Dissimilarity measures in detrital geochronology

Pieter Vermeesch
Department of Earth Sciences
University College London
Gower Street, London WC1E 6BT
p.vermeesch@ucl.ac.uk

November 22, 2017

Abstract

The ability to quantify the (dis)similarity between detrital age distributions is an essential aspect of sedimentary provenance studies. This paper reviews three different ways to do this. A first class of dissimilarity measures is based on parametric hypothesis tests such as the t- or chi-square test. These are designed to objectively decide whether two samples were derived from a common population. Appealing though such tests may appear in theory, in practice they offer little value to sedimentary geologists because their outcome depends on sample size. In contrast, the effect size of said tests is independent of sample size and can be used as an objective point of comparison between detrital age distributions. The main limitation of this approach is that it requires binning or averaging, which discards valuable information. A second class of dissimilarity measures is based on non-parametric hypothesis tests such as the Kolmogorov-Smirnov test. These do not require pre-treatment of the data and are able to capture more subtle differences between age distributions. Unfortunately, non-parametric tests do not have well defined sample effect sizes and so it is not possible to use them as an absolute point of comparison that is independent of sample size. Nevertheless, non-parametric dissimilarity measures can be used to quantify the relative differences between samples. A third class of dissimilarity measures aims to account for the analytical uncertainties of the age determinations. The likeness and cross-correlation coefficients are ad-hoc dissimilarity measures that are based on Probability Density Plots (PDPs). These apply a narrow smoothing kernel to precise data, and a wide smoothing kernel to imprecise data. In contrast, the Sircombe-Hazelton L2-norm uses Kernel Functional Estimates (KFEs), which use exactly the opposite strategy as PDPs. They apply a wide smoothing kernel to precise data, and a narrow smoothing kernel to imprecise data. This paper shows that the KFE-based approach produces sensible results, whereas the likeness and cross-correlation methods do not. The added complexity of the KFE approach is only worth the effort in studies that combine data acquired on equipment with hugely variable analytical precision. In most cases, there is no need to account for the analytical uncertainty of detrital age distributions. The sample effect size, non-parametric statistics, or L2-norm can be used to graphically compare samples by Multidimensional Scaling (MDS). In contrast with previous claims, there is no need for these measures to be independent of sample size.

keywords: statistics, geochronology, sediment

1 Introduction

Siliciclastic sediments and sedimentary rocks cover an estimated 66% of the land surface (Blatt and Jones, 1975). Sedimentary deposits contain valuable archives of Earth history and host economically important mineral resources. Constraining the provenance of siliclastic sediments is key to understanding these geological environments. Sedimentary provenance may be recovered using a variety of chemical, mineralogical or isotopic properties. Detrital geochronology is one approach that has steadily gained popularity over the

years. In this approach, a representative number of clasts are dated by radiometric geochronology. Examples of this are zircon U-Pb dating in sand (Pell et al., 1997; Garzanti et al., 2013) or silt (Nie et al., 2015), $^{40}Ar/^{39}Ar$ dating of detrital mica (Copeland and Harrison, 1990), fission track dating of apatite (Vermeesch, 2007) and zircon (Hurford et al., 1984), U-Th-He dating of apatite (Stock et al., 2006) and zircon (Rahl et al., 2003), and cosmogenic ^{21}Ne measurements in quartz clasts (Codilean et al., 2008). The sampling distributions are then either compared with the geologic map to identify individual point sources of sediment (Pell et al., 1997) or, more commonly, they can be compared with each other in order to trace the flow of sediment through an entire sediment routing system (Stevens et al., 2013; Vermeesch and Garzanti, 2015).

In provenance studies that involve a small number of easily distinguishable age distributions, reliable interpretation may be possible by simple visual inspection. But this is seldom true for the general case involving many samples that are only subtlely different. In this general case, it is useful to have some numerical value to objectively express the (dis)similarity of two samples. Three kinds of approaches have been used to do this. A first group of dissimilarity measures is based on parametric statistical hypothesis tests. Section 2 presents a brief primer to these procedures, using the chi-square test as an example (Section 2.1). This Section will introduce basic concepts such as effect size (Section 2.2) and p-value (Section 2.3). It will show that the former parameter is the only way to quantify the difference between two distributions in absolute terms, independent of sample size. In contrast, p-values exhibit a strong dependence on sample size that limits their usefulness for detrital geochronology.

Although Section 2 touches on some fundamental concepts that are relevant to detrital geochronology, it is also quite technical. Readers who are more interested in the practical aspects of the subject may wish to skip to Section 3, which discusses a second group of dissimilarity measures that are based on non-parametric hypothesis tests such as Kolmogorov-Smirnov and variants thereof such as the Cramér-von-Mises, Anderson-Darling and Kuiper tests. This approach has intuitive appeal, as it is better able to capture the richness of real age distributions without the need for any pre-treatment of the data. Finally, Earth Scientists have invented a number of ad-hoc dissimilarity measures that aim to capture and undo the effect of variable measurement uncertainty (aka 'heteroscedasticity'). Examples of this are Satkoski et al. (2013)'s *likeness* parameter, Saylor et al. (2012)'s *cross-correlation* coefficient, and Sircombe and Hazelton (2004)'s *L2-distance*. We will see that only the latter method is built on statistically sound foundations, whereas the other two methods yield problematic results in some simple but realistic end-member scenarios.

Section 5 discusses p-value tables, in which multiple samples are compared with each other using statistical hypothesis tests such as chi-square or Kolmogorov-Smirnov. We will see that such multi-sample comparisons are problematic due to the aforementioned sample size dependency of p-values, which is amplified by the increased occurrence of so-called 'Type-I' errors (as defined in Section 2.1). Multidimensional Scaling (MDS) is presented as an alternative approach to multi-sample comparison that is immune to these problems. In contrast with earlier claims by Vermeesch (2013), we will show that MDS does not require dissimilarity measures to be independent of sample size (Section 5).

The different dissimilarity measures discussed in this paper will be illustrated with nine synthetic populations:

- A consists of five uniformly distributed intervals of 10 Ma length each that are evenly spaced between 50 and 140 Ma and are separated by 10 Ma gaps.
- B is identical to A but has been offset by 20 Ma. In other words, distribution B misses the first uniform interval of distribution A and has an additional interval between 150 and 160 Ma.
- C is offset by a further 60 Ma with respect to distribution B.
- D is the convolution of distribution A with a Normal distribution with zero mean and 2 Ma standard deviation. This is the sampling distribution that would result from an infinite number of grains collected

from distribution A and dated with 2 Ma analytical precision.

$E - F$ are the sampling distributions of populations B and C, assuming 2 Ma analytical uncertainties just like in distribution D.

$G - I$ are the sampling distributions of populations $A - C$, assuming a 2 % *relative* age uncertainty.

The curves shown in Figure 1 are Probability Density Functions (PDFs). They represent the relative probability of any age or date in the population so that:

$$Prob(t_1 \leq t \leq t_2) = \int_{t_1}^{t_2} PDF(t)dt \tag{1}$$

The piecewise uniform populations $A - C$ represent the true age distributions, which are unknown and unobservable. The actual data that detrital geochronologists work with are not ages but *dates* that are affected by some degree of analytical uncertainty. As a result of this uncertainty, the PDF of the dates is always smoother than the PDF of the true ages (e.g., D and G vs. A; E and H vs. B; and F and I vs. C in Figure 1). Please note that a PDF is *not* the same as a PDP (Probability Density Plot), as will be explained in Section 4.1. This crucial distinction marks the key difference between the present paper and a recent review by Saylor and Sundell (2016), whose recommendations are in direct conflict with those presented herein.

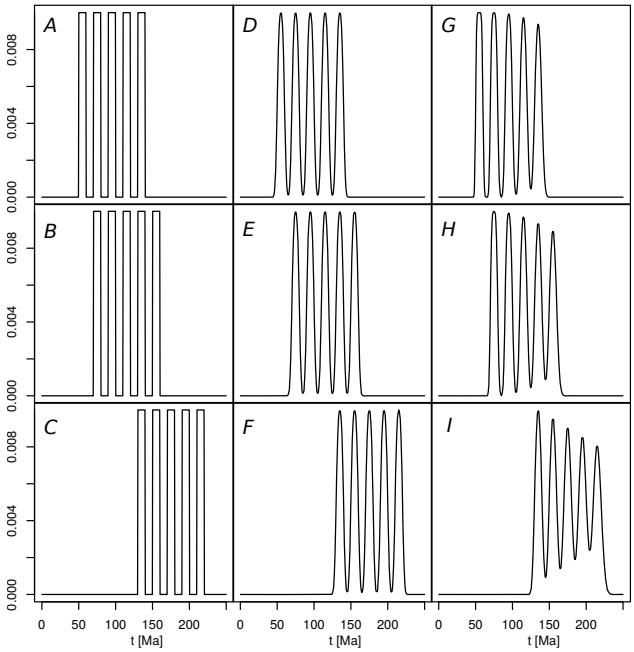

Figure 1: $A - C$: Probability Density Functions (PDFs) of three synthetic age distributions; $D - F$: PDFs of the dates obtained from $A - C$ assuming 2 Ma absolute measurement uncertainties; $G - I$: PDFs of the dates obtained from $A - C$ assuming 2 % relative measurement uncertainties.

