Received: 6 July 2020 | Accepted: 30 September 2020

DOI: 10.1111/2041-210X.13520



Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

# Unsupervised acoustic classification of individual gibbon females and the implications for passive acoustic monitoring

**Dena J. Clink** | **Holger Klinck**

Center for Conservation Bioacoustics, Cornell Laboratory of Ornithology, Cornell University, Ithaca, NY, USA

**Correspondence**
Dena J. Clink
Email: dena.clink@cornell.edu

**Funding information**
Fulbright Association

**Handling Editor:** Veronica Zamora-Gutierrez

## Abstract

1. Passive acoustic monitoring (PAM) has the potential to greatly improve our ability to monitor cryptic yet vocal animals. Advances in automated signal detection have increased the scope of PAM, but distinguishing between individuals—which is necessary for density estimation—remains a major challenge. When individual identity is known, supervised classification techniques can be used to distinguish between individuals. Supervised methods require labelled training data, whereas unsupervised techniques do not. If the acoustic signals of individuals are sufficiently different, the number of clusters might represent the number of individuals sampled. The majority of applications of unsupervised techniques in animal vocalizations have focused on quantifying species-specific call repertoires. However, with increased interest in PAM applications, unsupervised methods that can distinguish between individuals are needed.
2. Here we use an existing dataset of Bornean gibbon female calls with known identity from five sites on Malaysian Borneo to test the ability of three different unsupervised clustering algorithms (affinity propagation, *K*-medoids and Gaussian mixture model-based clustering) to distinguish between individuals. Calls from different gibbon females are readily distinguishable using supervised techniques. For internal validation of unsupervised cluster solutions, we calculated silhouette coefficients. For external validation, we compared clustering results with female identity labels using a standard metric: normalized mutual information. We also calculated classification accuracy by assigning unsupervised cluster solutions to females based on which cluster had the highest number of calls from a particular female.
3. We found that affinity propagation clustering consistently outperformed the other algorithms for all metrics used. In particular, classification accuracy of affinity propagation clustering was more consistent as the number of females increased, and when we randomly sampled females across sites.
4. We conclude that unsupervised techniques may be useful for providing additional information regarding individual identity for PAM applications. We stress that although we use gibbons as a case study, these methods will be applicable for any individually distinct vocal animal.

**KEYWORDS**

affinity propagation, Gaussian mixture models, *Hylobates*, *K*-medoids, normalized mutual information, passive acoustic monitoring, unsupervised clustering

  

## 1 | INTRODUCTION

Passive acoustic monitoring (PAM) of animals—which relies on the use of battery-operated autonomous recording devices—has been used extensively in marine systems (e.g. Davis et al., 2020; Fournet et al., 2018; Martin et al., 2013), and there has been a substantial increase in use in terrestrial systems in recent years (Sugai et al., 2019). Terrestrial applications of PAM include the assessment of occurrence (Vu & Tran, 2019) and movement patterns (Kalan et al., 2016), estimating population density (Dawson & Efford, 2009; Enari et al., 2017) and monitoring behaviour and activity patterns (e.g. Clink et al., 2020b; Wrege et al., 2017). PAM can provide data on vocal animals at ecologically relevant scales that are difficult to obtain when relying on human observers (Marques et al., 2013). In some cases, methods that rely on PAM have been shown to outperform human observers (e.g. Darras et al., 2019).

A significant amount of effort has been put into the development of different automated detection algorithms for terrestrial animals (Bardeli et al., 2010; Kalan et al., 2015; Katz et al., 2016; Keen et al., 2017; Zeppelzauer et al., 2015). For many PAM applications, such as population density estimation, effective discrimination between individuals is necessary. The ability to acoustically distinguish between individuals can also provide important insights into the behaviour and ecology of the focal animal(s) (Terry et al., 2005). For many applications a major hurdle for wide-scale implementation of PAM is the lack of robust techniques to effectively discriminate between calling individuals. Newer PAM applications of population density estimation, such as spatially explicit capture–recapture, generally require an input of individual identification (Augustine et al., 2018, 2019; Kidney et al., 2016), which can be provided either via acoustic localization or based on individual acoustic features of the detected 'signature' calls. Acoustic localization in terrestrial systems is possible (e.g. Spillmann et al., 2015) but requires time-aligned recording units and a substantial amount of analyst effort. Therefore, unsupervised methods that can provide information on the number of calling animals based on individual differences in call features are desirable.

In machine learning, supervised techniques are used when datasets are labelled, for example, in the case where the identity of each calling individual is known, and calls can be subsequently classified based on known class membership (Greene et al., 2008). Supervised methods of discriminating between primate individuals based on features of their calls have been well established (Clink et al., 2017; Leliveld et al., 2011; Mielke & Zuberbühler, 2013; Mitani et al., 1996; Rendall, 2003; Rendall et al., 1996; Terleph et al., 2015). An important caveat for the use of supervised methods is that individual identity must be known, which is often not the case with acoustic data collected autonomously over extended periods. Generally, unsupervised clustering algorithms are used to make inferences about unlabelled data, which is in contrast to supervised algorithms that require the input of labelled training data (Dinov, 2018; Greene et al., 2008).

Most unsupervised applications related to non-human primate vocalizations have investigated species-specific call repertoires (Keenan et al., 2013; Pozzi et al., 2009; Price et al., 2015; Valente et al., 2019; Wadewitz et al., 2015). One of the foundational applications of unsupervised clustering was used to investigate the vocal repertoire of Barbary macaques *Macaca sylvanus* (Hammerschmidt & Fischer, 1998). Since then, unsupervised approaches have been used to investigate vocal repertoires in indris *Indri indri* (Valente et al., 2019), black lemurs *Eulemur macaco* (Pozzi et al., 2009), *Eulemur* spp. (Gamba et al., 2015), marmosets *Callithrix jacchus* (Turesson et al., 2016), douc langurs *Pygathrix cinerea* (Riondato et al., 2017), Campbells's monkeys *Cercopithecus campbelli* (Keenan et al., 2013), chacma *Papio ursinus*, olive *P. anubis* and Guinea baboons *P. papio* (Hammerschmidt & Fischer, 2019) and gorillas *Gorilla gorilla* (Hedwig et al., 2014). Outside the Order Primates there have been a few applications of unsupervised classification of individuals, including bottlenose dolphins *Tursiops truncatus* (Kershenbaum et al., 2013) and Mexican ant thrushes *Formicarius moniliger* (Kirschel et al., 2009).

Validation of supervised classification is commonly done using *k*-fold cross-validation wherein the supervised algorithm (e.g. discriminant function analysis or support vector machine) is trained on a subset of the data (often for all observations but one, which is known as 'leave-one-out cross-validation'), and then the remaining observations are classified (Tan et al., 2016). Classification by the algorithm is compared with actual data labels (known as ground-truthing), and classification accuracy can be calculated. For unsupervised techniques, validation is less straight-forward, as the data for unsupervised methods are generally unlabelled. Cluster validation for unsupervised approaches includes: (a) determining the correct number of clusters; (b) evaluating how well the results of cluster analysis fit the data without reference to external information (internal validation) and (b) comparing the results of cluster analysis to external information (external validation; Tan et al., 2016).

Here we utilize an existing dataset of recordings of calls from Northern gray gibbon *Hylobates funereus* females recorded at five different sites in Malaysian Borneo to investigate the effectiveness of unsupervised clustering techniques to distinguish between individual females. Bornean gibbon females have a high degree of vocal individuality (Clink et al., 2017; Clink, Crofoot, & Marshall, 2018; Clink, Grote, et al., 2018), and supervised techniques have been shown to effectively discriminate 33 females with over 98% accuracy (Clink, Crofoot, & Marshall, 2018). Our main goal was to compare affinity propagation clustering (Frey & Dueck, 2007) with two other commonly used unsupervised clustering approaches: *K*-medoids (Kaufman & Rousseeuw, 1990) and Gaussian mixture model-based clustering (Fraley & Raftery, 2002). All three algorithms utilize a centroid-based clustering approach, but affinity propagation clustering has two notable differences: it does not require the user to input the number of clusters a priori and does not require a choice of initial starting points (Dueck, 2009; Frey & Dueck, 2007).

For internal validation of cluster solutions, we calculated silhouette coefficients (Rousseeuw, 1987). We then compared unsupervised clustering results with known female identity using a commonly used external validation metric: normalized mutual information (NMI; Bezdek, 1974; Wadewitz et al., 2015). In addition,

following Dueck (2009), we calculated the number of calls that were correctly classified by associating each cluster with the female that had the highest number of calls in that cluster. To compare the algorithms, we randomly chose 10, 20, 30, 40, and 50 females from our dataset over 100 iterations and calculated silhouette coefficients, NMI and per cent correct classifications. We also calculated the difference in the predicted number of clusters (individual females) and the actual number of individual females in the dataset. For more ecologically relevant comparisons, we also used a bootstrapping approach where we randomly chose 80% of the calls from each site in our dataset over 100 iterations and calculated internal, external and performance metrics as outlined above.

## 2 | MATERIALS AND METHODS

### 2.1 | Acoustic data collection

Gibbons are known for their coordinated, long-distance vocalizations (known as duets) between mated male and female pairs (Geissmann, 2002). The present analysis focused on the female contribution to the duet known as the great call; these calls follow a stereotyped, species-specific pattern (see Figure 1 for representative spectrograms) and have been shown to be individually distinct (Clink et al., 2017; Clink, Crofoot, & Marshall, 2018; Clink, Grote, et al., 2018). The original dataset contained 933 calls (range: 2–46 calls per female) collected from 66 different individual Northern gray gibbon *Hylobates funereus* females from five different sites in Malaysian Borneo: Maliau Basin Conservation Area, Deramakot Forest Reserve, Imbak Canyon Conservation Area, Danum Valley Conservation Area and Kalabakan Forest Reserve using a Marantz PMD 660 recorder (Marantz) equipped with a Røde NTG-2 directional condenser microphone (Røde Microphones); see Clink, Charif, et al. (2018) and Clink, Grote, et al. (2018) for details on data collection.

