Yamaguchi
Progress in Earth and Planetary Science (2023) 10:70
https://doi.org/10.1186/s40645-023-00601-y

Progress in Earth and Planetary Science

METHODOLOGY

Open Access

Validation of appropriate estimation criteria for the number of components for separating a polymodal grain-size distribution into lognormal distributions

Naofumi Yamaguchi[1*]

Abstract

Polymodal particle size distributions are generally analyzed by separating them into lognormal distributions, but estimating the precise number of lognormal components required remains a considerable problem. In the present study, appropriate evaluation criteria for the estimation of the number of components were examined by using artificial data for which the true number of components was known. The characteristics of estimations of the number of components by four evaluation criteria, the mean square error (MSE), Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R-squared (ARS), were investigated. The results showed that the MSE and ARS were less sensitive to the true number of components and tended to overestimate the number of components. By contrast, the AIC and BIC tended to underestimate the number of components, and their correct answer rates decreased as the true number of components increased. The BIC tended to include the true number of components among its higher ranked models. The present evaluation results suggest that the MSE, although frequently used, is not necessarily the most appropriate evaluation criterion, and that the AIC and ARS may be more appropriate criteria. Furthermore, checking whether the number of components estimated by the AIC or ARS is included among higher ranked BIC models might prevent overestimation and thereby allow for more valid estimation of the number of components. When the criteria were applied to grain-size distributions of lacustrine sediments, it was possible to estimate the number of components that reflected differences in grain-size distribution characteristics.

Keywords Polymodal grain-size distribution, Lognormal distribution fitting, Estimation criteria, Grain-size analysis

1 Introduction

The sediment grain-size distribution is fundamental information for various types of sediments and is used in many different fields because it reflects the origin of the sediments, the transport processes that acted on them, and the strength of the experienced transport forces. Sediment grain-size characteristics have been used, for example, to reconstruct past climate events and changes of depositional environments. The grain-size distributions of lake sediments have been used as an indicator of hydrological conditions associated with climatic and environmental changes in the region where the lake is located (e.g. Håkanson and Jansson 1983; Xiao et al. 2009; Dietze et al. 2014; Lu et al. 2018). Grain-size characteristics of aeolian sediments, including loess, have been regarded as a meaningful paleoclimatic proxy (e.g. Vandenberghe et al. 1997; Sun et al. 2002; Sun 2004; Qin et al. 2005; Lim and Matsumoto 2006; Machalett et al. 2008; Antoine et al. 2009; Vandenberghe 2013; Lin et al. 2016; Schulte et al. 2018). Tephra grain-size distributions

*Correspondence:
Naofumi Yamaguchi
naofumi.yamaguchi.sci@vc.ibaraki.ac.jp
[1] Global and Local Environment Co-creation Institute, Ibaraki University, Ohu 1375, Itako, Ibaraki 311-2402, Japan

have been used to estimate the characteristics of volcanic eruptions and the dispersal of particles in the atmosphere (e.g. Rose and Durant 2009; Burden et al. 2011; Engwell and Eychenne 2016; Rossi et al. 2019; Miwa et al. 2020). In addition, several methods have been proposed to estimate the transport pathways of modern surface clastic sediments based on spatial trend analyses of their grain-size characteristics (McCave 1978; McLaren 1981; McLaren and Bowles 1985; Gao and Collins 1992; Le Roux and Rojas 2007; Yamashita et al. 2018). Therefore, accurate methods of grain-size analysis and suitable interpretation of grain-size distributions are an important topic in Earth science.

One major issue in the analysis of sediment grain-size data is how to deal with a complex distribution. Previous studies have commonly assumed a unimodal distribution and used representative statistical parameters such as median diameter, sorting, skewness, and kurtosis to characterize the sediment grain size (Folk 1966). However, the grain-size distribution of sediments in various natural environments is frequently complex and polymodal rather than unimodal. In particular, with the advent of laser grain-size analyzers, which reveal grain-size distributions over a wide size range at high resolution, it has become clear that polymodal distributions are frequent for various types of sediments. For example, lake sediments, which may have multiple sources and be affected by multiple sedimentary processes, frequently have polymodal grain-size distributions. Such distributions complicate the analysis and interpretation of the representative parameters calculated by assuming unimodality. Aeolian deposits, including loess, also frequently have complicated polymodal grain-size distributions. Such polymodal grain-size distributions are thought to reflect multiple sediment sources and sedimentary processes, and each mode is thought to preserve environmental information associated with them (Tanner 1964; Visher 1969; Middleton 1976; Ashley 1978). Therefore, it is not appropriate to use just a few representative parameters based on assumed unimodality to interpret polymodal grain-size distributions. Appropriate analysis of such complex grain-size distributions can potentially allow each of the multiple origins and transport processes that produced the sediment deposit to be reconstructed and to provide additional valuable information.

Several methods have been proposed for analyzing complex polymodal distributions of grain size, including end-member mixing analysis (e.g. Weltje and Prins 2007; Parris et al. 2010; Dietze et al. 2012, 2013; Ijmker et al. 2012; Yu et al. 2016) and discrete parametric curve fitting. The latter method of separating grain-size distributions includes the use of prescribed distribution functions such as the Weibull distribution (Lim and Matsumoto 2006; Park et al. 2014; Wu et al. 2020; Peng et al. 2022a) and lognormal distributions as components of the polymodal distribution. In particular, a number of previous studies have used lognormal distributions (Qin et al. 2005; Xiao et al. 2009, 2012, 2013, 2015; Fettweis et al. 2012; Wang et al. 2014; Gammon et al. 2017; Lu et al. 2018; Miwa et al. 2020). The curve-fitting method with lognormal distributions is based on the assumption that various grain-size distributions can be represented as a single lognormal distribution or a mixture of several: for example, a unimodal grain-size distribution with an asymmetrical or skewed shape can be regarded as a mixture of several lognormal distributions (e.g. Tanner 1964; Ashley 1978). Thus, a given polymodal grain-size distribution can be decomposed into several components consisting of normal distributions on a logarithmic scale. Methods based on parametric curve fitting, including the lognormal distribution function fitting method, have the advantage that they can be used for a single sample, whereas end-member modeling analysis requires multiple samples. The lognormal distribution function fitting method has been successfully used to obtain information on past environmental changes from both lacustrine and aeolian sediments (Qin et al. 2005; Xiao et al. 2009, 2012, 2013, 2015; Gammon et al. 2017; Lu et al. 2018).

