Stat Papers
DOI 10.1007/s00362-016-0774-7

REGULAR ARTICLE

k-Boxplots for mixture data

Najla M. Qarmalah[1] · **Jochen Einbeck**[1] · **Frank P. A. Coolen**[1]

Received: 24 April 2015 / Revised: 18 April 2016

Abstract This article introduces a new graphical tool to summarize data which possess a mixture structure. Computation of the required summary statistics makes use of posterior probabilities of class membership which can be obtained from a fitted mixture model. Real and simulated data are used to highlight the usefulness of this tool for the visualization of mixture data in comparison to the traditional boxplot.

Keywords Mixture models · Boxplot · Posterior probability · EM-algorithm

1 Introduction

Visualization tools play an essential role for analysing, investigating, understanding, and communicating data, and the development of novel graphical tools continues to be a topic of interest in the statistical literature. For example, Wang and Bellhouse (2014) have recently introduced a new graphical approach called the shift function plot to evaluate the goodness-of-fit of a parametric regression model. A boxplot is one of the most popular graphical techniques used. It was proposed as a unimodal data display by Tukey (1977) who called it a "schematic plot" or "box-and-whisker plot" but it is now customarily called boxplot. A boxplot, in its simplest form, aims at summarizing a univariate data set by displaying five main statistical features which are the median, first quartile, third quartile, minimum value and maximum value.

The boxplot has become one of the most frequently used graphical techniques for analysing data because it gives information about the location, spread, skewness, and longtailedness of a data set at a quick glance. The median in a boxplot serves as a

✉ Jochen Einbeck
jochen.einbeck@durham.ac.uk

[1] Durham University, Durham, UK

Published online: 09 May 2016

measure of location. The dispersion of a data set could be assessed by observing the length of a box or the distance between the ends of the whiskers. The skewness can be observed by the deviation of the median line from the center of the box or by the length of the upper whisker relative to the length of the lower one. In addition, the distance between the ends of the whiskers compared to the length of the box displays longtailedness (Benjamini 1988). Alternative specifications of the ends of the whiskers are being used with a particular view to outlier detection. Specifically, boundaries $Q_1 - 1.5IQR$ and $Q_3 + 1.5IQR$ can be computed where Q_1, Q_3 and IQR are the first quartile, third quartile and interquartile range, respectively. Then, any observations smaller than $Q_1 - 1.5IQR$ or greater than $Q_3 + 1.5IQR$ are labelled as "outliers" (for more details, see e.g. Frigge et al. 1989). Finally, whiskers are drawn from the box to the furthest non outlying observations. Additionally, notches can be added which approximate a 95 % confidence interval for the median (Krzywinski and Altman 2014).

Further variants of the boxplot have been developed to analyse special kinds of data. For example, Abuzaid et al. (2012) have proposed a boxplot for circular data which is called a circular boxplot. Hubert and Vandervieren (2008) have presented an adjustment of the boxplot to tackle the outliers in skewed data by modifying the whiskers. Recently, Bruffaerts et al. (2014) have developed a "generalized" boxplot which is more appropriate for skewed distributions and distributions with heavy tails.

It was already observed by McGill et al. (1978) that the traditional boxplot is not able to adequately display data which is divided into certain groups or classes. Therefore, they developed a version of the boxplot for grouped data which sets the widths of each group—wise boxplot proportional to the square root of the group sizes. However, this technique requires that the groups are defined a priori, and that the group membership of each observation is known. In practice, one will often deal with data sampled from heterogeneous subpopulations for which the group membership is a latent variable, and, hence, unknown. To our knowledge, there does not exist an appropriate plot which represents such mixture data properly. Consequently, we have developed a new plot tailored to mixture data to which we refer as a k-boxplot, where k is the number of mixture components. Compared to a boxplot, the k-boxplot is able to display important additional information regarding the structure of the data set. Both k-boxplots and boxplots have a similar construction: they contain boxes and display extreme values. However, the k-boxplot visualizes the k components of mixture models by k different boxes, compared to a boxplot which has only one box. Then, a boxplot is a special case of a k-boxplot with $k = 1$.

Figure 1 provides a schematic display of (what we will refer to as a 'full') k-boxplot in the special case $k = 3$, which describes the main features of k-boxplots in general. The k-boxplot displays k rectangles oriented with the axes of a coordinate system in which one of the axes has the scale of a data set. The key features which appear in a k-boxplot are the weighted median, the first weighted quartile and the third weighted quartile in each box, which are found as the respective weighted quantiles using the posterior probabilities of group membership as weights (as will be explained in more detail later). Bottom and top of the boxes are drawn at the weighted first and third quartiles of the data in each group respectively. Weighted medians are displayed as horizontal lines drawn inside the boxes. Additional information is pro-

