ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2018)
67, *Part* 1, *pp.* 165–183

Bat echolocation call identification for biodiversity monitoring: a probabilistic approach

Vassilios Stathopoulos,

University of Warwick, Coventry, UK

Veronica Zamora-Gutierrez,

University of Cambridge, and University College London, UK, and CONACYT–Instituto Politécnico Nacional Centro Interdisciplinario de Investigación para el Desarrollo Integral Regional Unidad Durango, Mexico

Kate E. Jones

University College London and Zoological Society of London, UK

and Mark Girolami

Imperial College London and Alan Turing Institute for Data Science, London, UK

[Received August 2015. Final revision January 2017]

Summary. Bat echolocation call identification methods are important in developing efficient cost-effective methods for large-scale bioacoustic surveys for global biodiversity monitoring and conservation planning. Such methods need to provide interpretable probabilistic predictions of species since they will be applied across many different taxa in a diverse set of applications and environments. We develop such a method using a multinomial probit likelihood with independent Gaussian process priors and study its feasibility on a data set from an on-going study of 21 species, five families and 1800 bat echolocation calls collected from Mexico, a hotspot of bat biodiversity. We propose an efficient approximate inference scheme based on the expectation propagation algorithm and observe that the overall methodology significantly improves on currently adopted approaches to bat call classification by providing an approach which can be easily generalized across different species and call types and is fully probabilistic. Implementation of this method has the potential to provide robust species identification tools for biodiversity acoustic bat monitoring programmes across a range of taxa and spatial scales.

Keywords: Acoustic monitoring; Approximate Bayesian inference; Classification; Gaussian processes

1. Introduction

In the face of severe declines in populations of many wildlife species (Butchart *et al.*, 2010; Tittensor *et al.*, 2014) monitoring changes in ecological communities through time and space is critically important for conservation planning and decision making (Magurran *et al.*, 2010). Bats (order *Chiroptera*) with over 1200 species are the second-largest order of mammals (Simmons, 2005). Bats are considered to be important indicators of wider environmental health as they play key roles in ecosystems, providing many ecosystem services such as pollination and

Address for correspondence: Vassilios Stathopoulos, Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK.
E-mail: stathv@gmail.com

 0035–9254/18/67165
Published by John Wiley & Sons Ltd on behalf of the Royal Statistical Society.

regulation of insect populations (Jones *et al*., 2009). Bats also can be monitored non-invasively by acoustic sensors as many species actively use calls and the interpretation of their echos to detect, localize and classify objects (echolocation) (Griffin, 1944). Acoustic monitoring schemes are therefore useful conservation tools to monitor and determine the effect of anthropogenic activity on biodiversity (Barlow *et al*., 2015; Newson *et al*., 2015; Jones *et al*., 2013; Amorim *et al*., 2014). However, there are many challenges to implementing such schemes, not least of which is the paucity of software tools to detect and classify bat calls to species (Walters *et al*., 2013). Scaling up existing schemes from local and regional levels to international and global scales requires substantial expansion in call reference libraries for echolocating species alongside more sophisticated methods for detecting and classifying species.

Although echolocation is used by other taxa, the complexity and diversity of the design of echolocation calls is unparalleled within bats. Bats echolocate by emitting frequencies between 9 and 212 kHz and show a considerable diversity in call design. Calls can generally be categorized by their duration, bandwidth and use of harmonics (Maltby *et al*., 2009). Some of the variation in calls can be explained by a shared evolutionary history (closely related species have similar call designs) (Jones and Teeling, 2006). Whereas other call variation is due to adaptation to particular tasks (for example call frequency is influenced by the size of objects to be detected), call duration by how far away objects are, and bandwidth (the range of the frequencies of the call) influences how well the bat species deals with extremely cluttered environments (Jones and Holderied, 2007; Maltby *et al*., 2009). However, their calls also show great intraspecific variation and flexibility caused by habitat, geography, sex, gender and age, and in some cases call structures designs greatly overlap between species which makes species identification very challenging (Murray *et al*., 2001; Schnitzler *et al*., 2003). Previous research on bat echolocation call identification has approached the problem from a data classification perspective. The most studied methods use call parameters extracted from spectrograms and then discriminant function analysis, support vector machines (SVMs) or artificial neural networks are employed for supervised classification (Walters *et al*., 2012; Parsons and Jones, 2000; Fenton and Bell, 1981). These existing classification tools typically cover a small set of species and point estimates for the model parameters are obtained by using high quality recordings from well-curated collections of bat calls. Such methods do not associate species identification probabilities and since point estimates are used they do not generalize well to bat calls recorded in a diverse set of environments. Other studies, e.g. Skowronski and Harris (2006), have used hidden Markov and Gaussian mixture models. Although such methods provide probabilistic outputs they are generative in nature, trying to model the distribution of the bat calls in the collection. Ng and Jordan (2002) have shown that classification methods modelling the discriminant function directly generalize better than methods modelling the distribution of the training data unless there is substantial domain knowledge, something that is the case in the speech recognition community but not for bat echolocation calls. There is an opportunity to tackle this problem by applying other methods from the statistics and machine learning literature, such as generalized linear models (McCullagh and Nelder, 1989), Gaussian processes (GPs) (Rasmussen and Williams, 2005) and random forests (Dietterich, 2000). However, there are several factors which need to be considered before choosing a particular methodology.

Firstly, it is important that the uncertainty of species identification is determined correctly as this has conservation and planning implications as well as being a useful input for a number of habitat suitability analyses, such as species distribution models (Dudik *et al*., 2007). Therefore the output of the method should assign a species identification probability for each new sample to allow researchers to control accuracy by setting acceptance thresholds. Moreover, models that are trained on a call reference library with a data set of high quality recordings will have to

generalize to noisy unstructured data collected under a diverse set of environments. To achieve this any method should quantify and take into account the uncertainty that is associated with model parameters estimated on a specific data set when making predictions. Generalized linear models and GPs are theoretically well understood with well-studied methods for parameter inference by using Markov chain Monte Carlo (MCMC) (Dobson and Barnett, 2008) and approximate Bayesian inference methods (Rasmussen and Williams, 2005). Moreover, GPs *a priori* do not require specification of the functional form of the predictor function, e.g. linear or polynomial. Rather a prior distribution on functions is specified via a GP prior. This is an important characteristic since bat calls are represented by their two-dimensional spectrograms which exhibit highly non-linear features (Fig. 1 in Section 7).

In this study we use 1800 bat calls collected from 449 bats and 21 different species in five families from Mexico, which is a hotspot of global bat biodiversity representing a highly diverse call assemblage (Walters *et al.*, 2013). The data are presented in Section 6 and the signal processing methodology is presented in Section 7. We apply a multinomial probit regression model with a GP prior (Girolami and Rogers, 2006), Section 2, and we discuss how inference for such a model can be performed in Section 3. The nature of the model and the large number of classes as well as the number of samples pose several problems on well-established methods for exact MCMC inference using the pseudomarginal algorithm of Andrieu and Roberts (2009). We discuss these issues in detail in Section 4. For the particular application we resort to approximate Bayesian inference methods, and particularly the expectation propagation (EP) algorithm of Minka (2001), in Section 5. Experiments and results are presented in Section 8 and conclude with a discussion and remarks on the challenges of Bayesian inference in large problems in Section 9.

2. Multinomial probit with latent Gaussian processes

Let $\mathbf{y}$ and X be observed data where $\mathbf{y} = (y_1, \ldots, y_N)^{\mathrm{T}}$ with $y_n \in \{1, \ldots, C\}$ and $X = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ with $\mathbf{x}_n \in \mathbb{R}^D$. The observed variables y_n indicate bat species; for example $y_n = c$ indicates that the nth observation is from the cth species in the data set, and $\mathbf{x}_n$ are respective covariates, usually D measurements from the nth call's spectrogram. Also let $\mathbf{f} : \mathbb{R}^D \to \mathbb{R}^C$ with *latent* values $\mathbf{f}(\mathbf{x}_n) = \mathbf{f}_n = (f_n^1, f_n^2, \ldots, f_n^C)^{\mathrm{T}}$ such that when transformed by a sigmoid-like function give the class probabilities $p(y_n|\mathbf{f}_n)$.