2 chi-square

Consider the following four sets ('samples') of values ('dates'):

a = {115, 115, 88, 71, 89, 76, 110, 74, 58, 135, 90, 98, 114, 121, 90, 50, 91, 53, 136, 95}

b = {115, 55, 59, 114, 132, 98, 70, 74, 133, 57, 94, 100, 60, 135, 139, 119, 89, 51, 113}

c = {98, 133, 92, 110, 77, 77, 131, 97, 96, 149, 135, 74, 120, 130, 137, 130, 77, 78, 151, 77, 120}

d = {196, 221, 214, 129, 218, 136, 176, 150, 135, 171, 177, 158, 137, 138, 172, 195, 133, 193, 196, 177}

One way to compare these four samples is to bin them into two categories:

	a	b	c	d
$\leq$ 125 Ma	18	15	13	0
$>$ 125 Ma	2	4	8	20

To compare the bin counts in pairs, define the chi-square statistic as follows:

$$X^2_{stat} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

where O_i is the number of observed counts and E_i the number of expected counts in the i^{th} bin, and k represents the number of bins (e.g., 4 in the case of a 2-sample comparison of 2 bins each). The E_is are calculated from the marginal probabilities of the data table. For example, the expected bin counts for the comparison of samples a and b:

	a	b
$\leq$ 125 Ma	$\frac{(18+15)(18+2)}{18+15+2+4} = 16.9$	$\frac{(18+15)(15+4)}{18+15+2+4} = 16.1$
$>$ 125 Ma	$\frac{(2+4)(18+2)}{18+15+2+4} = 3.1$	$\frac{(2+4)(15+4)}{18+15+2+4} = 2.9$

Applying Equation 2 to all sample pairs yields a symmetric matrix of chi-square statistics (Table 1.i). This matrix indicates that samples a and b are the most similar ($\chi^2_{stat} = 0.26$) while samples a and d are the most dissimilar ($\chi^2_{stat} = 29.2$). This makes sense when we know that samples a and b were both derived from distribution D, whereas samples c and d were derived from E and F, respectively. So it appears that the chi-square statistic has adequately captured the differences between the three underlying populations. However, increasing the sample sizes from ~20 to ~200 changes all the chi-squared values (Table 1.ii). Most of them have increased, except for the dissimilarity between a and b, which has slightly decreased in value. It is not immediately clear what the actual X^2_{stat}-value means, or how to interpet the X^2_{stat}-values in a mixture of small and large samples, as is common in detrital geochronology (Stevens et al., 2013; Vermeesch, 2013). It is also not clear how 'significant' the differences between the various samples are. It would be useful if we could somehow 'prove' that samples a, b, e and f were drawn from the same population, and that the other samples were drawn from different populations.

2.1 hypothesis testing

We can formalise the two-sample comparison problem as a statistical 'null hypothesis' ($H_\circ$) by proposing that:

> $H_\circ$: 'two samples (x and y) were drawn from the same population.'

If $H_\circ$ is correct, then X^2_{stat} is expected to follow a chi-square distribution with $df = (n_s - 1)(n_c - 1)$ degrees of freedom, where n_s is the number of samples (i.e. $n_s = 2$ for a two-sample comparison) and n_c is the number of classes/bins (i.e., $n_c = 2$ as well). The probability of observing a X^2_{stat}-value at least as extreme as the value obtained from Equation 2 under $H_\circ$ is called the 'p-value'. A pre-defined cutoff value α may be used to evaluate $H_\circ$ on a 100(1-α)% confidence level. For example, if $\alpha = 0.05$ and $p < \alpha$, then $H_\circ$ is rejected in favour of the 'alternative hypothesis':

(i)	a	b	c	d
a	0.00	0.26	2.99	29.2
b	0.26	0.00	0.69	22.4
c	2.99	0.69	0.00	15.4
d	29.2	22.4	15.4	0.00

(ii)	e	f	g	h
e	0.00	0.22	18.5	266
f	0.22	0.00	23.6	278
g	18.5	23.6	0.00	172
h	266	278	172	0.00

Table 1: Chi-square statistics for the comparison of (i) samples a, b, c and d, which contain ∼20 dates each; and (ii) samples e, f, g and h, which contain ∼200 dates each, collected from populations D (samples a, b, e and f), E (samples c and g) and F (samples d and h).

H_a: 'two samples (x and y) were drawn from different populations.'

Applying this procedure to our synthetic data yields two tables of p-values:

	a	b	c	d
a	1.00	0.61	0.084	6×10^{-8}
b	0.61	1.00	0.41	2×10^{-6}
c	0.084	0.41	1.00	9×10^{-5}
d	6×10^{-8}	2×10^{-6}	9×10^{-5}	1.00

	e	f	g	h
e	1.00	0.64	1.7×10^{-5}	8.0×10^{-60}
f	0.64	1.00	1.2×10^{-6}	1.7×10^{-62}
g	1.7×10^{-5}	1.2×10^{-6}	1.00	3.4×10^{-39}
h	8.0×10^{-60}	1.7×10^{-62}	3.4×10^{-39}	1.00

All the p-values for comparison with sample d are well below the 0.05 cutoff, and so we can confidently reject the hypothesis that samples a, b and c were drawn from the same population as d. It is important to note that failure to reject the null hypothesis does not mean that $H_\circ$ has been accepted. Consider, for example, sample c in the first half of the table. Recall that this sample was collected from a different population (E) than samples a and b (which were sampled from distribution D), and so $H_\circ$ is most definitely false in this case. Nevertheless, the p-values for comparison of a and b with c are 0.084 and 0.41, respectively. Both of these are above the 0.05 threshold, resulting in a failure to reject the false null hypothesis. In statistical terms, we have committed a 'Type-II error' (a Type-I error occurs when we have accidentally rejected a true null hypothesis). It is only when sample size is increased from ∼20 to ∼ 200 dates per sample that the chi-square test gains sufficient 'power' to detect the relatively subtle difference between populations D and E. Thanks to this gain in power, the p-values for comparison of samples e and f (which are derived from population D) and sample g (which is derived from population E) have dropped to 1.7×10^{-5} and 1.2×10^{-6}, respectively. This is well below the cutoff value of 0.05 and so we have successfully rejected the null hypothesis and avoided the Type-II error. The concept of statistical power is very important but has received little attention in the context of detrital geochronology.

2.2 effect size

Statistical power is defined as 1-β, where β is the probability of committing a Type-II error (the probability of committing a Type-I error is simply α). If $H_\circ$ is false, then β decreases (and power increases) predictably

with sample size n (Cohen, 1977, 1992). The minimum sample size required to reject a false null hypothesis with 100(1-α)% confidence at least 100(1-β)% of the time crucially depends on 'the degree to which the null hypothesis is false' (Cohen, 1977). The latter quantity is better known as the 'effect size', ϵ. For the chi-square test, the effect size is defined as follows (Cohen, 1977, 1992):

$$\epsilon = \sqrt{\sum_{i=1}^{k} \frac{(p_i^a - p_i^\circ)^2}{p_i^\circ}} \tag{3}$$

where p_i° and p_i^a are the true proportions of each bin under the null hypothesis and actual populations, respectively. For the synthetic data of Figure 1, we can calculate the exact values of ϵ for comparison of populations D, E and F:

	D	E	F
D	0.00	0.31	1.15
E	0.31	0.00	0.93
F	1.15	0.93	0.00

Table 2: Chi-square population effect sizes (ϵ) of populations D–F.

Recall that, if $H_\circ$ is true (which is equivalent to saying that $\epsilon = 0$), then X^2_{stat} is expected to follow a chi-square distribution with *df* degrees of freedom. However, if $H_\circ$ is false (i.e., $\epsilon > 0$), then X^2_{stat} will follow a different distribution, namely the *non-central* chi-square distribution. This distribution shifts to higher values as a function of the three parameters: effect size, degrees of freedom and sample size (Appendix A). Thus, when sample size increases, the probability of rejecting the null hypothesis increases as well (Figure 2). Equation 3 defines the effect size ϵ as a population property which, by definition, is an unknown quantity. However, given two samples of finite size, we can *estimate* the effect size using Cramér's V parameter:

$$V = \sqrt{\frac{X^2_{stat}}{n}} \tag{4}$$

Computing Cramér's V for samples a-h yields two new dissimilarity matrices:

	a	b	c	d
a	0.00	0.11	0.39	1.21
b	0.12	0.00	0.19	1.09
c	0.38	0.18	0.00	0.86
d	1.21	1.06	0.88	0.00

	e	f	g	h
e	0.00	0.03	0.30	1.15
f	0.03	0.00	0.34	1.19
g	0.30	0.34	0.00	0.91
h	1.15	1.18	0.93	0.00

Two observations stand out from these tables. First, the values comparing the small samples (a–d) do not vary much from those comparing the large samples (e–h). Second, recalling that a, b, e and f were sampled from population D, samples c and g from E and samples d and h from F, there is an excellent agreement between the sample effect size V and the population effect sizes ϵ listed in Table 2. We will refer to the sample effect size as the chi-square *distance* in the remainder of this paper (McCune and Grace, 2002).