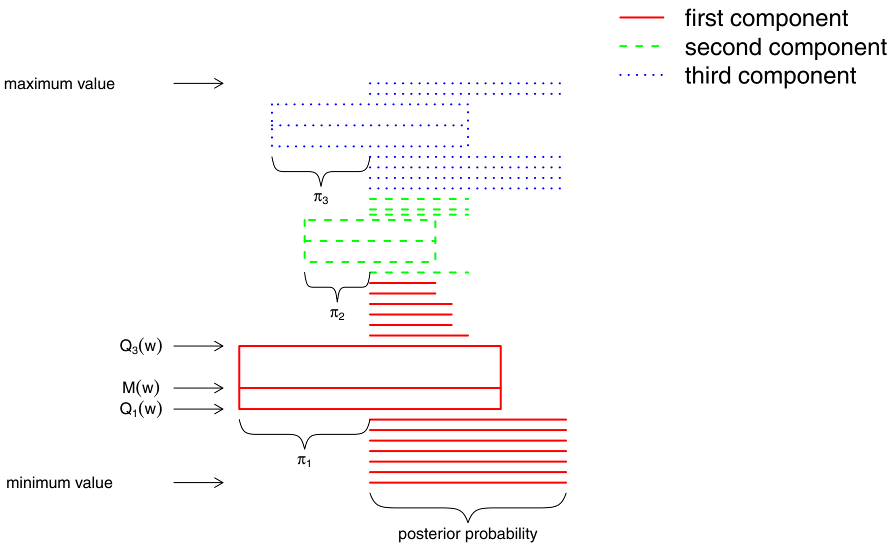

Fig. 1 Summary of information provided by a 3-boxplot in its 'full' form. Here, $M(w)$ denotes the weighted median, and $Q_j(w)$ the j–th weighted quartile, using the notation formally introduced in Sect. 2.2

vided through the widths of the boxes, which depend on the mixing proportions of the mixture.

Just as for usual boxplots, data points falling out of the boxes can be displayed in several ways. Here, any points fully outside the boxes are displayed individually through horizontal lines and can so be used to identify outliers. The length of these lines corresponds to posterior probabilities of group membership which will be explained in more detail by real data examples later. Some variants of the *k*-boxplot which display points outside the boxes in different ways will be implicitly introduced in Sect. 3.1.

By using *k*-boxplots for mixture data, the location, spread and skewness for each component in a mixture will be displayed transparently to the viewers. Each of the component—wise boxplots can be interpreted in the same way as traditional boxplots with respect to these measures, allowing for a detailed appraisal of the data. The required information in order to draw a *k*-boxplot can be estimated by different methods, for example by the EM-algorithm. It is emphasized that we do not consider *k*-boxplots to be an *inferential* tool. That is, *k*-boxplots will not make any automated decision on the choice of the mixture distributions or the number of components but they visualize the result of these inferential decisions made by the data analyst. Since the data analyst will be able to identify the impact of their model choices at a glance, *k*-boxplots will support them in making such choices in an informed manner.

The structure of the remainder of the paper is as follows. In Sect. 2 we describe the computational elements of a *k*-boxplot, which are the posterior probabilities derived from mixture models, as well as weighted quartiles. In Sect. 3, we discuss three real data examples and present the results of a small simulation. Finally, we provide conclusions in Sect. 4. Code to execute *k*-boxplots is provided in the statistical programming language R (R Core Team 2015) in form of function `kboxplot` in package **UEM**.

2 Computational elements of *k*-boxplots

2.1 Posterior probabilities

Mixture models play a vital role in the statistical analysis of data thanks to their flexibility to model a wide variety of random phenomena. They have been successfully employed for a wide range of applications in the biological, physical, and social sciences, including astronomy, medicine, psychiatry, genetics, economics, engineering, and marketing. In addition, mixture models have direct relevance for cluster and latent class analyses, discriminant analysis, image analysis and survival analysis (McLachlan and Peel 2004).

Assume a random variable Y with density $f(y)$ is described as a finite mixture of k probability density functions $f_j(y)$, $j = 1, \dots, k$, such that

$$f(y) = \sum_{j=1}^{k} \pi_j f_j(y) \tag{1}$$

with masses (or mixing proportions) $\pi_1, \dots, \pi_k$ with $0 < \pi_j < 1$ and $\sum_{j=1}^{k} \pi_j = 1$. We refer to $f_j(\cdot)$, which may depend on a parameter vector θ_j, as the j–th component of the mixture of probability density functions. Just to clarify terms, when speaking of 'mixture data' in this manuscript, we mean data y_i, $i = 1, \dots, n$ for which it is plausible to assume that they have been independently generated from, or at least can be represented by, a model of type (1).

Interpreting the π_j as 'prior' probability of class membership, then posterior probabilities of class membership are produced via Bayes theorem, that is, for the i–th observation y_i, $i = 1, \dots, n$ one has

$$r_{ij} = P(\text{observation } i \text{ belongs to comp. } j) = \frac{\pi_j f_j(y_i)}{\sum_{\ell=1}^{k} \pi_\ell f_\ell(y_i)}. \tag{2}$$

These posterior probabilities, which we combine into a weight matrix $R = (r_{ij})_{1 \le i \le n, 1 \le j \le k}$, form the key ingredient of k-boxplots. They will be used to compute the component—wise medians and quartiles, and furthermore they enable immediately computation of the estimate

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} r_{ij}, \tag{3}$$

which will be used to determine the width of the j–th k-boxplot. Note also, by assigning each data point y_i to the component j which maximizes r_{ij} for fixed i, posterior probabilities can be used as a classification tool. This is known as the maximum a posteriori (MAP) rule.

The estimates of θ_j are *not* needed for the construction of the k-boxplot in itself. However, computation of (2) involves the densities f_j and hence θ_j. So, the θ_j need to be computed along the way as well. Most commonly, mixture models will be estimated through the EM algorithm. In this case, the values θ_j will get updated in the M-step, and (2) corresponds exactly to the E-step, using the current estimates of π_j and θ_j. That is, in practice, the r_{ij} can be conveniently extracted from the output of the last EM iteration.