Recordings were taken at 44.1 kHz sampling rate, a sample size of 16-bits, and were saved as Waveform Audio (WAV) files. A previous analysis of this dataset indicated that most of the variation in call features occurs at the level of the individual, and that only one call feature (trill rate) varied across sites (Clink, Grote, et al., 2018). All data were collected via focal recordings at a distance of ~150 m or less, and calls included in the present analysis had a signal-to-noise ratio (SNR) of 10 dB or higher. As data were collected from wild, unhabituated gibbons, there was the possibility of misidentification of some females. But, the supervised classification accuracy of all females in our dataset was quite high (see below), which indicates that resampling the same female and classifying her as two separate females had a minimal impact on our results. See Clink et al. (2017) and Clink, Grote, et al. (2018) and the discussion (this paper) for more details.

### 2.2 | Acoustic data processing

Mel-frequency cepstral coefficients (MFCCs) are useful features for distinguishing between gibbon females (Clink, Crofoot, & Marshall, 2018) and were also used for this analysis. To calculate MFCCs, we used the package 'tuneR' (Ligges et al., 2016) in the R programming environment (R Core Team, 2019). Calls in our dataset varied in duration from 7.6 to 20.9 s. Although call duration varies between individuals, there were not site-level patterns of variation

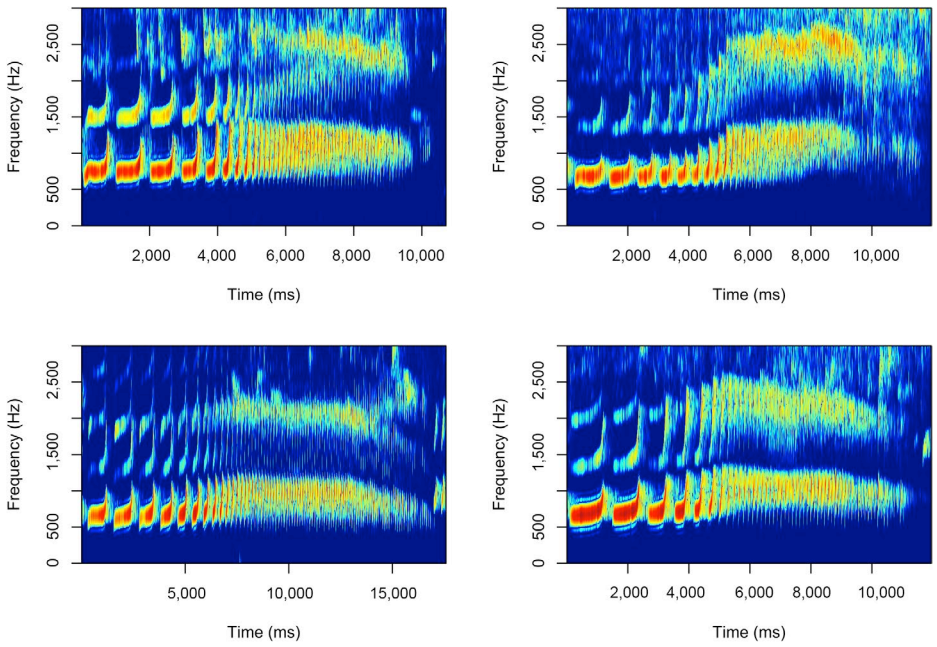


**FIGURE 1** Representative spectrograms of calls from four gibbon females. Spectrograms were made with the 'phonTools' (Barreda, 2015) package with a 512-point (11.6 ms) Hann window (3 dB bandwidth = 124 Hz), with 75% overlap and a 1024-point DFT, yielding time and frequency measurement precision of 2.9 ms and 43.1 Hz respectively

in this feature (Clink, Grote, et al., 2018). As most machine learning algorithms require feature vectors of equal length for each observation (in our case each call), we calculated MFCCs over a standardized number of windows (8) for each call, and the size of time windows we used to calculate MFCCs varied depending on the total duration of the call. For each of the eight windows, we calculated 12 Mel-filters (or bandpass filters; Davis & Mermelstein, 1980) between 500 and 1,500 Hz, which corresponds with the fundamental frequency range of gibbon female great calls. The first MFCC for each time window corresponds to the amplitude or loudness of the signal (Muda et al., 2010); this will vary depending on the recording distance to the calling animal and is therefore not appropriate for discriminative tasks so we only used 11 MFCCs for each time window.

In addition, MFCCs describe the spectral envelope at particular points in time, but do not capture temporal variation in the signal. Therefore, we also calculated delta-cepstral coefficients which provide a measure of change from one frame to the next, and provide information about the temporal dynamics of the signal (Kumar et al., 2011). As we estimated 11 MFCCs for each time window, we also had 11 delta coefficients. We also included duration, which resulted in a final feature vector of 177 parameters describing each call. Given the previous success with using MFCCs, the delta coefficients, and call duration as features for distinguishing between gibbon females (Clink, Crofoot, & Marshall, 2018), we used only these features for our unsupervised clustering experiments.

## 2.3 | Supervised classification

To investigate how the number of calls included per female influenced our classification results, we ran an experiment where we randomly chose 3–15 calls per female over 100 iterations. We used a support vector machine (SVM; Cortes & Vapnik, 1995) implemented in the R package 'E1071' (Meyer et al., 2017) for supervised classification of female calls. We used a 'radial basis' kernel type and set the *k*-fold cross-validation parameter to 5, which means that 80% of the data were used for training and 20% were used for testing. We ran the experiment over 100 iterations for each number of calls (3–15). There was variation in the number of calls per female in our original dataset, which means that the number of females included necessarily decreased as the number of calls increased. To determine the minimum number of calls per female needed for stabilization of classifier performance we plotted the results using the R package 'GGPUBR' (Kassambara, 2017).

## 2.4 | Unsupervised clustering

We compared three unsupervised clustering methods: affinity propagation clustering (Frey & Dueck, 2007) using the package 'APCLUSTER' (Bodenhofer et al., 2011), *K*-medoids clustering (Macqueen, 1967) using the 'CLUSTER' package (Maechler et al., 2019) and Gaussian mixture model-based clustering (Duda & Hart, 1973) using the 'MCLUST' package (Scrucca et al., 2016). All three algorithms can be used to cluster unlabelled data, but there are some fundamental differences. In *K*-means, the user defines the target number of clusters, *k*. The algorithm randomly assigns *k* data points to be used as the starting centroid(s). All data points are assigned to a cluster based on closest proximity to the randomly placed centroid, and the algorithm calculates a new centroid for each cluster. The cluster centroid is then iteratively optimized, and this process continues until the centroids have stabilized, or the defined number of iterations has been achieved (Hamerly & Drake, 2015).

We used *K*-medoids clustering (also known as partitioning around medoids (Kaufman & Rousseeuw, 1990), which is a more robust version of *K*-means (Maechler et al., 2019). The main difference is that *K*-medoids clustering chooses a data point as the centre of each cluster, whereas in *K*-means the centroid may not represent an actual data point (Madhulatha, 2011). Unlike *K*-means, *K*-medoids does not calculate the mean for each cluster, but chooses a representative data point or medoid, which makes it more robust to noise and outliers (Reynolds et al., 2006). Gaussian mixture model-based clustering can be considered conceptually as an extension of *K*-medoids with two differences. First, *K*-medoids 'hard-assigns' points to a cluster, whereas Gaussian mixture model-based clustering provides a probability that a point belongs to each of the possible clusters. Second, *K*-medoids clustering lacks flexibility in cluster shape, and assumes clusters have a circular shape (or hypersphere in higher dimensions), which is often not the case with real-world data; Gaussian mixture model-based clustering allows for oblong or elliptical shaped clusters (Fraley & Raftery, 2002).

Unlike many other commonly used unsupervised clustering algorithms, affinity propagation clustering does not require the number of clusters to be predetermined (Frey & Dueck, 2007). Another proposed benefit of affinity propagation clustering is that it identifies exemplar observations for each cluster type (Brusco et al., 2019); however, other algorithms such as *k*-medoids also do this (Scrucca et al., 2016; Macqueen, 1967; Madhulatha, 2011). Affinity propagation clustering relies on what is known as a message-passing algorithm that takes similarities between data points as the input (Brusco et al., 2019). The data points can be considered as occurring in a network, and because of the structure of the network, affinity propagation clustering considers all data points simultaneously, which means that unlike other algorithms, the results will not be influenced by choosing the initial set of points (Dueck, 2009). The network is based on a factor graph (a type of graphical model) wherein real-valued functions (known as messages) are passed between the data points (Frey & Dueck, 2007). Graphical models are an effective way to express and visualize the structure of a network of data points (Frey & Dueck, 2007). Two types of messages are sent between points: the first are termed 'responsibilities' which are sent from data points to candidate exemplars, and the second are termed 'availabilities' which are sent from candidate exemplars to the data points; these messages represent the suitability of one point to act as an exemplar of the

other (Dueck, 2009). As mentioned above, the suitability of data points to be exemplars is considered while simultaneously considering the suitability of other data points to be exemplars (Frey & Dueck, 2007). After convergence, a small set of exemplars is used to describe the dataset (Pedregosa et al., 2011). To-date, affinity propagation clustering has been used in a few bioacoustics applications, including anomaly detection in forest soundscapes (Sethi et al., 2020) and unsupervised clustering of male gibbon solo phrases (Clink et al., 2020a).

## 2.5 | Internal validation

For internal validation of cluster solutions, we compared each of the three clustering algorithms using silhouette coefficients (Maechler et al., 2019; Rousseeuw, 1987). Silhouette coefficients provide a measure of how similar an object is relative to the established clusters; values range from −1 to 1, and higher silhouette coefficients indicate a more appropriate clustering solution (Rousseeuw, 1987). Both *K*-medoids and Gaussian mixture model-based clustering require the user to input the number of clusters, so we ran cluster solutions for a range of numbers of clusters (see details below). For *K*-medoids, we calculated the silhouette coefficient to assess the discreteness of cluster solutions; we chose the cluster solution, which had the highest silhouette coefficient.