2.3 The trouble with p-values

The ability to reject a false null hypothesis steadily increases with sample size. Conversely, the sample size required to consistently reject a false null hypothesis is inversely proportional to the 'degree of falseness' of

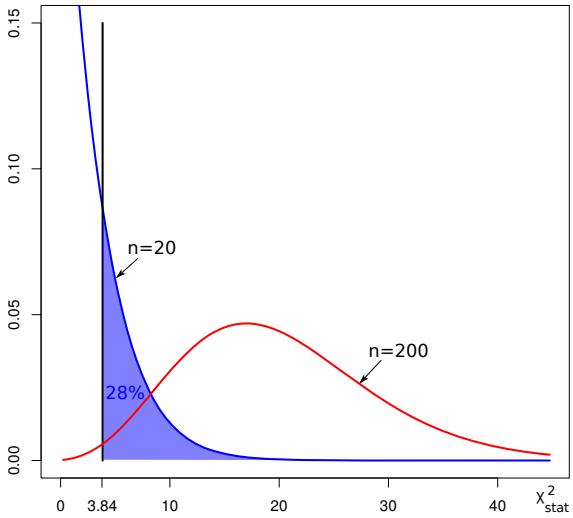

Figure 2: The non-central chi-square distribution for comparison of two samples collected from populations D and E using sample sizes of 20 and 200 dates; the vertical line marks the 95% cutoff for the null hypothesis ($H_\circ$) that both samples were derived from the same distribution. Small samples are more likely to yield low X^2_{stat}-values and fail to reject $H_\circ$, incurring a Type-I error. For the 20-grain dataset, this occurs 78% of the time.

said hypothesis. In other words, no matter how small the difference between distributions may be, there is always a sample size able to detect this difference. This observation is very important for the application of statistical hypothesis testing to detrital geochronology. No two geological samples are ever exactly the same. Even two samples collected from the same river bed, or from the same alluvial fan surface, are likely to be slightly different for any number of reasons, such as hydraulic sorting (Garzanti et al., 2009), sample heterogeneity (DeGraaff-Surpless et al., 2003), or sample preparation (Sircombe and Stern, 2002). The geological record is extremely heterogeneous and no sedimentary mixing process is able to completely erase this inherent heterogeneity. There is always a point, which may only occur after hundreds or even thousands of analyses, where this heterogeneity becomes detectable. This simple observation casts doubt on the scientific value of formalised hypothesis tests (Vermeesch, 2009; Ziliak and McCloskey, 2008). The key point is that there exists a fundamental difference between a mathematical ('null') hypothesis and a scientific hypothesis. Whereas a mathematical hypothesis is either true or false, scientific hypotheses are always false (at some decimal place, Tukey, 1991). There are just two situations where statistical hypothesis tests and p-values do serve a purpose:

1. To prevent statistically unjustified interpretations. Visual comparison of random samples and empirical distributions can be misleading. The human brain has a tendency to see clusters and patterns where there are none. It is very difficult to assess the 'significance' of random sampling fluctuations by mere visual inspection. Statistical hypothesis tests can be useful for verifying whether perceived differences may simply result from this randomness. If the p-value of such a test exceeds 0.05, then any differences between two samples are most likely due to random sampling fluctuations. Put in a different way, a p-value exceeding 0.05 means that the sample size is insufficient to detect the true difference between the two samples of interest. The situation is slightly more complex when more than two samples are involved, as will be discussed in Section 5.

2. In powered hypothesis tests. If the effect size is known, or is postulated *a priori*, then one can calculate the sample size required to detect this effect at least 80% of the time, say. Such *powered* tests are commonly used in the medical sciences, where the 'effect' may correspond to the gain in life expectancy of patients, and the 'sample size' is the number of patients enrolled in a clinical trial. Unfortunately,

in detrital geochronology it is rarely possible to quantify the effect size in advance. Hence the concept of 'statistical significance' ($p < 0.05$) has little scientific value and should probably be abandoned in this context. This point of view is gradually gaining acceptance in many fields of science (e.g., Gelman and Stern, 2006; Stang et al., 2010; Head et al., 2015). In fact, some journals have even gone so far as banning the use of p-values from their pages altogether (Trafimow and Marks, 2015). A case can be made that the Earth Sciences should follow suit.

In conclusion, when comparing two age distributions, the scientifically relevant question is how different two samples are, and not how *significantly* different they are. Only the effect size is able to objectively quantify the difference between two samples independent of sample size. The widely documented utility of the effect size (Cohen, 1977, 1992) is in direct contradiction with Saylor and Sundell (2016)'s assessment that sample size dependence is actually a desirable quality.

3 Non-parametric statistics

The chi-square distance discussed in the previous section is naturally suited to categorical variables (e.g., sandstone petrography, Weltje, 2002), and not so much to continuous variables such as geochronological dates. Although it is certainly possible to plot geochronological data as a histogram and treat the resulting bin counts as categorical variables, this procedure requires the analyst to make a number of rather arbitrary design decisions. First of all, (s)he must choose an appropriate number of bins. Second, (s)he must decide where to place these bins. As a general rule of thumb, one should strive to do this in such as way that the expected counts for the majority of bins is least 10 items, although acceptable results are obtained even when the minimum expected number of counts is as low as 1-4 (Fienberg, 1979). Although this binning procedure yielded good results in Section 2, these would have been (slightly) different had a different set of bins been selected. The requirement to bin the data also reduces the ability of the chi-square distance to detect all but the crudest differences between age distributions. For example, in the two-bin example of Section 2, the chi-square approach hinges on the gradual shift of the three distributions towards older ages. But it would be unable to detect more subtle changes in shape. Consider distribution B as an example (Figure 1.B). This distribution consists of five piecewise uniform 'blocks'. Three of these blocks fall below the 125 Ma cutoff value used to divide the population into two bins. Now, suppose that these three sub-populations were merged into one big block, a single bell curve, or any other shape. The chi-square test would be unable to 'see' the difference between those different options.

These problems can be avoided by using non-parametric dissimilarity measures that do not require binning. The oldest and most widely used of these is the Kolmogorov-Smirnov (KS) statistic (Massey, 1951). Given two samples $x = \{x_1, ..., x_n\}$ and $y = \{y_1, ..., y_m\}$, the KS statistic is defined as the maximum vertical difference between two empirical cumulative distribution functions (ecdf, also known as Cumulative Age Distributions or CADs in the context of detrital geochronology, Vermeesch, 2007):

$$KS(x, y) = \max_{t \in \{x,y\}} |F_x(t) - F_y(t)| \tag{5}$$

where $F_x(t)$ and $F_y(t)$ are the proportion of dates in samples x and y that are younger than t. The KS-statistic spans values between zero (perfect overlap between the two distributions) and one (no overlap between the two distributions). Thus, the KS-dissimilarity can be thought of as the fractional difference between two distributions. The KS statistic, like the chi-square statistic, also changes with increasing sample size. But unlike the chi-square test, the Kolmogorov-Smirnov test does not have a well-defined effect size (Vermeesch, 2016). It is therefore not possible to make analytical predictions about statistical power, and there is no KS-equivalent to the non-central chi-square distribution. However, it is possible to assess power by simulation. Figure 3 shows the KS-distribution for two samples drawn from the same population ($H_\circ$) and for two samples drawn from populations D and F (H_a), respectively, using sample sizes of $n = 20$ and $n = 200$. The increase in sample size results in a narrowing of the KS-distribution, and a shift towards

smaller values, which is most pronounced for the null hypothesis (Figure 3).

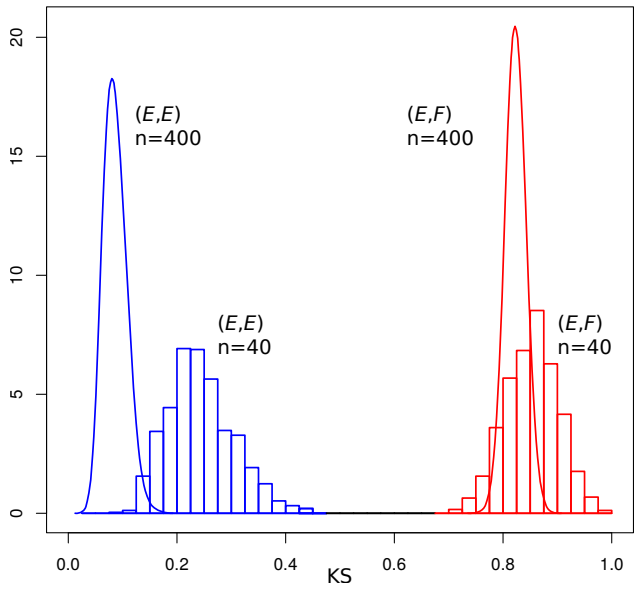

Figure 3: The expected distribution (based on 1000 simulations) of the Kolmogorov-Smirnov (KS) statistic comparing one sample collected from population E with another sample from the same population ($H_\circ$, blue distributions), or with a sample from population F (H_a, red distributions), for sample sizes of $n = 20$ (histograms) and $n = 200$ dates per sample (continuous curves).