The multinomial probit model assumes that the species indicator variables y_n have a multinomial distribution with probabilities given by a transformation of the latent function values $\mathbf{f}_n$:

$$p(y_n|\mathbf{f}_n) = \int \mathcal{N}(u_n|0,1) \prod_{j=1, j\neq y_n}^{C} \Phi(u_n + f_n^{y_n} - f_n^j)\, \mathrm{d}u_n. \tag{1}$$

In equation (1) $\mathcal{N}$ denotes the Gaussian density function and Φ its cumulative distribution function. There are other alternatives sigmoid functions to equation (1), such as the softmax function $\exp(f_n^{y_n})/\Sigma_{j=1}^{C} \exp(f_n^j)$. However, equation (1) is convenient for obtaining a Gibbs sampler (Albert and Chib, 1993) or approximate Bayesian inference algorithms (Girolami and Rogers, 2006; Riihimaki *et al.*, 2013).

For the *latent* function values we assume independent zero-mean GP priors for each species similarly to Rasmussen and Williams (2005), i.e. *a priori* we assume that latent variables from different species are independent. However, in the posterior, latent variables will not necessarily be independent because of the interactions of the latent variables in the likelihood (1). Collecting all *latent* function values in $\mathbf{f} = (f_1^1, \ldots, f_N^1, f_1^2, \ldots, f_N^2 \ldots, f_1^C, \ldots, f_N^C)^{\mathrm{T}}$ the GP prior is

$$p(\mathbf{f}|X, \boldsymbol{\theta}) = \mathcal{N}\{\mathbf{f}|\mathbf{0}, K(\boldsymbol{\theta})\}, \tag{2}$$

where $K(\boldsymbol{\theta})$ is a $CN \times CN$ block covariance matrix with block matrices $K^1(\boldsymbol{\theta}), \ldots, K^C(\boldsymbol{\theta})$, each of size $N \times N$, on its diagonal. Elements $K^c_{i,j}$ define the prior covariance between the *latent* function values f^c_i and f^c_j governed by a covariance function $k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta})$ with unknown parameters $\boldsymbol{\theta}$. A different covariance can be used for each diagonal block of the covariance matrix corresponding to different values of the species indicator variables.

The only requirement for the covariance function $k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta})$ is that it is a positive definite function. Here we describe the two types of covariance functions that we used in this study in more detail. One of the most frequently used covariance functions is the squared exponential function

$$k(\mathbf{x}_i, \mathbf{x}_j\boldsymbol{\theta}) = \sigma^2 \exp\left\{ -\frac{\sum_{d=1}^{D} (x_{i,d} - x_{j,d})^2}{2l^2} \right\},$$

where σ^2 and l are the magnitude and length scale parameters and $\boldsymbol{\theta} = (\sigma^2, l)^{\mathrm{T}}$. The length scale parameter controls the smoothness of the function and measures, in units, how far on the input space one must move for the function to change value drastically. The magnitude parameter controls the magnitude of this change.

The squared exponential function can be generalized to have one length scale parameter for each dimension of the input space, i.e.

$$k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta}) = \sigma^2 \exp\left\{ -\sum_{d=1}^{D} \frac{(x_{i,d} - x_{j,d})^2}{2l_d^2} \right\},$$

where $\boldsymbol{\theta} = (\sigma^2, l_1, \ldots, l_D)^{\mathrm{T}}$. Such covariance functions have a very interesting interpretation. The larger the length scale parameter the smoother the functions are. In the limit of $l_d \to \infty$ the function becomes constant and therefore informs us that there is no information in the particular input dimension x_d for the task at hand. Thus once estimates or posteriors of the parameter values have been found we can inspect their values or distributions to assess their relevance in the problem at hand.

Finally, we can combine covariance functions that are obtained on a different set of covariates or measurements for the same problem to form new valid covariance functions. For example, given two sets of measurements for the same data $X^{(1)}$ and $X^{(2)}$ we can construct a composite covariance by $w_1 k_1(\mathbf{x}^{(1)}_i, \mathbf{x}^{(1)}_j|\boldsymbol{\theta}^{(1)}) + w_2 k_2(\mathbf{x}^{(2)}_i, \mathbf{x}^{(2)}_j|\boldsymbol{\theta}^{(2)})$, where the relative weights w_1 and w_2 sum to 1. The parameters of the covariance function then become $\boldsymbol{\theta} = (w_1, w_2, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})^{\mathrm{T}}$. This is an important property of GPs for applications such as that considered here since we often can have different representations of the data which would not be necessarily compatible. For example, in our application we have measurements directly from spectrograms of calls but we also have the whole spectrogram. These two measures capture different aspects of the information in the call and we would like to include in our analysis both representations of the data. Information that is contained in the measurements is also contained in the spectrogram but it is transformed in a non-linear way by using expert knowledge whereas some information from the spectrogram is not in covered by the measurements at all. Using a weighted combination of the kernel functions allows inference about which representation is more suitable for the problem after we obtain estimates or posterior samples for the weights.

3. Parameter and predictive inference

Inference for multinomial probit regression with latent GPs entails two aspects: parameter

inference and prediction. The only free parameters in the model are the covariance function parameters $\boldsymbol{\theta}$ which, as discussed in the previous section, are interpretable. Given observed data $\mathbf{y}$ and X, and a prior distribution $p(\boldsymbol{\theta})$ inference about the parameters can be made by calculating the posterior

$$p(\boldsymbol{\theta}|X,\mathbf{y}) \propto p(\boldsymbol{\theta}) \int p(\mathbf{y}|\mathbf{f})\, p(\mathbf{f}|X,\boldsymbol{\theta})\, \mathrm{d}\mathbf{f}, \tag{3}$$

where $p(\mathbf{y}|\mathbf{f}) = \Pi_{n=1}^{N} p(\mathbf{y}_n|\mathbf{f}_n)$ and involves integrating latent function values $\mathbf{f}$. By inspecting the posterior or by obtaining estimates $\hat{\boldsymbol{\theta}}$, e.g. by taking $\hat{\boldsymbol{\theta}}$ to be the maximum of the posterior, we can infer the most likely parameters for the data and interpret the results according to the previous section.

For making predictions for new, previously unobserved, data $\mathbf{x}_*$ we need to obtain the predictive distribution for the unobserved species indicator variable y_*,

$$p(y_* = c|\mathbf{x}_*, X, \mathbf{y}) = \int p(y_* = c|\mathbf{f}_*)\, p(\mathbf{f}_*|\mathbf{x}_*, X, \mathbf{f}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}, \mathbf{f}|X, \mathbf{y})\, \mathrm{d}\mathbf{f}_*\, \mathrm{d}\mathbf{f}\, \mathrm{d}\boldsymbol{\theta}, \tag{4}$$

which involves integrating both latent variables $\mathbf{f}_*$ and $\mathbf{f}$ and parameters $\boldsymbol{\theta}$. In the predictive distribution (4) the second term in the integral is the GP predictive distribution for the new latent function values $\mathbf{f}_*$. For simplicity of the notation let K be the covariance matrix evaluated at $\boldsymbol{\theta}$, Q_* be a $CN \times C$ matrix

$$Q_* = \begin{pmatrix} \mathbf{k}_*^1 & 0 & \dots & 0 \\ 0 & \mathbf{k}_*^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{k}_*^C \end{pmatrix},$$

where $\mathbf{k}_*^c$ is the vector such that the nth element is equal to $k^c(\mathbf{x}_n, \mathbf{x}_*|\boldsymbol{\theta})$, with c indexing values of the species indicator variables $\mathbf{y}$.

From the properties of multivariate normal variables $\mathbf{f}_*$ is also normally distributed with mean $\boldsymbol{\mu}_* = Q_*^{\mathrm{T}} K^{-1} \mathbf{f}$ and variance $\Sigma_* = K_{**} - Q_*^{\mathrm{T}} K^{-1} Q_*$, where K_{**} is a $C \times C$ diagonal matrix such that the cth element is equal to $k^c(\mathbf{x}_*, \mathbf{x}_*|\boldsymbol{\theta})$. Thus $p(\mathbf{f}_*|\mathbf{x}_*, \mathbf{f}, X, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}_*|\boldsymbol{\mu}_*, \Sigma_*)$. The last term in the integral (4) is the joint posterior of the latent function values $p(\boldsymbol{\theta}, \mathbf{f}|X, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{f})\, p(\mathbf{f}|X, \boldsymbol{\theta})\, p(\boldsymbol{\theta})$.