The lognormal distribution function fitting method involves two processes: (1) estimation of the parameters of the appropriate lognormal distributions (i.e. the mean and standard deviation of each, and their mixing proportions) and (2) determination of the number of components. Commercial data analysis software (e.g. Igor Pro, PeakFitNagashima et al. 2004; Wang et al. 2014; Gammon et al. 2017; Miwa et al. 2020), original programs (e.g. Sun et al. 2002; Sasaki and Kiyono 2003; Xiao et al. 2012, 2013), and R platform packages (e.g. Buckland et al. 2021) have all been used to estimate the lognormal distribution parameters. In particular, several easy-to-use R packages that have appeared in recent years have made parameter estimation easier. By contrast, a method for determining the number of components has not yet been established so this process remains arbitrary. The number of components has been estimated from the number of peaks in the grain-size distribution (e.g. Xiao et al. 2012), or by using the mean square error (MSE) as an evaluation criterion (e.g. Xiao et al. 2013). When the MSE is used as an evaluation criterion, the model for the number of components with the smallest MSE is adopted, or that model is adopted having the smallest number of components with the MSE value below a certain threshold value. Using the number of peaks in the grain-size distribution may lead to underestimation of the number of components because it can fail to find a small component hidden by a larger adjacent component. In addition, it

has been generally pointed out that use of the MSE has the statistical problem that the number of components may be overestimated in evaluations based solely on the agreement of the curve fitting to the original data using the MSE value (Bishop 2006). Although several other evaluation criteria, including the Akaike information criterion (AIC; Akaike 1973) and the Bayesian information criterion (BIC; Schwarz 1978), have seen generally wide use for the estimation of the number of components, they have not yet been applied to the separation of grain-size distributions, nor has their validity been examined. In the present study, therefore, appropriate criteria for evaluating the number of components and their estimation characteristics when separating a grain-size distribution into lognormal distributions were investigated. Separation tests were carried out on artificial grain-size distributions for which the number of components and their parameters were known, and the estimation characteristics of each of four commonly used evaluation criteria were investigated to clarify the points that need to be considered when they are used. In addition, the four evaluation criteria were applied to an actual lake sediment sample, and the results were evaluated.

2 Method

2.1 Test procedure

The tests were carried out through the following procedure.

(i) Given an artificial grain-size data at phi scale ($=-\log_2(D/D_0)$; D is particle size in milimeters and D_0 is a reference diameter, equal to 1 mm) consisting of a mixture of normal distributions, for an assumed number of components n of 1–9, the parameters of each component (mean, standard deviation and mixing proportion) were estimated.

(ii) The values of the criteria to be validated were obtained for each of the cases in (i).

The artificial data used in this procedure, the method for estimating parameters by fitting, and the evaluation criteria considered are described in detail below.

2.2 Artificial datasets used for testing

A total of five artificial datasets were generated, each consisting of a different number (ranging from 2 to 6) of mixed normal distributions (components). Each dataset consisted of 1000 cases of mixed normal distributions of grain size at phi scale. For each case, the parameters of the normal distribution components used for the mixing were chosen as follows: The mean value of one component was fixed at 0 phi, and the distance between the mean values of adjacent components was set to a uniform random number ranging from 0.4 to 2 phi. The standard deviation of each mixed normal distribution was set to a uniform random number ranging from 0.2 to 0.8 phi. The mixing proportions were set by generating uniform random numbers that summed to 100% in increments of 1%. The minimum value of a mixing proportion was set at 5%.

2.3 Separation into normal distributions

To decompose the artificial grain-size distributions (i.e. mixed normal distributions at phi scale) and to estimate their parameters, the expectation–maximization (EM) algorithm (Dempster et al. 1977) was employed. The EM algorithm is a frequently used iterative method for finding (local) maximum likelihood estimates of unknown parameters in statistical models (McLachlan and Krishnan 2007). The EM iteration alternates between an expectation step and a maximization step: the expectation step creates a function for the expectation of the log-likelihood using the current estimates of the parameters, and the later maximization step calculates parameters that maximize the expected log-likelihood found in the previous expectation step. This algorithm yields the mixing proportions, means, and standard deviations of the normal distributions that compose a given grain-size distribution. The analyses were performed in R (version 4.1.2; R Core Team 2021), using the package 'mixR' (version 0.2.0; Yu 2022) to run the EM algorithm.

The quality of the fitting of the mixture of normal distributions to the artificial distributions was evaluated by calculating the figure-of-merit (FOM) value (Peng et al. 2022b) as follows:

$$FOM = \frac{\sum_{i=1}^{m} \left| y_i - \widehat{y_i} \right|}{\sum_{i=1}^{m} \widehat{y_i}} \times 100\%$$

where y_i is the volume percentage of the measured grain size (i.e. the given artificial grain size in the present study) in the i-th grain-size interval, $\widehat{y_i}$, is the fitted grain-size volume percentage in the i-th grain-size interval, and m is the number of grain-size intervals.

2.4 Criteria for determining the number of components

A total of four evaluation criteria were tested: the mean square error (MSE), which has been used in lake sediment studies (e.g. Xiao et al. 2013), and three other evaluation criteria commonly used for model selection in various fields: the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the adjusted R-squared (ARS). Each of these criteria were calculated as follows:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \widehat{y_i}\right)^2$$

$$AIC = -2\ln L + 2K$$

$$BIC = -2\ln L + K\ln m$$

$$ARS = 1 - \frac{\frac{\sum_{i=1}^{m}\left(y_i-\widehat{y_i}\right)^2}{m-3n-1}}{\frac{\sum_{i=1}^{m}\left(y_i-\bar{y}\right)^2}{m-1}}$$

where L is the maximized value of the likelihood function, K is the number of degrees of freedom, $\bar{y}$ is the mean of y_i, and n is the number of components. In the present study, the number of degrees of freedom was $3n-1$. Note that for the criteria MSE, AIC, and BIC, the preferred model is the one for which the value of the criterion is smaller, whereas for ARS, the preferred model is the one for which the value of the criterion is larger. MSE is simply the goodness of fit of the estimated mixture of fitting curves to the original data. Therefore, MSE tends to select excessively complex models, as described above (Bishop 2006). By contrast, the other three criteria are proposed to select the appropriate model by penalizing the increase in the number of components and the complexity of the model.

3 Results

The FOM for the optimal fitting curve obtained by the EM algorithm for each mixed normal distribution case was less than 2.9%. The parameters of the normal distribution components were not always estimated correctly even when the number of true components was given, and errors in parameter estimation were more frequent the higher the true number of mixture components in the artificial grain-size distribution (Fig. 1, Additional file 1: Fig. S1). Note that the accuracy of the parameter estimation depended on the degree of overlap of the adjacent components in the artificial data. Details on this point are described in Additional file 2.