Like the chi-square statistic, the Kolmogorov-Smirnov statistic also forms the basis of a statistical hypothesis test. This proceeds in exactly the same fashion as the chi-square test. Given some data, we calculate the KS-statistic and the probability (p-value) of observing a value at least as extreme as this statistic under the null distribution. Again, the p-value depends not only on the true difference between the populations from which the two samples were drawn, but also on sample size. And again, the power of the KS-test to resolve even the tiniest difference between two detrital populations monotonically increases with sample size. In the absence of a sample effect size, the KS-test is not particularly useful for detrital geochronology. The KS-statistic, however, is very useful for multi-sample comparisons by Multidimensional Scaling (MDS, Vermeesch, 2013) analysis. Further details of this will be provided in Section 5. The KS-statistic and -test are relatively insensitive to differences between the tails of distributions (Figure 4). A host of alternative nonparametric statistics have been proposed to address this issue. Two examples of this are the Kuiper statistic[1] (Kuiper, 1960):

$$Kuip(x,y) = \max_{t\in\{x,y\}} \left[F_x(t) - F_y(t), 0\right] + \max_{t\in\{x,y\}} \left[F_y(t) - F_x(t), 0\right] \tag{6}$$

and the Cramér-von-Mises statistic (Anderson, 1962):

$$CvM(x,y) = \frac{n_x n_y}{n_x + n_y} \int_{-\infty}^{+\infty} \left[F_x(t) - F_y(t)\right]^2 dF_{\{x,y\}} \tag{7}$$

where $F_{\{x,y\}}$ is the CAD of the pooled samples x and y, and n_x and n_y are the number of dates contained in them.

[1] In addition to begin more sensitive to the tails of distributions, the Kuiper-statistic and -test are also invariant to changes of scale and origin. This makes this statistic particularly useful for circular data such as compass bearings.

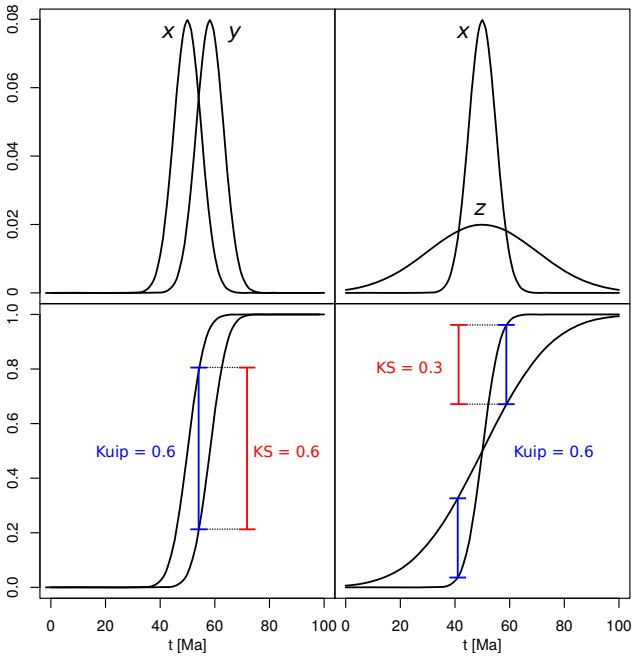

Figure 4: The Kolmogorov-Smirnov (red) and Kuiper (blue) statistics are equally sensitive to the difference between sample distributions x and y, which are laterally offset with respect to each other. The Kuiper statistic is more sensitive than the KS-statistic to the difference between sample distributions x and z, which lies in their tails.

4 Ad-hoc dissimilarity measures

Geochronological dates are associated with analytical uncertainties, which can range from <0.1% (TIMS U-Pb dating) to >100% (fission tracks). None of the dissimilarity measures proposed thus far *explicitly* take into account these uncertainties. In a way, doing so is not necessary for reasons that were already briefly touched on in Section 1. Any distribution of dates represents the convolution of a true age distribution with an error distribution. If the latter is identical for all the samples under consideration, then the dissimilarity measures will only encode differences between the age distributions (plus random sampling fluctuations). Thus, it is generally not necessary to explicitly account for the analytical uncertainties, even when they are large and vary between grains. Nevertheless, several ad-hoc approaches have been developed by geologists to achieve this very goal. This Section will discuss three of these approaches: likeness and cross-correlation (Section 4.1), and the Sircombe-Hazelton L2 norm (Section 4.2).

4.1 likeness and cross-correlation

In many samples, the analytical uncertainty varies greatly between grains. In statistical terminology, this is known as 'heteroscedasticity'. The heteroscedasticity of geochronological data generally has a physical origin. For example, the chemical concentration of the parent nuclide may vary significantly within a sample, making some grains easier to date precisely than others (Galbraith and Green, 1990). Whatever the origin of the heteroscedasticity may be, it seems reasonable to consider the precise dates to be more 'valuable' than the imprecise ones. One popular way to visually express this point of view is the 'Probability Density Plot' (PDP), which is not to be confused with the PDF of Equation 1. A PDP is constructed by summing a number of Gaussian distributions (one for each grain) whose means and standard deviations correspond to the individual dates and their analytical uncertainties, respectively:

$$PDP_x(t) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{N}\left(t|\mu = x_i, \sigma = s_i\right) \tag{8}$$

where $\mathcal{N}\left(t|\mu,\sigma\right)$ is the probability of t under a Normal distribution wih mean μ and standard deviation σ, x_i is the i^{th} date and s_i is its analytical uncertainty. The most precise dates in a sample stand out as sharp peaks in a PDP, whereas imprecise dates are smoothed out. The idea behind this procedure is to emphasize the 'good' data and reduce the prominence of the 'bad' data (Hurford et al., 1984; Brandon, 1996). Satkoski et al. (2013)'s likeness parameter, which is based on an approach used by Amidon et al. (2005a,b), is calculated by taking two PDPs and computing the area overlap between them[2]:

$$L(x,y) = 1 - \frac{1}{2}\int_0^\infty \left|PDP_x(t) - PDP_y(t)\right| dt \tag{9}$$

Saylor et al. (2012, 2013)'s cross-correlation coefficient is obtained by evaluating the two PDPs at equally spaced points t_i (for $1 \leq i \leq n_t$) and computing Pearson's correlation coefficient for the scatterplot of $PDP_x(t_i)$ against $PDP_y(t_i)$. This procedure was inspired by similar approaches in image processing (Lewis, 1995; Pan et al., 2009). But whereas digital images are naturally divided into pixels, cross-correlating PDPs requires discretisation of continuous functions. This first point of criticism will be revisited later in this Section. Both the likeness and cross-correlation coefficients take on values between 0 and 1. In this respect they behave similarly to the KS and Kuiper statistics. Unfortunately this is where the similarity ends, for neither the likeness nor the cross-correlation coefficient stand up to further scrutiny. The main problem with these approaches is their reliance on PDPs, which are fundamentally flawed as a data visualisation tool (Galbraith, 1998; Vermeesch, 2012). In spite of their name, PDPs do not qualify as bona fide probability density estimators. For very precise data, they break down into sequences of spikes. For large samples of imprecise data, PDPs are curves that fit neither the true age distributions, nor the distributions of the measured dates (Vermeesch, 2012). These problems undermine the reliability of any derived quantities such as likeness and cross-correlation, and further mathematical procedures that are based on these dissimilarity measures (e.g., Sundell and Saylor, 2017). Let us illustrate this with two synthetic yet realistic examples (Figure 5). The first example compares two samples of 20 dates that were drawn from population A with a $\sigma = 0.05$ Ma Gaussian analytical error (Figure 5). This example emulates the case of a TIMS U-Pb dataset with better than 1 permil precision. As predicted before and shown in Figure 5, the corresponding PDPs consist of a succession of sharp spikes that do not overlap between the two samples, resulting in a likeness value of 0.21% and a cross-correlation of 0.00077. These extremely low values imply that the two samples were collected from extremely dissimilar populations. This is an absurd conclusion given that they were, in fact, sampled from exactly the same population. For comparison, the KS and Kuiper distances between the two samples are 0.15 and 0.25, respectively, which correctly identifies them as being very similar.

As a second example, let us now investigate what happens when the analytical uncertainties are not small but large. Consider two uniform age distributions J and K that range from 40 to 50 Ma and from 100 to 110 Ma, respectively (Figure 6.i). Suppose that we collect one large sample ($n > 100$ dates, say) from each of these populations, with 10 Ma measurement uncertainties at 1σ. The probability distributions of the dates then range from ∼15 Ma to ∼75 Ma for the first sample and from ∼75 to ∼135 Ma for the second sample. Thus, there is no overlap between the PDFs of either the ages or the dates (Figure 6.ii). However, when we construct the PDPs of the two samples, this double-smooths the age distribution by adding a second convolution with the error distribution. As a result, the PDPs range from ∼0 Ma to ∼90 Ma for the first sample and from ∼60 Ma to ∼150 for the second one (Figure 6.iii). Thus, there exists a substantial overlap between the PDPs, which results in non-zero likeness and cross-correlation coefficients. Again, likeness and (particularly) cross-correlation are producing incorrect results that do not improve by increasing sample size.

[2]The observant reader might note that Equation 9 does not quite agree with Equation 1 of Satkoski et al. (2013). It does, however, describe the procedure used in the spreadsheet calculation of the online supplement to that paper.

