present at every child location. The model assumes the system is closed so that the sediment at each child comes entirely from the parent sources. Therefore, each child sediment is composed of minerals from the parent sources that are created by an unknown number of mineral formation events at source locations.

A common choice for modeling a mixture of unknown distributions is the Dirichlet process. Figure 2 demonstrates visually an example simulation from the Dirichlet process model. Mathematically, a Dirichlet Process is constructed over an infinite-dimensional mixture of base measures $G(\boldsymbol{\theta})$ for a family of probability measures $G(\cdot)$ and parameters $\boldsymbol{\theta}$. However, despite the Dirichlet Process being infinite-dimensional, the expected number of clusters scales logarithmically with respect to sample size (Ghosal, 2010) which induces sparsity and makes it easy to numerically approximate the infinite mixture with a finite number of clusters K much, much less that the sample size n. Thus, even though the true number of mineral formation events K recorded by a detrital sample is unknown, the *a priori* expected number of clusters can be determined using the sample size with a fixed K chosen much larger than the expected number of clusters to ensure a good finite approximation.

Figure 2a shows an example simulation from the Dirichlet process prior using a finite approximation with a mixture of K Gaussian distributions. Each of



(a) Distribution of simulated mineral formation events. Each color represents a different formation event. Notice that some formation events have wider standard deviations (i.e., resulting from longer durations of mineral formation), while other formation events are shorter.

(b) The mineral formation events from Fig. 2a are re-weighted to account for relative abundance of potentially observable mineral ages for the parent.

(c) Discrete, colored formation events in the parent distribution in Fig. 2b are replaced by gray because the formation events are unknown. The observed data are shown as a rug plot along the y-axis.

Fig. 2: A cartoon of the mixing model over hypothetical mineral formation events for a single parent distribution. The y-axis of each plot is the age of formation and the x-axis is the probability density of the hypothetical parent distribution.

the shaded colors is a Gaussian distribution where the each distinct color shade depicts a discrete mineral formation event. The means and variances differ for each of the distributions reflecting that formation events can arise from various geologic processes (Fig. 2a). The transition from Fig. 2a to Fig. 2b represents the statistical process model which accounts for geological characteristics and processes including aerial extent, differential erosion, the abundance of minerals of different ages within different rocks, and other factors that influence the proportion of minerals of a given age in sediment at a site (Amidon et al., 2005a,b). Figure 2b shows the age distributions in Fig. 2a that have been re-weighted to account for all of the geological factors determining the mixing proportions of the possible formation events for the parent.

Figure 2c depicts a realization from the process model for a given parent. The labels ($k = 1, \ldots, K$) for each mineral formation event are not observed. Thus, the simulated observation distribution only records the age of the detrital minerals in sediment and the color shading is dropped representing this lack of knowledge about the labels. The data are not observations of the mixture density in Fig. 2c but are actually a finite sample taken from from the

mixture which is shown as a rug plot where each tick on the y axis represents an observed detrital mineral grain date. Thus, the number of mineral formation events is potentially challenging to extract from the data as neither the true age dating density nor the labels that identify the underlying formation events are known.

3.2.1 Modeling parent sediment ages

Let the $i = 1, \ldots, n_b$ latent detrital sediment grain ages from parent b be represented by $\tilde{z}_{ib}$. The sediment grain from which we estimate the latent age $\tilde{z}_{ib}$ from the observed age z_{ib} is assumed to come from a single mineral formation event implying the mixture distribution over mineral formation events

$$\tilde{z}_{ib}|\boldsymbol{\mu}_b, \boldsymbol{\sigma}_b^2, \gamma_{ib} \sim \begin{cases} \mathrm{N}\left(\tilde{z}_{ib}\middle|\mu_{b1}, \sigma_{b1}^2\right) & \text{if } \gamma_{ib} = 1 \\ \vdots & \vdots \\ \mathrm{N}\left(\tilde{z}_{ib}\middle|\mu_{bK}, \sigma_{bK}^2\right) & \text{if } \gamma_{ib} = K, \end{cases} \tag{2}$$

where γ_{ib} is a categorical random variable whose value indicates which formation event k the detrital mineral comes from. We assume the probability of a detrital mineral coming from formation event k is $p_{bk} \equiv P(\gamma_{ib} = k)$. Then, we write the joint distribution over all mineral grains from parent b as

$$\tilde{\mathbf{z}}_b|\boldsymbol{\mu}_b, \boldsymbol{\sigma}_b^2, \boldsymbol{\gamma}_b \sim \prod_{i=1}^{n_b} \mathrm{N}\left(\tilde{z}_{ib}\middle|\mu_{b1}, \sigma_{b1}^2\right)^{I\{\gamma_{ib}=1\}} \mathrm{N}\left(\tilde{z}_{ib}\middle|\mu_{b2}, \sigma_{b2}^2\right)^{I\{\gamma_{ib}=2\}} \cdots \mathrm{N}\left(\tilde{z}_{ib}\middle|\mu_{bK}, \sigma_{bK}^2\right)^{I\{\gamma_{ib}=K\}} \tag{3}$$

where $I\{\gamma_{ib} = k\}$ is an indicator function that takes the value 1 if $\gamma_{ib} = k$ and 0 otherwise. Because there are a large number of indicator functions, we integrate them out of the process model to improve mixing and model fit. The integrated age distribution model is

$$\tilde{\mathbf{z}}_b|\boldsymbol{\mu}_b, \boldsymbol{\sigma}_b^2, \boldsymbol{p}_b \sim \prod_{i=1}^{n_b} \sum_{k=1}^{K} p_{bk} \, \mathrm{N}\left(\tilde{z}_{ib}\middle|\mu_{bk}, \sigma_{bk}^2\right) \tag{4}$$

where $\mathbf{p}_b = (p_{b1}, \ldots, p_{bK})'$ is a vector of positive mixing probabilities with $\sum_{k=1}^{K} p_{bk} = 1$. To account for uncertainty in the number of formation events K, the probabilities $\mathbf{p}_b$ can be modeled using the Dirichlet process as described in detail in Sect. 3.3. In the model as written, the assumption is that there could be a different set of mineral formation events for each parent. However, as these different events could be jointed to form a superset of all events common across locations, we will assume that all formation events are potentially shared among parents to simplify the fitting algorithm.

3.2.2 Modeling children sediment ages

The process model specifies the proportion of each parent distribution in the child distribution. We represent the mixing proportions of the B parent distributions for the child of interest with the parameter $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_B)'$, with $\sum_{b=1}^{B} \phi_b = 1$ and $\phi_b > 0$. The parameter ϕ_b is the proportion of the child distribution that comes from parent b and accounts for differential mixing of parents. As the statistical model is not mechanistic, the specific, geologic interpretation of $\boldsymbol{\phi}$ changes based on the context of the sediment transport system. For example, when the parents that are comprised of bedrock, $\boldsymbol{\phi}$ is a function of each parent's relative aerial extent in the drainage catchment, average erosion rate, and average concentration of the detrital mineral of interest (Amidon et al., 2005a). If parents are sediment inputs (e.g., rivers), then $\boldsymbol{\phi}$ is a function of each parent's relative sediment supply and the average concentration of the detrital mineral of interest within the sediment.