The rate at which the number of components estimated as optimal matched the true number of components (N_t) (i.e. correct answer rate) differed among the criteria and depended on the value of N_t for the given grain-size distribution (Fig. 2). For all criteria, the correct answer rate tended to decrease as N_t increased (Fig. 2). When N_t was relatively small, the AIC and BIC tended to have higher correct answer rates than the other criteria, whereas when N_t was larger, the ARS and MSE tended to have higher correct answer rates than the AIC and BIC. When N_t was two, the BIC had the highest correct answer rate (89.6%), followed by the AIC (81.3%), ARS (69.6%), and MSE (62.9%). However, compared to the other criteria, the BIC showed a greater decrease in the correct answer rate as N_t increased, and its correct answer rate was lowest when N_t was larger than four (44.9% for $N_t=4$; 30.7% for $N_t=5$; 22.7% for $N_t=6$). By contrast, the MSE did not decrease as rapidly as N_t increased, and its correct answer rate was second highest among the criteria when N_t was 5 or more (49.0% for $N_t=5$; 41.4% for $N_t=6$). The correct answer rate of AIC was high for N_t counts up to 4 (75.1% for $N_t=3$; 59.6% for $N_t=4$), and it was comparable to those of the MSE for N_t of 5 and above (46.2% for $N_t=5$; 37.5% for $N_t=6$). The ARS showed a similar trend to the MSE, but its correct answer rate was higher than that of the MSE regardless

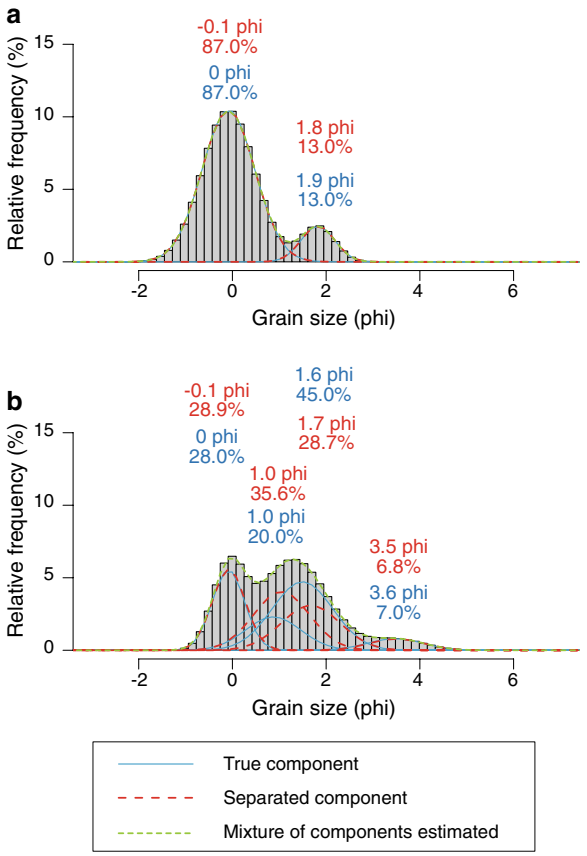

Fig. 1 Examples of the artificial grain-size distributions and separation results. Examples of the artificial grain-size distributions used in the present study and separation results when the true numbers of components were **a** two and **b** four. The blue solid lines show the true normal distribution components. The red and green dashed lines show the components separated by the EM algorithm and their mixing curves, respectively. The values shown above each component are the mean grain sizes and the mixing proportions of the true and separated components (blue and red, respectively)

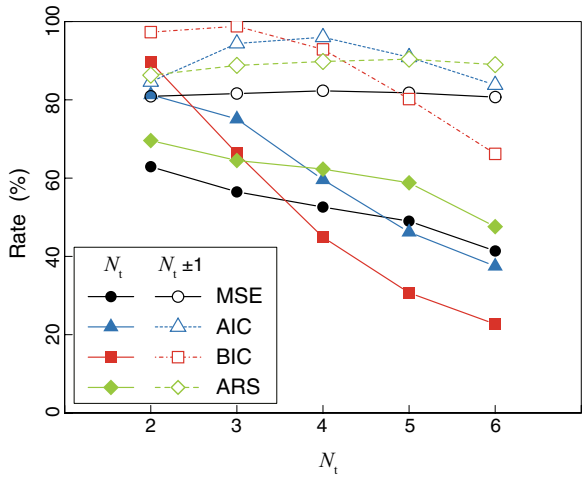

Fig. 2 Correct answer rates for each criterion relative to the true number of components. Rates at which the true number of components N_t (filled symbols) or the true number of components within $N_t \pm 1$ (white symbols) were estimated by each of the criteria in relation to N_t

of N_t, and its correct answer rate was highest for true component counts of 4 and above (62.3% for $N_t = 4$; 58.8% for $N_t = 5$; 47.6% for $N_t = 6$). The trend in the rate at which the difference between the estimated number of components was within $N_t \pm 1$ also differed among criteria and depended on N_t. For the BIC, that rate decreased markedly when N_t was 4 or above, whereas for the MSE and ARS, it remained approximately constant at around 81% and 88%, respectively, regardless of N_t. The rates for the AIC were roughly comparable to those for the ARS, but the AIC rate was highest when N_t was 4.

The difference between the number of components estimated as optimal (N_e) and N_t (N_e minus N_t) tended to differ among the criteria (Fig. 3). The MSE and ARS often overestimated the number of components regardless of N_t, whereas the AIC and BIC often underestimated the number of components, especially as N_t increased. The BIC more often estimated the number of components to be one fewer than the true number of components when N_t was 4 or more. When the AIC was used to estimate the number of components for data with low N_t, it rarely overestimated the number of components by more than five.

Among the nine models (i.e. with an assumed number of components of 1–9), the frequency with which the model with the true number of components was selected in the higher ranks of the criterion also tended to vary among the criteria (Fig. 4). In the present study, the percentage of cases with the true component number was highest for the BIC, exceeding 95%. Even when the true number of components was 6, the top four ranked models contained the true number of components in 95% of the cases (Fig. 4e).

4 Discussion

From the FOM values, it can be assumed that the parameter estimation in the present study resulted in good fitting. However, the parameter estimation of the normal distributions was not always correct, even when the true number of components was given (Fig. 1; Additional file 1: Fig. S1). In the case of estimation using actual natural grain-size distributions, the parameters of the true components are not known. For this reason, the discussion here is based on the values of the evaluation criteria obtained for each case, irrespective of the correctness of the parameter estimates. Note that the accuracy of parameter estimation and the correct answer rate for each criterion decreased with the degree of overlap of adjacent components, while it did not affect the characteristics of each criterion, discussed below. Caution should be paid with regard to the accuracy of the estimation of parameters and the number of components for cases with significant overlap of adjacent components, but this effect of the degree of overlap does not need to be taken into account in the comparison of the characteristics of each evaluation criterion in the following discussion. Details are described in Additional file 2.

The present results suggest that it is important to estimate the number of components by taking into account the estimation characteristics of each criterion: the AIC and BIC can be calculated with high accuracy when the actual number of components is small, but they tend to be slightly underestimated as the true number of components increases (Figs. 2 and 3). For this reason, caution should be exercised in their use when the grain size distribution is complex and may consist of more than five normal distributions. In contrast, the present results suggest that MSE and ARS are less sensitive to the value of the true number of components than AIC and BIC (Fig. 2); therefore, they may be effectively used for complex grain-size distributions. It should be noted, however, that the MSE and ARS may be overestimated.