The necessary integrations for the parameter posterior and predictive distribution cannot be computed in closed form since the likelihood function $p(\mathbf{y}|\mathbf{f})$ is not conjugate with the normal distribution. Therefore, we must resort either to Monte Carlo integration or approximate algorithms for GPs. Both options are discussed in the subsequent sections.

4. Markov chain Monte Carlo inference for the multinomial probit model

To perform posterior parameter inference for the multinomial probit model with latent GPs we must obtain samples from the posterior distribution (3). Sampling directly from the marginal posterior for the parameters $\boldsymbol{\theta}$ is not possible since the integration cannot be carried out analytically but we can set up a Markov chain to sample from the joint posterior $p(\boldsymbol{\theta}, \mathbf{f}|\mathbf{X}, \mathbf{y})$. Once converged we can obtain the marginal for $\boldsymbol{\theta}$ by simply discarding samples for $\mathbf{f}$. We can then inspect the histograms and posterior plots for the parameters and interpret as discussed above.

For predictive inference we require N_S samples from the joint posterior $p(\boldsymbol{\theta}, \mathbf{f}|\mathbf{X}, \mathbf{y})$. Denoting

the ith sample from the joint posterior by $\boldsymbol{\theta}^{(i)}, \mathbf{f}^{(i)}$ a Monte Carlo estimate of distribution (4) can be obtained by

$$p(y_* = c|\mathbf{x}_*, X, \mathbf{y}) \approx \frac{1}{N_S} \sum_{i=1}^{N_S} \int p(y_* = c|\mathbf{f}_*)\, p(\mathbf{f}_*|\mathbf{x}_* X, \mathbf{f}^{(i)}, \boldsymbol{\theta}^{(i)})\, \mathrm{d}\mathbf{f}_*. \tag{5}$$

An important aspect for inference is to design efficient MCMC sampling schemes to obtain samples from the joint posterior. Obtaining samples with joint proposals on both $\mathbf{f}$ and $\boldsymbol{\theta}$ is problematic since it is extremely unlikely to propose a set of latent function values $\mathbf{f}$ and parameters $\boldsymbol{\theta}$ that are compatible. This is due to the strong dependence of $\mathbf{f}$ on $\boldsymbol{\theta}$ in the covariance function as well as the high dimensionality of $\mathbf{f}$. Such proposals will therefore have very high rejection rates, making the chain very slow to converge.

An alternative is to design a Gibbs sampler (Robert and Casella, 2005) which at each iteration samples from the conditionals $p(\mathbf{f}|X, \mathbf{y}, \boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|X, \mathbf{y}, \mathbf{f})$. Gibbs samplers for latent Gaussian models have been extensively studied in the literature (Papaspiliopoulos *et al.*, 2007; Yu and Meng, 2011; Murray and Adams, 2010; Filippone and Girolami, 2014) and an extensive comparison with regard to their efficiency and properties can be found in Filippone *et al.* (2013). Despite the large interest for MCMC inference for latent Gaussian models there is little research on multinomial models for classification and especially for applications with large numbers of classes and observations such as the application that we consider in this paper.

A common problem with Gibbs samplers for latent Gaussian models is that there is a strong dependence of the parameters and the latent function values which results in slow converging and poorly mixing chains (Filippone *et al.*, 2013). Also Filippone *et al.* (2012) reported chains of 5 million samples even with the reparameterizations of Papaspiliopoulos *et al.* (2007) and Yu and Meng (2011). This renders Gibbs sampling a non-viable approach for our application where the number of observations is so high that even a single sample from the posterior requires long computation time.

This strong dependence can be avoided by sampling $\boldsymbol{\theta}$ from its marginal distribution by integrating $\mathbf{f}$. Unfortunately this is not possible when the likelihood function $p(\mathbf{y}|\mathbf{f})$ is not conjugate with the multivariate normal distribution. Andrieu and Roberts (2009) showed that we can obtain samples from the exact posterior by using an estimate of the marginal provided that it is unbiased. This has been exploited by Filippone and Girolami (2014) for binary probit classification with GPs and has been shown to be the most efficient method compared with other Gibbs sampling algorithms. In the remainder of this section we show how we can design such a sampler for multinomial probit classification and discuss its feasibility on the application that is considered in this paper.

4.1. Feasibility of the pseudomarginal Gibbs sampling

To sample from the marginal posterior for the parameters $\boldsymbol{\theta}$ we can design a Metropolis–Hastings sampler (Robert and Casella, 2005). Using a proposal distribution $\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})$, we can generate proposed samples $\boldsymbol{\theta}'$ which are accepted with probability

$$\alpha = \frac{p(\mathbf{y}|X, \boldsymbol{\theta}')\, p(\boldsymbol{\theta}')\, \pi(\boldsymbol{\theta}|\boldsymbol{\theta}')}{p(\mathbf{y}|X, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta}'|\boldsymbol{\theta})},$$

where $p(\mathbf{y}|X, \boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})\, p(\mathbf{f}|X, \boldsymbol{\theta}) \mathrm{d}\mathbf{f}$. Andrieu and Roberts (2009) showed that we can obtain a chain that converges to the correct posterior by replacing the marginal $p(\mathbf{y}|X, \boldsymbol{\theta})$ with an unbiased estimate $\hat{p}(\mathbf{y}|X, \boldsymbol{\theta})$. Such an estimate can be computed by an importance sampling algorithm with N_I samples from an approximating distribution $q(\mathbf{f}|\mathbf{y}, X, \boldsymbol{\theta})$. An importance sampling estimate for the marginal can be computed by

$$\hat{p}(\mathbf{y}|X,\boldsymbol{\theta}) = \frac{1}{N_I}\sum_{i=1}^{N_I}\frac{p(\mathbf{y}|\mathbf{f}^{(i)})\,p(\mathbf{f}^{(i)}|X,\boldsymbol{\theta})}{q(\mathbf{f}^{(i)}|\mathbf{y},X,\boldsymbol{\theta})}.$$

An important aspect for the efficiency of the pseudomarginal algorithm is the variance of the estimate $\hat{p}(\mathbf{y}|X,\boldsymbol{\theta})$, since the larger the variance the larger the rejection rate for proposed samples. The variance of the estimate is reduced to 0 when $q(\mathbf{f}|\mathbf{y},X,\boldsymbol{\theta})$ is proportional to $p(\mathbf{y}|\mathbf{f})\,p(\mathbf{f}|X,\boldsymbol{\theta})$ (Robert and Casella, 2005). This can be achieved when $q(\mathbf{f}|\mathbf{y},X,\boldsymbol{\theta})$ is the posterior of $\mathbf{f}$ which we cannot sample directly for the multinomial probit model. Therefore we need to find an easy-to-sample distribution which approximates as closely as possible the posterior of $\mathbf{f}$. Both the EP (Minka, 2001) algorithm and the Laplace approximation (LA) (Williams and Barber, 1998) approximate the posterior for the latent variables with a multivariate normal distribution. Kuss and Rasmussen (2005) evaluated different approximations for the binary probit classification with GPs whereas Filippone and Girolami (2014) evaluated pseudomarginal algorithms based on the LA and EP approximations also for the binary probit model. The results from both studies indicate that the EP algorithm better approximates the posterior and it also results in an efficient pseudomarginal algorithm compared with the LA.

The variance of the pseudomarginal estimate is greatly affected by the total number of species, C, and the number of observations, N, as both affect the dimension of the latent function values $\mathbf{f}$. The larger the dimension of $\mathbf{f}$ the more samples from the approximate distribution will be needed to reduce the variance. An additional complication that arises in our application is that the one-dimensional integral in the multinomial probit function (1) is also not analytically tractable. In the binary probit function the integral can be computed analytically since there is only one normal cumulative distribution function term in the integral. Since this is a one-dimensional integral we use Gaussian quadrature to obtain an estimate. However, this increases the computational overhead of the overall algorithm. We should mention here that instead of the multinomial probit function we could have used the softmax function; however, this will not allow us to derive the EP approximation for the posterior of $\mathbf{f}$ and will restrict us to the LA.