For Gaussian mixture model-based clustering, the 'Mclust' function comes with an internal capability to compare cluster solutions using Bayesian information criterion (BIC), so we chose the clustering solution with the lowest BIC (Scrucca et al., 2016). For affinity propagation clustering, the number of clusters returned by the algorithm can be influenced by the input preferences (Bodenhofer et al., 2011), so we systematically varied the input preferences using the '*q*' input from 0 to 1 (in increments of 0.1), returned the cluster solutions and calculated a silhouette coefficient; this is known as adaptive affinity propagation clustering (Wang et al., 2008). Our early experiments indicated that using the median of similarities ($q = 0.5$) as input preference led to an optimal number of clusters, so the results we report here are based on using the median values. This is the default in the 'apcluster' function and is based on strategies outlined in Frey and Dueck (2007).

## 2.6 | External validation

For external validation, we used two different approaches. First, we aimed to measure how well the unsupervised clustering results matched the labelled data using standard metrics. The two most commonly used metrics for measuring agreement between clustering results and ground-truth (labelled) data are the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). We calculated ARI (Hubert & Arabie, 1985) using the 'MCLUST' package (Scrucca et al., 2016), and we calculated NMI (Bezdek, 1974) using the 'ARICODE' package (Chiquet & Rigaill, 2019). Both of these metrics provide a measure of how well the results of two overlapping clusters match; a value close to 1 indicates almost perfect agreement, whereas if the compared clusters have little conformity, the values will be close to zero. We found that there was a very high correlation (>0.95) between the two metrics, so report only the NMI.

Unlike supervised classification, unsupervised algorithms do not provide a class label for each observation, which makes it less straight-forward to assess their accuracy in assigning observations to their respective classes. However, one approach is to associate the unsupervised cluster with the true cluster that accounts for the largest number of data points in the unsupervised cluster (Dueck, 2009). For both *K*-medoids and Gaussian mixture model-based clustering, we assigned a class label (female identity) to the cluster that had the highest number of calls from that female. For affinity propagation clustering, we used the class label for the exemplars of each cluster (which are provided by the algorithm) to determine which cluster was associated with which female. Importantly, if there is no penalty for over-splitting the ideal scenario would be to place each data point into its own category or cluster. To avoid this pitfall, we identified all instances in which a female was placed in two clusters or categories. We then identified which cluster had the higher number of calls, assigned that cluster to the corresponding female and assigned the other clusters a 'NA' label so that it would be calculated as an incorrect classification. We then calculated the per cent of observations correctly assigned to their respective classes by the algorithm.

## 2.7 | Bootstrapping using the entire dataset

To assess the performance of the three different algorithms, we conducted a set of experiments wherein we randomly chose calls from 10, 20, 30, 40 or 50 females in our dataset over 100 iterations. For both *K*-medoids and Gaussian mixture model-based clustering, we allowed the number of clusters to vary from 2 to 53 clusters (the maximum number of females in our dataset; see below). We chose the cluster solution with the highest silhouette coefficient for *K*-medoids clustering and chose the cluster solution with the lowest BIC for Gaussian mixture model-based clustering. For each iteration, we calculated the silhouette coefficient, NMI, and the per cent of observations correctly assigned to their respective class.

## 2.8 | Bootstrapping across sites

Our dataset included gibbon females from five different sites, which provided a natural experiment for us to investigate how the three algorithms varied across sites with a known number of females. For each site we randomly selected 80% of the calls. As outlined above, we calculated the silhouette coefficient, NMI and the per cent of observations correctly assigned to the respective class over 100 iterations.

### 2.9 | Visualization of results

To visualize our results we used a uniform manifold learning technique (UMAP; McInnes et al., 2018) implemented in the R package 'UMAP' to embed the 177 features from each gibbon female call into a two-dimensional space. UMAP is an effective dimensionality technique that has been used to visualize differences in forest soundscapes (Sethi et al., 2020) and two distinct taxonomic groups of a neotropical passerine (Parra-Hernández et al., 2020). We used the package 'GGPLOT2' (Wickham, 2016) to plot the UMAP projections. To calculate classification accuracies used in the UMAP plots, we implemented the three unsupervised clustering algorithms as described above using the entire dataset.

## 3 | RESULTS

### 3.1 | Supervised classification

To investigate how the number of calls included per female influenced our classification accuracy we randomly selected 3–15 calls per female and classified them using SVM over 100 iterations. We found that SVM classification performance stabilized (and classification accuracy was >90%) when we used at least seven calls per female (Figure 2). We omitted any females with fewer than seven calls ($N = 13$ females) for our subsequent tests of the unsupervised classification algorithms, which reduced our sample size to 883 calls from 53 gibbon females. See Table 1 for a summary of sample size by site for the reduced dataset.

### 3.2 | Bootstrapping over 100 iterations

We randomly chose calls from 10, 20, 30, 40 or 50 females over 100 iterations. For all three algorithms, the silhouette coefficient (which we used as an internal validation metric) decreased as the number of clusters increased (Table 2). We found that affinity propagation clustering outperformed *K*-medoids and Gaussian mixture model-based clustering in both external validation metrics—NMI and per cent correct classification—notably when the number of randomly chosen clusters (females) was high. Affinity propagation clustering had a substantially higher mean classification accuracy and lower deviation from the actual number of females in the dataset than the other two algorithms, particularly for experiments with greater than 30 randomly chosen females (Figure 3; Table 2).

### 3.3 | Unsupervised clustering across sites

We conducted a second experiment where we divided our dataset by site and randomly chose 80% of calls from each site over 100 iterations and compared the three algorithms as outlined above. For both

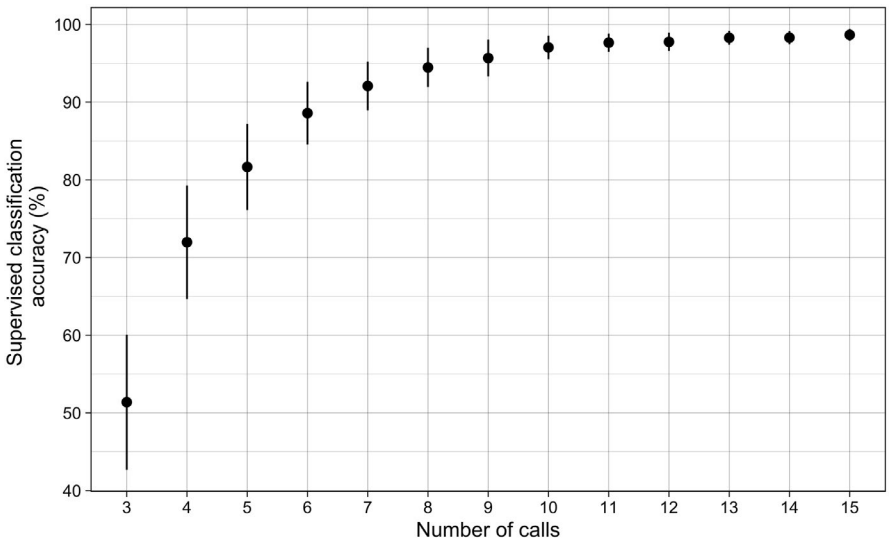


**FIGURE 2** Supervised classification accuracy ($M \pm SD$) as a function of number of randomly selected calls included per female. We randomly selected 3–15 calls per female and tested the performance of SVM classification over 100 iterations

**TABLE 1** Summary of sample size of calls used for tests of the three unsupervised clustering algorithms

| Site | Total females | Total calls | Mean calls | Min calls | Max calls |
|---|---|---|---|---|---|
| Deramakot Forest Reserve (Site 1) | 8 | 118 | 14.75 | 8 | 25 |
| Danum Valley (Site 2) | 12 | 199 | 16.58 | 7 | 32 |
| Imbak Canyon (Site 3) | 8 | 157 | 19.62 | 7 | 46 |
| Maliau Basin (Site 4) | 3 | 65 | 21.67 | 11 | 41 |
| Kalabakan Forest Reserve (Site 5) | 22 | 344 | 15.64 | 7 | 46 |
| All sites | 53 | 883 | ~ | ~ | ~ |

**TABLE 2** Comparison of three unsupervised clustering approaches (affinity propagation, *K*-medoids and Gaussian mixture-model based clustering) of bootstrapping over 100 iterations for randomly chosen calls from 10, 20, 30, 40 or 50 females. For each number of females, we include the number of calls analyzed (range) along with the mean ± *SD* of the number of clusters returned by the algorithm, the per cent of observations correctly classified, the silhouette coefficient and normalized mutual information index (NMI)