For a single latent child sediment mineral date $\tilde{y}_i$, the sediment grain for that mineral comes from only one parent. Using a categorical variable, the distribution of the child sediment grain can be written as a weighted mixture of components (similar to (2)) where each component is a weighted sum of the parent distributions (i.e., the sum is over the K weighted densities in (4) but the mixture density is evaluated with child observations rather than the parent observations). Then, the latent indicator variables can be integrated out of the mixture, similar to (4), giving the integrated child age distribution model

$$\tilde{\mathbf{y}}_i | \{\boldsymbol{\mu}_b, \boldsymbol{\sigma}_b^2, \boldsymbol{p}_b\}_{b=1}^{B}, \boldsymbol{\phi} \sim \prod_{i=1}^{n_y} \sum_{b=1}^{B} \phi_b \sum_{k=1}^{K} p_{bk} \, \mathrm{N}\left(\tilde{y}_i \middle| \mu_{bk}, \sigma_{bk}^2\right),$$

where the notation $\{\boldsymbol{\theta}_b\}_{b=1}^{B}$ denotes the set of parameters $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_B\}$.

3.3 Top-down mixing prior model

The conceptual process model assumes the number of mineral formation events K is known. In practice, the number of formation events is unknown and is a parameter to be estimated. In fact, it is likely that the different parent sites will have different numbers of mineral formation events based on site-specific history. The prior model addresses the fundamental question of estimating the number of mineral formation events.

There are a variety of potential approaches to model the unknown number of formation events. First, one can treat the number of formation events as a fixed parameter, perform a grid search over the different number of formation events, and choose the model that best fits the data based on some information theoretic criteria (Miller and Harrison, 2018). A second approach is to sample over the latent unknown number of formation events using a reversible jump

algorithm (Green, 1995). The third approach is to assign a Dirichlet process prior over the number of formation events. The Dirichlet process estimates an unknown number of components without *a priori* specifying the number.

The Dirichlet process is an infinite dimensional stochastic process which is a distribution over distributions (Ferguson, 1973). We assign the range of mineral ages for the kth formation event the base probability distribution $G(\boldsymbol{\theta}_{bk})$ which depends on parameters $\boldsymbol{\theta}_{bk}$. There are many possible choices for the base distribution $G(\boldsymbol{\theta}_{bk})$; we assume a normal distribution $N(\mu_{bk}, \sigma^2_{bk})$ with mean μ_{bk} and variance σ^2_{bk}, therefore $\boldsymbol{\theta}_{bk} = (\mu_{bk}, \sigma^2_{bk})'$. Other possible choices include a log-normal or gamma distribution that enforces a positive support on the observed age dates or a skew-t distribution that allows for asymmetry in the duration of formation events (Jasra et al., 2006). While other distributions are possible, and may better capture the effects of natural dispersion in geochronometers, we rely on normal distributions here to minimize the number of model parameters and emphasize inclusion of the multiple parent and children sediments withing the bottom-up unmixing framework. Because we assume the age distribution of a single mineral formation event is relatively short with respect to the overall time of interest, the variance parameters that model the duration of the mineral formation events σ^2_{bk} will be small relative to the scale of the observed age distribution. Note that the variance σ^2_{bk} represents the process variance due to mineral formation events and is different than the measurement process variances, $\boldsymbol{\sigma}^2_y$ and $\{\boldsymbol{\sigma}^2_{z_b}\}_{b=1}^B$, which are fixed and known.

We use the stick-breaking representation of a Dirichlet process

$$\sum_{k=1}^{\infty} p_{bk} G(\boldsymbol{\theta}_{bk}), \tag{5}$$

where p_{bk} are the positive mixing weights with $\sum_{k=1}^{\infty} p_{bk} = 1$. In practice, $p_{bk} \approx 0$ for large k, therefore, the infinite sum is well approximated by the finite sum $\sum_{k=1}^{K} p_{bk}$ for a large enough K (for most problems K=10 or K=20 is sufficiently large). The stick-breaking representation for $\mathbf{p}_b$ is constructed by transforming auxiliary variables $\tilde{\mathbf{p}}_b = (\tilde{p}_{b1}, \ldots, \tilde{p}_{bK-1})'$ using the stick-breaking representation

$$p_{bk} = \begin{cases} \tilde{p}_{b1} & \text{for } k = 1, \\ \tilde{p}_{bk} \prod_{k'=1}^{k-1} (1 - \tilde{p}_{bk'}) & \text{for } k = 2, \ldots, K-1, \\ \prod_{k'=1}^{k-1} (1 - p_{bk'}) & \text{for } k = K. \end{cases}$$

Priors on the $\tilde{p}_{bk}$ are assigned exchangeable Beta$(1, \alpha_b)$ priors giving rise to the stick-breaking Dirichlet process. The hyperparameters α_b are given exchangeable Gamma$(1, 1)$ priors that control the Dirichlet process concentration (i.e., smaller α_b give fewer formation events, larger α_b give more formation events). Because our study site is constrained geographically, the parent and child sites contain mineral grains derived from common formation events. As

such, we follow Lock and Dunson (2015) and used shared kernels by letting $\boldsymbol{\theta}_{bk} \equiv \boldsymbol{\theta}_k = (\mu_k, \sigma_k^2)'$ for all $b = 1, \ldots, B$. In theory, there could be no overlap at all among the formation events, although in this situation it would be possible to choose a K that is large enough such that the number of shared kernels is larger than the total number of formation events among the parents and would result in equivalent inference.

The mixing kernel means μ_k are assigned vague, independent $N(\mu_\mu, \sigma_\mu^2)$ priors with $\mu_\mu = 150$ Million years (Myr) and $\sigma_\mu^2 = 150^2$ Myr^2. The standard deviations for the ages of formation are assigned truncated half-Cauchy priors $\sigma_k \sim \text{Cauchy}^+(0, s)I\{0 < \sigma_k < \omega\}$, where we choose s to be small relative to the range of dates observed and ω provides an upper limit to the duration of formation events. For the case study where the majority of dates span the range of 0 to about 300 Myr, we set s to be 25 Myr and set ω to be 50 Myr. The truncation is important to prevent the Dirichlet process mixture from generating unrealistically long formation events which does not match our *a priori* geologic knowledge.

The mixing parameter $\boldsymbol{\phi}$ is assigned a Dirichlet($\alpha_\phi \mathbf{1}$) prior where $\mathbf{1}$ is a vector of ones and the hyperparameter α_ϕ is assigned a Gamma$(1, 1)$ prior. When α_ϕ is small the mixing proportions concentrate with a large probability on a single parent component. When α_ϕ is one $\boldsymbol{\phi}$ will be uniformly distributed over all possible mixing proportions, When α_ϕ is large the mixing proportion will be concentrated at equal mixing proportions $\left(\frac{1}{B}, \ldots, \frac{1}{B}\right)$.