In Table 1 we show the variance of the log-pseudomarginal for various sample sizes. We also compare the variance on the full data set that is used in this study (21 species and 1432 observations) with the variance on a subset with only three species and 200 observations. The approximating distribution was obtained by the EP algorithm that is described in the next section. We can see that even with 1000 samples the variance is significantly high, leading to very high rejection rates for our sampler. Although there are several methods for reducing the variance of Monte Carlo estimates, e.g. Assaraf and Caffarel (1999) and Mira *et al.* (2013), we do not pursue this further. The reason is that such methods will further complicate implementation

Table 1. Comparison of variance of the log-pseudomarginal estimate for various sample sizes from the approximating distributions on two data sets

N_I	*Results for* $C=3$, $N=200$	*Results for* $C=21$, $N=1432$
10	0.0213	0.5763
50	0.0038	0.3316
100	0.0025	0.2779
500	0.0003	0.0653
1000	0.0003	0.0295

and increase an additional computational overhead per iteration which as discussed below is already prohibitive.

The computation time for calculating a single estimate of the pseudomarginal for $N_I = 500$ using the full data set of 21 species and 1432 observations is approximately 4.9 min on a four-core Intel i7 processor with 16 Gbytes of random-access memory. The computation time is largely dominated by the EP approximation which scales as $O(CN^3)$ with the sampling of the approximate posterior and the Monte Carlo estimate of the likelihood adding significant constant terms. We have tried parallel and graphics processing unit implementations for the EP approximation. However, the benefits of both implementations were diminished by the large communication overheads, either between different nodes in the cluster or between the main memory and graphics processing unit memory. Additionally, for predictive inference we must sum over all posterior samples and also integrate latent function values $\mathbf{f}_*$ for the new observations $\mathbf{x}_*$ which is also performed by using Monte Carlo sampling. All the above indicate that full Bayesian inference for high dimensional latent variable models with large data sets remains a very challenging problem both computationally and statistically. Moreover, recently Chopin and Ridgway (2015) showed that for binary classification problems the EP approximation provides very good results and is sometimes preferable to full MCMC inference because of its low computational cost and ease of implementation.

5. Approximate Bayesian inference with expectation propagation

Approximate Bayesian inference for latent variable models relies on point estimates of the parameters $\boldsymbol{\theta}$ instead of obtaining samples from the posterior. Such estimates are usually obtained by finding the parameters that maximize the posterior or some approximation to the posterior, i.e. $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|X, \mathbf{y})$. Predictive inference is then conditional on the parameter estimate:

$$p(y_* = c|\mathbf{x}_*, X, \mathbf{y}) = \int p(y_* = c|\mathbf{f}_*)\, p(\mathbf{f}_*|\mathbf{x}_*, X, \mathbf{f}, \hat{\boldsymbol{\theta}})\, p(\mathbf{f}|X, \mathbf{y}, \hat{\boldsymbol{\theta}})\, \mathrm{d}\mathbf{f}_*\, \mathrm{d}\mathbf{f}. \tag{6}$$

As discussed in the previous section the posterior (3) for the parameters cannot be obtained analytically since it requires integration of the latent variables $\mathbf{f}$. Similarly, we must also integrate over $\mathbf{f}$ in equation (6) where the last term in the integral is the posterior of the latent variables conditioned on the parameters:

$$p(\mathbf{f}|X, \mathbf{y}, \boldsymbol{\theta}) = \frac{1}{Z} p(\mathbf{f}|X, \boldsymbol{\theta}) \prod_{n=1}^{N} p(y_n|\mathbf{f}_n). \tag{7}$$

Z is the normalizing constant or the marginal likelihood:

$$Z = p(\mathbf{y}|X, \boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})\, p(\mathbf{f}|X, \boldsymbol{\theta})\, \mathrm{d}\mathbf{f},$$

which is proportional to the posterior of the parameters. Thus if we find a convenient approximation of the posterior for $\mathbf{f}$ and its normalizing constant we can obtain a parameter estimate by maximizing the marginal and analytically integrate $\mathbf{f}$ in the predictive distribution.

Following Riihimaki *et al.* (2013), using the EP method the posterior is approximated by

$$q_{\mathrm{EP}}(\mathbf{f}|X, \mathbf{y}, \boldsymbol{\theta}) = \frac{1}{Z_{\mathrm{EP}}} p(\mathbf{f}|X, \boldsymbol{\theta}) \prod_{n=1}^{N} \tilde{t}_n(\mathbf{f}_n|\tilde{Z}_n, \tilde{\boldsymbol{\mu}}_n, \tilde{\Sigma}_n), \tag{8}$$

where $\tilde{t}_n(\mathbf{f}_n|\tilde{Z}_n, \tilde{\boldsymbol{\mu}}_n, \tilde{\Sigma}_n) = \tilde{Z}_n \mathcal{N}(\mathbf{f}_n|\tilde{\boldsymbol{\mu}}_n \tilde{\Sigma}_n)$ are local *likelihood* approximate terms with parameters $\tilde{Z}_n, \tilde{\boldsymbol{\mu}}_n$ and $\tilde{\Sigma}_n$. In other words, each *multinomial probit likelihood term* in equation (7) is approximated by a scaled Gaussian function.

After initializing the parameters of the approximate terms, each term is then updated sequentially by first removing the ith term from the marginal posterior to give the *cavity* distribution

$$q_{-n}(\mathbf{f}_n) := q_{\mathrm{EP}}(\mathbf{f}_n|X, \mathbf{y}, \boldsymbol{\theta})\, \tilde{t}_n(\mathbf{f}_n|\tilde{Z}_n, \tilde{\boldsymbol{\mu}}_n, \tilde{\Sigma}_n)^{-1}. \tag{9}$$

Distribution (9) is then combined with the exact ith likelihood term to form the *tilted* distribution

$$\hat{p}_n(\mathbf{f}_n) := \hat{Z}_n^{-1} q_{-n}(\mathbf{f}_n) p(y_n|\mathbf{f}_n). \tag{10}$$

Then a Gaussian approximation $\hat{q}_n(\mathbf{f}_n)$ to $\hat{p}_n(\mathbf{f}_n)$ is obtained by matching their moments. The parameters of the ith approximate term are then updated such that the mean and covariance of the approximate marginal $q_{\mathrm{EP}}(\mathbf{f}_n)$ are consistent with $\hat{q}_n(\mathbf{f}_n)$, i.e.

$$\tilde{t}_n^{\mathrm{new}} \propto \hat{q}_n(\mathbf{f}_n) q_{-n}(\mathbf{f}_n)^{-1}.$$

At every update step the approximate posterior (8) is also updated.

More specifically the tilted distribution for the multinomial probit models is

$$\hat{p}_n(\mathbf{f}_n) = \hat{Z}_n^{-1} \mathcal{N}(\mathbf{f}_n|\boldsymbol{\mu}_{-n}\Sigma_{-n}) \int \mathcal{N}(u_n|0,1) \prod_{j=1, j\neq y_n}^{C} \Phi(u_n + f_n^{y_n} - f_n^j)\, \mathrm{d}u_n, \tag{11}$$

where $\Sigma_{-n} = (\Sigma_n^{-1} - \tilde{\Sigma}_n^{-1})^{-1}$ and $\boldsymbol{\mu}_{-n} = \Sigma_{-n}(\Sigma_n^{-1}\boldsymbol{\mu}_n - \tilde{\Sigma}_n^{-1}\tilde{\boldsymbol{\mu}}_n)$ are the mean and covariance of the cavity.

Unlike the binary probit case, where the tilted distribution is univariate and thus its moments are easy to compute, the tilted distribution for the multinomial probit model is C dimensional. Previous work on EP approximations for the multinomial probit model (Girolami and Zhong, 2007) further approximated the moments of the tilted distribution using the LA. This assumes that the distributions can be closely approximated by a multivariate normal distribution. Seeger and Jordan (2004) used C-dimensional numerical integration which for large values of C is problematic. Kim and Ghahramani (2006) replaced the multinomial probit function with a threshold function which results in an EP algorithm that is similar to the EP algorithm for the binary problem. However, the algorithm scales as $O\{(C-1)^3N^3\}$ which again for large C becomes prohibitively expensive.

The recent work of Riihimaki *et al.* (2013) proposed to approximate the tilted distribution also with EP. They called their method nested EP since an inner EP algorithm is used at each iteration of the outer EP algorithm to approximate the moments of the tilted distribution. However, they showed that the algorithm can be seen as a standard EP algorithm with $N(C-1)$ approximate terms and thus the inner EP algorithm requires only a single iteration provided that the approximate parameters are not initialized and they are started from their previous values. Furthermore they showed that the structure of the site precision matrices $\tilde{\Sigma}_n^{-1}$ is such that an efficient implementation which scales as $O\{(C-1)N^3\}$ is possible. This is similar to the computational complexity of the LA even though EP gives a better approximation to the posterior. Details of the approximation to the tilted distribution and the EP algorithm can be found in the on-line supplementary material of this paper or Riihimaki *et al.* (2013).