The application of k-boxplots is not restricted to a certain choice of component densities. In principle, k-boxplots can be used to visualize the results of fitting a mixture of any (combination of) densities f_j, provided that one is able to compute the parameters θ_j in the M-step. The choice of the f_j is down to the data analyst. In the absence of strong motives to use a different distribution, the normal distribution will often be a convenient choice for the component densities. In this case,

$$f_j(y) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(y-\mu_j)^2}{2\sigma_j^2}\right)$$

where μ_j are the component means and σ_j the component standard deviations. Maximizing the complete log-likelihood in the M-step then gives the estimates

$$\begin{aligned} \hat{\mu}_j &= \frac{\sum_{i=1}^n r_{ij} y_i}{\sum_{i=1}^n r_{ij}}; \\ \hat{\sigma}_j^2 &= \frac{\sum_{i=1}^n r_{ij}(y_i - \hat{\mu}_j)^2}{\sum_{i=1}^n r_{ij}}. \end{aligned} \tag{4}$$

The EM-algorithm consists of iterating Eqs. (2) and (4) until convergence (Dempster et al. 1977). Initial values $\theta_j^{(0)}, \pi_j^{(0)}, j = 1, \ldots, k$, are required for the first E-step.

It is well known that different starting points can lead to different solutions, corresponding to different local maxima of the log-likelihood. See McLachlan and Peel (2004) for a detailed discussion of this problem. Possible strategies for choosing starting points include random initialization, quantile–based initialization, scaled Gaussian Quadrature points, or short EM runs (Biernacki et al. 2003). The matter continues to be the content of current discussion and research; with a recent contribution on the topic provided by Baudry and Celeux (2015).

2.2 Weighted quartiles

Suppose $y_1 \leq \ldots \leq y_n$ indicate the ordered observations and $w = \{w_1, \ldots, w_n\}$ are a set of corresponding non-negative weights. Then

$$m(w) = \max\left\{\ell : \sum_{i=\ell}^n w_i \geq \frac{1}{2}\sum_{i=1}^n w_i\right\},$$

gives the maximal index ℓ so that the total weight of observations larger or equal than y_ℓ is at least 50 %. Hence, the weighted median of $y_1, \ldots, y_n$ is defined to be

$$M(w) \equiv y_{m(w)}$$

Table 1 Illustration of computation of weighted quantiles.

ℓ	1	2	3	4	5
y_ℓ	1	3	4	7	9
w_ℓ	0.2	0.25	0.3	0.05	0.2
$\sum_{i=\ell}^{n} w_i$	1	0.80	0.55	0.25	0.2

(Fried et al. 2007). There is no unique definition for quartiles, but in analogy to the above, one can define the first weighted quartile of $y_1, \ldots, y_n$ as $Q_1(w) = y_{q_1(w)}$, where

$$q_1(w) = \max\left\{\ell : \sum_{i=\ell}^{n} w_i \geq \frac{3}{4}\sum_{i=1}^{n} w_i\right\},$$

and the third weighted quartile of $y_1, \ldots, y_n$ as $Q_3(w) = y_{q_3(w)}$, where

$$q_3(w) = \max\left\{\ell : \sum_{i=\ell}^{n} w_i \geq \frac{1}{4}\sum_{i=1}^{n} w_i\right\}.$$

For example, the weighted median of 1, 3, 4, 7 and 9 with weights 0.2, 0.25, 0.3, 0.05, and 0.2 is $y_3 = 4$, because $0.2 + 0.05 + 0.3 \geq 0.5$. In addition, the first and third weighted quartile of the data are $y_2 = 3$ and $y_4 = 7$ respectively because $0.2 + 0.05 + 0.3 + 0.25 \geq 0.75$ and $0.2 + 0.05$=0.25. An illustration of this process is provided in Table 1.

In the case of a k-boxplot, the box corresponding to the j–th component is fully determined by the observations y_i and the weights $w_i \equiv r_{ij}$, $i = 1, \ldots, n$. Note that these weights, for fixed j, generally do not sum to 1.

3 Examples

In this section three real data examples are presented to illustrate the usefulness of the k-boxplots for mixture data, especially compared to boxplots.

3.1 Example 1: energy use data

The data discussed in this example come from the International Energy Agency (IEA)[1]. They give the annual energy use (in kg oil equivalent per capita) for 134 countries around the world between 1971 and 2012. Due to the nature of the data, which are restricted to the positive range and feature several countries with extremely large energy use, a log-transformation will be applied in all further analyses.

[1] International Energy Agency, available at: http://www.iea.org/.