The results of the present validation using artificial data suggested that the AIC or ARS would be preferable for estimating the number of components by using a single criterion for natural grain-size data where the true number of components is unknown. The present results indicate that the correct answer rate of the MSE, which has been used in previous studies of sediment grain size, is not higher than the rates of the other criteria; thus, it is not necessarily the most suitable criterion for estimating the correct number of components (Fig. 2). The correct answer rate of the ARS, which similarly tends to be less

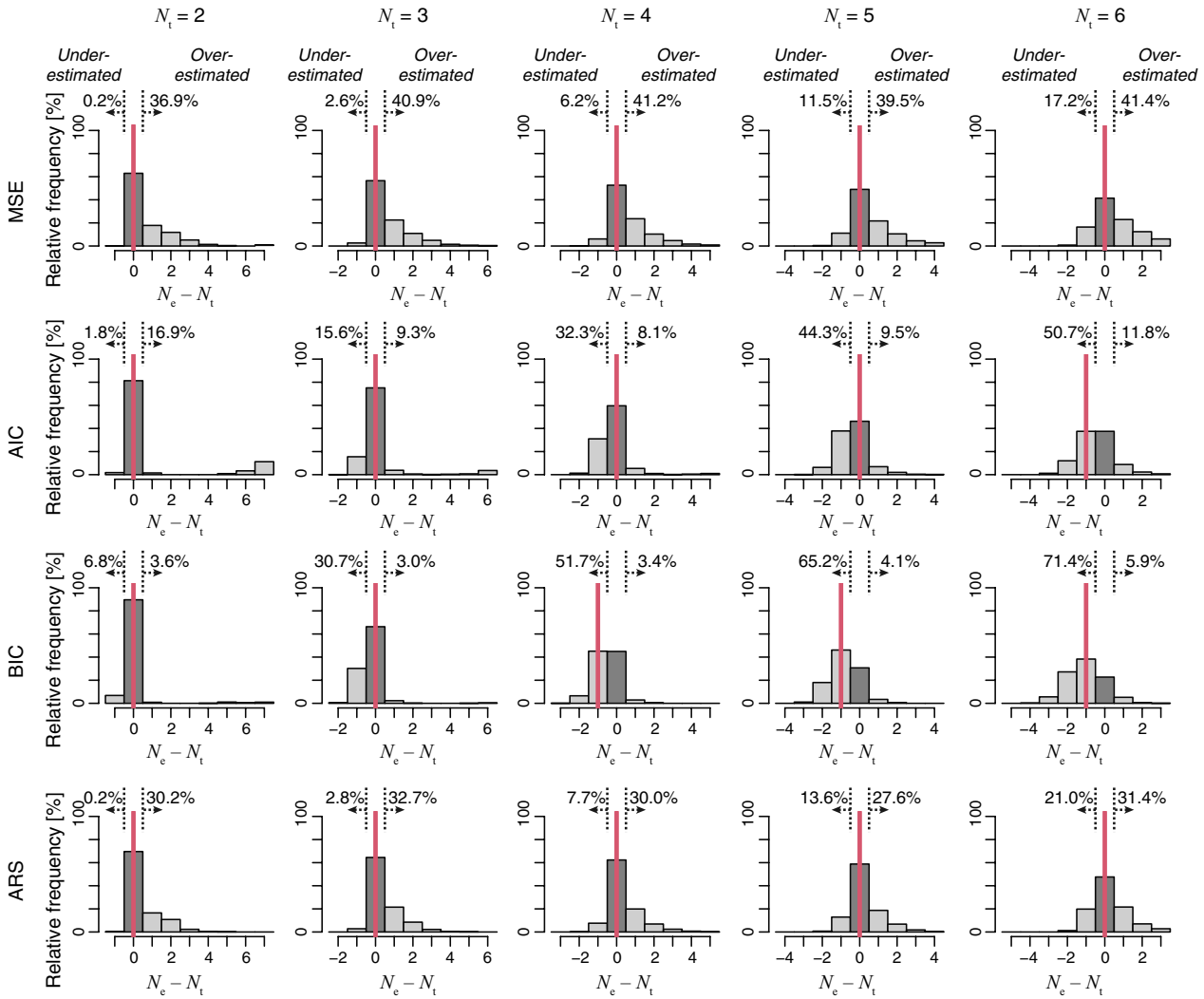

Fig. 3 Histograms of the difference between the estimated and the true number of components. Histograms of the difference between the number of components estimated as optimal (N_e) and the true number of components (N_t) (N_e–N_t) for each criterion and each true component number. Dark gray bars indicate cases when the true number of components was estimated ($N_e = N_t$). The red lines are the median values

dependent on the value of the true number of components, is higher in all component numbers from 2 to 6. Although the correct answer rate of the AIC is lower than that of the ARS when the true number of components is large, the AIC successfully estimates the true component number ± 1 at a higher rate overall; this finding suggests that use of the AIC is less likely to result in large estimation failures. Use of the AIC would be also preferred when analysing less complex grain-size distributions. Furthermore, more appropriate component number estimation might be achieved by taking into account the estimation characteristics of each criterion and using ones with complementary characteristics. For example, for the BIC, the top four models contained the correct number of components with a probability of more than 95% (Fig. 4). Considering this feature of the BIC, it should be possible to avoid the overestimation problem of, for example, the AIC or ARS, by checking whether the number of components they selected was among the higher ranked models of the BIC.

5 Application to natural grain-size data: an example

5.1 Sample and method of analysis

The four criteria tested in the present study were applied to the grain-size distribution of sediment samples from Lake Kitaura (Fig. 5). Lake Kitaura is a freshwater lake in central Japan, with an area of 35.2 km^2, a maximum depth