Given the EP approximation to the posterior of $\mathbf{f}$ and its marginal, we can use them to estimate parameters and to obtain the predictive distribution (6). The integration over $\mathbf{f}$ in expression (6) is now analytically tractable if we replace the last term in the integral with the EP approximation. However, we still need to integrate out $\mathbf{f}_*$. We can approximate the integral by using the EP approximation of the tilted distribution and its equivalent with obtaining the moment $Z_{\hat{q}_*}$ or the normalizing constant of the tilted distribution. More details and an efficient implementation are provided in the supplementary material.

6. Description of the data set

Bat echolocation calls were recorded across north and central Mexico from June to November 2012 and from February to May 2013. We used 10 mist nets erected at ground level (0–3 m) set from sunset to sunrise to capture bats. Live trapped bats were measured and identified to species level by using field keys (Medellín *et al*., 2008; Ceballos and Oliva, 2005) and bat taxonomy followed (Simmons, 2005). We constructed an echolocation call library by recording the calls of captured individuals using two different techniques:

(a) bats were recorded while released from the hand about 6–10 m from the bat detector in open areas and away from vegetation and
(b) bats were tied to a zip line and recorded while flying along the zip flight path.

Echolocation calls were recorded with a Pettersson 1000x bat detector (Pettersson Elektronik AB, Uppsala, Sweden). The bat detector was set to record calls manually in realtime, full spectrum at 500 kHz. Each recording consists of multiple calls from a single individual bat.

In total our data set consists of 21 species in five families, 449 individual bats and 8429 calls; Table 2. Care must be taken when splitting the data to training and test sets during cross-validation

Table 2. Data set statistics†

Species	*Samples*	*Calls*
Family: Emballonuridae		
1 *Balantiopteryx plicata*	16	384
Family: Molossidae		
2 *Nyctinomops femorosaccus*	16	311
3 *Tadarida brasiliensis*	49	580
Family: Mormoopidae		
4 *Mormoops megalophylla*	10	135
5 *Pteronotus davyi*	8	106
6 *Pteronotus parnellii*	23	313
7 *Pteronotus personatus*	7	51
Family: Phyllostomidae		
8 *Artibeus jamaicensis*	11	82
9 *Desmodus rotundus*	6	38
10 *Leptonycteris yerbabuenae*	26	392
11 *Macrotus californicus*	6	53
12 *Sturnira ludovici*	12	71
Family: Vespertilionidae		
13 *Antrozous pallidus*	58	1937
14 *Eptesicus fuscus*	74	1589
15 *Idionycteris phyllotis*	6	177
16 *Lasiurus blossevillii*	10	90
17 *Lasiurus cinereus*	5	42
18 *Lasiurus xanthinus*	8	204
19 *Myotis volans*	8	140
20 *Myotis yumanensis*	5	89
21 *Pipistrellus hesperus*	85	2445

†'Samples' refers to recordings from individual bats though there can be multiple calls from the same bat in each recording.

to ensure that calls from the same individual are not in both sets. For this reason we split our data set by using recordings instead of calls. For species with fewer than 100 recordings we include as many calls as possible up to a maximum of 100 calls per species. After this selection process each fold has approximately 1400 calls for training and 200 calls for testing from 21 species. The raw data as well as the post-processed and fivefold cross-validation sets are available to download as supplementary material for this paper from `http://www.engage-project.org/`.

7. Signal processing and data representation

The vector representation $\mathbf{x}_n$ for each call is constructed by extracting call parameters from the spectrograms following Walters *et al.* (2012) using Sonobat version 3.0 software (Szewczak, 2010). The spectrogram of a call is calculated by using a Hamming window of size 256 with 95% overlap and a fast Fourier transform length of 512; Fig. 1. The frequency range of the spectrogram is thresholded by removing frequencies below 5 kHz and above 210 kHz. This is done to reduce the size of the spectrogram and also to remove low and high frequency noise since we know that there are no bats echolocating outside this frequency range. In total 31 parameters are calculated including call duration in milliseconds, the highest and lowest frequencies of the call, total frequency spread, the frequency with maximum amplitude and the frequencies at the start and end of the call (see the on-line supplementary material section 6 for full details). All 31

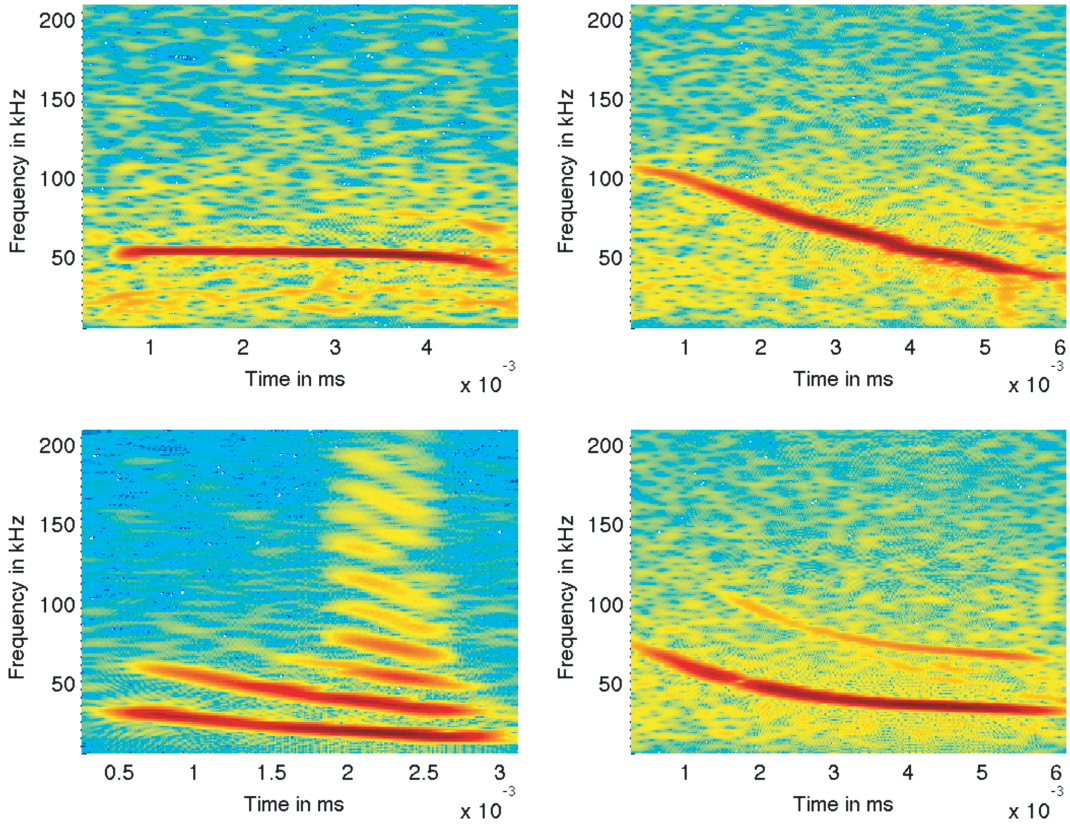

Fig. 1. Example of spectrograms of calls from four different species: see the text for details on spectrogram computation

call parameters are extracted by using the Sonobat version 3.0 software. All 31 call parameters are concatenated in the vector $\mathbf{x}_n$ and a squared exponential kernel with individual length scales for each parameter is used for the GP classifier.

Although extracting call parameters from call spectrograms captures some of the call characteristics and shape, a large amount of information is discarded, e.g. harmonics. An alternative to characterizing a call by using predefined parameters is to utilize its spectrogram directly. However, owing to the differences in call duration the spectrograms will need to be normalized to have the same length by using some form of interpolation. In this work we borrow ideas from speech recognition (Sakoe and Chiba, 1978) and previous work on bird call classification (Damoulas *et al.*, 2010) and employ the dynamic time warping (DTW) kernel to compare two calls' spectrograms directly.