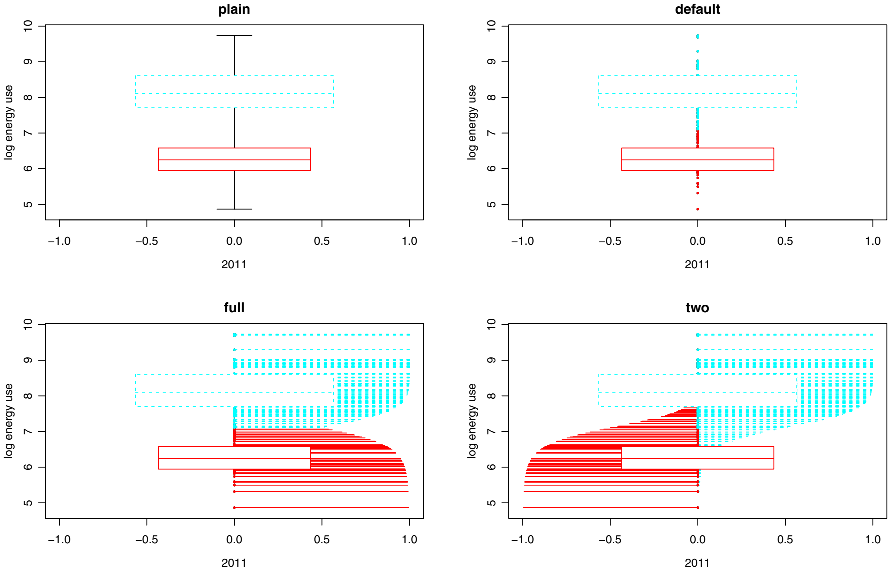

Fig. 2 Four variants of 2-Boxplots of log energy use data in 2011

We consider only the year 2011 initially, for which Fig. 2 presents four different types of 2-boxplots for the log energy use (The bimodal character of country-wise log-energy data has already been reported in Einbeck and Taylor 2013). The boxplots are labelled in the title area by the corresponding option which needs to be specified as `type` argument in R function `kboxplot`. All four versions carry the main feature of a 2-boxplot—the two boxes which indicate the location, spread and size of the two components. We see that lower box represents the group of low energy use countries and the upper box visualises high energy use countries. One can observe from these figures that the number of high energy use countries is higher than the number of low energy use countries according to the widths of the boxes which are determined by the fitted mixing proportions $\hat{\pi}_j$ (we use the convention that the π_j correspond exactly to the half-width). Further, one gets information on the spread and location of groups by observing the bottom, top and cut lines of the boxes which represent the weighted first and third quartiles and the weighted median respectively.

The four types of k-boxplots differ in how the individual observations are presented. The 'plain' version of the 2-boxplot in the top left corner is most closely resembling a traditional boxplot in its simplest form: there are two boxes representing the mixture components, with whiskers drawn up to the overall maximum and minimum. For k-boxplots, we do not consider it a sensible option to draw the whiskers up to a certain multiple of the interquartile range. The reason is that this range would have to be calculated with respect to the corresponding maximum or minimum box, which would be little informative especially if the range of this box is small.

The 'default' option (top right) provides slightly more information. Here data points falling outside the boxes are plotted explicitly, hence making this representation particularly suitable to identify outlying cases. Furthermore, the points are coloured

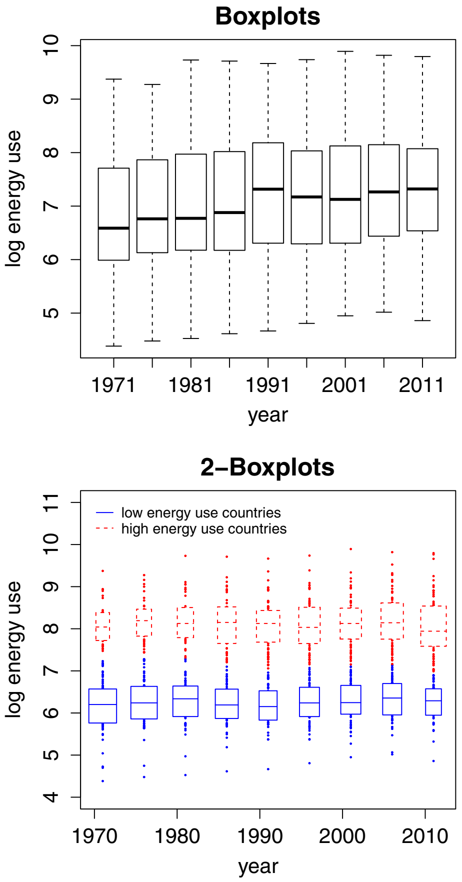

Fig. 3 Boxplots [*top*] and 2-Boxplots [*bottom*] of log energy use data between 1971 and 2012

according to the MAP classification rule; that is, for country i one identifies the component j for which the posterior probability r_{ij} is maximal and then colours the point in the same colour as the box for that component. The 'full' version in the bottom left—corresponding to the representation from Fig. 1—provides another layer of detail, by giving explicitly the posterior probabilities of belonging to 'their' component (to which they were assigned according to the MAP rule). The lines have maximum length 1 in which case a country is classified with 100 % posterior probability to one of the two groups. Finally, in the bottom right panel, yet another variant is offered which gives a full picture of all posterior probabilities, represented by lines of length 1 which are split-coloured around the ordinate axis according to the values of r_{ij}, $j = 1, 2$. This variant is only supported for $k = 2$ as there are presentational difficulties otherwise.

Figure 3 [top] presents boxplots of log energy use data of the countries in selected years between 1971 and 2012. The five main features of a boxplot are obvious in each year. The median of log energy use data increased till the early 90's. It should be noted at this occasion that until 1989 only data for 112 countries were available, and that the sharp increase in 1991, and the subsequent decrease, can be explained by the inclusion of many new countries in 1991 after the fall of the iron curtain, and the subsequent political and economical developments in those countries.

Overall, the boxplots convey the impression that, taking the 1990 effect aside, there has been a relatively steady increase of energy use throughout all countries over time. The sequence of 2-boxplots of log energy use data shown in Fig. 3 [bottom] shows that this interpretation is actually not accurate. We see that the data form two groups, where one group corresponds to high energy use (supposedly so-called 'developed') countries, and one group corresponding to low energy use countries. The median as a measure of location almost does not change at all in either of the two groups, which appears to be in conflict with the information transmitted by the boxplots. However, what did change over time is that the low-energy-use group got smaller, and the high-energy-use group got larger, represented by the boxes getting slimmer and wider, respectively. This can be interpreted as that, over the years, more and more countries have managed to make the transition from a low to a high energy use country. This example demonstrates how misinterpretations based on traditional boxplots can be avoided when using the proposed graphical representation which takes the mixture character of the data into account. It is noted for completeness that, due to the non-linearity of the logarithm, the preceding analysis is not equivalent to fitting a mixture of log-normal distributions to the original data.