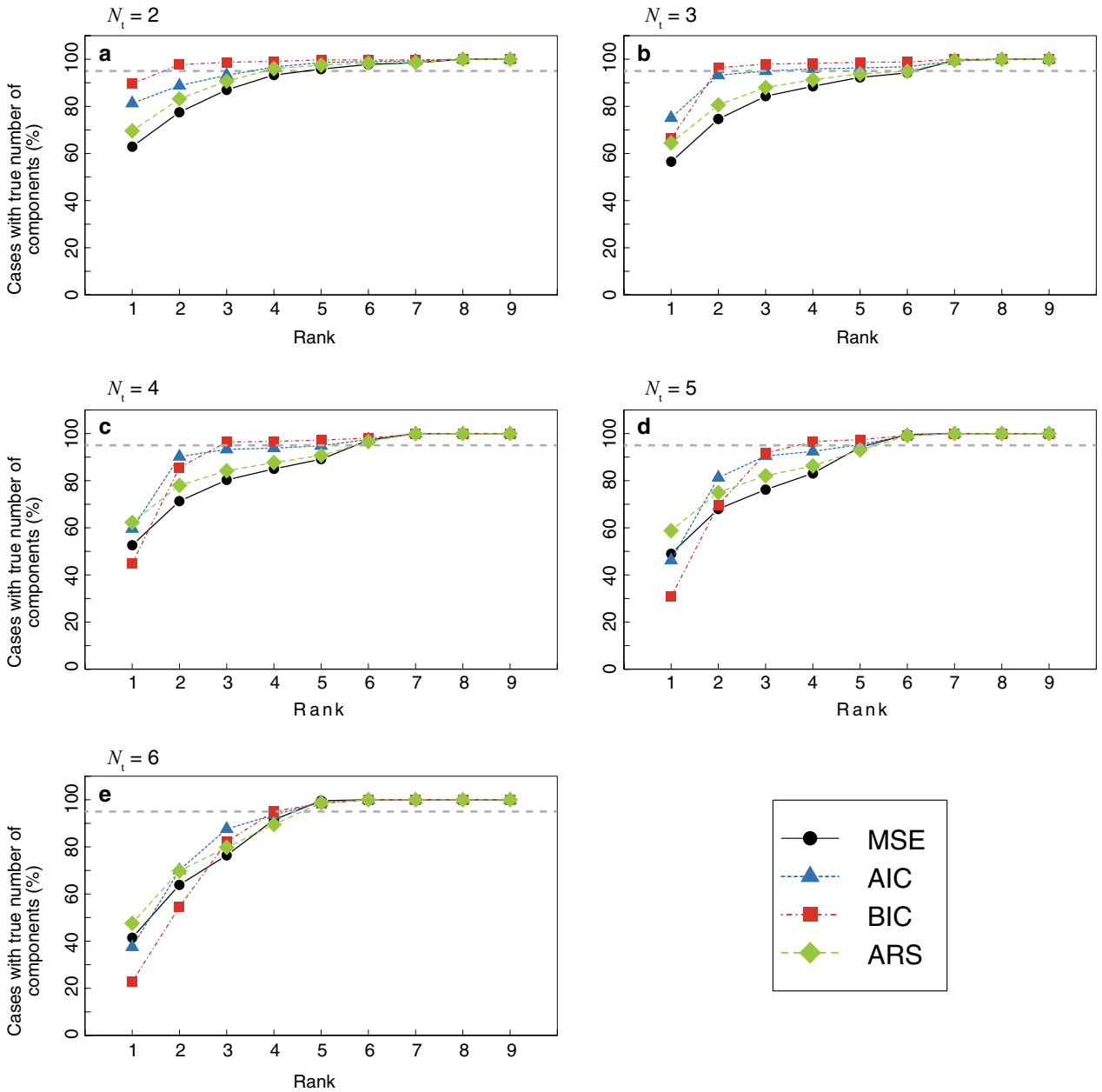

Fig. 4 Percentage of cases in which the true component number was included above the rank. Percentage of cases in which the true component number was included above the rank for datasets with **a**–**e** 2–6 true components. The horizontal dashed line indicates 95%

of 7 m (except where it has been artificially deepened by dredging), and a mean depth of 4 m (Fig. 5). Samples of bottom surface sediment were collected with an Ekman–Birge grab sampler (15 cm × 15 cm) from two sites close to the center of the lake, KT2104-06 and KT2104-07 (Fig. 5c), in April 2021, where the water depth was 6.5 and 5.9 m, respectively. Sites KT2104-06 and KT2104-07 were 916 and 424 m, respectively, from the nearest shore. Approximately 2 g of each sample was placed in a beaker and pre-treated with 50 mL of 10% H_2O_2 to remove organic matter. Then, each sample was rinsed with purified water and dispersed with 30 mL of 5.5 g/L $(NaPO_3)_6$. The grain-size distributions were measured with a laser grain-size analyzer (SALD-2300, Shimadzu Corporation, Kyoto, Japan).

The resulting grain-size distributions of the two samples were polymodal and exhibited their highest values at around 6.0 phi (Fig. 5d). The two distributions shared

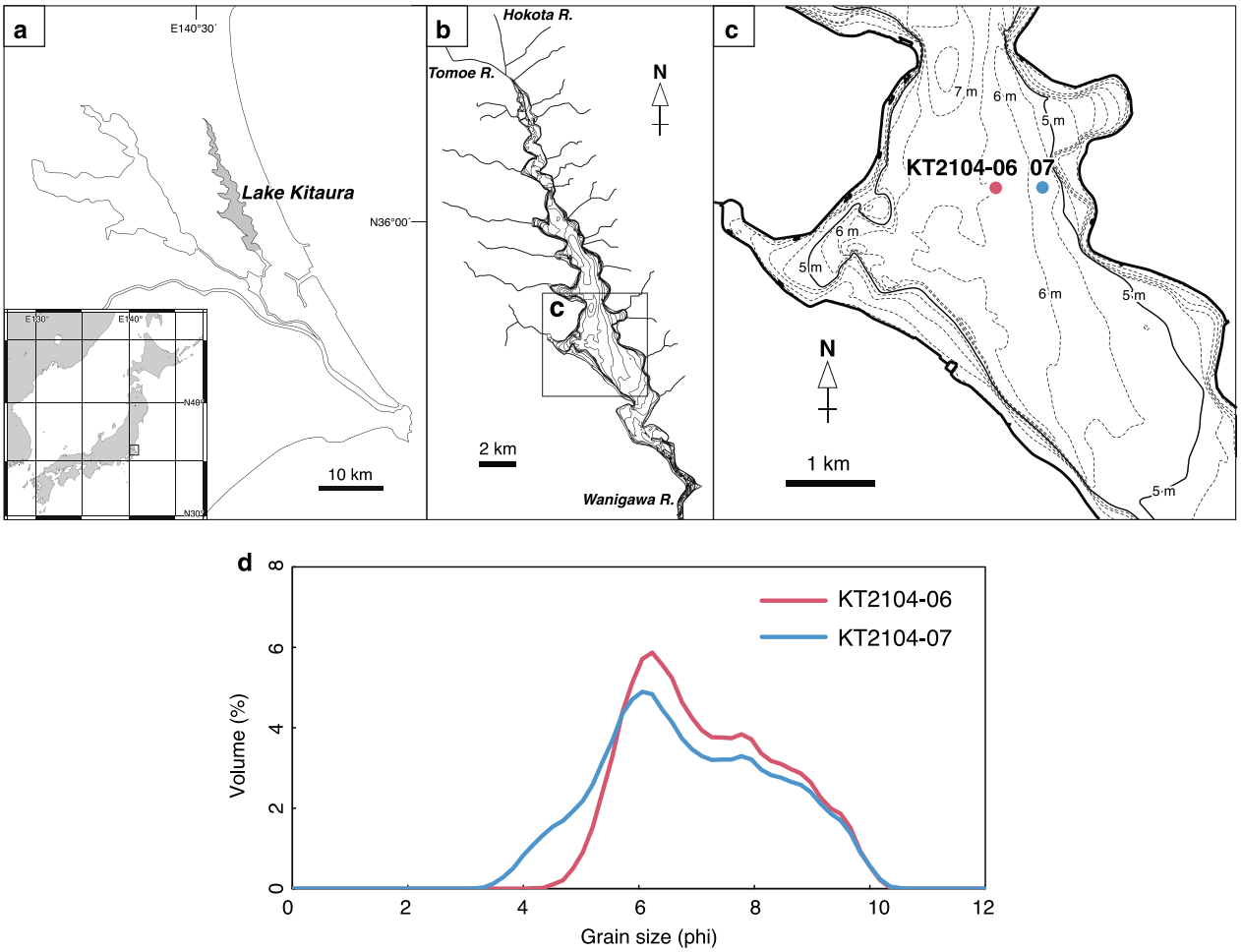

Fig. 5 Study site location and grain-size distributions of the two samples. **a** Location and **b** bathymetry of Lake Kitaura, and **c** locations of the sampling sites (KT2104-06 and -07) and **d** grain-size distributions of the sediment samples from those two sites. The depth isobath interval is 0.5 m. Bathymetric data are from the Geospatial Information Authority of Japan (2018)

some peaks and inflexion points; for example, both distributions had peaks at around 6.0 and 7.6 phi and an inflexion point at around 8.6 phi (Fig. 5d). The sample from site KT2104-07, which was closer to the shore and at a shallower water depth, also had a slope inflexion point at around 4.3 phi, whereas the sample from site KT2104-06 lacked this feature.