Given two calls i and j from the library and their spectrograms S_i and S_j, where $S_i \in \mathbb{C}^{F\times W}$ with F being the number of frequency bands and W the number of windows, the dissimilarity matrix $D^{i,j} \in \mathbb{R}^{W\times W}$ is constructed such that

$$D^{i,j}(w,v) = 1 - \frac{S_i(\cdot,w)^{\mathrm{T}} S_j(\cdot,v)}{\sqrt{\{S_i(\cdot,w)^{\mathrm{T}} S_i(\cdot,w) S_j(\cdot,v)^{\mathrm{T}} S_j(\cdot,v)^{\mathrm{T}}\}}}. \tag{12}$$

DTW uses the dissimilarity matrix to stretch or expand spectrogram S_i over time to match S_j by calculating the optimal warping path with the smallest alignment cost by $c_{i,j}$ by

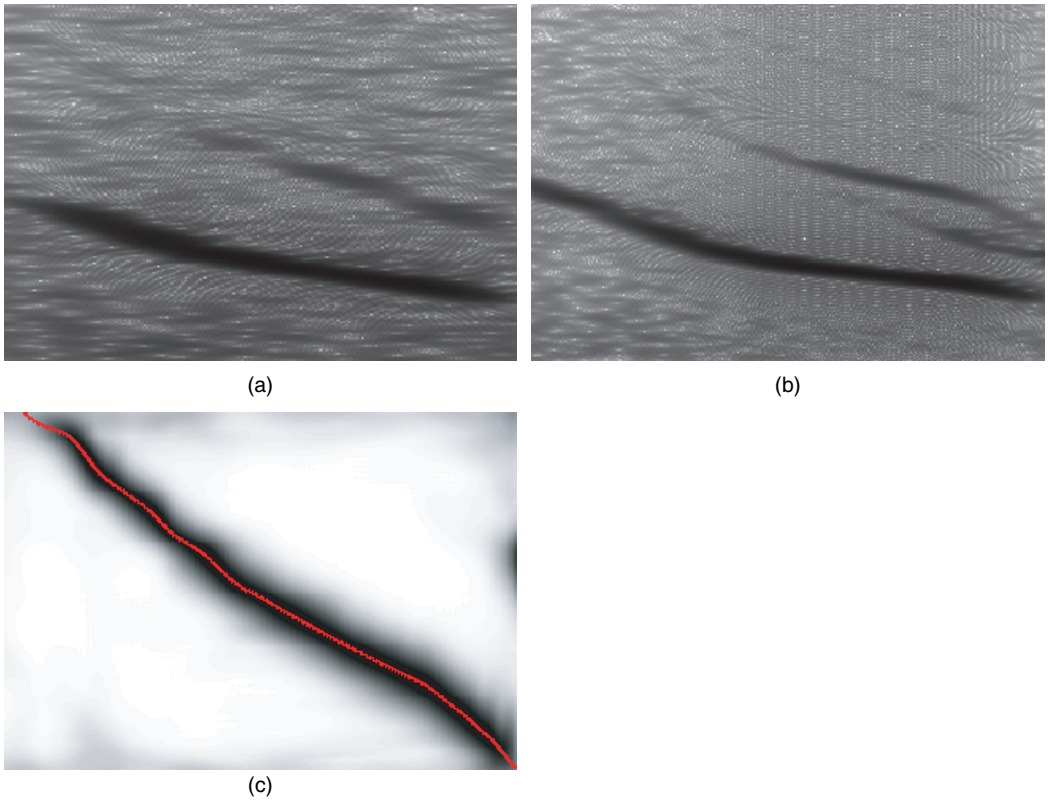

Fig. 2. Example of the DTW optimal warping path for two call spectrograms from the same species: (a), (b) call spectrograms S_i; (c) dissimilarity matrix $D_{i,j}$ and optimal warping path

using dynamic programming. Fig. 2 illustrates the optimal warping path for two calls in the library.

For each call we construct a vector representation $\mathbf{x}_n$ by computing the optimal warping paths with all N calls from the library and concatenating the alignment costs such that $\mathbf{x}_n = (c_{n,1}, \ldots, c_{n,N})$. We then use the squared exponential covariance function for the covariance matrix of the GP classifier.

8. Results and interpretation

We compare the classification accuracy of the multinomial probit regression with GP prior classifier by using the two representations that were discussed in Section 7. The values of the call parameters are normalized to have zero mean and standard deviation 1 by subtracting the mean and dividing by the standard deviation of the call parameters in the training set. For the 32 covariance function parameters, σ^2 and $\lambda_1, \ldots, \lambda_{31}$ we use independent gamma priors with shape parameter 1.5 and scale parameter 10. For the DTW representation each call vector of optimal alignment costs is normalized to unit length and independent gamma (1.5, 10) priors are used for the magnitude and length–scale covariance function parameters. We also combine the covariance functions for both representations by using a linear combination as discussed in Section 2. The weights for the linear combination of the DTW and call parameters kernel functions are restricted to be positive and to sum to 1 and a flat Dirichlet prior is used.

As a baseline we also compare with a multiclass SVM classifier by using the LibSVM software library (Chang and Lin, 2011). For the SVM we use the call parameters and the DTW representations with automatic relevance determination and squared exponential kernels respectively, and we also combine the two kernels similarly to the GP model that was discussed above. We use the same MATLAB code to precompute the kernels for the SVM as that used for the GP models. In that way we ensure as fair a comparison as possible. All kernel parameters and the relaxation parameter C of the SVM were optimized with Bayesian optimization (Snoek *et al.*, 2012) by using the MATLAB code that was provided by Gardner *et al.* (2014). Bayesian optimization for SVMs allows us to exploit kernels with many parameters, such as automatic relevance determination, in contrast with the commonly used method of grid search over a prespecified set of values. We have also trained SVMs with a squared exponential kernel for call shape parameters and DTW by using a grid search. However, the cross-validated error rate is almost the same albeit with higher variance.

Table 3. Classification error rates for bat species identification from echolocation calls by using various classification methods and data representations†

Method	*Error rate*	*Standard deviation*
SVM call parameters	0.26	±0.037
SVM DTW	0.25	±0.038
SVM DTW + call parameters	0.22	±0.054
GP call parameters	0.24	±0.052
GP DTW	0.21	±0.026
GP DTW + call parameters	*0.20*‡	± *0.037*

†See the text for details
‡Lowest error rate.

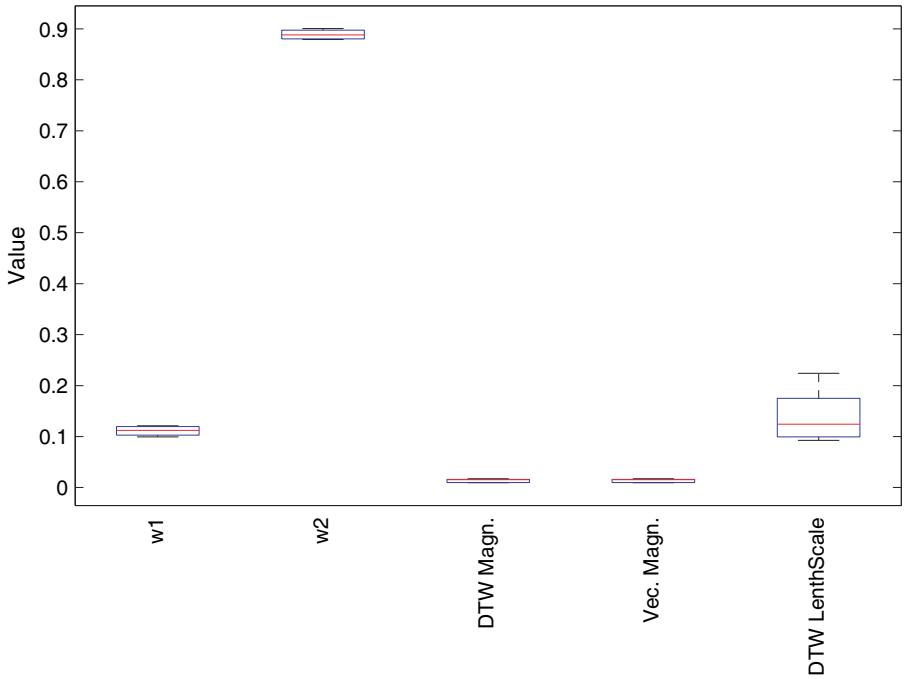

Fig. 3. Model parameters estimated on five folds: w1 and w2 are the respective weights for the call parameters and DTW representations; DTW Magn. and Vec. Magn. are the magnitudes of the squared exponential covariance function parameters for the DTW and call parameters representations; DTW LengthScale is the length scale of the DTW squared exponential covariance function

Table 3 compares the missclassification rate by using various data representations and two different classification methods. Results are averages of a fivefold cross-validation. In all cases multinomial probit regression with GP priors is superior to SVMs. Therefore, although we sacrifice in computational time and complexity of the algorithm we do not sacrifice in accuracy whereas we gain significantly with regard to interpretation. We can see that the DTW representation is significantly better for characterizing the species variations achieving a better classification accuracy irrespectively of the classifier. However, results are improving by also considering information from the call parameters. This highlights the importance of combining information from different representations while it highlights the importance of prior knowledge in constructing call parameters. Additionally, the optimized weights for the kernel combination significantly favour the DTW covariance function with a weight of about 0.9 in contrast with the call parameters with weight about 0.1: Fig. 3. If we fix the weight parameters to equal values we obtain a classification error rate of 0.22 ± 0.031, highlighting the importance of the DTW kernel matrix.