3.2 Example 2: internet users data

In this example, we consider a data set of size $n = 100$ which was originally given in the form of a time series of the numbers of users connected to the internet through a server every minute. The data are available in the R package **datasets** under the name `WWWusage` and visualized by a boxplot and a histogram in Fig. 4.

The histogram suggests that distributions with either $k = 3$ or $k = 4$ may be adequate. Considering firstly $k = 3$, we have produced 3-boxplots of the log(`WWWusage`) data using a mixture of three normal distributions, where two different cases have been considered. In the first case, we have allowed the components of the normal mixture to have unequal variances σ_j^2. In the second case, we assumed equal variances $\sigma_j^2 \equiv \sigma^2$, in which case the second of the estimators in (4) has to be adapted to become

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}(y_i - \hat{\mu}_j)^2.$$

In Fig. 5a, the 3-boxplot of log(`WWWusage`) for the unequal variance case is presented. There are three boxes which represent three categories in terms of the number of the internet users in different periods. It can be observed that the majority of the data fall into the central box, representing the large majority of time points for which a medium number of internet users was observed. There are additionally two smaller clusters corresponding to low and high internet usage, respectively. The 3-boxplots in the equal variance case are presented in Fig. 5b. We see that there is not much differ-

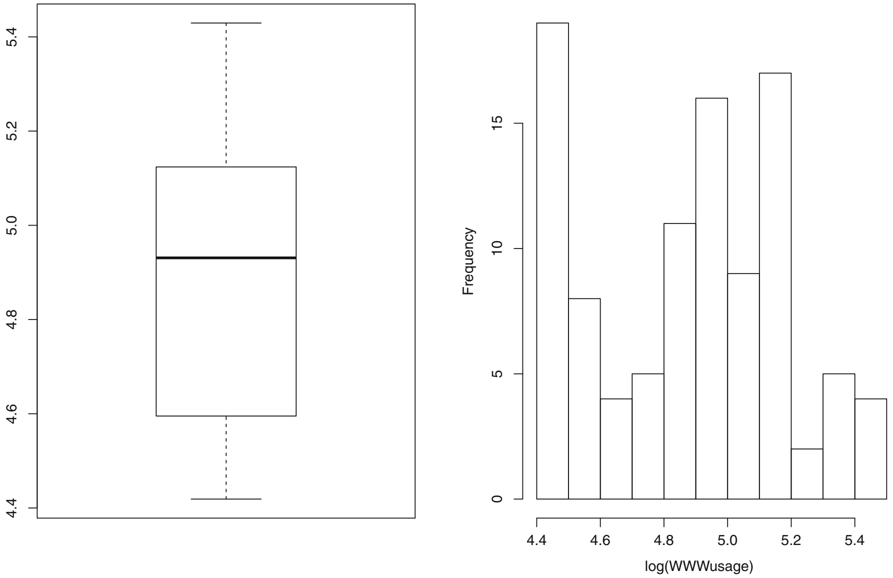

Fig. 4 Boxplot and histogram of log of the numbers of internet users

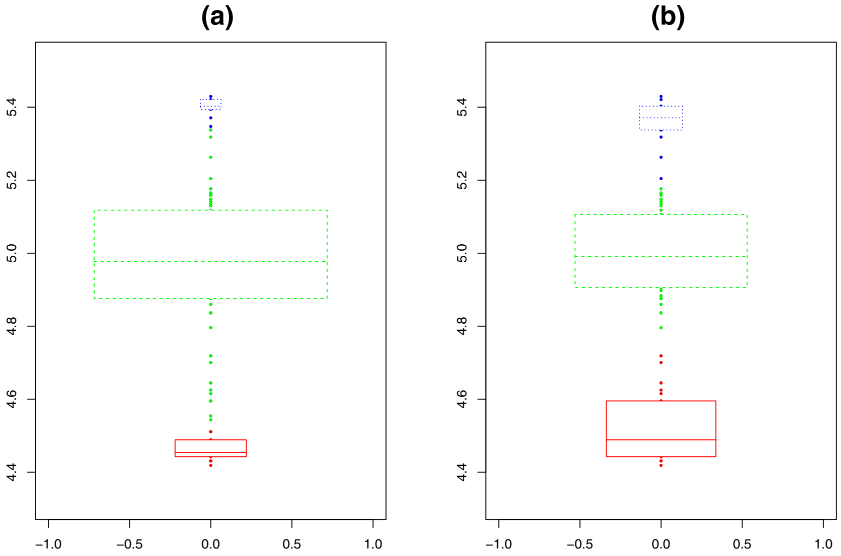

Fig. 5 3-Boxplots of log of the numbers of internet users, **a** with unequal variances, **b** with equal variances

ence between the plots at this instance, though, expectedly, the spread of the smaller boxes in the equal variance case is a bit larger than for the unequal variance case. All this information on size and structure of clusters cannot be observed by a traditional boxplot.