3.4 Top-down mixing posterior distribution

The top-down mixing model posterior distribution is

$$\begin{aligned}\left[\tilde{\mathbf{y}}, \{\tilde{\mathbf{z}}_b\}_{b=1}^B, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \{\boldsymbol{p}_b\}_{b=1}^B, \boldsymbol{\phi}, \alpha_\phi, \boldsymbol{\alpha} \middle| \mathbf{y}, \boldsymbol{\sigma}_y, \{\mathbf{z}_b, \boldsymbol{\sigma}_b^2\}_{b=1}^B\right] \propto \\ [\mathbf{y}|\tilde{\mathbf{y}}, \boldsymbol{\sigma}_y] \prod_{b=1}^B [\mathbf{z}_b|\tilde{\mathbf{z}}_b, \boldsymbol{\sigma}_b] \times \\ \left[\tilde{\mathbf{y}} \middle| \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \{\boldsymbol{p}_b\}_{b=1}^B, \boldsymbol{\phi}\right] \prod_{b=1}^B \left[\tilde{\mathbf{z}}_b \middle| \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{p}_b\right] \times \\ [\boldsymbol{\mu}] \left[\boldsymbol{\sigma}^2\right] [\boldsymbol{\phi}|\alpha_\phi] [\alpha_\phi] \left(\prod_{b=1}^B [\boldsymbol{p}_b|\alpha_b] [\alpha_b] \right),\end{aligned} \tag{6}$$

where each line on the right hand side of the proportional symbol is the data, process, and prior model, respectively. We estimate the posterior using Markov Chain Monte Carlo (MCMC) with the *R* package *NIMBLE* (de Valpine et al., 2017) using an adaptive block Metropolis-Hastings algorithm (Haario et al., 2001). The constrained auxiliary variables $\tilde{\mathbf{p}}_b$ and standard deviations $\boldsymbol{\sigma}^2$ are transformed to unconstrained support (logit- and log-scale

transformations) for tuning the Metropolis-Hastings block proposals, with corresponding Jacobian adjustments to the acceptance probabilities. The sampling of the sum-to-one mixing proportion $\boldsymbol{\phi}$ is performed by introducing auxiliary variables $\tilde{\boldsymbol{\phi}}$, assigning a stick-breaking prior on $\tilde{\boldsymbol{\phi}}$, then sampling on a logit-scale after correcting for the transformation using the Jacobian to induce a Dirichlet($\alpha_\phi \mathbf{1}$) prior on $\boldsymbol{\phi}$.

4 Bottom-up unmixing model

The second research question is: can we reconstruct unobserved parent age distributions from multiple child observations? In previous work, this analysis has been variably termed "end-member mixing analysis", "end-member modeling", or "end-member analysis" as applied to unmixing grain size or detrital age distributions (Sharman and Johnstone, 2017; Saylor et al., 2019). The end-member unmixing model has two components. First, the number of parents B is unknown and needs to be estimated. Second, given the number of parents B, what are the unobserved mineral formation age distributions for the B parents? For this paper, we assume the number of parents B is known. There are a number of criteria for selecting the number of parents including using Bayesian information criteria, reversible jump MCMC (Jasra et al., 2006), assuming a Dirichlet process over the number of parents, or fitting a mixture of finite mixtures (Miller and Harrison, 2018). Rather than explore these ideas, we devote our effort on developing the unmixing model for a fixed number of parents. The end-member model uses the same general framework presented in the mixture of Gaussians model (6) with some modifications.

4.1 Bottom-up unmxing data model

Let $d = 1, \ldots, D$ index the D child sediments that are each composed of $i = 1, \ldots, n_d$ samples. As before, we assume a Gaussian measurement distribution for child d given by

$$\mathbf{y}_d | \tilde{\mathbf{y}}_d, \boldsymbol{\sigma}_d^2 \sim \mathrm{N}(\tilde{\mathbf{y}}_d, \mathrm{diag}(\boldsymbol{\sigma}_d^2)),$$

where $\tilde{\mathbf{y}}_d$ is the true, unobserved n_d-vector of sediment dates and $\boldsymbol{\sigma}_d$ is a fixed and known n_d-vector of reported dating standard deviations.

Unlike in the top-down mixing model above, none of the parent $\mathbf{z}$s are observed. Hence, the parent distributions are estimated entirely using child sediment observations. Assuming a fixed and known number of parents B, the bottom-up process model for a mineral grain age arising from the dth child is

$$\tilde{y}_{id} | \boldsymbol{\mu}, \boldsymbol{\sigma}^{\mathbf{2}}, \{\boldsymbol{p}_b\}_{b=1}^B, \boldsymbol{\phi}_d \sim \sum_{b=1}^{B} \phi_{db} \sum_{k=1}^{K} p_{bk} \, \mathrm{N}(\mu_k, \sigma_k^2), \tag{7}$$

where, like before, we assume a Gaussian mixing distribution using shared kernels across the B parents. The B-dimensional mixture proportion $\boldsymbol{\phi}_d = (\phi_{d1}, \ldots, \phi_{dB})'$ models the proportion of the dth child sediment that can be attributed to each of the B parents. Like the top-down mixing model, these equations can be derived by introducing categorical random variables then marginalizing out the latent indicator variables from the model. The bth unknown parent age distribution is given by $\sum_{k=1}^{K} p_{bk}\, \mathrm{N}(\mu_k, \sigma_k^2)$. The prior model for the bottom-up unmixing model is the same as for the top-down mixing model, except for the variables have different dimensions.

4.2 Bottom-up unmixing posterior distribution

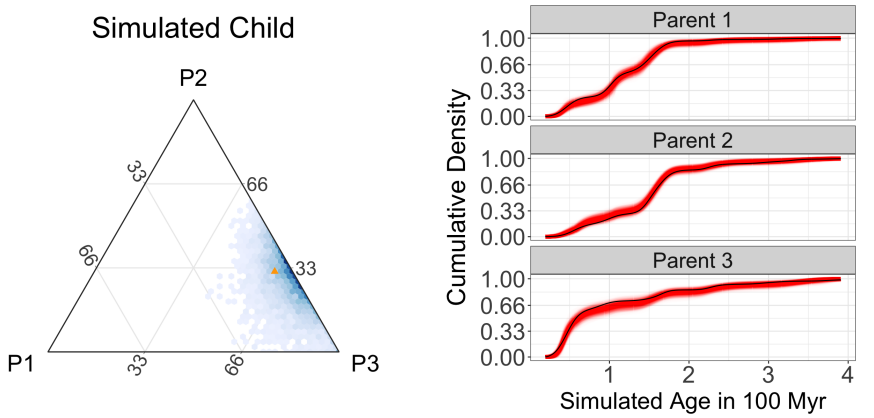
The posterior distribution that we estimate with the end member unmixing model is