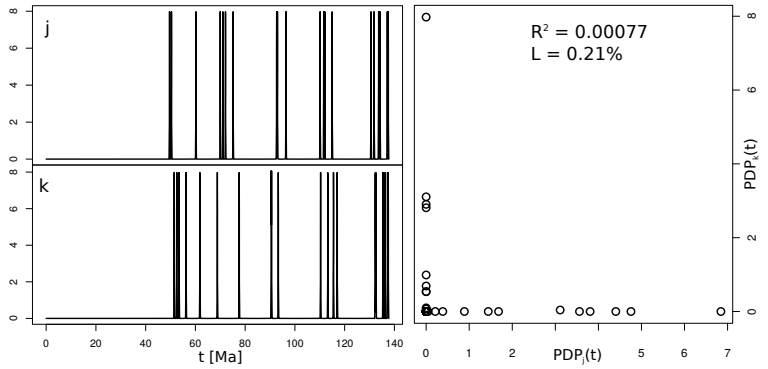

Figure 5: Two samples j and k were drawn from population A with 0.05 Ma analytical uncertainties (1σ). Left: due to the high precision, the Probability Density Plots (PDPs) of the two samples consist of a number of spikes resulting in an extremely low value of 0.21% for Satkoski et al. (2013)'s likeness parameter. Right: evaluating the two PDPs at 200 equally spaced points and plotting the corresponding values against each other yields a cross-correlation coefficient (Saylor et al., 2013) of only 7.7×10^{-4}. These results demonstrate that neither likeness nor cross-correlation are viable dissimilarity measures.

The above two examples demonstrate that likeness and cross-correlation of PDPs are built on shaky foundations and do not qualify as valid dissimilarity measures for detrital geochronology. Replacing PDPs with Kernel Density Estimates (KDEs, Vermeesch, 2012), as briefly mentioned by Satkoski et al. (2013), alleviates some of these problems but also introduces two new problems. First, recall that the main advantage of KDEs over histograms is that they are continuous functions that do not require binning. It is awkward to then discretise these smooth curves into regular time intervals for cross-correlation. Using KDEs instead of PDPs also produces another problem, which is how to select the bandwidth. This issue will be discussed in more detail in the following Section, which introduces our final dissimilarity measure, the Sircombe-Hazelton L2-norm.

4.2 Sircombe-Hazelton

A Kernel Density Estimate (KDE) is defined as:

$$KDE_x(t) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{K}(t|x_i, bw) \tag{10}$$

where $\mathcal{K}$ is the 'kernel' and bw is the 'bandwidth' (Silverman, 1986; Vermeesch, 2012). In this paper we will assume the kernel to be Gaussian, in which case

$$KDE_x(t) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{N}(t|\mu = x_i, \sigma = bw) \tag{11}$$

Note the similarity in form between the definition of a KDE (Equation 11) and that of a PDP (Equation 8). Herein lies the source of much confusion. For small samples, the bandwidth (bw) is greater than the analytical precision (s_i) whereas for large samples the opposite is true. Thus, the PDP does not converge to a KDE at large sample sizes, contrary to claims by Pullen et al. (2014). Sircombe and Hazelton (2004) define a Kernel Functional Estimate (KFE) as follows:

$$KFE_x(t) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{N}\left(t|\mu = x_i, \sigma = \sqrt{c_1^2 - s_i^2}\right) \tag{12}$$

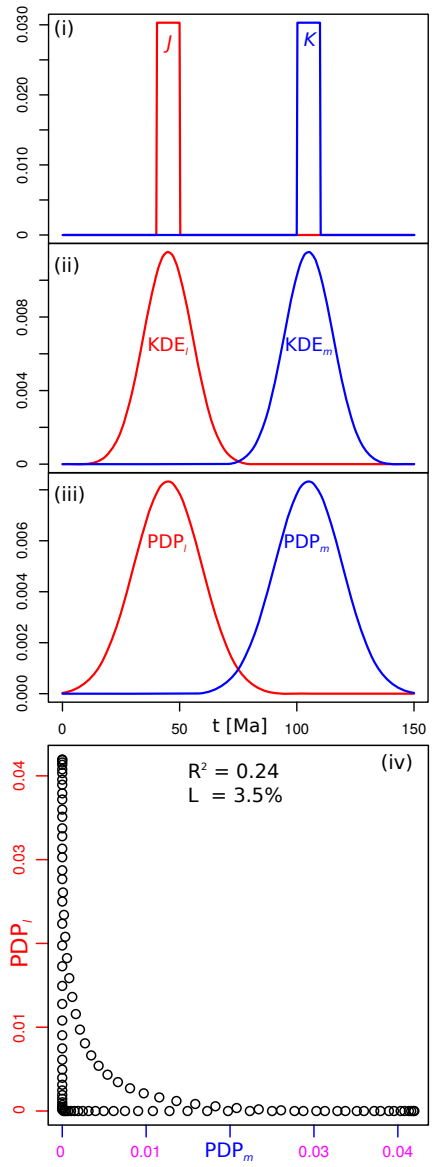

Figure 6: (i) Two piecewise uniform populations J (red) and K (blue); (ii) two samples l (red) and m (blue), drawn from J and K, respectively, and affected by 10 Ma analytical uncertainties (1σ); (iii) the PDPs of the two samples are smoothed versions of the distributions of the dates shown in the previous panel; this results in a partial overlap of the two curves, causing (iv) Saylor et al. (2013)'s cross-correlation coefficient to be significantly greater than zero. This is simply an artifact of the use of PDPs and says nothing about the data.

where $c_1 > \max(s_1, ..., s_n)$. A KFE is a special case of a KDE in which the bandwidth is allowed to vary between sample points so as to smooth the precise data more than the imprecise data. Recall that a PDP smooths the precise data less than the imprecise data. Thus, KFEs and PDPs work in *opposite* ways. KFEs are useful when comparing two datasets that were acquired under different analytical conditions resulting in widely differing analytical uncertainties. As an example, let us compare sample m, which was drawn from population K (Section 4.1), with a second sample (o) drawn from this same population but with a 10 times smaller analytical uncertainty (1 Ma instead of 10 Ma). Despite being drawn from the same population, the dates in samples m and o are distributed differently. The sample distribution of m ranges from 80-130 Ma, whereas that of sample o only ranges from 98-112 Ma (Figure 7). The PDPs of the two samples look even more different, with ranges of 72-138 Ma and 97-113 Ma, respectively. The KFEs, however, are identical (Figure 7), which makes them an effective tool to separate geologically meaningful differences from analytically induced differences between samples. KFEs can be used as the basis of a dissimilarity measure (Sircombe and Hazelton, 2004):

$$SH(x,y) = \sqrt{\int_{-\infty}^{\infty} \left[KFE_x(t) - KFE_y(t)\right]^2 dt} \tag{13}$$

Note that the actual value of this distance is rather arbitrary as it depends on the choice of c_1 in Equation 12. In this respect, the SH-distance is similar to the chi-square dissimilarity, which depends on the choice of bin width. But as long as the same value is used for all samples, the relative differences between the SH-distances do carry meaningful information about the (dis)similarities of geochronological data. Nevertheless, for datasets that do not exhibit great inter-sample differences in analytical precision, the additional smoothing required by the SH-method is arguably not justified.

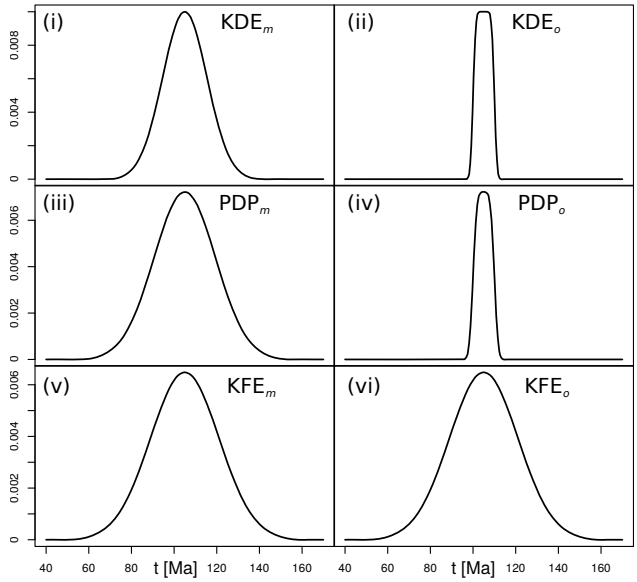

Figure 7: (i) the Kernel Density Estimate (KDE) of sample m, which was drawn from population K with σ=10 Ma analytical uncertainties; (ii) the KDE of sample o, which was also drawn from population K, but with σ=1 Ma analytical uncertainties; (iii) the PDP of sample m, which is double smoothed with respect to K; (iv) the PDP of sample o, which is smoothed less than that of m due to the smaller analytical uncertainties of o compared to m; (v) the Kernel Functional Estimate (KFE) of sample m with $c_1^2 = 250$ in Equation 12; (vi) the KFE of sample o using the same value of c_1^2, which smooths this precise dataset more than sample m.