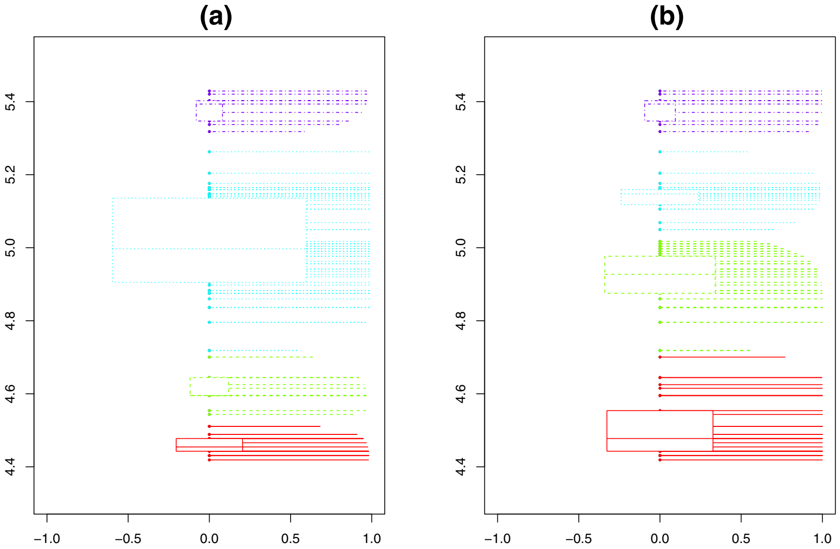

Fig. 6 4-Boxplots of log of the numbers of internet users, **a** with unequal variances, **b** with equal variances

Proceeding now to the case $k = 4$, Fig. 6a, b provide 4-boxplots of log(`WWWusage`) in the unequal and equal variance case, respectively. We see that, as compared to the 3-boxplots, the boxes have been split differently: In the unequal variance case (a), the low-usage box has been split, while in the equal variance case (b) the medium box has been split. Furthermore, we have provided these 4-boxplots in their 'full' form, which allows insights into the MAP classification of data points to clusters, as well as the posterior probability of belonging to that cluster (symbolized by the length of the horizontal line drawn to the right). We see that an appreciable number of observations is allocated to each cluster. If classification is the main purpose of the study, then this graphical information may be very useful.

Summarizing, while the most suitable working assumption (in terms of the choice of k and the choice of equal or unequal component variances) will depend on the particular application, the point that we want to make here is that the impact of this choice on the fitted model may be quite large, and that the k-boxplots allow the data analyst to visualize the consequence of their choice at a glance, which will be helpful to support their decision process on which model to choose. A k-boxplot is a tool to be used to visualize the different clusters in mixture data however it is not an inference method in itself. Consequently, as like any other graphical tool, the data analyst should not solely rely on a k-boxplot to determine the distribution of data.

3.3 Example 3: rainfall data

This final example will illustrate that the k-boxplot can be applied to a variety of statistical models as long as the output provides access to the matrix of posterior

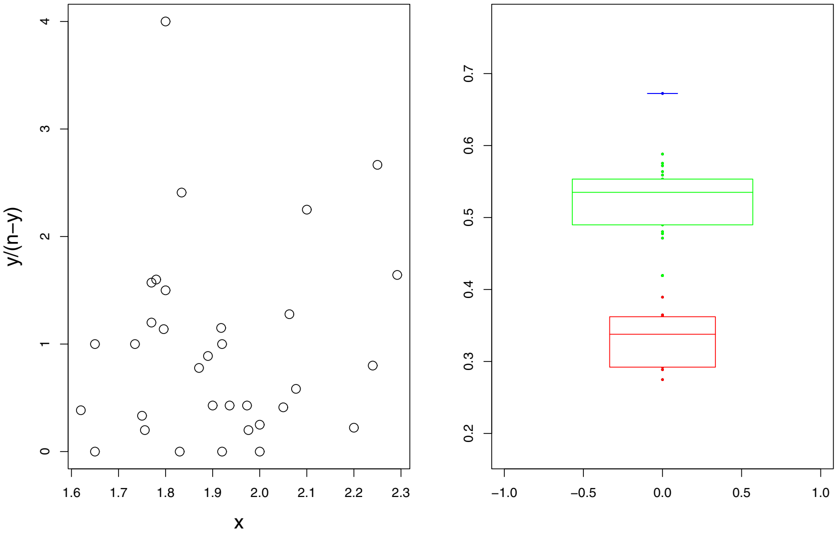

Fig. 7 *Left* Rainfall data. *Right* 3-boxplot of fitted probabilities $\hat{p}_i$ using model (5)

probabilities, $R = (r_{ij})$. We use for this illustration a data set giving the number of subjects y_i out of n_i testing positively for toxoplasmosis in $i = 1, \ldots, 34$ cities in El Salvador. The data set is available in the R package **npmlreg** under the name `rainfall`. It had previously been suggested in the literature that the annual rainfall, x_i (in 1000 mm) impacts on the occurrence of toxoplasmosis via a quadratic logistic regression model, such as $\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$, where $p_i = E(y_i)/n_i$ is the probability for an individual of contracting toxoplasmosis in city i. The data are displayed in Fig. 7 (left). However, Aitkin and Francis (1995) demonstrated that the dependence on rainfall becomes insignificant when a random effect term, z_i, is introduced which accounts for overdispersion, that is

$$\log \frac{p_i}{1-p_i} = z_i, \tag{5}$$

or, equivalently, $p_i = e^{z_i}/(1 + e^{z_i})$. Under this model, the mixture probability p_i is assumed to be driven only by randomness.