$$\begin{aligned}\left[\{\tilde{\mathbf{y}}_d\}_{d=1}^{D}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \{\boldsymbol{p}_b\}_{b=1}^{B}, \{\boldsymbol{\phi}_d\}_{d=1}^{D}, \boldsymbol{\alpha}_\phi, \boldsymbol{\alpha} \middle| \{\mathbf{y}_d, \boldsymbol{\sigma}_d^2\}_{d=1}^{D}\right] \propto \\ \prod_{d=1}^{D} \left[\mathbf{y}_d | \tilde{\mathbf{y}}_d, \boldsymbol{\sigma}_d\right] \times \\ \prod_{d=1}^{D} \left[\tilde{\mathbf{y}}_d \middle| \boldsymbol{\phi}_d, \{\boldsymbol{p}_b\}_{b=1}^{B}, \boldsymbol{\mu}, \boldsymbol{\sigma}\right] \times \\ \left[\boldsymbol{\mu}\right] \left[\boldsymbol{\sigma}^2\right] \left(\prod_{b=1}^{B} \left[\boldsymbol{p}_b | \alpha_b\right] \left[\alpha_b\right]\right) \left(\prod_{d=1}^{D} \left[\boldsymbol{\phi}_d | \alpha_{\phi d}\right] \left[\alpha_{\phi d}\right]\right),\end{aligned} \tag{8}$$

where the priors are the same as those in (6) except that there are now D children which implies there are now D $\alpha_{\phi d}$s. Likewise, the MCMC algorithm is the same as presented in Sect. 3 except for a change in dimension for some parameters.

5 Simulation of synthetic detrital age distributions

We explore the performance of the model using a synthetic detrital age distribution dataset. The aim of the simulation study is to understand how the model performs using realistic data and verify the model is capable of recovering the simulated parameters. The simulation study framework can also be used to understand how uncertainty in estimation varies with respect to sample size, variability in the data, and consequences of prior assumptions, although these details are not explored in this work (Vehtari et al., 2017). For example, by simulating data with smaller sample sizes than observed, the impact on the uncertainty estimates can be explored and perhaps used to guide sample size recommendations for the collection of new data. Similarly, simulations with different observation errors can be used to quantify the amount of analytic



(a) Ternary plot showing posterior mixing proportion estimates shaded relative to posterior denisty in blue and the simulated true mixing proportions as an orange triangle.

(b) Plot of simulated parents with fitted posterior CDF estimates in red and the simulated true CDF in black. Each red line represents a posterior sample of the cumulative parent age density.

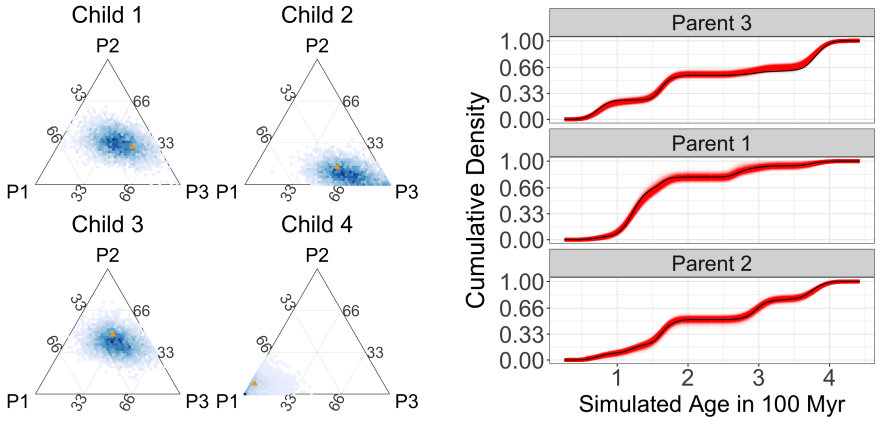
Fig. 3: Results for the top-down mixture modeling approach based on simulated data. As can be seen in the figures, the top-down mixture model is performing well in estimating the simulated mixing distributions.

precision needed for an inferential question analogous to how sample size and power calculations can be used in experimental design.

For the simulation study, a synthetic dataset is created using the top-down mixing model in (6) for $B = 3$ parents and a single child (Fig. 1a). The parent distributions were composed of 200, 250, and 150 simulated sediment grains, respectively, and the child distribution was composed of 150 sediment grains. In simulation, we used dating uncertainties $(\boldsymbol{\sigma}_y, \boldsymbol{\sigma}_z)$ that were about 1-3% of the total range of the simulated age distribution of 0-400 Myr. These are similar to measurement uncertainties in the case study and demonstrate the model is capable of accounting for measurement uncertainty.

The posterior samples for the mixing proportion $\boldsymbol{\phi}$ are shown in Fig. 3a in hexagonal bins with blue shading proportional to the posterior density (Hamilton, 2018), and the simulated mixing proportion is represented by the orange triangle. The simulated mixing proportion (orange triangle) lies within the region of high posterior density demonstrating that the model is accurately estimating the mixing proportions. Figure 3b shows the estimated cumulative distribution functions (CDFs) with posterior samples in red and the simulated CDF in black. The results in Fig. 3 demonstrate that the model is accurately estimating the simulated mixing proportions $\boldsymbol{\phi}$ as well as the parent age distributions, validating the effectiveness of the top-down mixing model to recover simulated parameters of interest.

The second simulation generated data from the bottom-up, end-member unmixing model (Fig. 1b) to test how well the proposed framework can re-



(a) Posterior estimates of mixing proportions for 4 of the 20 children from the unmixing model shown. The blue shading is relative to posterior density and the simulated true mixing proportions are shown as orange triangles.

(b) Posterior estimates of the unobserved parent CDFs in red. The simulated parent CDF is shown in black.

Fig. 4: Results for the bottom-up, end-member unmixing model using simulated data. The bottom-up unmixing model does a good job of estimating the true, unobserved parent age CDFs despite the model not using any simulated parent data.

construct unobserved parent distributions from a set of child observations. For the simulation, we used $B = 3$ parents and $D = 20$ children where each child consisted of 250 measured sediment grains following the model in (8). The range of simulated mixing proportions was simulated uniformly over the three-dimensional simplex resulting in some end members being close to pure end members (e.g., close to 90% of sediment grains coming from a single parent). The dating uncertainties ($\boldsymbol{\sigma}_y$) were set at about 1-3% of the total range of the simulated age distribution of 0-400 Myr.