5 Multi-sample comparison

So far we have been mainly concerned with the comparison of pairs of samples. However, most geological studies require the simultaneous comparison of multiple samples. Such multi-sample studies become increasingly common as the analytical cost of detrital geochronology steadily drops with time (Vermeesch and Garzanti, 2015). Multi-sample comparison adds a second layer of statistical complexity to the interpretation of detrital age spectra. A common approach to making this exercise more objective is to tabulate the p-values of statistical tests such as Kolmogorov-Smirnov and highlight all those values that are less than 0.05, say. Besides the usual caveats regarding statistical hypothesis tests (Section 2.3), such p-value tables also increase the occurrence of Type-I errors. To illustrate this point, consider a 9×9 table of p-values obtained by comparing 10 samples of 100 ages drawn from population A, using the KS-test. In this synthetic case, the null hypothesis is true and therefore we would expect the p-values in our table to fall above the $\alpha = 0.05$ cutoff. For any given pair of samples, there is a 1/20 chance of incurring a Type-I error. But in a 9×9 table

	2	3	4	5	6	7	8	9	10
1	0.81	0.28	0.47	0.58	0.37	0.70	0.81	0.58	0.70
2		0.47	0.58	0.97	0.58	0.47	0.97	0.91	0.37
3			0.81	0.58	0.02	0.08	0.15	0.05	0.04
4				0.97	0.11	0.37	0.81	0.15	0.11
5					0.47	0.21	0.81	0.58	0.37
6						0.02	0.47	0.70	0.47
7							0.47	0.28	0.37
8								0.81	0.47
9									0.47

Table 3: Table of Kolmogorov-Smirnov p-values comparing 10 random samples of 100 ages from population A with each other. Three of these values fall below the $\alpha = 0.05$ cutoff that is typically used for a statistical hypothesis test. In other words, three Type-I errors have been committed in this multiple comparison exercise.

of pairwise comparisons, there are $9 \times 8/2 = 36$ such comparisons. This increases the chance of incurring a Type-I error to $100(1 - [19/20]^{36}) = 84\%$ (Table 3). In other words, the occurrence of 'significant' differences between age distributions in p-value tables does not necessarily mean that the underlying population are different as well. One way to address this issue is to simply reduce the p-value cutoff from α to α/N, where N is the number of pairwise comparisons. This is known as the 'Bonferroni correction' (Rice, 1995). Unfortunately, the Bonferroni correction is overly conservative. It reduces statistical power and increases the chance of committing a Type-II error (Nakagawa, 2004). It is probably better to abandon p-value tables altogether.

Simply tabulating the dissimilarities instead of the p-values works better but requires that sample sizes do not vary much between samples. Furthermore, visual interpretation of such tables becomes progressively more challenging with increasing number of samples. The number of pairwise comparisons between N samples is $N(N-1)/2$, so that $N = 10$ samples may be compared in 45 possible ways, whereas $N = 100$ samples represent no fewer than 4500 pairwise comparisons. It is more productive in such cases to explore alternative, graphical means of interpreting the data. Multidimensional Scaling (MDS) is one example of such a method that has proven to be quite useful in this respect (Vermeesch, 2013). MDS takes a table of pairwise dissimilarities as input, and produces a set of two (or more) dimensional coordinates as output. The scatterplot of these coordinates represents a 'map' in which similar samples plot close together and dissimilar samples plot far apart. If the dissimilarity measure fulfils the metric requirements (nonnegativity, symmetry and triangle inequality), then the configuration can be obtained by straightforward matrix algebra (Young and Householder, 1938). All the dissimilarity measures discussed in this paper fulfil the nonnegativity and symmetry requirements. The triangle inequality is fulfilled by the chi-square distance (McCune and Grace, 2002) and by the KS-statistic:

$$\begin{aligned} KS(x,z) &= \max_{t\in\{x,y\}} |F_x(t) - F_z(t)| \\ &= \max_{t\in\{x,y\}} |F_x(t) - F_y(t) + F_y(t) - F_z(t)| \\ &\leq \max_{t\in\{y,z\}} |F_y(t) - F_y(t)| + \max_{t\in\{y,z\}} |F_y(t) - F_z(t)| \\ &= KS(x,y) + KS(y,z) \end{aligned} \tag{14}$$

for any three samples x, y and z. The Kuiper statistic obeys the triangle inequality for similar reasons, but the Cramér-von-Mises statistic and variants thereof such as the Anderson-Darling statistic (Anderson and Darling, 1954) does not. For example, when inspecting the CvM-dissimilarities of samples b, c and d, it turns out that:

$$CvM(b,c) + CvM(c,d) = 0.153 + 4.044 = 4.197 < 30.196 = CvM(b,d) \tag{15}$$

Thus, the Cramér-von-Mises statistic does not behave as a *bona fide* distance measure or 'metric'. However, Torgerson (1952) showed that it is easy to fix this violation of the triangle inequality by simply adding a constant (c_2, say) to the dissimilarities. For example, for our Cramér-von-Mises example we can define:

$$\delta(x, y) = c_2 + CvM(x, y) \tag{16}$$

Setting $c_2 = 50$, it is easy to see that:

$$\delta(b, c) + \delta(c, d) = 50.153 + 54.044 = 104.197 > 80.196 = \delta(b, d) \tag{17}$$

which fulfils the triangle inequality. Equation 16 is the simplest example of a transformation that converts dissimilaries into *disparities*. Generalising this concept to any monotonically increasing function further increases the flexibility of MDS (Shepard, 1962; Kruskal and Wish, 1978; Shepard, 1980). In some cases, it even allows the condition of symmetry to be violated. Examples of this are travel distances by airplane or ship, which may depend on the wind direction or whether one sails up or down a river; or the level of physical attraction between two people, which is rarely symmetric. The Shepard (1962) approach is referred to as *nonmetric* MDS. The Young and Householder (1938) approach is called *classical* MDS, although confusingly this term is sometimes also used for the Torgerson (1952) method, or even for any MDS algorithm, including nonmetric approaches, that are based on a single dissimilarity matrix (as opposed to 3-way MDS, Carroll and Chang, 1970; Vermeesch and Garzanti, 2015).

Vermeesch (2013) proposed that, for the application of MDS to detrital geochronology, it is desirable for the dissimilarities to be independent of sample size. Only the chi-square effect size fulfils this requirement. Contrary to claims by Vermeesch (2013), the KS-dissimilarity is actually not independent of sample size, as discussed in Section 3 and shown in Figure 3. Fortunately, Vermeesch (2016) pointed out that Vermeesch (2013) was not only wrong about the KS-distance being independent of sample size, but also about the necessity of this independence requirement itself. In fact, such a condition would be incompatible with the non-negativity requirement for noisy data. This is because the error distribution of any strictly positive dissimilarity measure is inevitably skewed towards positive values. Consider, for example, the Euclidean distance between two points that are drawn from Normal distributions with zero mean and a standard deviation σ. The expected value for this distribution is $\sigma\sqrt{\pi/2}$ (i.e., the mean of a Rayleigh distribution). If σ is the standard error of the mean, then it is easy to introduce a sample size dependence on the average Euclidean distance between points. Thus, even the Euclidean distance, which is the archetypal example of an MDS-worthy dissimilarity measure, does not fulfil the requirement of sample size independence. To see why this requirement is not necessary, let us consider the simple case of a three point comparison. Suppose that we have three samples x_1, y_1 and z_1, let $d(x_1, y_1)$ and $d(x_1, z_1)$ be the dissimilarities between x_1 and y_1 or z_1, respectively. Further suppose that $d(x_1, y_1) < d(x_1, z_1)$. Next, consider a fourth sample x_2 drawn from the same population as x_1. Assume that x_2 contains a different number of dates than x_1, so that $d(x_1, y_1) \neq d(x_2, y_1)$ and $d(x_1, z_1) \neq d(x_2, z_1)$. In that case MDS will still work if $d(x_2, y_1) < d(x_2, z_1)$. This is a much weaker requirement than sample size independence. It is a requirement that is fulfilled by the Euclidean distance, but also by the KS- and related dissimilarity measures.

As an example, Figure 8 shows the MDS configuration of 14 samples $a - h$ and $q - w$, whose sampling distributions are shown in Figure 1. Recall that samples a, b, e and f were drawn from population D, samples c and g from population E, and samples d and h from population F. Additional samples q and u were derived from population G, samples r and v from population H, and samples s and w from population I. Samples $a - d$ and $q - s$ each contain 20 single grain ages, whereas samples $e - h$ and $u - w$ contain an order of magnitude more dates. Despite the huge range in sample sizes, the MDS map shows a sensible configuration, correctly grouping all the samples that share the same population, irrespective of sample size. Nevertheless, it is also important to note that the differences between populations $D - F$ and $G - I$ are small in comparison with the statistical 'noise' in the KS-dissimilarities of the small samples. This is the reason why 200-grain sample g plots closer to 200-grain sample y (which comes from a different population) than it does to 20-grain sample c (which comes from the same population as g).

6 Conclusions

The effect size of parametric tests such as chi-square is the only way to compare the dissimilarity between two detrital age distributions in absolute terms and independent of sample size. Unfortunately, the chi-square effect size requires binning data and lacks the resolution to 'see' the difference between subtlely different samples. Dissimilarity measures based on non-parametric statistics such as Kolmogorov-Smirnov or Kuiper are much better at recognising these differences, but do exhibit a dependence on sample size. This reduces their usefulness as an absolute point of comparison between samples of greatly different size. Fortunately, most detrital studies only rely on the relative differences between age distributions. The rank order of, say, the KS and Kuiper dissimilarities between samples, is independent of sample size and can be visualised by MDS analysis.