In the nonparametric maximum likelihood approach, the distribution of the random effect z_i can be left unspecified and is represented through a finite (discrete) mixture, the parameters of which are estimated via the EM-algorithm. Following the analysis by Aitkin and Francis (1995), we use $k = 3$, yielding a weight matrix $R \in \mathbb{R}^{34\times 3}$ which can be used to produce a 3-boxplot of fitted values $\hat{p}_i$ of model (5). The result is provided in Fig. 7 (right)—one clearly sees here the ‘unobserved heterogeneity’ in form of three subpopulations which is responsible for the overdispersion. It is clear from the k-boxplot that one observation has captured a component by its own, which may suggest to consider this observation as an outlier, and to reduce the number of components by one. Following up this route more thoroughly, we refit the model

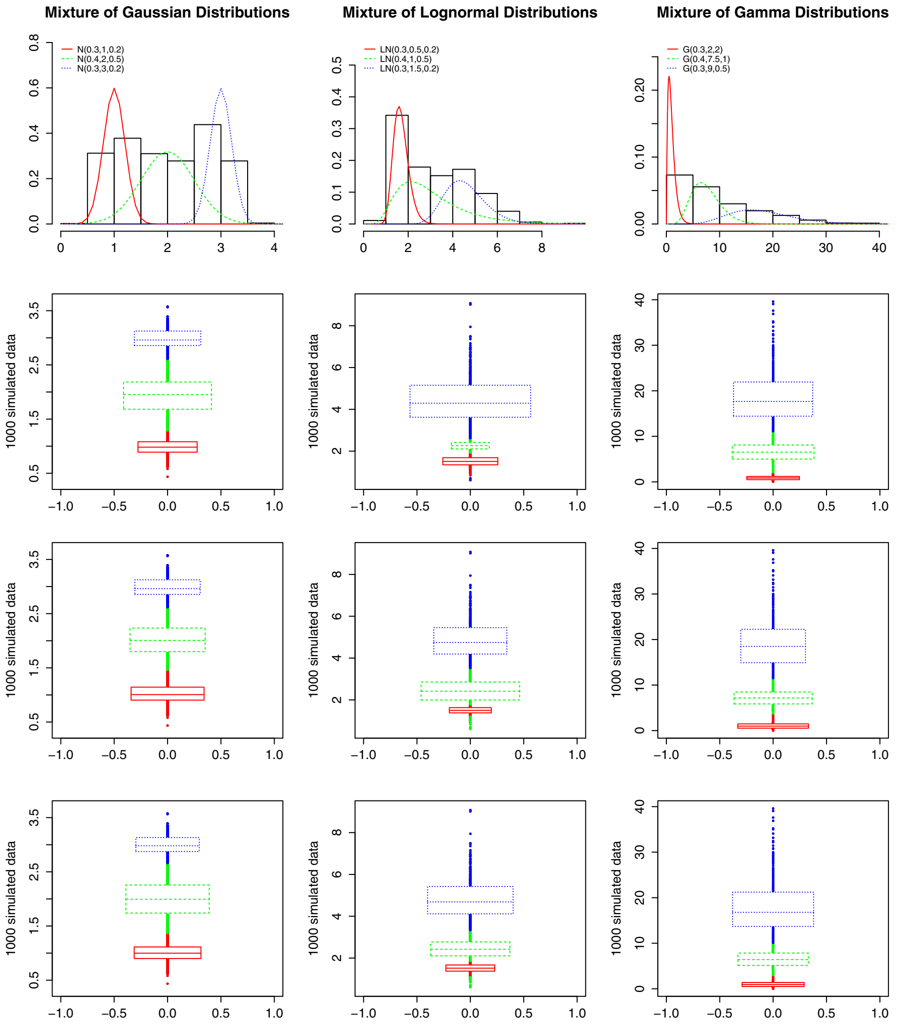

Fig. 8 3-boxplots simulated from scenarios **a**, **b**, **c** [from *left* to *right*] and fitted through Gaussian, lognormal, and Gamma component densities [from *top* to *bottom*]

using $k = 2$, yielding a decrease in disparity (i.e., $-2 \log L$, with L being the model likelihood) of 1.74. This number corresponds just to the test statistic of the likelihood ratio test for $H_0 : k = 2$ versus $H_1 : k = 3$. A bootstrapped null distribution can be obtained by resampling data from the fitted model for $k = 2$, refitting both models for $k = 2$ and $k = 3$, and computing the corresponding likelihood ratio in each case (Polymenis and Titterington 1998). Carrying out this procedure for 999 bootstrap replicates, we find that the value 1.74 would be ranked in 909th position among the bootstrapped likelihood ratios. The resulting p-value of 0.091 gives borderline strong evidence for the existence of the third component. It is finally worth noting that, if a component is captured by a single outlier, this increases the robustness of other components to this outlier.

3.4 Simulation

In order to get some insight into the behaviour of the k-boxplots under the use of component distributions other than Gaussian, and in particular under component misspecification, we have carried out a small–scale simulation in which data sets are simulated from three scenarios. Under all three simulated scenarios we use $k = 3$, $\pi_1 = 0.3$ and $\pi_2 = 0.4$, but the component densities differ as follows:

(a) a mixture of three Gaussian component densities with $\mu_j = j$, $\sigma_1 = \sigma_3 = 0.2$ and $\sigma_2 = 0.5$;
(b) a mixture of three log-normal densities with $\mu_j = j/2$, and σ_j, $j = 1, 2, 3$ as in (a);
(c) a mixture of three Gamma densities with shape parameters $2, 7.5, 9$ and scale parameters 2, 1, 0.5, respectively.

The true underlying densities are provided in the top row of Fig. 8 along with histograms of the simulated data sets. The panels below show 3-boxplots fitted to the simulated data using a mixture of three Gaussian distributions, log-normal distributions and Gamma distributions, respectively. That is, the component distributions are correctly specified along the diagonal of the 3×3 panel of 3-boxplots but misspecified off the diagonal.