Figure 4a shows the posterior samples for the mixing proportion of each child in hexagonal bins with shading in blue proportional to posterior density and the orange triangle shows the simulated mixing proportion. Based on the simulation study, the model can recover the mixing proportions in this simulation example with high accuracy as the orange triangle is within regions of high posterior mass. Even though the model uses none of the data from the parents, the bottom-up unmixing model produces reasonable end-member parent age distribution estimates. Figure 4b shows the estimated CDF produced by the bottom-up unmixing model which shows the model is estimating the unobserved parent distributions. However, the precision for the bottom-up unmixing model is lower in the bottom-up unmixing simulation relative to the the top-down mixture simulation despite the bottom-up umixing simulated data containing many more sediment grain samples.

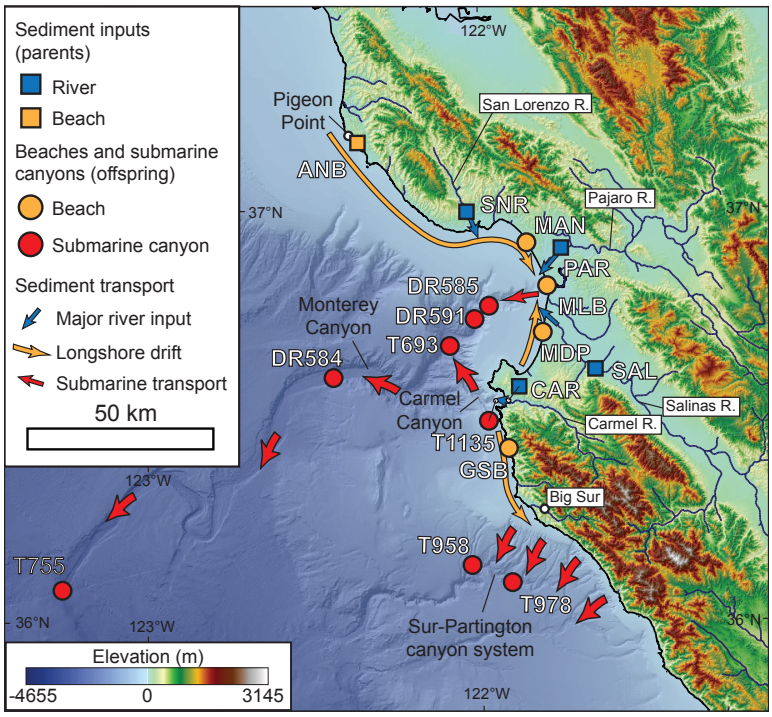


Fig. 5: Locations of the parents and children data for the study region in California, USA (Sharman and Johnstone, 2017).

6 Application to a Natural Case Study

We apply the top-down mixing and bottom-up unmixing models to a well-constrained modern dataset from the central California coast (Sickmann et al., 2016) shown in Fig. 5. Following the same mixing framework presented in Sharman and Johnstone (2017), there are five samples (river and beach sediment) used to characterize three distinct sediment inputs (parents) to the region, each with a distinct detrital age distribution. Parents 1 and 2 (P1 and P2) are comprised of river samples (CAR and SAL) that represent sediment sources along the Big Sur coastline and Salinas River drainage, respectively. Parent 3 (P3) is comprised of two river samples (SNR and PAR) and one beach sample (ANB) that represent northern sediment sources in the Santa Cruz Mountains and western Diablo Range (Sickmann et al., 2016; Sharman and Johnstone, 2017). Twelve child samples (beach and submarine canyon sediment) are used to characterize how these parents are mixed in littoral and marine environments. In total, this dataset (Fig. 6) consists of 4,026 individual detrital zircon U-Pb analyses, with individual samples having 82 to 316 analyses each (median of 290 analyses per sample) (Sickmann et al., 2016). The age range of the case study data covered approximately 80-3000 Myr; however, there were very

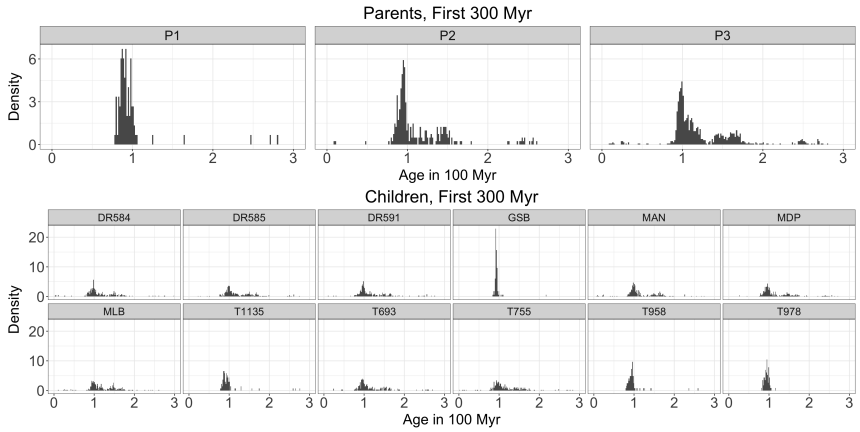
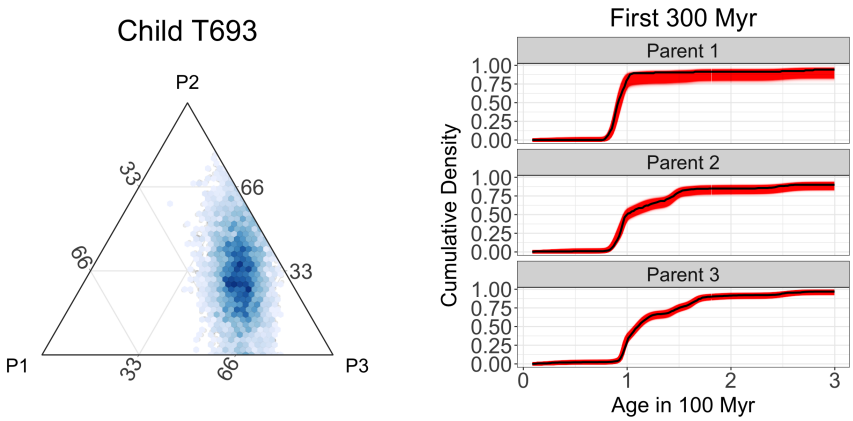


Fig. 6: The sediment age data for the study region in California, USA used for the mixing and unmixing models (Sickmann et al., 2016). The three parent age distributions are shown in the top plot and the 12 child age distributions are shown in the bottom plot. The x-axes represent the measured age in 100 Myr and the y-axes shows the empirical density.

few observation older than 300 Myr (approximately 6% of the data) so even though the model was fit to the entire age range, the presentation of results focuses on the period 0-300 Myr to better interpret the results.

We first examine the top-down mixture model (Fig. 1a). Figure 7a shows the reconstruction of the mixing proportions for a sample from a submarine canyon (T693) modeled as a mixture of the three specified parent distributions (P1-P3). Visual inspection of the histograms of the data (Fig. 6) would suggest that this child sample is a mixture composed mostly of P3 with some contribution from P2. The posterior estimates of the mixing proportion of each parent for child T693 confirms that the primary component of the mixture is from the P3 with parent P2 as the secondary component (Fig. 7a). Figure 7b shows the model is capturing the basic patterns in the parent CDFs as the estimated CDFs are very close to the empirical CDFs for the parents.