| Number of females | Clustering method | Number of calls (range) | Number of clusters ($M \pm SD$) | Percent correct ($M \pm SD$) | Silhouette ($M \pm SD$) | NMI ($M \pm SD$) |
|---|---|---|---|---|---|---|
| 10 | Affinity | 102–238 | 12.49 ± 2.02 | 75.91 ± 12.76 | 0.33 ± 0.05 | 0.72 ± 0.14 |
| | Gaussian | 100–253 | 15.52 ± 8.24 | 70.98 ± 20.46 | 0.28 ± 0.07 | 0.73 ± 0.24 |
| | *K*-medoids | 106–274 | 7.18 ± 3.92 | 53.24 ± 18.25 | 0.37 ± 0.04 | 0.55 ± 0.28 |
| 20 | Affinity | 224–445 | 22.88 ± 2.31 | 72.55 ± 9.31 | 0.28 ± 0.02 | 0.59 ± 0.14 |
| | Gaussian | 267–436 | 25.34 ± 9.3 | 50.78 ± 9.78 | 0.18 ± 0.04 | 0.45 ± 0.12 |
| | *K*-medoids | 240–444 | 9.10 ± 8.61 | 34.59 ± 15.25 | 0.32 ± 0.03 | 0.33 ± 0.31 |
| 30 | Affinity | 403–601 | 32.31 ± 2.53 | 68.59 ± 5.44 | 0.25 ± 0.02 | 0.43 ± 0.09 |
| | Gaussian | 425–588 | 31.42 ± 9.47 | 50.36 ± 7.38 | 0.17 ± 0.02 | 0.43 ± 0.11 |
| | *K*-medoids | 405–602 | 10.89 ± 11.87 | 26.65 ± 12.13 | 0.28 ± 0.03 | 0.25 ± 0.27 |
| 40 | Affinity | 592–738 | 40.91 ± 2.29 | 66.80 ± 3.45 | 0.23 ± 0.01 | 0.34 ± 0.06 |
| | Gaussian | 585–732 | 36.07 ± 7.38 | 49.39 ± 6.14 | 0.16 ± 0.02 | 0.39 ± 0.09 |
| | *K*-medoids | 581–737 | 9.50 ± 12.96 | 19.51 ± 8.74 | 0.26 ± 0.02 | 0.17 ± 0.23 |
| 50 | Affinity | 784–860 | 50.15 ± 1.29 | 64.83 ± 2.05 | 0.21 ± 0.01 | 0.27 ± 0.03 |
| | Gaussian | 793–856 | 38.7 ± 5.14 | 49.17 ± 3.86 | 0.15 ± 0.01 | 0.39 ± 0.06 |
| | *K*-medoids | 782–857 | 5.72 ± 9.38 | 16.26 ± 4.26 | 0.25 ± 0.01 | 0.08 ± 0.13 |

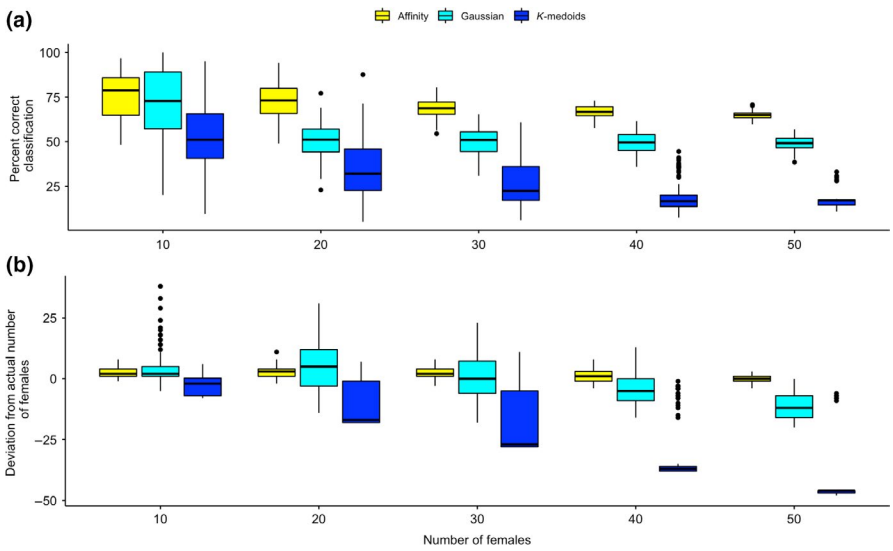


**FIGURE 3** Results of bootstrapping over 100 iterations for randomly chosen calls from 10, 20, 30, 40 or 50 females. Per cent correct classification (a) and deviation from the actual number of females (b) of randomly sampled number of females using affinity propagation, Gaussian mixture model-based or *K*-medoids clustering. The boxplots show the median, first and third percentiles, and range for each algorithm and sample size

of our external validation metrics, we found that affinity propagation clustering outperformed *K*-medoids and Gaussian mixture model-based clustering (Table 3). We also found that when comparing the predicted number of clusters (or individuals) to the actual number of individuals, affinity propagation clustering showed substantially less deviation from the actual number of females than the other two methods (Figure 4). Lastly, we found that the per cent of calls that were classified to the correct female was higher for affinity propagation clustering, and this was consistent across sites (Table 3; Figure 4). We also found that classification accuracy was substantially higher for affinity propagation clustering when we included all the females in our dataset (Figure 5).

## 4 | DISCUSSION

Here we compare the ability of three unsupervised clustering algorithms to distinguish between calls of Northern gray gibbon female individuals. We show that affinity propagation clustering substantially outperformed *K*-medoids and Gaussian model-based clustering. In particular, classification accuracy of affinity propagation clustering was more consistent as the number of randomly sampled females increased and when we randomly selected calls from each of our sites. We conclude that affinity propagation clustering may be a useful tool for determining the number of calling individuals (with some error) in PAM applications. Our results indicate that

**TABLE 3** Summary of results of unsupervised clustering female gibbon calls from five sites on Malaysian Borneo. For each site we include the mean ± *SD* of the number of clusters returned by the algorithm, the per cent of observations correctly classified, silhouette coefficient and normalized mutual information index (NMI)

| Site | Number of calls | Clustering method | Number of clusters ($M \pm SD$) | Percent correct ($M \pm SD$) | Silhouette ($M \pm SD$) | NMI ($M \pm SD$) |
|---|---|---|---|---|---|---|
| Deramakot ($N = 8$) | 94 | Affinity | 9.04 ± 1.20 | 70.65 ± 10.42 | 0.36 ± 0.03 | 0.77 ± 0.07 |
| | | Gaussian | 51.29 ± 8.45 | 13.21 ± 12.9 | 0.53 ± 0.06 | 0.15 ± 0.15 |
| | | *K*-medoids | 49.56 ± 7.18 | 10.43 ± 7.20 | 0.54 ± 0.06 | 0.12 ± 0.08 |
| Danum Valley ($N = 12$) | 159 | Affinity | 12.39 ± 1.43 | 66.46 ± 8.07 | 0.36 ± 0.03 | 0.74 ± 0.08 |
| | | Gaussian | 19.24 ± 9.15 | 57.99 ± 17.08 | 0.33 ± 0.06 | 0.72 ± 0.18 |
| | | *K*-medoids | 47.96 ± 14.56 | 23.12 ± 2.90 | 0.49 ± 0.04 | 0.27 ± 0.03 |
| Imbak Canyon ($N = 8$) | 125 | Affinity | 8.91 ± 0.91 | 62.84 ± 12.41 | 0.29 ± 0.03 | 0.75 ± 0.07 |
| | | Gaussian | 15.70 ± 10.79 | 55.57 ± 19.96 | 0.28 ± 0.08 | 0.73 ± 0.2 |
| | | *K*-medoids | 34.26 ± 24.33 | 20.29 ± 26.03 | 0.5 ± 0.03 | 0.14 ± 0.17 |
| Maliau Basin ($N = 3$) | 52 | Affinity | 6.28 ± 1.46 | 51.27 ± 22.48 | 0.33 ± 0.06 | 0.69 ± 0.15 |
| | | Gaussian | 34.86 ± 2.64 | 5.15 ± 4.54 | 0.55 ± 0.08 | 0.09 ± 0.09 |
| | | *K*-medoids | 32.52 ± 5.39 | 0.91 ± 4.42 | 0.56 ± 0.06 | 0.02 ± 0.11 |
| Kalabakan ($N = 22$) | 275 | Affinity | 23.90 ± 1.73 | 64.71 ± 7.98 | 0.33 ± 0.02 | 0.77 ± 0.05 |
| | | Gaussian | 36.62 ± 8.60 | 45.21 ± 10.11 | 0.31 ± 0.03 | 0.62 ± 0.11 |
| | | *K*-medoids | 50.23 ± 4.99 | 37.85 ± 6.05 | 0.38 ± 0.02 | 0.48 ± 0.08 |

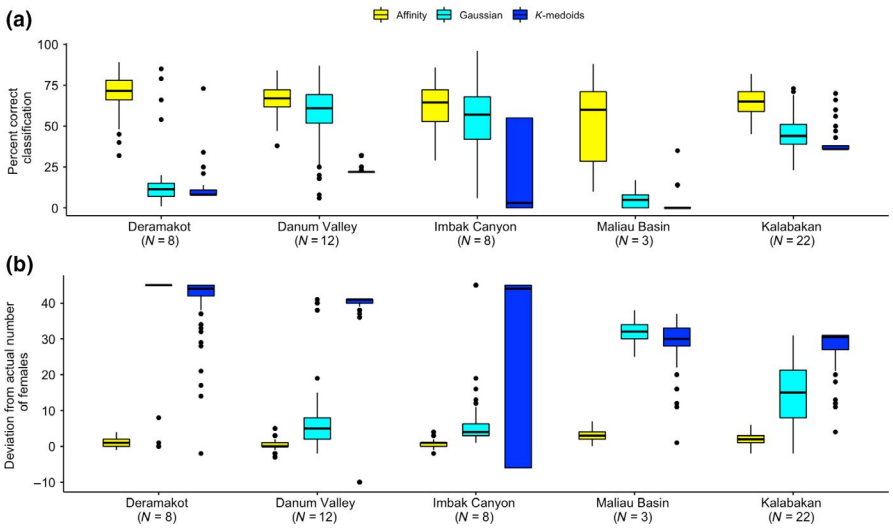


**FIGURE 4** A comparison of bootstrapped unsupervised clustering results across five sites in Malaysian Borneo. We randomly chose 80% of the calls from each site over 100 iterations and ran each of the three unsupervised clustering algorithms. The top panel (a) shows the per cent of observations that were correctly classified, and the bottom panel (b) shows the deviation of the predicted number of clusters (individual females) from the actual number of females recorded. The boxplots show the median, first and third percentiles and range of the data for each site and algorithm

affinity propagation clustering will be particularly useful for monitoring of vocal animals that have relatively stereotyped and individually distinct calls; further studies on taxa with diverse signal types will be informative. Importantly, future applications that combine data on individual identity collected via PAM with spatially explicit capture–recapture density estimation models (Kidney et al., 2016; Stevenson et al., 2015) have the possibility to revolutionize how we monitor populations of gibbons and other vocal animals.