The independent length scales allow us also to interpret the discriminatory power of the 31 call parameters (Fig. 4). We believe that such an analysis is not only useful for validating the design of acoustic classification tools but also helps to understand how different call designs have evolved within bats (Maltby *et al.*, 2009). The call parameters with estimated values below 0.5 are parameters which describe call slope (PrcntMaxAmpDur, LedgeDuration, PrcntKneeDur, TimeFromMaxToFc and EndSlope; on-line supplementary material section 6). Call slope reflects the relative amount of the call that is at a constant frequency (a horizontal or narrowband call) or is frequency modulated (a vertical or broadband call). Different echolocating bat species

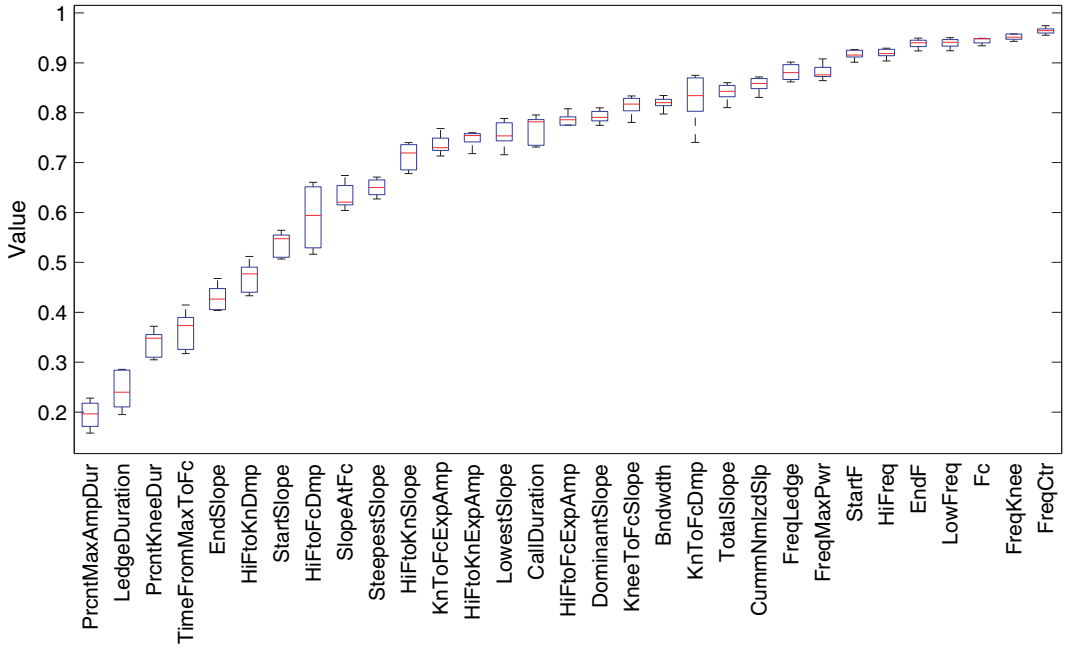

Fig. 4. Estimated length scale parameters on five folds for 31 call parameters; parameters are shown sorted on the basis of their mean value over the five folds to aid interpretation; see the on-line supplementary material Section 6 for a description of the parameters

have calls that are a mixture of constant frequency and frequency-modulated components which are adaptive to their particular environments. For example, one way of distinguishing targets from other objects in cluttered environments is to emit a narrowband call with a long duration, enabling detection of the fluttering wings of an insect against a still background. Alternatively, the bandwidth of the echolocation call can be increased (either through the bandwidth of a single call or through the use of harmonics), which helps to resolve different size classes (Maltby *et al.*, 2009). Thus slope parameters might reflect the interspecific evolution of bat calls for different habitats and different targets and may provide an effective way to distinguish between species in acoustic classification tools. Recent studies of bat classification have also suggested that call shape might be a potential focus for new classification tools (Obrist *et al.*, 2007; Walters *et al.*, 2013; MacLeod *et al.*, 2013; Lundy *et al.*, 2011) and our study adds further support to this idea.

In Fig. 5 the confusion matrix from the best of the fivefold classification results, 15% misclassification rate, are shown. Similar results are obtained by the remaining folds but are not shown. There is an overall high accuracy for all classes. Misclassification rates are higher for species within different families than for species of different families, which is consistent with the idea that some variation in call design is due to evolutionary constraints (closely related species are more similar) (Jones and Teeling, 2006; Jung *et al.*, 2014). This suggests that including an evolutionary constraint, e.g. determined by phylogenetic relationships, might aid with bat classification tools. We found higher misclassification rates with species within *Vespertilionidae*, which are well known to have species with similar calls (Walters *et al.*, 2012). For example, the *vespertilionid Lasiurus xanthinus*, class 18, had a relatively high misclassification rate and was often misclassified as *Antrozous pallidus*, class 13, which needs to be further investigated. However, we also found that the very similar calls of other species within *Vespertilionidae*,