There exists some disagreement within the detrital geochronology community about the best way to deal with analytical uncertainty. This paper has reiterated the point previously made by Vermeesch (2012) that, as long as all the samples in a provenance study were analysed using similar laboratory conditions there is no real need to explicitly account for the analytical uncertainty. Deconvolution of the age distribution and the distribution of analytical uncertainties is notoriously difficult. But for the vast majority of provenance studies such deconvolution is not necessary, and it suffices that we directly compare the distribution of the dates. It is only when a provenance study combines samples analysed under different laboratory conditions exhibiting widely different precision that extra care must be taken. The Sircombe-Hazelton L2-norm has been specifically designed for this situation (Section 4.2).

Detrital geochronology, like other areas within the Earth Sciences, is a field of research that increasingly depends on computing and statistics for data interpretation. It is not always easy to adapt existing statistical methods to Earth Science applications. The author of the present paper is not immune to these difficulties, as is evident from the Errata of his paper on Multidimensional Scaling (Vermeesch, 2013, 2014, 2016) which, fortunately, do not seem to undermine the validity of that method. More serious examples of misused statistics are unpowered statistical hypothesis tests (Ziliak and McCloskey, 2008; Trafimow and Marks, 2015), and the misunderstanding that PDPs are the same as KDEs due to their similar appearance (Equations 8 and 10). As difficult as it may be to get existing statistical methods right, inventing new ones is even more difficult. Likeness and cross-correlation are examples of statistical devices that appear sensible at first but do not stand up to further scrutiny. When assessing the performance of such *ad hoc* statistical methods, it is useful to explore their behaviour under asymptotic conditions. For example, one would expect statistical estimators of some unknown quantity to converge to the correct solution in the limit of infinite sample size and infinitessimal analytical uncertainty. This paper has shown that PDPs, likeness and cross-correlation do not fulfil this requirement. Given the difficulty of inventing new statistical methods, it is possibly best to avoid doing so altogether, unless this happens in collaboration with a mathematician. One successful example of this is the aforementioned L2-norm, which is the fruit of a collaboration between a geoscientist and a statistician (Sircombe and Hazelton, 2004).

Appendix: power analysis and confidence intervals

As the name suggests, the non-central chi-square distribution is shifted towards higher values with respect to the ordinary ('central') chi-square distribution (Section 2.2). It is not described by one but two parameters: the degrees of freedom, df, and the 'non-centrality parameter', λ, which is a function of the effect size (ϵ) and sample size (n):

$$\lambda = n\epsilon^2 \tag{18}$$

The expected value of the non-central chi-square distribution is given by:

$$\overline{X}^2 = df + \lambda \tag{19}$$

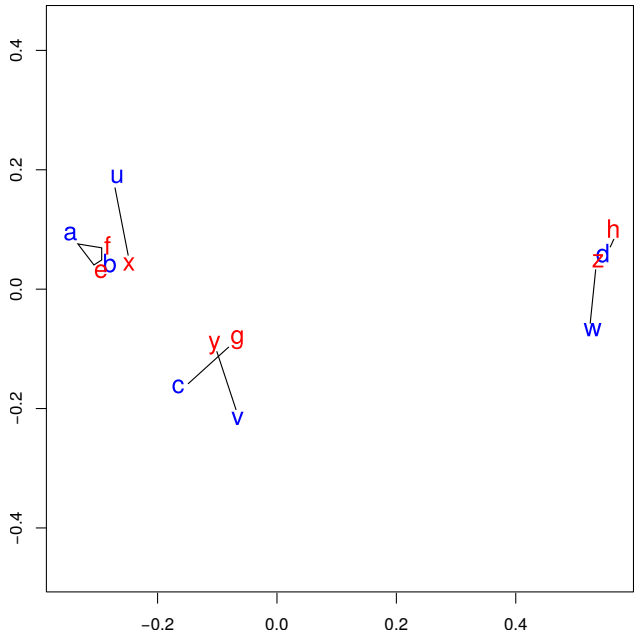

Figure 8: Kolmogorov-Smirnov-based Multidimensional Scaling (MDS) configuration of samples $a-h$ and $q-w$, whose sampling distributions are shown in Figure 1. black lines connect samples drawn from identical detrital populations. Despite an order of magnitude difference in sample size between samples $a-e$, $q-s$ (20 grains each, blue) and $d-h$, $u-w$ (200 grains each, red), the MDS configuration correctly groups the samples according to provenance. This illustrates that the sample-size dependence of the KS-dissimilarity shown in Figure 3 does not pose a major problem for MDS analysis. The x- and y-scale are normalised to the sum of the squared Euclidean distances within the MDS configuration.

This explains some important trends and patterns in Table 1. For example, $X^2_{stat}(a,b) < X^2_{stat}(a,c) < X^2_{stat}(a,d)$ because $\epsilon(a,b) < \epsilon(a,c) < \epsilon(a,b)$. And $X^2_{stat}(c,d) < X^2_{stat}(g,h)$ because $n(c), n(d) < n(g), n(h)$. Using the non-central chi-square distribution, it is possible to calculate the probability β of committing a Type-II error for any given effect size and sample size. This is done as follows. First, we look up the cutoff value of the central chi-square distribution for the confidence level and degrees of freedom of interest. For example, to compare two samples drawn from populations D and E using two histogram bins, we can choose $\alpha = 0.05$ and $df = (2-1)(2-1) = 1$. The resulting cutoff value for χ^2_{stat} is 3.84. Next, we look up the percentile corresponding to this value under a non-central distribution with one degree of freedom and a non-centrality parameter given by Equation 18. For example, when comparing samples and a and c, $\lambda = 20 \times 0.31^2 = 1.9$. The percentile corresponding to $X^2_{stat} = 3.84$ under a non-central chi-square distribution with $\lambda = 1.9$ is 0.72. In other words, there is a 72% chance of committing a Type-II error and the false null hypothesis going undetected. A tenfold increase of sample size from 20 to 200 would result in a tenfold increase of the non-centrality parameter. Under the corresponding non-central chi-square distribution, the cumulative chance of falling below the cutoff value of 3.84 drops to only 0.8%. So increasing sample size from 20 to 200 increases the power (defined as $1-\beta$) of the chi-square test from 0.28 to 0.99. The increase in statistical power with sample size is a universal fact of hypothesis testing. The population ϵ is unknown but can be estimated using Cramér's V. We can construct a 100(1-α)% confidence interval for ϵ by finding the subset of population effect sizes that are more than 100(1-α/2)% likely to yield a chi-square value greater than (for the upper limit of the confidence interval) or less than (for its lower limit) X^2_{stat}. See Figure 9 for an example. Increasing sample size tightens the confidence interval for ϵ but essentially does not change the value of V.

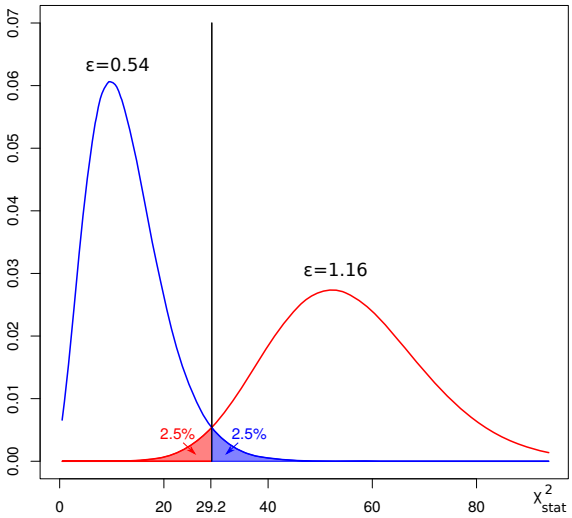

Figure 9: The vertical line marks the observed X^2_{stat}-value for comparison between samples a and d (Table 1.i); the two non-central chi-square distributions mark a 95% confidence interval [0.54,1.16] for the population effect size ϵ.

Acknowledgments

The author would like to thank Dr. Joel Saylor and an anonymous reviewer for detailed feedback on the submitted manuscript.

References

Amidon, W. H., Burbank, D. W., and Gehrels, G. E. Construction of detrital mineral populations: insights from mixing of U-Pb zircon ages in Himalayan rivers. *Basin Research*, 17:463–485, 2005a.

Amidon, W. H., Burbank, D. W., and Gehrels, G. E. U–Pb zircon ages as a sediment mixing tracer in the Nepal Himalaya. *Earth and Planetary Science Letters*, 235(1):244–260, 2005b.

Anderson, T. W. On the distribution of the two sample Crámer - Von Mises criterion. *Annals Mathematical Statistics*, 33, 1962. doi: 10.1214/aoms/1177704477.

Anderson, T. W. and Darling, D. A. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):pp. 765–769, 1954.

Blatt, H. and Jones, R. L. Proportions of exposed igneous, metamorphic, and sedimentary rocks. *Geological Society of America Bulletin*, 86(8):1085–1088, 1975.

Brandon, M. Probability density plot for fission-track grain-age samples. *Radiation Measurements*, 26(5): 663 – 676, 1996. ISSN 1350-4487. doi: DOI: 10.1016/S1350-4487(97)82880-6.