The bottom-up, end-member unmixing model results for the case study data are shown in Fig. 8. The posterior density for the mixing proportions are shown for four children in Fig. 8a as hexagonal bins with shading in blue proportional to posterior density. The posterior distribution of CDFs for the reconstructed parents in Fig. 8b show that the model is very uncertain about the age distribution for parent P2. This is not a totally unsurprising result as the age distributions for parent P2 and parent P3 are very similar and thus are difficult to tease apart in an unsupervised learning situation like that in the bottom-up unmixing model. In Fig. 8a the mixture probability for child T693 is more concentrated at P3 when compared to the top-down mixing proportion estimates in Fig. 7a. Because the model expresses uncertainty about the distribution of parent P2, the difference in estimation between the top-



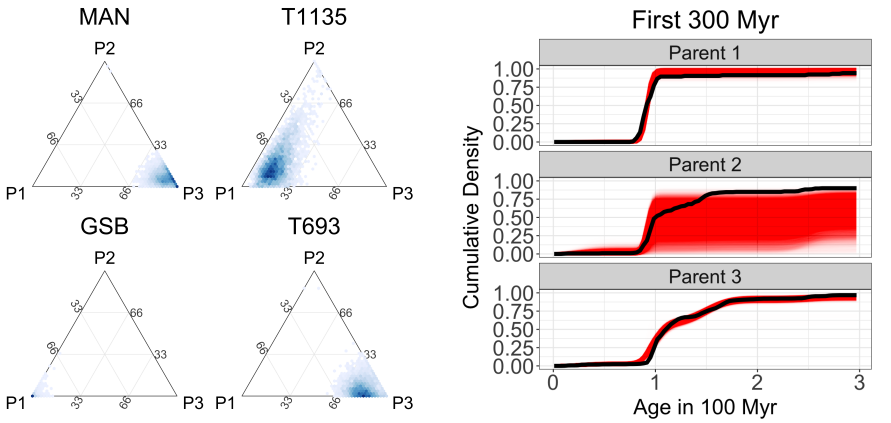
(a) Ternary plot showing posterior density estimates of mixing proportions. The relative posterior density is shown in blue shading.

(b) Posterior estimates of the parent and child CDFs shown in red. The empirical CDFs calculated from the raw data are shown in black.

Fig. 7: Results from the top-down mixing model using data from the study region in California, USA. The model results, when applied to sample T693, demonstrates that the top-down mixing model is able to accurately reconstruct the parent and child distributions and produce estimates of the mixing proportions with associated uncertainty.

down and bottom-up mixing proportions for child T693 can be attributed to this lack of identifiability among parents P2 and P3. Similar to how the important variables in a linear regression can change with small perturbations of the data when the covariates are highly co-linear (i.e., in models with a high variance inflation factor), the attribution of the child sediment to a given parent would change which parent distribution (P2 or P3) is estimated with accuracy and which parent shows evidence of non-identifiability. Thus, the non-identifiability manifests as large posterior uncertainty in the CDF for parent P2 and provides a caution about being overly confident in the inference about the bottom-up unmixing model.

A large overlap in the distribution of parent ages is a feature that often occurs in detrital zircon geochronology studies. The preservation of zircons through multiple cycles of erosion and re-sedimentation means that overlapping zircon ages will be present in many rocks. For parent age distributions that are quite similar to one another, the reconstruction of the unknown parent distributions suffers from weak identifiability. In these situations, the estimated parents jointly contain all of the formation events, but the model is unable to attribute the formation events to the correct parents. In other words, while the model identifies the correct age components, the model sometimes struggles to correctly group these components into the correct parent distributions. This aliasing effect is not an unexpected result because Bayesian



(a) Posterior estimates for the mixing proportions of each parent for four child sediments. Notice that without observing the parents, the posterior distribution of mixing proportions for child T693 is generally similar to the top-down mixing model in Fig. 7a but has a slightly different shape.

(b) Posterior estimates for the unobserved parent cumulative distribution functions shown in red over 0-300 Myr. The black lines show the empirical cumulative distribution functions.

Fig. 8: Results of the end-member unmixing model fit to data from the study region in California, USA. These figures show that the end-member unmixing model is estimating the parameters of interest, but with some inaccuracies due to a lack of identifiability. However, these issues are easily identified by the end user due to the large amount of posterior uncertainty which provides a check on overly strong inferential claims.

nonparametric models are well understood to suffer from non-identifiability issues (Ferguson, 1983; Diebolt and Robert, 1994; Richardson and Green, 1997; Frühwirth-Schnatter, 2006).

Non-identifiability is inherent in all end-member unmixing models (Weltje and Prins, 2007). To overcome the non-identifiability in other modeling frameworks, a potential solution is to impose constraints on the end-members and/or provide informative initial conditions for maximum likelihood optimization algorithms (Donoho and Stodden, 2004; Miao and Qi, 2007; Chen and Guillaume, 2012). Therefore, any end-member unmixing model that uses only child age distributions will have issues in accurately reconstructing the parent distributions if the assumption of the constraints is not met (i.e., the parent age distributions are structurally similar). Bottom-up unmixing models provide a useful way to explore large detrital datasets with unknown sedimentary sources. Our proposed model framework provides a way to identify those datasets that either are or are not susceptible to non-identifiability by producing uncertainty estimates that are larger when the model is weakly identifiable (Fig. 8b, parent 2). Thus, the uncertainty intervals are a useful diagnostic check for identifiability.

Direct, probabilistic estimates of uncertainty and the ability to calculate derived quantities with uncertainty is a benefit of the proposed method and of Bayesian methods in general. Thus, we can answer questions like what is the probability that at least 50% of child sediment T693 comes from parent P3 using the top-down mixing model applied to the case study data? The answer is calculated directly from the posterior samples using the Monte Carlo approximation $\frac{1}{L}\sum_{\ell=1}^{L} I\{\phi_3^{(\ell)} \geq 0.5\} = 0.672$, where $\ell = 1, \ldots, L$ are the indices of the MCMC samples and $\phi_{\text{P3}}^{(\ell)}$ is the estimated mixing proportion for the ℓth MCMC iteration. The probability that at least 50% of the child sediment comes from parent P3 and at least 25% comes from parent P2 is $\frac{1}{L}\sum_{\ell=1}^{L} I\{\phi_{\text{P3}}^{(\ell)} \geq 0.5\} \times I\{\phi_{\text{P2}}^{(\ell)} \geq 0.25\} = 0.283$. Another question of interest that can be expressed in term of the model is "How many mineral formation events occurred for each parent?" We assumed that there was an upper bound of K possible mineral formation events for each parent distribution modeled by the K-dimensional probability vector $\mathbf{p}_b$ for each of the B parents. For a given parent b and a fixed threshold τ where we conclude that a potential formation event is realized (say $\tau = 0.01$ or $\tau = 0.05$), the estimated number of formation events is $\frac{1}{L}\sum_{\ell=1}^{L}\sum_{k=1}^{K} I\{p_{bk}^{(\ell)} \geq \tau\}$. For example, the posterior mean estimate for the number of formation events for parent P1 under the top-down mixing model is 4.37 (95% CI 4-6) when the threshold $\tau = 0.05$ and 8.34 (95% CI 7-10) when $\tau = 0.01$. Because the model produces a posterior probability, any other such probabilistic questions like those above can be calculated as derived quantities. For example, we can ask questions like: what proportion of a given sample contains grains older than a given age? or what is the probability that an unobserved parent contains grains within a particular age range. Once the posterior samples have been calculated, any questions that can be evaluated using derived quantities can be answered probabilistically.