Affinity propagation clustering has been shown to outperform other unsupervised algorithms in diverse applications including unsupervised classification of images (Dueck & Frey, 2007), high-dimensional gene expression data (Kiddle et al., 2010) and functional magnetic resonance imaging data (Zhang et al., 2011). The superior performance of affinity propagation clustering can be attributed to the following. First, affinity propagation clustering utilizes a message-passing algorithm (Frey & Dueck, 2007) that allows it to consider all data points as exemplars simultaneously. This is in contrast to both *K*-medoids (Madhulatha, 2011) and Gaussian mixture model-based clustering (Shireman et al., 2017) wherein initial starting exemplars must be chosen. These algorithms are sensitive to initial starting points, and tend to work better when number of clusters is small and the initial selection is close to a good clustering solution; increasing the number of initializations generally leads to only slight improvements of performance (Dueck, 2009). Second, affinity propagation clustering does not require the user to input a predetermined number of clusters. In most unsupervised clustering applications the number of groups or classes is not known a priori, so specifying the number of clusters beforehand is not ideal. There

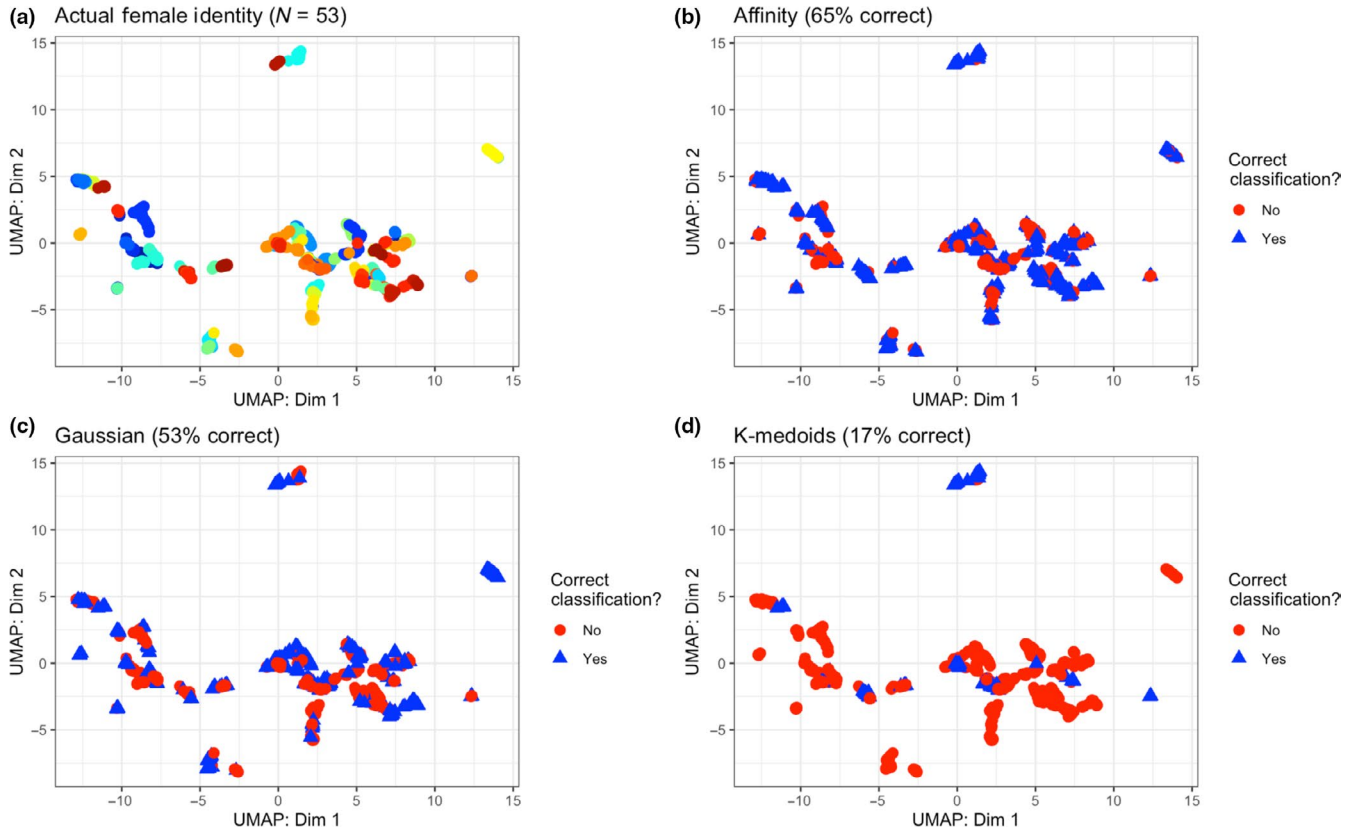


**FIGURE 5** UMAP projections for all gibbon female calls included in our dataset (*N* females = 53; *N* calls = 883). Each point represents a two-dimensional embedding of a single call. For plot (a), the colour of the points represents an individual female, and for plots (b)–(d) the colour and shape indicate whether the indicated unsupervised classification algorithm assigned that call to the correct female

are ways to address this limitation (e.g. loop over a set number of clusters and choose a cluster solution based on the best internal validation score) but this approach is computationally costly, and often leads to suboptimal results (this study). Lastly, the fact that the affinity propagation clustering algorithm takes a similarity matrix as an input means that it is broadly generalizable to many different types of datasets, and particularly for the multivariate datasets that are common in bioacoustics and PAM applications (Dueck, 2009).

For the present study, we chose to use MFCCs because our previous work indicated that these features lead to a higher supervised classification accuracy of gibbon females than features extracted from the spectrogram (Clink, Crofoot, & Clink, Crofoot, & Marshall, 2018). It seems likely that other types of feature extraction methods [e.g. automated estimation of spectral and temporal features (Araya-Salas & Smith-Vidaurre, 2017) or convolutional neural network feature embeddings (Sethi et al., 2020; Simonyan & Zisserman, 2014)] should work with affinity propagation clustering, assuming there is sufficient inter-class variation in the signals of interest. Importantly, our analysis focused on female gibbon calls which are highly stereotyped, yet individually distinct calls that can be effectively classified with high accuracy using supervised methods. Oftentimes, acoustic signals of interest are highly variable within individuals of the same call type [e.g. elephant rumbles (Hedwig et al., 2019); male gibbon solos (Clink et al., 2020a); humpback whale song (Noad et al., 2000)]. Although affinity propagation clustering has been shown to be effective in a variety of different applications (see above), it is unclear how it will perform on calls with levels of higher intra-individual variability. Future studies that use affinity propagation clustering on less stereotyped signals will be informative.

A caveat regarding our results is that we used calls taken from focal recordings, which means that the calls are relatively high quality (SNR > 10 dB) compared to many that would be captured using autonomous recorders deployed at fixed locations. Like all methods of feature extraction, the ability to use MFCCs for classification will be influenced by the signal-to-noise ratio (SNR) (Spillmann et al., 2017). MFCCs may be less robust to changes in SNR than other types of feature extraction such as estimating more noise-robust features from the spectrogram, e.g. Mellinger & Bradbury, 2007). However, the detection distance of a signal (and subsequent ability to use that signal for classification) will depend on many factors (such as source level of the calling animal, propagation loss and topography). Therefore, the detection distance and the acceptable cut-off for the SNR of a signal that can be used for classification needs to be determined empirically. As automated detectors are influenced by SNR, it may be possible to determine some threshold or SNR-cut-off in which the ability to detect and classify individuals is relatively high. In the future, we hope to test how the choice of features, along with quality of the recordings, influences the performance of affinity propagation clustering.

Interestingly, in many cases, when the algorithm(s) misclassified a female, the call was classified as another female that was recorded in

close spatial proximity to the actual female. There are a few possible explanations for this. Although we know little about dispersal in gibbons, what we do know indicates that gibbons tend not to disperse far from their natal home ranges (Matsudaira et al., 2018). There is evidence that closely related primates have similar call features (Kessler et al., 2012; Levréro et al., 2015). Therefore, it is possible that females that were recorded in close spatial proximity are closely related, and that either due to genetics, behavioural drift (Mundinger, 1980) or learning (Koda et al., 2013) the calls of neighbouring females were similar enough in some cases to be included in the same cluster. Another possibility is that during data collection, as recordings were taken from wild, unhabituated gibbons, there were particular instances in which a single female was recorded on two separate occasions and classified as two different females. Gibbons are territorial, and for censusing, it is generally agreed that groups that are mapped >500 m apart should be considered separate groups (Brockelman & Srikosamatara, 1993). However, there is substantial variation in documented home-range size (Bartlett et al., 2016; Cheyne et al., 2019), which means that in some cases, animals that were recorded >500 m on different days may be the same animal.

There still remain outstanding questions that need to be answered before unsupervised clustering can be fully utilized for PAM applications, and the answers are likely to be taxa specific. First, the few studies that have investigated temporal stability in gibbon individual vocal signatures indicate that signatures are stable over time (>2 years; Fan et al., 2011, Feng et al., 2014), but there is evidence that temporal and spectral features of gibbon calls change with age (Terleph et al., 2016). Therefore, it is essential to determine the stability of vocal signatures relative to the duration of the PAM study. Second, the results of unsupervised clustering need to be 'ground-truthed' using PAM data where the identity (and timing) of calling individuals is known. Although these data are difficult to collect (hence the reason for wanting to use PAM), it will be crucial to validate PAM data with data labeled by a human observer. Lastly, although we show that affinity propagation clustering substantially outperformed the other two algorithms, there was still error in the predicted versus actual number of females, as well as in the per observation classification. The acceptable amount of error in unsupervised classification for PAM applications will depend on the research question, and for cases wherein assumptions regarding individual identity allow little room for error, these approaches may not be appropriate.