Carroll, J. D. and Chang, J.-J. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. *Psychometrika*, 35(3):283–319, 1970.

Codilean, A. T., Bishop, P., Stuart, F. M., Hoey, T. B., Fabel, D., and Freeman, S. P. H. T. Single-grain cosmogenic ^{21}Ne concentrations in fluvial sediments reveal spatially variable erosion rates. *Geology*, 36: 159–162, 2008.

Cohen, J. *Statistical power analysis for the behavioral sciences*. Academic Press New York, 1977.

Cohen, J. A power primer. *Psychological bulletin*, 112(1):155, 1992.

Copeland, P. and Harrison, T. M. Episodic rapid uplift in the Himalaya revealed by $^{40}Ar/^{39}Ar$ analysis of detrital K-feldspar and muscovite, Bengal fan. *Geology*, 18(4):354–357, 1990.

DeGraaff-Surpless, K., Mahoney, J. B., Wooden, J. L., and McWilliams, M. O. Lithofacies control in detrital zircon provenance studies: Insights from the Cretaceous Methow basin, southern Canadian Cordillera. *Geological Society of America Bulletin*, 115:899–915, 2003.

Fienberg, S. E. The use of chi-squared statistics for categorical data problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 54–64, 1979.

Galbraith, R. F. and Green, P. F. Estimating the component ages in a finite mixture. *Nuclear Tracks and Radiation Measurements*, 17:197–206, 1990.

Galbraith, R. The trouble with "probability density" plots of fission track ages. *Radiation Measurements*, 29:125–131., 1998.

Garzanti, E., Andò, S., and Vezzoli, G. Grain-size dependence of sediment composition and environmental bias in provenance studies. *Earth and Planetary Science Letters*, 277:422–432, 2009. doi: 10.1016/j.epsl.2008.11.007.

Garzanti, E., Vermeesch, P., Andò, S., Vezzoli, G., Valagussa, M., Allen, K., Kadi, K. A., and Al-Juboury, A. I. Provenance and recycling of Arabian desert sand. *Earth-Science Reviews*, 2013.

Gelman, A. and Stern, H. The difference between significant and not significant is not itself statistically significant. *The American Statistician*, 60(4):328–331, 2006.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. The extent and consequences of p-hacking in science. *PLoS Biol*, 13(3):e1002106, 2015.

Hurford, A., Fitch, F., and Clarke, A. Resolution of the age structure of the detrital zircon populations of two Lower Cretaceous sandstones from the Weald of England by fission track dating. *Geological Magazine*, 121:269–396, 1984.

Kruskal, J. B. and Wish, M. *Multidimensional scaling*, volume 07-011 of *Sage University Paper series on Quantitative Application in the Social Sciences*. Sage Publications, Beverly Hills and London, 1978.

Kuiper, N. H. Tests concerning random points on a circle. In *Indagationes Mathematicae (Proceedings)*, volume 63, pages 38–47. Elsevier, 1960.

Lewis, J. P. Fast normalized cross-correlation. In *Vision interface*, volume 10, pages 120–123, 1995.

Massey, F. J. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

McCune, B. and Grace, J. *Analysis of ecological communities*. MjM Software Design, 2002.

Nakagawa, S. A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology*, 15(6):1044–1045, 2004.

Nie, J., Stevens, T., Rittner, M., Stockli, D., Garzanti, E., Limonta, M., Bird, A., Andò, S., Vermeesch, P., Saylor, J., et al. Loess plateau storage of northeastern Tibetan plateau-derived Yellow River sediment. *Nature communications*, 6, 2015.

Pan, B., Qian, K., Xie, H., and Asundi, A. Two-dimensional digital image correlation for in-plane displacement and strain measurement: a review. *Measurement science and technology*, 20(6):062001, 2009.

Pell, S. D., Williams, I. S., and Chivas, A. R. The use of protolith zircon-age fingerprints in determining the protosource areas for some Australian dune sands. *Sedimentary Geology*, 109:233–260., 1997.

Pullen, A., Ibáñez-Mejía, M., Gehrels, G. E., Ibáñez-Mejía, J. C., and Pecha, M. What happens when n= 1000? Creating large-n geochronological datasets with LA-ICP-MS for geologic investigations. *Journal of Analytical Atomic Spectrometry*, 29(6):971–980, 2014.

Rahl, J. M., Reiners, P. W., Campbell, I. H., Nicolescu, S., and Allen, C. M. Combined single-grain (U-Th)/He and U/Pb dating of detrital zircons from the Navajo Sandstone, Utah. *Geology*, 31:761–764, 2003. doi: 10.1130/G19653.1.

Rice, J. A. *Mathematical Statistics and Data Analysis*. Duxbury, Pacific Grove, California, 1995.

Satkoski, A. M., Wilkinson, B. H., Hietpas, J., and Samson, S. D. Likeness among detrital zircon populations – An approach to the comparison of age frequency data in time and space. *Geological Society of America Bulletin*, 125(11-12):1783–1799, 2013.

Saylor, J. E. and Sundell, K. E. Quantifying comparison of large detrital geochronology data sets. *Geosphere*, pages GES01237–1, 2016.

Saylor, J. E., Stockli, D. F., Horton, B. K., Nie, J., and Mora, A. Discriminating rapid exhumation from syndepositional volcanism using detrital zircon double dating: Implications for the tectonic history of the Eastern Cordillera, Colombia. *Geological Society of America Bulletin*, 124(5-6):762–779, 2012.

Saylor, J. E., Knowles, J. N., Horton, B. K., Nie, J., and Mora, A. Mixing of source populations recorded in detrital zircon U-Pb age spectra of modern river sands. *The Journal of Geology*, 121(1):17–33, 2013.

Shepard, R. N. The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27(2):125–140, 1962.

Shepard, R. N. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468):390–398, 1980.

Silverman, B. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

Sircombe, K. N. and Hazelton, M. L. Comparison of detrital zircon age distributions by kernel functional estimation. *Sedimentary Geology*, 171:91–111, 2004. doi: 10.1016/j.sedgeo.2004.05.012.

Sircombe, K. N. and Stern, R. A. An investigation of artificial biasing in detrital zircon U-Pb geochronology due to magnetic separation in sample preparation. *Geochimica et Cosmochimica Acta*, 66(13):2379–2397, 2002.

Stang, A., Poole, C., and Kuss, O. The ongoing tyranny of statistical significance testing in biomedical research. *European journal of epidemiology*, 25(4):225–230, 2010.

Stevens, T., Carter, A., Watson, T., Vermeesch, P., Andò, S., Bird, A., Lu, H., Garzanti, E., Cottam, M., and Sevastjanova, I. Genetic linkage between the Yellow River, the Mu Us desert and the Chinese Loess Plateau. *Quaternary Science Reviews*, 78:355–368, 2013.

Stock, G. M., Ehlers, T. A., and Farley, K. A. Where does sediment come from? Quantifying catchment erosion with detrital apatite (U-Th)/He thermochronometry. *Geology*, 34:725–728, 2006. doi: 10.1130/G22592.1.

Sundell, K. and Saylor, J. E. Unmixing detrital geochronology age distributions. *Geochemistry, Geophysics, Geosystems*, 2017.

Torgerson, W. S. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419, 1952.

Trafimow, D. and Marks, M. Editorial. *Basic and Applied Social Psychology*, 37(1):1–2, 2015. doi: 10.1080/01973533.2015.1012991.

Tukey, J. W. The philosophy of multiple comparisons. *Statistical science*, pages 100–116, 1991.

Vermeesch, P. Quantitative geomorphology of the White Mountains (California) using detrital apatite fission track thermochronology. *Journal of Geophysical Research (Earth Surface)*, 112(F11):3004, 2007. doi: 10.1029/2006JF000671.

Vermeesch, P. Lies, Damned Lies, and Statistics (in Geology). *EOS Transactions*, 90:443–443, 2009. doi: 10.1029/2009EO470004.

Vermeesch, P. On the visualisation of detrital age distributions. *Chemical Geology*, 312-313:190–194, 2012. doi: 10.1016/j.chemgeo.2012.04.021.

Vermeesch, P. Multi-sample comparison of detrital age distributions. *Chemical Geology*, 341:140–146, 2013.

Vermeesch, P. Corrigendum to "Multi-sample comparison of detrital age distributions" [Chem. Geol. 341 (11 March 2013)140-146]. *Chemical Geology*, (380):191, 2014.

Vermeesch, P. Corrigendum to: "Multi-sample comparison of detrital age distributions (vol 341, pg 140, 2013)". *Chemical Geology*, 425:145–145, 2016.

Vermeesch, P. and Garzanti, E. Making geological sense of 'Big Data' in sedimentary provenance analysis. *Chemical Geology*, 409:20–27, 2015.

Weltje, G. Quantitative analysis of detrital modes: statistically rigorous confidence regions in ternary diagrams and their use in sedimentary petrology. *Earth-Science Reviews*, 57(3-4):211 – 253, 2002. ISSN 0012-8252. doi: DOI: 10.1016/S0012-8252(01)00076-9.

Young, G. and Householder, A. S. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3(1):19–22, 1938.

Ziliak, S. T. and McCloskey, D. N. *The cult of statistical significance: how the standard error costs us jobs, justice, and lives*. University of Michigan Press, 2008.