The grain-size data were separated into normal distributions at phi scale, and the values of the four criteria were obtained in the same manner as in the validation on the artificial data described above (Sects. 2.1, 2.3 and 2.4).

5.2 Estimating the number of components using the four evaluation criteria

For the sample from site KT2104-06, which was further away from the shore, different numbers of components were estimated by each of the four criteria (Fig. 6a–d): in the MSE-, AIC-, BIC-, and ARS-based estimations, the optimal number of components was 9, 4, 3, and 6 respectively. Both the AIC-preferred model of 4 components and ARS-preferred model of 6 components were among the top four BIC models (Fig. 6c). In contrast, the MSE-preferred model of 9 components was not included among the higher ranked BIC models; this result suggests that the MSE-preferred model may be an overestimation. Different numbers of components were also estimated by the four criteria for the sample from site KT2104-07, which was closer to shore (Fig. 6f–i): in the MSE-, AIC-, BIC-, and ARS-based estimations, 8, 5, 4, and 8 components, respectively, were found to be optimal. The MSE-preferred and ARS-preferred model of 8 components was not included among the higher ranked BIC models (Fig. 6h); this result suggests overestimation. When the grain-size distributions of the samples from sites KT2104-06 and -07 were separated into normal distributions at phi scale with 4 and 5 components, respectively,

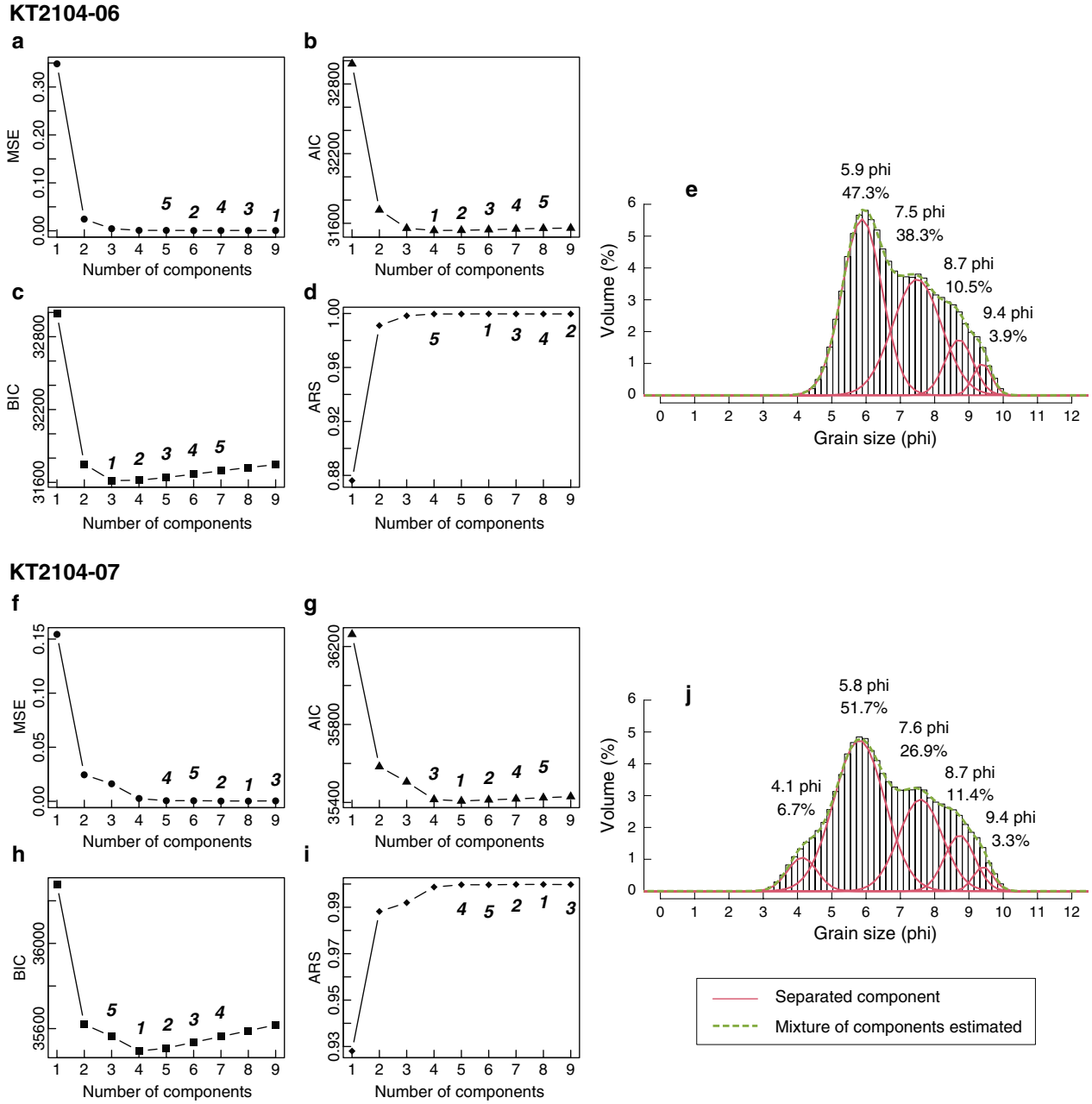

Fig. 6 Values of each criterion for the two lake sediment samples and examples of component separation. Values of **a** MSE, **b** AIC, **c** BIC, and **d** ARS for 1–9 assumed components, and **e** component separation of the sample from KT2104-06. Values of **f** MSE, **g** AIC, **h** BIC, and **i** ARS for 1–9 assumed components, and **j** component separation of the sample from KT2104-07. The boldface numbers above component number show the ranking of those models

based on the AIC criterion results given above, normal distribution components with shared mean values at 5.8–5.9, 7.5–7.6, 8.7, and 9.4 phi were estimated (Fig. 6e and j). Of these shared components, those at 5.8–5.9, 7.5–7.6, and 8.7 phi were generally consistent with the peaks and slope inflexion points visually observed in the grain-size distributions (Fig. 5d). A component with a mean value of 4.1 phi, which was estimated only for the sample from site KT2104-07, corresponds to the slope inflexion point at around 4.3 phi visually observed in the grain size distribution (Fig. 6j). Although one advantage of the method of separation into lognormal distributions for finding the number of component is that it can be applied to a single sample (Peng et al. 2022b), as in the present example,

the number of components can be estimated with greater validity and geological meaning by analyzing samples from the same vicinity and comparing the results.