In addition, the ability to include prior information in the Bayesian framework is a useful tool that can be used to improve estimation and test geologic hypotheses. For example, certain geologic events, such as the Grenville orogeny, produced large amounts of zircon that have since been broadly dispersed and recycled in sedimentary rocks (Moecher and Samson, 2006). Priors that account for the likelihood of observing zircons of Grenville-age (or other known zircon-producing events) can be introduced into this model framework to improve performance. In addition, our framework can accommodate a variety of detrital data with different magnitudes of uncertainty. As analytic techniques for dating minerals improve, it is important to account for dating uncertainties that might have orders of magnitude difference, making our method more robust to future improvements in analytic laboratory techniques.

7 Conclusion

Starting from a conceptual model of how sediments mix over a landscape, we developed a generative Bayesian nonparametric statistical model for detri-

tal mineral age data. This model allows us to characterize the uncertainty in the age distributions of parents and children and the mixing proportions for sediments while explicitly accounting for the uncertainties in measured dates (Jasra et al., 2006; Tye et al., 2019). Because the model can generate sediment age distributions, we can directly explore the assumptions of the model by simulating synthetic data. Running a simulation experiment demonstrated the model is capable of recovering simulated distributions which supports the usefulness of the framework when applied to observed data.

We proposed two frameworks to model the sediment mixing mechanisms: the top-down mixing model where mineral dates are measured for both parent and child sediments and a bottom-up unmixing framework where mineral dates are only measured for the children. The top-down model estimated the parent and child distributions and the mixing proportions with high precision and accuracy. The bottom-up model occasionally showed evidence of non-identifiability in the simulation experiments and did show non-identifiability in the case study, suggesting the inference for the bottom-up model is less precise than for the top-down mixing model. Because the variances of these estimates are larger in our bottom-up unmixing model, the user is provided with feedback about the potential pitfalls in being overly confident about the reconstructed parent distributions.

Obtaining correct inference is vitally important for any statistical model. However, many models make identifying when inference is suspect challenging. The explicit modeling of uncertainty presented in this manuscript provides a check on overly confident inference. As such, the inference in the model presented herein provides a useful diagnostic on the quality of model fit. In the case study, the posterior distribution for one of the parent distributions was estimated with a large amount of uncertainty. Therefore, the inference based on the model fit is less reliable than the model that includes the parent data. Thus, explicit modeling of uncertainty is critical in providing information about the quality of the inference.

While this manuscript focuses on estimating the mixture of geochronological measurements from sediments, the methods discussed can be applied to mixtures of any univariate variable of interest. For example, the top-down mixing and bottom-up unmixing models can be applied to mixtures of sediment grain size (Weltje and Prins, 2007). In addition to applying the model frameworks to other variables, the extension of the mixing and unmixing models to multivariate data would allow for the inclusion of more nuanced and detailed data. In the cases where the distributions are only weakly identified based on one variable, the mixing distributions might be identifiable using other variables. Thus, the results presented in this manuscript can provide a roadmap for future development and extension to better characterize geologic landscapes.

8 Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Support for SAJ came from the FEDMAP component of the US Geological Survey National Cooperative Geologic Mapping Program. This draft manuscript is distributed solely for purposes of scientific peer review. Its content is deliberative and predecisional, so it must not be disclosed or released by reviewers. Because the manuscript has not yet been approved for publication by the U.S. Geological Survey (USGS), it does not represent any official USGS finding or policy. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Code and data for replication of results presented in this manuscript can be found freely available under the permissive MIT license on GitHub at `https://github.com/jtipton25/mixing-manuscript`.

Keywords Detrital sediment age distributions · Sediment unmixing · Bayesian nonparametrics · Uncertainty quantification

References

Amidon WH, Burbank DW, Gehrels GE (2005a) Construction of detrital mineral populations: insights from mixing of U-Pb zircon ages in Himalayan rivers. Basin Research 17(4):463–485

Amidon WH, Burbank DW, Gehrels GE (2005b) U–Pb zircon ages as a sediment mixing tracer in the Nepal Himalaya. Earth and Planetary Science Letters 235(1-2):244–260

Berliner LM (2003) Physical-statistical modeling in geophysics. Journal of Geophysical Research: Atmospheres 108(D24)

Blake WH, Boeckx P, Stock BC, Smith HG, Bodé S, Upadhayay HR, Gaspar L, Goddard R, Lennard AT, Lizaga I, et al. (2018) A deconvolutional Bayesian mixing model approach for river basin sediment source apportionment. Scientific reports 8(1):1–12

Chang W, Haran M, Applegate P, Pollard D (2016) Calibrating an ice sheet model using high-dimensional binary spatial data. Journal of the American Statistical Association 111(513):57–72

Chen JH, Moore JG (1982) Uranium-lead isotopic ages from the Sierra Nevada batholith, California. Journal of Geophysical Research: Solid Earth 87(B6):4761–4784

Chen W, Guillaume M (2012) HALS-based NMF with flexible constraints for hyperspectral unmixing. EURASIP Journal on Advances in Signal Processing 2012(1):54

Cooper RJ, Krueger T (2017) An extended Bayesian sediment fingerprinting mixing model for the full bayes treatment of geochemical uncertainties. Hydrological Processes 31(10):1900–1912

de Valpine P, Turek D, Paciorek C, Anderson-Bergman C, Temple Lang D, Bodik R (2017) Programming with models: writing statistical algorithms for general model structures with NIMBLE. Journal of Computational and Graphical Statistics 26:403–413, DOI 10.1080/10618600.2016.1172487

Diebolt J, Robert CP (1994) Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society Series B (Methodological) pp 363–375

Donoho D, Stodden V (2004) When does non-negative matrix factorization give a correct decomposition into parts? In: Advances in Neural Information Processing Systems, pp 1141–1148

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. The Annals of Statistics pp 209–230

Ferguson TS (1983) Bayesian density estimation by mixtures of normal distributions. In: Recent Advances in Statistics, Elsevier, pp 287–302