Female gibbon calls exhibit a higher degree of vocal individuality than male gibbon calls (Clink et al., 2020a; Lau et al., 2018), which makes them more suitable for these types of classification problems. But, focusing on female calls also has relevance for density estimation, as female gibbons generally call only if they are in a mated pair, whereas males will call irrespective of their mated status (Brockelman & Srikosamatara, 1993). Gibbon acoustic surveys rely on the female call to indicate the presence of a gibbon group and necessarily report group (as opposed to individual) density (Hamard et al., 2010; Kidney et al., 2016; Yin et al., 2016). We propose that for gibbons—which have individually distinct calls (Clink et al., 2017; Clink, Crofoot, & Marshall, 2018; Clink, Grote, et al., 2018), occur at relatively low densities (Brockelman & Srikosamatara, 1993; Kidney et al., 2016) and are highly territorial (Mitani, 1984)—unsupervised clustering via affinity propagation clustering may be a useful addition for density estimation using PAM. In addition, future applications that investigate individual turnover in disturbed habitats may also benefit from unsupervised classification of individuals.

## 5 | CONCLUSIONS

Given the current biodiversity conservation crisis (Barnosky et al., 2011), effective monitoring of the status and trends of populations is critical for conservation efforts across taxa. We show that unsupervised clustering may provide an important additional tool for the monitoring of endangered, vocal animals. Whether unsupervised clustering of individuals can or should be incorporated into PAM programmes depends on a multitude of factors including the number of recording units and spacing of the array, along with the behavioural ecology of the target animals, including source level of the calling animal, home-range size of the animal relative to the array spacing, tendency for animals to call at the same time and whether the focal animals are territorial. And most importantly, the degree of vocal individuality and ability to distinguish between individuals will dictate whether this approach is appropriate.

### ACKNOWLEDGEMENTS

We would like to acknowledge Marie A. Roch for her helpful comments on earlier versions of this manuscript. We thank Sarab S. Sethi for insightful conversations about affinity propagation clustering. We gratefully acknowledge Sabah Biodiversity Centre for providing permission for data collection for this project, and we thank Abdul Hamid Ahmad for serving as our local collaborator. D.J.C. thanks the Fulbright U.S. Student Program for funding field work that allowed for collection of data for this project.

### CONFLICT OF INTEREST

The authors confirm that there are no potential conflicts of interest.

### AUTHORS' CONTRIBUTIONS

D.J.C. collected the data for previous work. Both authors contributed to analysis and interpretation of the data, along with manuscript writing.

### ETHICS STATEMENT

All data used in the present study come from previously published sources (Clink, Charif, et al., 2018; Clink, Grote, et al., 2018). Approval to the collected data was granted by the Sabah Biodiversity Council (access license number: JKM/MBS 1000-2/2 JLD.3 (42)) and data were collected in accordance with the University of California, Davis, Institutional Animal Care and Use Committee (IACUC) Protocol 29-30.

### PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13520.

## DATA AVAILABILITY STATEMENT

All data and R code needed to recreate analyses are provided on Dryad Digital Repository via https://datadryad.org/stash/dataset/doi:10.5061/dryad.sbcc2fr4p (Clink & Klinck, 2020). Raw sound files are available upon request to the corresponding author.

## ORCID

*Dena J. Clink* https://orcid.org/0000-0003-0363-5581
*Holger Klinck* https://orcid.org/0000-0003-1078-7268

## REFERENCES

Araya-Salas, M., & Smith-Vidaurre, G. (2017). warbleR: An R package to streamline analysis of animal acoustic signals. *Methods in Ecology and Evolution*, *8*(2), 184–191.

Augustine, B. C., Royle, J. A., Kelly, M. J., Satter, C. B., Alonso, R. S., Boydston, E. E., & Crooks, K. R. (2018). Spatial capture–recapture with partial identity: An application to camera traps. *The Annals of Applied Statistics*, *12*(1), 67–95. https://doi.org/10.1214/17-AOAS1091

Augustine, B. C., Royle, J. A., Murphy, S. M., Chandler, R. B., Cox, J. J., & Kelly, M. J. (2019). Spatial capture–recapture for categorically marked populations with an application to genetic capture–recapture. *Ecosphere*, *10*(4), e02627. https://doi.org/10.1002/ecs2.2627

Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K. H., & Frommolt, K. H. (2010). Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters*, *31*(12), 1524–1534. https://doi.org/10.1016/j.patrec.2009.09.014

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O. U., Swartz, B., Quental, T. B., Marshall, C., McGuire, J. L., Lindsey, E. L., Maguire, K. C., Mersey, B., & Ferrer, E. A. (2011). Has the Earth's sixth mass extinction already arrived? *Nature*, *471*(7336), 51–57.

Barreda, S. (2015). *phonTools: Functions for phonetics in R*. R Package Version 0.2-2.1.

Bartlett, T. Q., Light, L. E. O., & Brockelman, W. Y. (2016). Long-term home range use in white-handed gibbons (*Hylobates lar*) in Khao Yai National Park. Thailand. *American Journal of Primatology*, *78*(2), 192–203. https://doi.org/10.1002/ajp.22492

Bezdek, J. C. (1974). Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology*, *1*(1), 57–71. https://doi.org/10.1007/BF02339490

Bodenhofer, U., Kothmeier, A., & Hochreiter, S. (2011). APCluster: An R package for affinity propagation clustering. *Bioinformatics*, *27*(17), 2463–2464. https://doi.org/10.1093/bioinformatics/btr406

Brockelman, W. Y., & Srikosamatara, S. (1993). Estimation of density of gibbon groups by use of loud songs. *American Journal of Primatology*, *29*(2), 93–108. https://doi.org/10.1002/ajp.1350290203

Brusco, M. J., Steinley, D., Stevens, J., & Cradit, J. D. (2019). Affinity propagation: An exemplar-based tool for clustering in psychological research. *British Journal of Mathematical and Statistical Psychology*, *72*(1), 155–182. https://doi.org/10.1111/bmsp.12136

Cheyne, S. M., Capilla, B. R., Abdulaziz, K., Supiansyah, Adul, Cahyaningrum, E., & Smith, D. E. (2019). Home range variation and site fidelity of Bornean southern gibbons [*Hylobates albibarbis*] from 2010–2018. *PLoS ONE*, *14*(7), 1–13. https://doi.org/10.1371/journal.pone.0217784

Chiquet, J., & Rigaill, G. (2019). *aricode: Efficient computations of standard clustering comparison measures*. Retrieved from https://cran.r-project.org/package=aricode

Clink, D. J., Bernard, H., Crofoot, M. C., & Marshall, A. J. (2017). Investigating individual vocal signatures and small-scale patterns of geographic variation in female Bornean Gibbon (*Hylobates muelleri*) great calls. *International Journal of Primatology*, *38*(4), 656–671. https://doi.org/10.1007/s10764-017-9972-y

Clink, D. J., Charif, R. A., Crofoot, M. C., & Marshall, A. J. (2018). Evidence of vocal performance constraints in a female non-human primate. *Animal Behaviour*, *141*, 85–94. https://doi.org/10.1016/j.anbehav.2018.05.002

Clink, D. J., Crofoot, M. C., & Marshall, A. J. (2018). Application of a semi-automated vocal fingerprinting approach to monitor Bornean gibbon females in an experimentally fragmented landscape in Sabah, Malaysia. *Bioacoustics*, *28*, 193–209. https://doi.org/10.1080/09524622.2018.1426042

Clink, D. J., Grote, M. N., Crofoot, M. C., & Marshall, A. J. (2018). Understanding sources of variance and correlation among features of Bornean gibbon (*Hylobates muelleri*) female calls. *Journal of the Acoustical Society of America*, *144*, 698–708.

Clink, D. J., Hamid Ahmad, A., & Klinck, H. (2020a). Brevity is not a universal in animal communication: Evidence for compression depends on the unit of analysis in small ape vocalizations. *Royal Society Open Science*, *7*(4), 200151. https://doi.org/10.1098/rsos.200151

Clink, D. J., Hamid Ahmad, A., & Klinck, H. (2020b). Gibbons aren't singing in the rain: Presence and amount of rainfall influences ape calling behavior in Sabah, Malaysia. *Scientific Reports*, *10*(1), 1282. https://doi.org/10.1038/s41598-020-57976-x

Clink, D. J., & Klinck, H. (2020). Data from: Unsupervised acoustic classification of individual gibbon females and the implications for passive acoustic monitoring. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.sbcc2fr4p

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Darras, K., Batáry, P., Furnas, B. J., Grass, I., Mulyani, Y. A., & Tscharntke, T. (2019). Autonomous sound recording outperforms human observation for sampling birds: A systematic map and user guide. *Ecological Applications*, *29*(6), e01954. https://doi.org/10.1002/eap.1954

Davis, G. E., Baumgartner, M. F., Corkeron, P. J., Bell, J., Berchok, C., Bonnell, J. M., Bort Thornton, J., Brault, S., Buchanan, G. A., Cholewiak, D. M., Clark, C. W., Delarue, J., Hatch, L. T., Klinck, H., Kraus, S. D., Martin, B., Mellinger, D. K., Moors-Murphy, H., Nieukirk, S., … Van Parijs, S. M. (2020). Exploring movement patterns and changing distributions of baleen whales in the western North Atlantic using a decade of passive acoustic data. *Global Change Biology*, *26*(9), 4812–4840. https://doi.org/10.1111/gcb.15191

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366. https://doi.org/10.1109/TASSP.1980.1163420

Dawson, D. K., & Efford, M. G. (2009). Bird population density estimated from acoustic signals. *Journal of Applied Ecology*, *46*(6), 1201–1209. https://doi.org/10.1111/j.1365-2664.2009.01731.x

Dinov, I. D. (Ed.). (2018). k-Means clustering. In *Data science and predictive analytics: Biomedical and health applications using R* (pp. 443–473). Springer International Publishing. https://doi.org/10.1007/978-3-319-72347-1_13

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis* (Vol. 3). Wiley.