6 Conclusions

To investigate appropriate estimation criteria for the number of components when separating a grain-size distribution into lognormal distributions, four estimation criteria were evaluated by using artificial grain-size data where the number of components was known. The criteria mean square error (MSE), Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R-squared (ARS) were evaluated, and the estimation characteristics of each were investigated. The results showed that estimation characteristics of the number of components differed among the criteria: the MSE and ARS were less affected by the true number of components and tended to overestimate the number of components. In comparison, the AIC and BIC showed a tendency to underestimate the number of components, with the correct answer rate decreasing as the true number of components increased. The BIC tended to include the true number of components among its higher ranked models. These evaluation results suggest that the frequently used criterion MSE is not necessarily the most appropriate evaluation criterion and that the AIC and ARS may be more appropriate. Furthermore, by checking whether the number of components estimated by the AIC or ARS is included among the higher ranked BIC models, the possibility of overestimation can be avoided, and the estimation of the number of components may be more valid. As an example, the four evaluation criteria were applied to the estimation of the number of components in grain-size distributions of bottom surface sediments from Lake Kitaura. The number of components obtained with the AIC appeared to reflect observed differences in the characteristics of the grain-size distributions.

Abbreviations

AIC	Akaike information criterion
ARS	Adjusted R-squared
BIC	Bayesian information criterion
EM algorithm	Expectation–maximization algorithm
FOM	Figure-of-merit
MSE	Mean square error

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40645-023-00601-y.

Additional file 1: **Fig. S1**. Relationship between the mean, mixing proportion and standard deviation of the true normal distribution component s obtained during artificial dataset prepara tion versus those of the estimated distribution components.

Additional file 2. Detailed descriptions of the effects of the degree of overlap of adjacent components in the artificial grain size data in the present test on the estimation of the distribution parameters and the number of components.

Acknowledgements
I thank two anonymous reviewers for their helpful comments, which improved this paper.

Author contributions
NY designed the study, analyzed the data, and wrote the manuscript. The author read and approved the final manuscript.

Funding
This work was partly supported by a Grant-in-Aid for Scientific Research (C) (No. 21K03677) from the Japan Society for the Promotion of Science to NY.

Availability of data and material
The data analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The author declares that he has no competing interest.

Received: 29 May 2023 Accepted: 9 December 2023
Published online: 13 December 2023

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Caski F (eds) Proceedings of the 2nd international symposium on information theory. Akadémiai Kiadó, Budapest, pp 267–281

Antoine P, Rousseau DD, Fuchs M, Hatté C, Gauthier C, Marković SB, Jovanović M, Gaudenyi T, Moine O, Rossignol J (2009) High-resolution record of the last climatic cycle in the southern Carpathian Basin (Surduk, Vojvodina, Serbia). Quat Int 198:19–36

Ashley GM (1978) Interpretation of polymodal sediments. J Geol 86:411–421

Bishop CM (2006) Pattern recognition and machine learning. Springer-Verlag, New York

Buckland HM, Saxby J, Roche M, Meredith P, Rust AC, Cashman KV, Engwell SL (2021) Measuring the size of non-spherical particles and the implications for grain size analysis in volcanology. J Volcanol Geoth Res 415:107257. https://doi.org/10.1016/j.jvolgeores.2021.107257

Burden RE, Phillips JC, Hincks TK (2011) Estimating volcanic plume heights from depositional clast size. J Geophys Res Solid Earth 116:B11206. https://doi.org/10.1029/2011JB008548

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B (methodol) 39:1–22

Dietze E, Hartmann K, Diekmann B, IJmker J, Lehmkuhl F, Opitz S, Stauch G, Wünnemann B, Borchers A (2012) An end-member algorithm for deciphering modern detrital processes from lake sediments of Lake Donggi Cona, NE Tibetan Plateau, China. Sediment Geol 243:169–180

Dietze E, Wünnemann B, Hartmann K, Diekmann B, Jin H, Stauch G, Yang S, Lehmkuhl F (2013) Early to mid-Holocene lake high-stand sediments at Lake Donggi Cona, northeastern Tibetan Plateau, China. Quat Res 79:325–336

Dietze E, Maussion F, Ahlborn M, Diekmann B, Hartmann K, Henkel K, Kasper T, Lockot G, Opitz S, Haberzettl T (2014) Sediment transport processes across the Tibetan Plateau inferred from robust grain-size end members in lake sediments. Clim past 10:91–106

Engwell S, Eychenne J (2016) Contribution of fine ash to the atmosphere from plumes associated with pyroclastic density currents. In: Mackie S, Cashman