Fosdick JC, Grove M, Graham SA, Hourigan JK, Lovera O, Romans BW (2015) Detrital thermochronologic record of burial heating and sediment recycling in the Magallanes foreland basin, Patagonian Andes. Basin Research 27(4):546–572

Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Springer Science & Business Media

Gehrels G (2012) Detrital zircon U-Pb geochronology: Current methods and new opportunities. Tectonics of Sedimentary Basins: Recent Advances pp 45–62

Gehrels G (2014) Detrital zircon U-Pb geochronology applied to tectonics. Annual Review of Earth and Planetary Sciences 42:127–149

Ghosal S (2010) The Dirichlet process, related priors and posterior asymptotics. Bayesian nonparametrics 28:35

Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82(4):711–732

Guan Y, Haran M, Pollard D (2018) Inferring ice thickness from a glacier dynamics model and multiple surface data sets. Environmetrics 29(5-6):e2460

Haario H, Saksman E, Tamminen J (2001) An adaptive Metropolis algorithm. Bernoulli 7(2):223–242

Hamilton N (2018) ggtern: An Extension to ggplot2, for the creation of ternary diagrams. URL `https://CRAN.R-project.org/package=ggtern`, R package version 2.2.2

Hefley TJ, Brost BM, Hooten MB (2017) Bias correction of bounded location errors in presence-only data. Methods in Ecology and Evolution 8(11):1566–1573

Irwin WP, Wooden JL (1999) Plutons and accretionary episodes of the Klamath Mountains, California and Oregon. Tech. rep., US Geological Survey

Jasra A, Stephens DA, Gallagher K, Holmes CC (2006) Bayesian mixture modelling in geochronology via Markov chain Monte Carlo. Mathematical Geology 38(3):269–300

Kimbrough DL, Grove M, Gehrels GE, Dorsey RJ, Howard KA, Lovera O, Aslan A, House PK, Pearthree PA (2015) Detrital zircon U-Pb provenance

713 of the Colorado River: A 5 my record of incision into cover strata overlying
714 the Colorado Plateau and adjacent regions. Geosphere 11(6):1719–1748
715 Lock EF, Dunson DB (2015) Shared kernel Bayesian screening. Biometrika
716 102(4):829–842
717 Mason CC, Fildani A, Gerber T, Blum MD, Clark JD, Dykstra M (2017)
718 Climatic and anthropogenic influences on sediment mixing in the Missis-
719 sippi source-to-sink system using detrital zircons: Late Pleistocene to recent.
720 Earth and Planetary Science Letters 466:70–79
721 Miao L, Qi H (2007) Endmember extraction from highly mixed data using min-
722 imum volume constrained nonnegative matrix factorization. IEEE Transac-
723 tions on Geoscience and Remote Sensing 45(3):765–777
724 Miller JW, Harrison MT (2018) Mixture models with a prior on the number of
725 components. Journal of the American Statistical Association 113(521):340–
726 356
727 Moecher DP, Samson SD (2006) Differential zircon fertility of source terranes
728 and natural bias in the detrital zircon record: Implications for sedimentary
729 provenance analysis. Earth and Planetary Science Letters 247(3-4):252–266
730 Paterson GA, Heslop D (2015) New methods for unmixing sediment grain size
731 data. Geochemistry, Geophysics, Geosystems 16(12):4494–4506
732 Puetz SJ, Ganade CE, Zimmermann U, Borchardt G (2018) Statistical anal-
733 yses of global U-Pb database 2017. Geoscience Frontiers 9(1):121–145
734 Reiners PW, Brandon MT (2006) Using thermochronology to understand oro-
735 genic erosion. Annual Review Earth Planetary Sciences 34:419–466
736 Richardson S, Green PJ (1997) On Bayesian analysis of mixtures with an
737 unknown number of components (with discussion). Journal of the Royal
738 Statistical Society: series B (Statistical Methodology) 59(4):731–792
739 Romans BW, Castelltort S, Covault JA, Fildani A, Walsh J (2016) Environ-
740 mental signal propagation in sedimentary systems across timescales. Earth-
741 Science Reviews 153:7–29
742 Saylor JE, Sundell K, Sharman G (2019) Characterizing sediment sources by
743 non-negative matrix factorization of detrital geochronological data. Earth
744 and Planetary Science Letters 512:46–58
745 Sharman GR, Johnstone SA (2017) Sediment unmixing using detrital
746 geochronology. Earth and Planetary Science Letters 477:183–194
747 Sharman GR, Sylvester Z, Covault JA (2019) Conversion of tectonic and cli-
748 matic forcings into records of sediment supply and provenance. Scientific
749 reports 9(1):4115
750 Sickmann ZT, Paull CK, Graham SA (2016) Detrital-zircon mixing and parti-
751 tioning in fluvial to deep marine systems, Central California, USA. Journal
752 of Sedimentary Research 86(11):1298–1307
753 Stock GM, Ehlers TA, Farley KA (2006) Where does sediment come from?
754 Quantifying catchment erosion with detrital apatite (U-Th)/He ther-
755 mochronometry. Geology 34(9):725–728
756 Sundell K, Saylor JE (2017) Unmixing detrital geochronology age distribu-
757 tions. Geochemistry, Geophysics, Geosystems

Tipton J, Hooten M, Goring S (2017) Reconstruction of spatio-temporal temperature from sparse historical records using robust probabilistic principal component regression. Advances in Statistical Climatology, Meteorology and Oceanography 3(1):1–16

Tipton JR, Hooten MB, Pederson N, Tingley MP, Bishop D (2016) Reconstruction of late Holocene climate based on tree growth and mechanistic hierarchical models. Environmetrics 27(1):42–54

Tipton JR, Hooten MB, Nolan C, Booth RK, McLachlan J (2019) Predicting paleoclimate from compositional data using multivariate Gaussian process inverse prediction. Annals of Applied Statistics 13(4):2363–2388, DOI 10.1214/19-AOAS1281

Tye A, Wolf A, Niemi N (2019) Bayesian population correlation: A probabilistic approach to inferring and comparing population distributions for detrital zircon ages. Chemical Geology 518:67–78

Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing 27:1413–1432, DOI 10.1007/s11222-016-9696-4

Vermeesch P (2012) On the visualisation of detrital age distributions. Chemical Geology 312:190–194

Ward EJ, Semmens BX, Schindler DE (2010) Including source uncertainty and prior information in the analysis of stable isotope mixing models. Environmental Science & Technology 44(12):4645–4650

Weltje GJ (1997) End-member modeling of compositional data: Numerical-statistical algorithms for solving the explicit mixing problem. Mathematical Geology 29(4):503–549

Weltje GJ, Prins MA (2007) Genetically meaningful decomposition of grain-size distributions. Sedimentary Geology 202(3):409–424

Wotzlaw JF, Schaltegger U, Frick DA, Dungan MA, Gerdes A, Günther D (2013) Tracking the evolution of large-volume silicic magma reservoirs from assembly to supereruption. Geology 41(8):867–870