Dueck, D. (2009). *Affinity propagation: Clustering data by passing messages* (Dissertation). University of Toronto, Toronto.

Dueck, D., & Frey, B. J. (2007). Non-metric affinity propagation for unsupervised image categorization. In *2007 IEEE 11th International Conference on Computer Vision* (pp. 1–8). IEEE.

Enari, H., Enari, H., Okuda, K., Yoshita, M., Kuno, T., & Okuda, K. (2017). Feasibility assessment of active and passive acoustic monitoring of sika deer populations. *Ecological Indicators*, *79*, 155–162. https://doi.org/10.1016/j.ecolind.2017.04.004

Fan, P.-F., Xiao, W., Feng, J.-J., & Scott, M. B. (2011). Population differences and acoustic stability in male songs of wild western black crested gibbons (*Nomascus concolor*) in Mt. Wuliang, Yunnan. *Folia Primatologica*, *82*(2), 83–93. https://doi.org/10.1159/000329128

Feng, J.-J., Cui, L.-W., Ma, C.-Y., Fei, H.-L., & Fan, P.-F. (2014). Individuality and stability in male songs of cao vit gibbons (*Nomascus nasutus*) with potential to monitor population dynamics. *PLoS ONE*, *9*(5), e96317. https://doi.org/10.1371/journal.pone.0096317

Fournet, M., Matthews, L., Gabriele, C., Haver, S., Mellinger, D., & Klinck, H. (2018). Humpback whales Megaptera novaeangliae alter calling behavior in response to natural sounds and vessel noise. *Marine Ecology Progress Series*, *607*, 251–268. https://doi.org/10.3354/meps12784

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*(458), 611–631. https://doi.org/10.1198/016214502760047131

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972–976. https://doi.org/10.1126/science.1136800

Gamba, M., Friard, O., Riondato, I., Righini, R., Colombo, C., Miaretsoa, L., Torti, V., Nadhurou, B., & Giacoma, C. (2015). Comparative analysis of the vocal repertoire of Eulemur: A dynamic time warping approach. *International Journal of Primatology*, *36*(5), 894–910. https://doi.org/10.1007/s10764-015-9861-1

Geissmann, T. (2002). Duet-splitting and the evolution of gibbon songs. *Biological Reviews*, *77*(1), 57–76. https://doi.org/10.1017/S1464793101005826

Greene, D., Cunningham, P., & Mayer, R. (2008). Unsupervised learning and clustering. In M. Cord & P. Cunningham (Eds.), *Machine learning techniques for multimedia*. Cognitive Technologies (pp. 51–90). Springer. https://doi.org/10.1007/978-3-540-75171-7_3

Hamard, M., Cheyne, S. M., & Nijman, V. (2010). Vegetation correlates of gibbon density in the peat-swamp forest of the Sabangau catchment, Central Kalimantan, Indonesia. *American Journal of Primatology*, *72*(7), n/a–n/a. https://doi.org/10.1002/ajp.20815

Hamerly, G., & Drake, J. (2015). Accelerating Lloyd's algorithm for k-means clustering. In M. Celebi (Ed.), *Partitional clustering algorithms* (pp. 41–78). Springer. https://doi.org/10.1007/978-3-319-09259-1_2

Hammerschmidt, K., & Fischer, J. (1998). The vocal repertoire of Barbary macaques: A quantitative analysis of a graded signal system. *Ethology*, *104*(3), 203–216. https://doi.org/10.1111/j.1439-0310.1998.tb00063.x

Hammerschmidt, K., & Fischer, J. (2019). Baboon vocal repertoires and the evolution of primate vocal diversity. *Journal of Human Evolution*, *126*, 1–13. https://doi.org/10.1016/j.jhevol.2018.10.010

Hedwig, D., Hammerschmidt, K., Mundry, R., Robbins, M. M., & Boesch, C. (2014). Acoustic structure and variation in mountain and western gorilla close calls: A syntactic approach. *Behaviour*, *151*(8), 1091–1120.

Hedwig, D., Verahrami, A. K., & Wrege, P. H. (2019). Acoustic structure of forest elephant rumbles: A test of the ambiguity reduction hypothesis. *Animal Cognition*, *22*(6), 1115–1128. https://doi.org/10.1007/s10071-019-01304-y

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

Kalan, A. K., Mundry, R., Wagner, O. J. J., Heinicke, S., Boesch, C., & Kühl, H. S. (2015). Towards the automated detection and occupancy estimation of primates using passive acoustic monitoring. *Ecological Indicators*, *54*, 217–226

Kalan, A. K., Piel, A. K., Mundry, R., Wittig, R. M., Boesch, C., & Kühl, H. S. (2016). Passive acoustic monitoring reveals group ranging and territory use: A case study of wild chimpanzees (*Pan troglodytes*). *Frontiers in Zoology*, *13*(1), 34. https://doi.org/10.1186/s12983-016-0167-8

Kassambara, A. (2017). *ggpubr: 'ggplot2' based publication ready plots*. R package version 0.1.

Katz, J., Hafner, S. D., & Donovan, T. (2016). Tools for automated acoustic monitoring within the R package monitoR. *Bioacoustics*, *25*(2), 197–210. https://doi.org/10.1080/09524622.2016.1138415

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

Keen, S. C., Shiu, Y., Wrege, P. H., & Rowland, E. D. (2017). Automated detection of low-frequency rumbles of forest elephants: A critical tool for their conservation. *The Journal of the Acoustical Society of America*, *141*(4), 2715–2726. https://doi.org/10.1121/1.4979476

Keenan, S., Lemasson, A., & Zuberbühler, K. (2013). Graded or discrete? A quantitative analysis of Campbell's monkey alarm calls. *Animal Behaviour*, *85*(1), 109–118. https://doi.org/10.1016/j.anbehav.2012.10.014

Kershenbaum, A., Sayigh, L. S., & Janik, V. M. (2013). The encoding of individual identity in dolphin signature whistles: How much information is needed? *PLoS ONE*, *8*(10), e77671. https://doi.org/10.1371/journal.pone.0077671

Kessler, S. E., Scheumann, M., Nash, L. T., & Zimmermann, E. (2012). Paternal kin recognition in the high frequency/ultrasonic range in a solitary foraging mammal. *BMC Ecology*, *12*(1), 26. https://doi.org/10.1186/1472-6785-12-26

Kiddle, S. J., Windram, O. P. F., McHattie, S., Mead, A., Beynon, J., Buchanan-Wollaston, V., Denby, K. J., & Mukherjee, S. (2010). Temporal clustering by affinity propagation reveals transcriptional modules in Arabidopsis thaliana. *Bioinformatics*, *26*(3), 355–362. https://doi.org/10.1093/bioinformatics/btp673

Kidney, D., Rawson, B. M., Borchers, D. L., Stevenson, B. C., Marques, T. A., & Thomas, L. (2016). An efficient acoustic density estimation method with human detectors applied to gibbons in Cambodia. *PLoS ONE*, *11*(5), e0155066. https://doi.org/10.1371/journal.pone.0155066

Kirschel, A. N. G., Earl, D. A., Yao, Y., Escobar, I. A., Vilches, E., Vallejo, E. E., & Taylor, C. E. (2009). Using songs to indentify individual Mexican antthrush *Formicarius moniliger*: Comparison of four classification methods. *Bioacoustics*, *19*, 1–20.

Koda, H., Lemasson, A., Oyakawa, C., Rizaldi, A. A. A., Pamungkas, J., & Masataka, N. (2013). Possible role of mother-daughter vocal interactions on the development of species-specific song in gibbons. *PLoS ONE*, *8*(8), e71432. https://doi.org/10.1371/journal.pone.0071432

Kumar, K., Kim, C., & Stern, R. M. (2011). Delta-spectral cepstral coefficients for robust speech recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings* (pp.4784–4787).IEEE.https://doi.org/10.1109/ICASSP.2011.5947425

Lau, A., Clink, D. J., Crofoot, M. C., & Marshall, A. J. (2018). Evidence for high variability in temporal features of the male coda in Müller's Bornean Gibbons (*Hylobates muelleri*). *International Journal of Primatology*, *39*(4), 670–684. https://doi.org/10.1007/s10764-018-0061-7

Leliveld, L. M. C., Scheumann, M., & Zimmermann, E. (2011). Acoustic correlates of individuality in the vocal repertoire of a nocturnal primate (*Microcebus murinus*). *The Journal of the Acoustical Society of America*, *129*(4), 2278–2288. https://doi.org/10.1121/1.3559680

Levréro, F., Carrete-Vega, G., Herbert, A., Lawabi, I., Courtiol, A., Willaume, E., Kappeler, P. M., & Charpentier, M. (2015). Social shaping of voices does not impair phenotype matching of kinship in mandrills. *Nature Communications*, *6*(May), 7609. https://doi.org/10.1038/ncomms8609

Ligges, U., Krey, S., Mersmann, O., & Schnackenberg, S. (2016). *{tuneR}: Analysis of music*. Retrieved from http://r-forge.r-project.org/projects/tuner/

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (Vol. 233, pp. 281–297).

Madhulatha, T. S. (2011). Comparison between K-means and K-medoids clustering algorithms. In D. C. Wyld, M. Wozniak, N. Chaki, N. Meghanathan, & D. Nagamalai (Eds.), *Advances in computing and information technology* (pp. 472–481). Springer, Berlin Heidelberg.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2019). *cluster: Cluster analysis basics and extensions*. R package version 2.1.0.

Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., Harris, D., & Tyack, P. L. (2013). Estimating animal

population density using passive acoustics. *Biological Reviews*, *88*(2), 287–309. https://doi.org/10.1111/brv.12001

Martin, S. W., Marques, T. A., Thomas, L., Morrissey, R. P., Jarvis, S., DiMarzio, N., Moretti, D., & Mellinger, D. K. (2013). Estimating minke whale (*Balaenoptera acutorostrata*) boing sound density using passive acoustic sensors. *Marine Mammal Science*, *29*(1), 142–158. https://doi.org/10.1111/j.1748-7692.2011.00561.x

Matsudaira, K., Ishida, T., Malaivijitnond, S., & Reichard, U. H. (2018). Short dispersal distance of males in a wild white-handed gibbon (*Hylobates lar* ) population. *American Journal of Physical Anthropology*, *167*(1), 61–71. https://doi.org/10.1002/ajpa.23603

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv Preprint*, ArXiv:1802.03426

Mellinger, D. K., & Bradbury, J. W. (2007). Acoustic measurement of marine mammal sounds in noisy environments. In *Proceedings of the International Conference on underwater acoustical measurements: Technologies and results* (pp. 273–280).

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2017). *e1071: Misc functions of the department of statistics*. Probability Theory Group.

Mielke, A., & Zuberbühler, K. (2013). A method for automated individual, species and call type recognition in free-ranging animals. *Animal Behaviour*, *86*(2), 475–482. https://doi.org/10.1016/j.anbehav.2013.04.017

Mitani, J. C. (1984). The behavioral regulation of monogamy in gibbons (*Hylobates muelleri*). *Behavioral Ecology and Sociobiology*, *15*(3), 225–229. https://doi.org/10.1007/BF00292979

Mitani, J. C., Gros-Louis, J., & Macedonia, J. M. (1996). Selection for acoustic individuality within the vocal repertoire of wild chimpanzees. *International Journal of Primatology*, *17*(4), 569–583. https://doi.org/10.1007/BF02735192

Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using Mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *Journal of Computing*, *2*(3), 2151–9617.

Mundinger, P. (1980). Animal cultures and a general theory of cultural evolution. *Ethology and Sociobiology*, *1*(3), 183–223. https://doi.org/10.1016/0162-3095(80)90008-4

Noad, M. J., Cato, D. H., Bryden, M. M., Jenner, M.-N., & Jenner, K. C. S. (2000). Cultural revolution in whale songs. *Nature*, *408*(6812), 537. https://doi.org/10.1038/35046199

Parra-Hernández, R. M., Posada-Quintero, J. I., Acevedo-Charry, O., & Posada-Quintero, H. F. (2020). Uniform manifold approximation and projection for clustering taxa through vocalizations in a neotropical passerine (Rough-Legged Tyrannulet, *Phyllomyias burmeisteri*). *Animals*, *10*(8), 1406. https://doi.org/10.3390/ani10081406

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine*, *12*, 2825–2830.

Pozzi, L., Gamba, M., & Giacoma, C. (2009). The use of Artificial Neural Networks to classify primate vocalizations: A pilot study on black lemurs. *American Journal of Primatology*, *72*(4), n/a–n/a. https://doi.org/10.1002/ajp.20786

Price, T., Wadewitz, P., Cheney, D., Seyfarth, R., Hammerschmidt, K., & Fischer, J. (2015). Vervets revisited: A quantitative analysis of alarm call structure and context specificity. *Scientific Reports*, *5*, 13220. https://doi.org/10.1038/srep13220

R Core Team. (2019). *R: A language and environment for statistical computing*. Austria. Retrieved from https://www.r-project.org/

Rendall, D. (2003). Acoustic correlates of caller identity and affect intensity in the vowel-like grunt vocalizations of baboons. *The Journal of the Acoustical Society of America*, *113*(6), 3390. https://doi.org/10.1121/1.1568942

Rendall, D., Rodman, P. S., & Emond, R. E. (1996). Vocal recognition of individuals and kin in free-ranging rhesus monkeys. *Animal Behaviour*, *51*(5), 1007–1015. https://doi.org/10.1006/anbe.1996.0103

Reynolds, A. P., Richards, G., de la Iglesia, B., & Rayward-Smith, V. J. (2006). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, *5*(4), 475–504. https://doi.org/10.1007/s10852-005-9022-1

Riondato, I., Cissello, E., Papale, E., Friard, O., Gamba, M., & Giacoma, C. (2017). Unsupervised acoustic analysis of the vocal repertoire of the gray-shanked Douc Langur (*Pygathrix cinerea*). *Journal of Computational Acoustics*, *25*(03), 1750018.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sethi, S. S., Jones, N. S., Fulcher, B. D., Picinali, L., Clink, D. J., Klinck, H., Orme, D., Wrege, P. H., & Ewers, R. M. (2020). Characterising soundscapes across diverse ecosystems using a universal acoustic feature-set. *Proceedings of the National Academy of Sciences of the United States of America*, *117*, 17049–17055.

Shireman, E., Steinley, D., & Brusco, M. J. (2017). Examining the effect of initialization strategies on the performance of Gaussian mixture modeling. *Behavior Research Methods*, *49*(1), 282–293. https://doi.org/10.3758/s13428-015-0697-6

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint*, ArXiv:1409.1556.

Spillmann, B., van Noordwijk, M. A., Willems, E. P., Mitra Setia, T., Wipfli, U., & van Schaik, C. P. (2015). Validation of an acoustic location system to monitor Bornean orangutan (*Pongo pygmaeus wurmbii*) long calls. *American Journal of Primatology*, *77*(7), 767–776. https://doi.org/10.1002/ajp.22398

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 205–233.

Spillmann, B., van Schaik, C. P., Setia, T. M., & Sadjadi, S. O. (2017). Who shall I say is calling? Validation of a caller recognition procedure in Bornean flanged male orangutan (*Pongo pygmaeus wurmbii*) long calls. *Bioacoustics*, *26*(2), 109–120. https://doi.org/10.1080/09524622.2016.1216802

Stevenson, B. C., Borchers, D. L., Altwegg, R., Swift, R. J., Gillespie, D. M., & Measey, G. J. (2015). A general framework for animal density estimation from acoustic detections across a fixed microphone array. *Methods in Ecology and Evolution*, *6*(1), 38–48. https://doi.org/10.1111/2041-210X.12291

Sugai, L. S. M., Silva, T. S. F., Ribeiro, J. W., & Llusia, D. (2019). Terrestrial passive acoustic monitoring: Review and perspectives. *BioScience*, *69*(1), 15–25. https://doi.org/10.1093/biosci/biy147

Tan, P.-N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education India.

Terleph, T. A., Malaivijitnond, S., & Reichard, U. H. (2015). Lar gibbon (*Hylobates lar*) great call reveals individual caller identity. *American Journal of Primatology*, *821*, 811–821. https://doi.org/10.1002/ajp.22406

Terleph, T. A., Malaivijitnond, S., & Reichard, U. H. (2016). Age related decline in female lar gibbon great call performance suggests that call features correlate with physical condition. *BMC Evolutionary Biology*, *16*(1), 4. https://doi.org/10.1186/s12862-015-0578-8

Terry, A. M., Peake, T. M., & McGregor, P. K. (2005). The role of vocal individuality in conservation. *Frontiers in Zoology*, *2*(1), 10. https://doi.org/10.1186/1742-9994-2-10

Turesson, H. K., Ribeiro, S., Pereira, D. R., Papa, J. P., & De Albuquerque, V. H. C. (2016). Machine learning algorithms for automatic classification of marmoset vocalizations. *PLoS ONE*, *11*(9), e0163041. https://doi.org/10.1371/journal.pone.0163041

Valente, D., De Gregorio, C., Torti, V., Miaretsoa, L., Friard, O., Randrianarison, R. M., Giacoma, C., & Gamba, M. (2019). Finding meanings in low dimensional structures: Stochastic neighbor

embedding applied to the analysis of *Indri indri* vocal repertoire. *Animals*, *9*(5), 243. https://doi.org/10.3390/ani9050243

Vu, T. T., & Tran, L. M. (2019). An application of autonomous recorders for gibbon monitoring. *International Journal of Primatology*, *40*(2), 169–186. https://doi.org/10.1007/s10764-018-0073-3

Wadewitz, P., Hammerschmidt, K., Battaglia, D., Witt, A., Wolf, F., & Fischer, J. (2015). Characterizing vocal repertoires—Hard vs. soft classification approaches. *PLoS ONE*, *10*(4), e0125785. https://doi.org/10.1371/journal.pone.0125785

Wang, K., Zhang, J., Li, D., Zhang, X., & Guo, T. (2008). *Adaptive affinity propagation clustering*. Retrieved from http://arxiv.org/abs/0805.1096

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.

Wrege, P. H., Rowland, E. D., Keen, S., & Shiu, Y. (2017). Acoustic monitoring for conservation in tropical forests: Examples from forest elephants. *Methods in Ecology and Evolution*, *8*(10), 1292–1301. https://doi.org/10.1111/2041-210X.12730

Yin, L. Y., Fei, H. L., Chen, G. S., Li, J. H., Cui, L. W., & Fan, P. F. (2016). Effects of group density, hunting, and temperature on the singing patterns of eastern hoolock gibbons (*Hoolock leuconedys*) in Gaoligongshan, Southwest China. *American Journal of Primatology*, *78*(8), 861–871. https://doi.org/10.1002/ajp.22553

Zeppelzauer, M., Hensman, S., & Stoeger, A. S. (2015). Towards an automated acoustic detection system for free-ranging elephants. *Bioacoustics*, *24*(1), 13–29. https://doi.org/10.1080/09524622.2014.906321

Zhang, J., Li, D., Chen, H., & Fang, F. (2011). Analysis of activity in fMRI data using affinity propagation clustering. *Computer Methods in Biomechanics and Biomedical Engineering*, *14*(03), 271–281. https://doi.org/10.1080/10255841003766829

**How to cite this article:** Clink DJ, Klinck H. Unsupervised acoustic classification of individual gibbon females and the implications for passive acoustic monitoring. *Methods Ecol Evol*. 2020;00:1–14. https://doi.org/10.1111/2041-210X.13520