The main conclusions from Fig. 8 are (i) the mixture proportions are in the most cases approximately correctly captured; (ii) if the data are simulated from Gaussian components (first column), then the 3-boxplots are quite robust to component misspecification; (iii) in the bottom right 2×2 panel, we see that the skewness of the original distribution is correctly represented by the fitted distribution; (iv) if a Gaussian mixture is fitted to ‘true’ lognormal or Gamma components, then the tail component tends to carry too much weight.

4 Conclusions

We have presented a new powerful graphical tool to visualize and analyse data stemming from a mixture of k distributions which we named a k-boxplot. This plot can be used to visualize the different k groups of mixture data which a boxplot is not able to achieve. It is a useful extension of the traditional boxplot especially for finding additional information regarding the location and spread of individual groups in mixture data which are ignored by a boxplot. Similar to a boxplot, a k-boxplot can visualize outliers in the data. The k-boxplot cannot be considered as an inference method which would be able to make automated decisions about the distribution or the number of components in mixture data. However, it is a useful tool to support the data analyst in this respect. For instance, overlapping or very small boxes may be a sign that the number of components should be reduced, or long one-sided tails outside the boxes may be a sign that the Gaussian component densities are not adequate.

k-boxplots are implemented in the function `kboxplot` which is made available as part of the R package **UEM**. The implemented R subroutine provides several graphical options for the data analyst, including a black-and white option. There are two ways in which this function can be used. The first option is to apply `kboxplot` directly onto the data itself, in which case the model will be fitted implicitly. The alternative option, which we would consider as the recommended option as it gives better control over the process, is to apply `kboxplot` onto a previously fitted model, for which subroutines provided within R package **UEM** could be used; but also functions from alternative R packages, or even alternative software, may be considered for this purpose, as long as they provide access to the weight matrix R. For instance, the computations in Example 3 in this paper have been carried out using the function `alldist` from R package **npmlreg**.

Given the matrix R, the computational complexity of producing a k-boxplot is of order $O(nk)$ as compared to $O(n)$ for a traditional boxplot. For all data sets, choices of k, and graphical variants considered in this paper, the computational time to produce a k-boxplot, given R, has been less than 0.02 seconds on an Intel® Core(TM) i7-3770 CPU @ 3.40GHz machine. The computations required for the underlying inferential mechanism will usually contribute the larger computational burden. For instance, for Example 2, which has been computed using the EM routines built into R package **UEM**, this computation required 0.11 seconds for the (unequal variance) 3-component model (28 EM iterations), and 0.26 seconds for the 4-component model (52 EM iterations), using Gaussian Quadrature points as starting points in each case. It is noted at this occasion that R code to reproduce the examples presented in this paper is provided in the R Documentation files of R package **UEM**.

One issue which we have given rather marginal attention is the selection of the number of components, k. This problem is inherent to the mixture fitting technique, and while there does exist a rich literature on suggested methods how to select this number k, this question is eventually still down to the subjective judgement of the data analyst. In order to arrive at this judgement, the data analyst will undoubtedly benefit from a simple graphical tool, as the proposed one, which visualizes the structure of the mixture model which is obtained under the hypothesized number k at a glance. In this sense, a k-boxplot could contribute to the question of selecting the number k, in conjunction with existing quantitative techniques such as the parametric bootstrap, as illustrated in the Example in Sect. 3.3.

Acknowledgments We thank an editor and two referees for the excellent comments which led to improved insights and conclusions from our results. The first author of this manuscript is grateful to the Saudi Arabian Cultural Bureau in London for the financial support of this work.

References

Abuzaid AH, Mohamed IB, Hussin AG (2012) Boxplot for circular variables. Comput Stat 27(3):381–392

Aitkin M, Francis BJ (1995) Fitting overdispersed generalized linear models by nonparametric maximum likelihood. GLIM Newsletter 25:37–45

Baudry JP, Celeux G (2015) EM for mixtures. Stat Comput 25(4):713–726

Benjamini Y (1988) Opening the box of a boxplot. Am Stat 42(4):257–262

Biernacki C, Celeux G, Govaert G (2003) Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Comput Stat Data Anal 41:561–575

Bruffaerts C, Verardi V, Vermandele C (2014) A generalized boxplot for skewed and heavy-tailed distributions. Stat Prob Lett 95:110–117

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39(1):1–38

Einbeck J, Taylor J (2013) A number-of-modes reference rule for density estimation under multimodality. Stat Neerland 67(1):54–66

Fried R, Einbeck J, Gather U (2007) Weighted repeated median smoothing and filtering. J Am Stat Assoc 102:1300–1308

Frigge M, Hoaglin DC, Iglewicz B (1989) Some implementations of the boxplot. Am Stat 43(1):50–54

Hubert M, Vandervieren E (2008) An adjusted boxplot for skewed distributions. Comput Stat Data Anal 52(12):5186–5201

Krzywinski M, Altman N (2014) Points of significance: visualizing samples with box plots. Nature Methods 11(2):119–120

McGill R, Tukey JW, Larsen WA (1978) Variations of box plots. Am Stat 32(1):12–16

McLachlan G, Peel D (2004) Finite mixture models, 3rd edn. John Wiley & Sons, New York

Polymenis A, Titterington DM (1998) On the determination of the number of components in a mixture. Stat Prob Lett 38(4):295–298

R Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL: https://www.R-project.org/

Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Boston

Wang Z, Bellhouse D (2014) A diagnostic tool for regression analysis of complex survey data. Stat Papers 56:1–13