K, Ricketts H, Rust A, Watson M (eds) Volcanic ash. Elsevier, Amsterdam, pp 67–85
Fettweis M, Baeye M, Lee BJ, Chen P, Yu JC (2012) Hydro-meteorological influences and multimodal suspended particle size distributions in the Belgian nearshore area (southern North Sea). Geo Mar Lett 32:123–137
Folk RL (1966) A review of grain-size parameters. Sedimentology 6:344–359
Gammon PR, Neville LA, Patterson RT, Savard MM, Swindles GT (2017) A log-normal spectral analysis of inorganic grain-size distributions from a Canadian boreal lake core: towards refining depositional process proxy data from high latitude lakes. Sedimentology 64:609–630
Gao S, Collins M (1992) Net sediment transport patterns inferred from grain-size trends, based upon definition of "transport vectors." Sediment Geol 81:47–60
Geospatial Information Authority of Japan (2018) Lake data of Kitaura and Sotonasakaura. https://www1.gsi.go.jp/geowww/lake/download/kitaura-sotonasakaura/kitaura-sotonasakaura-2018. Accessed 17 Feb 2022
Håkanson L, Jansson M (1983) Principles of lake sedimentology. Springer, Berlin
Ijmker J, Stauch G, Dietze E, Hartmann K, Diekmann B, Lockot G, Opitz S, Wünnemann B, Lehmkuhl F (2012) Characterisation of transport processes and sedimentary deposits by statistical end-member mixing analysis of terrestrial sediments in the Donggi Cona lake catchment, NE Tibetan Plateau. Sediment Geol 281:166–179
Le Roux JP, Rojas EM (2007) Sediment transport patterns determined from grain size parameters: overview and state of the art. Sediment Geol 202:473–488
Lim J, Matsumoto E (2006) Bimodal grain-size distribution of aeolian quartz in a maar of Cheju Island, Korea, during the last 6500 years: its flux variation and controlling factor. Geophys Res Lett 33:L21816. https://doi.org/10.1029/2006GL027432
Lin Y, Mu G, Xu L, Zhao X (2016) The origin of bimodal grain-size distribution for aeolian deposits. Aeolian Res 20:80–88
Lu Y, Fang X, Friedrich O, Song C (2018) Characteristic grain-size component-A useful process-related parameter for grain-size analysis of lacustrine clastics? Quatern Int 479:90–99
Machalett B, Oches EA, Frechen M, Zöller L, Hambach U, Mavlyanova NG, Marković SB, Endlicher W (2008) Aeolian dust dynamics in central Asia during the Pleistocene: driven by the long-term migration, seasonality, and permanency of the Asiatic polar front. Geochem Geophys Geosyst. https://doi.org/10.1029/2007GC001938
McCave IN (1978) Grain-size trends and transport along beaches: example from eastern England. Mar Geol 28:M43–M51
McLachlan GJ, Krishnan T (2007) The EM algorithm and extensions, 2nd edn. John Wiley & Sons, Hoboken
McLaren P (1981) An interpretation of trends in grain size measures. J Sediment Petrol 51:611–624
McLaren P, Bowles D (1985) The effects of sediment transport on grain-size distributions. J Sediment Res 55:457–470
Middleton GV (1976) Hydraulic interpretation of sand size distributions. J Geol 84:405–426
Miwa T, Iriyama Y, Nagai M, Nanayama F (2020) Sedimentation process of ashfall during a Vulcanian eruption as revealed by high-temporal-resolution grain size analysis and high-speed camera imaging. Prog Earth Planet Sci 7:3. https://doi.org/10.1186/s40645-019-0316-8
Nagashima K, Tada R, Matsui H (2004) Intensity variation in the Asian monsoon and the Westerly during the last 140kyr deduced from grain size analysis of Japan Sea sediments. Quat Res (daiyonki-Kenkyu) 43:85–97 (**in Japanese with English abstract**)
Park CS, Hwang S, Yoon SO, Choi J (2014) Grain size partitioning in loess–paleosol sequence on the west coast of South Korea using the Weibull function. CATENA 121:307–320
Parris AS, Bierman PR, Noren AJ, Prins MA, Lini A (2010) Holocene paleostorms identified by particle size signatures in lake sediments from the northeastern United States. J Paleolimnol 43:29–49
Peng J, Wang X, Yin G, Adamiec G, Du J, Zhao H, Kang S, Zheng Y (2022a) Accumulation of aeolian sediments around the Tengger Desert during the late Quaternary and its implications on interpreting chronostratigraphic records from drylands in north China. Quat Sci Rev 275:107288. https://doi.org/10.1016/j.quascirev.2021.107288
Peng J, Zhao H, Dong Z, Zhang Z, Yang H, Wang X (2022b) Numerical methodologies and tools for efficient and flexible unmixing of single-sample grain-size distributions: application to late Quaternary aeolian sediments from the desert-loess transition zone of the Tengger Desert. Sediment Geol 438:106211. https://doi.org/10.1016/j.sedgeo.2022.106211
Qin X, Cai B, Liu T (2005) Loess record of the aerodynamic environment in the east Asia monsoon area since 60,000 years before present. J Geophys Res 110:B01204. https://doi.org/10.1029/2004JB003131
R Core Team (2021) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
Rose WI, Durant AJ (2009) Fine ash content of explosive eruptions. J Volcanol Geoth Res 186:32–39
Rossi E, Bonadonna C, Degruyter W (2019) A new strategy for the estimation of plume height from clast dispersal in various atmospheric and eruptive conditions. Earth Planet Sci Lett 505:1–12
Sasaki T, Kiyono Y (2003) Development of a GUI program for the analysis of grain size distributions-–an automation for the procedure of Inokuchi and Mezaki (1974). J Sedimentol Soc Jpn 57:35–41
Schulte P, Sprafke T, Rodrigues L, Fitzsimmons KE (2018) Are fixed grain size ratios useful proxies for loess sedimentation dynamics? Experiences from Remizovka, Kazakhstan. Aeolian Res 31:131–140
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Sun D (2004) Monsoon and westerly circulation changes recorded in the late Cenozoic aeolian sequences of Northern China. Glob Planet Change 41:63–80
Sun D, Bloemendal J, Rea DK, Vandenberghe J, Jiang F, An Z, Su R (2002) Grain-size distribution function of polymodal sediments in hydraulic and aeolian environments, and numerical partitioning of the sedimentary components. Sediment Geol 152:263–277
Tanner WF (1964) Modification of sediment size distributions. J Sediment Petrol 34:156–164
Vandenberghe J (2013) Grain size of fine-grained windblown sediment: a powerful proxy for process identification. Earth Sci Rev 121:18–30
Vandenberghe J, Zhisheng A, Nugteren G, Huayu L, Van Huissteden K (1997) New absolute time scale for the Quaternary climate in the Chinese loess region by grain-size analysis. Geology 25:35–38
Visher GS (1969) Grain size distributions and depositional processes. J Sediment Res 39:1074–1106
Wang K, Zheng H, Tada R, Irino T, Zheng Y, Saito K, Karasuda A (2014) Millennial-scale East Asian Summer Monsoon variability recorded in grain size and provenance of mud belt sediments on the inner shelf of the East China Sea during mid-to late Holocene. Quat Int 349:79–89
Weltje GJ, Prins MA (2007) Genetically meaningful decomposition of grain-size distributions. Sediment Geol 202:409–424
Wu L, Krijgsman W, Liu J, Li C, Wang R, Xiao W (2020) CFLab: a MATLAB GUI program for decomposing sediment grain size distribution using Weibull functions. Sediment Geol 398:105590. https://doi.org/10.1016/j.sedgeo.2020.105590
Xiao J, Chang Z, Si B, Qin X, Itoh S, Lomtatidze Z (2009) Partitioning of the grain-size components of Dali Lake core sediments: evidence for lake-level changes during the Holocene. J Paleolimnol 42:249–260
Xiao J, Chang Z, Fan J, Zhou L, Zhai D, Wen R, Qin X (2012) The link between grain-size components and depositional processes in a modern clastic lake. Sedimentology 59:1050–1062
Xiao J, Fan J, Zhou L, Zhai D, Wen R, Qin X (2013) A model for linking grain-size component to lake level status of a modern clastic lake. J Asian Earth Sci 69:149–158
Xiao J, Fan J, Zhai D, Wen R, Qin X (2015) Testing the model for linking grain-size component to lake level status of modern clastic lakes. Quat Int 355:34–43
Yamashita S, Naruse H, Nakajo T (2018) Reconstruction of sediment-transport pathways on a modern microtidal coast by a new grain-size trend analysis method. Prog Earth Planet Sci 5:7. https://doi.org/10.1186/s40645-018-0166-9
Yu Y (2022) mixR: an R package for finite mixture modeling for both raw and binned data. J Open Source Softw 7:4031. https://doi.org/10.21105/joss.04031
Yu S-Y, Colman SM, Li L (2016) BEMMA: a hierarchical Bayesian end-member modeling analysis of sediment grain-size distributions. Mathematical Geosci 48:723–741

Publisher's Note