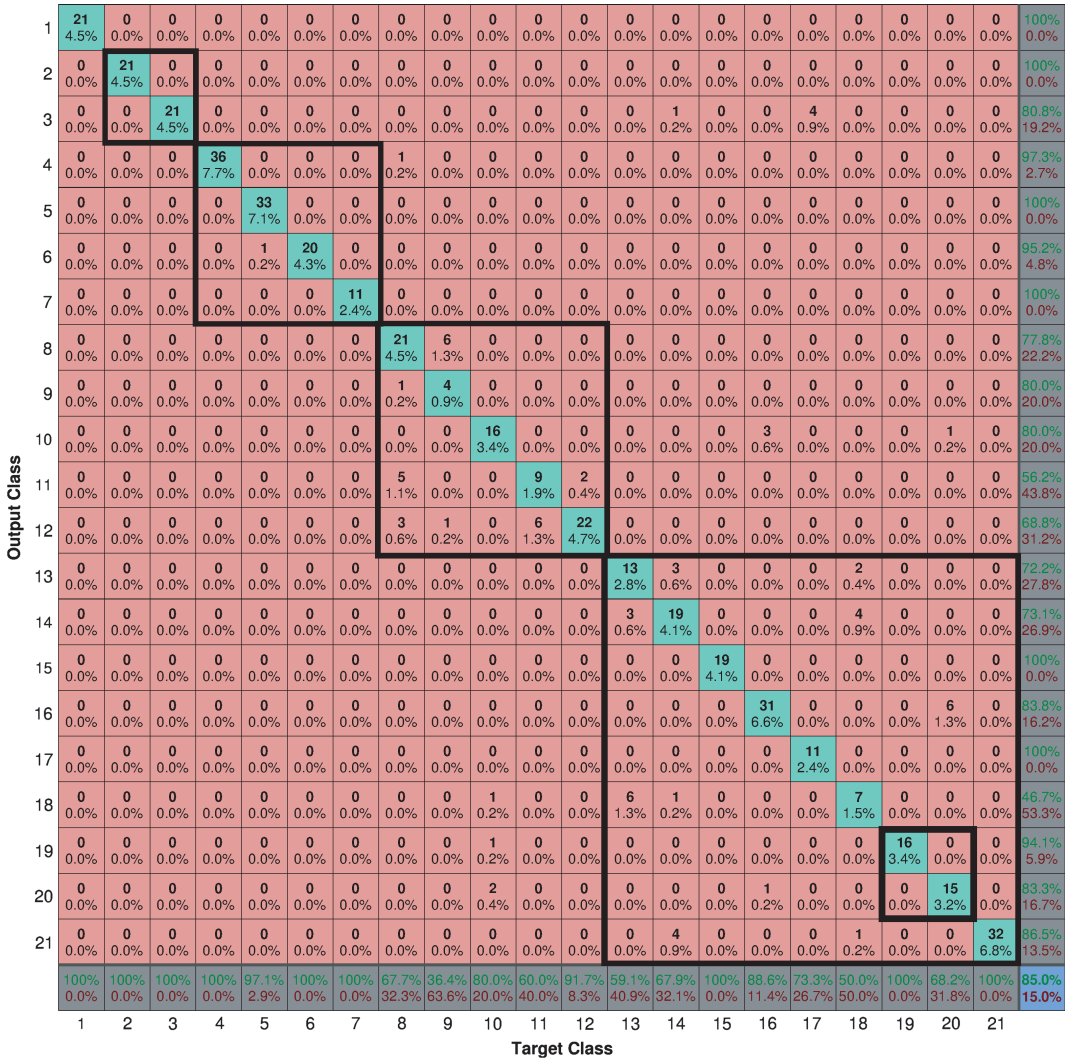

Fig. 5. Confusion matrix obtained from the best of the five folds: classes are in the same order and grouped as in Table 2; bold boxes group species of the same family; the two *Myotis* species are also grouped to highlight the good misclassification rates obtained; the number in the *i*th row and *j*th column is a count of the calls in the test set known to be in species *i* but predicted in species *j*

the *Myotis* species, are easily discriminated. This is a group whose species are traditionally extremely challenging to classify (Walters *et al.*, 2012). However, as we used only a limited number of species, the success of our approach should be explored with larger numbers of *Myotis* species.

9. Conclusion

Automatic bat call classification methods can significantly impact the way that bioacoustic surveys are designed and open up new opportunities for global, long-term biodiversity monitoring programmes. However, for such methods to be applicable they need to provide interpretable

probabilistic outputs and to characterize properly the underlying parameter uncertainty to be generalizable in different operational environments (e.g. different geographic regions or diverse bat assemblages). In this paper we show that multinomial probit regression with GP priors has such attributes. Although exact Bayesian inference is still a challenging problem for this model efficient approximate inference algorithms are available and show promising results. Also recent results (Chopin and Ridgway, 2015) show that the EP approximation that is used in this paper is a suitable alternative for cases where exact inference is not possible. The high accuracy that was obtained in this study to separate difficult-to-classify species (e.g. *Myotis*) sets the ground for a further development of an automatic identification tool and it suggests promising applications to a bigger set of species. Additionally, the fully probabilistic output better quantifies the uncertainty in classification, making species monitoring and subsequent conservation planning more accurate.

References

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass.*, **88**, 669–679.

Amorim, F., Carvalho, S. B., Honrado, J. and Rebelo, H. (2014) Designing optimized multi-species monitoring networks to detect range shifts driven by climate change: a case study with bats in the north of Portugal. *PLOS ONE*, **9**, article e87291.

Andrieu, C. and Roberts, G. O. (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, **37**, 697–725.

Assaraf, R. and Caffarel, M. (1999) Zero-variance principle for Monte Carlo algorithms. *Phys. Rev. Lett.*, **83**, 4682–4685.

Barlow, K. F., Briggs, P. A., Haysom, K. A., Hutson, A. M., Lechiava, N. L., Rarey, P. A., Walsh, A. L. and Langton, S. D. (2015) Citizen science reveals trends in bat populations: The national bat monitoring programme in Great Britain. *Biol. Conservn*, **182**, 14–26.

Butchart, S. H. M., Walpole, M., Collen, B., van Strien, A., Scharlemann, J. P., Almond, R. E., Baillie, J. E., Bonhard, B., Brown, C., Bruno, J., Carpenter, J. E., Carr, G. M., Chanson, J., Chenery, A. M., Csirke, J., Davidson, N. C., Dentener, F., Foster, M., Galli, A., Galloway, J. N., Genovesi, P., Gregory, R. D., Hockings, M., Kapos, V., Lamarque, J. F., Leverington, F., Loh, J., McGeoch, M. A., McRae, L., Minasyan, A., Hernández Morcillo, M., Oldfield, T. E., Pauly, D., Quader, S., Revenga, C., Sauer, J. R., Skolnik, B., Spear, D., Stanwell-Smith, D., Stuart, S. N., Symes, A., Tierney, M., Tyrrell, T. D., Vie, J. C. and Watson, R. (2010) Global biodiversity: indicators of recent declines. *Science*, **328**, 1164–1168.

Ceballos, G. and Oliva, G. (2005) *Los Mamíferos Silvestres de México*. Mexico City: Fondo de Cultura Economica.

Chang, C.-C. and Lin, C.-J. (2011) LIBSVM: a library for support vector machines. *ACM Trans Intell. Syst. Technol.*, **2**, article 27.

Chopin, N. and Ridgway, J. (2015) Leave Pima Indians alone: binary regression as a benchmark for Bayesian computation. *Preprint*. École Nationale de la Statistique et de l'Administration Économique, Paris. (Available from `http://arxiv.org/abs/1506.08640`.)

Damoulas, T., Henry, S., Farnsworth, A., Lanzone, M. and Gomes, C. (2010) Bayesian classification of flight calls with a novel dynamic time warping kernel. In *Proc. 9th Int. Conf. Machine Learning and Applications*, pp. 424–429. Washington DC: Institute of Electrical and Electronics Engineers Computer Society.

Dietterich, T. G. (2000) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Mach. Learn.*, **40**, 139–157.

Dobson, A. J. and Barnett, A. (2008) *An Introduction to Generalized Linear Models*, 3rd edn. London: Chapman and Hall–CRC.

Dudik, M., Phillips, S. J. and Schapire, R. E. (2007) Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *J. Mach. Learn. Res.*, **8**, 1217–1260.

Fenton, M. B. and Bell, G. P. (1981) Recognition of species of insectivorous bats by their echolocation calls. *J. Mammlgy*, **62**, 233–242.

Filippone, M. and Girolami, M. (2014) Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE Trans. Pattn Anal. Mach. Intell.*, **36**, 2214–2226.

Filippone, M., Marquand, A. F., Blain, C. R. V., Williams, S. C. R., Mouro-Miranda, J. and Girolami, M. (2012) Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities. *Ann. Appl. Statist.*, **6**, 1883–1905.

Filippone, M., Zhong, M. and Girolami, M. (2013) A comparative evaluation of stochastic-based inference methods for Gaussian process models. *Mach. Learn.*, **93**, 93–114.

Gardner, J., Kusner, M., Weinberger, K. Q., Cunningham, J. and Xu, Z. (2014) Bayesian optimization with

inequality constraints. In *Proc. 31st Int. Conf. Machine Learning* (eds T. Jebara and E. P. Xing), pp. 937–945. Cambridge: MIT Press.
Girolami, M. and Rogers, S. (2006) Variational bayesian multinomial probit regression with Gaussian process priors. *Neurl Computn*, **18**, 1790–1817.
Girolami, M. and Zhong, M. (2007) Data integration for classification problems employing Gaussian process priors. In *Advances in Neural Information Processing Systems*, vol. 19 (eds B. Schölkopf, J. Platt and T. Hoffman), pp. 465–472. Cambridge: MIT Press.
Griffin, D. R. (1944) Echolocation by blind men, bats and radar. *Science*, **100**, 589–590.
Jones, G. and Holderied, M. W. (2007) Bat echolocation calls: adaptation and convergent evolution. *Proc. R. Soc. Lond.* B, **274**, 905–912.
Jones, G., Jacobs, D., Kunz, T., Willig, M. and Racey, P. (2009) Carpe noctem: the importance of bats as bioindicators. *Endangrd Spec. Res.*, **8**, 93–115.
Jones, K. E., Russ, J. A., Bashta, A.-T., Bilhari, Z., Catto, C., Csősz, I., Gorbachev, A., Győrfi, P., Hughes, A., Ivashkiv, I., Koragina, N., Kurali, A., Langton, S., Collen, A., Margiean, G., Pandourski, I., Parson, S., Prokofev, I., Szodoray-Paradi, A., Szodoray-Paradi, F., Tilova, E., Walters, C. L., Weatherill, A. and Zavarzin, O. (2013) Indicator bats program: a system for the global acoustic monitoring of bats. In *Biodiversity Monitoring and Conservation: Bridging the Gaps between Global Commitment and Local Action* (eds B. Collen, J. E. M. Pettorelli, N. Baillie and S. M. Durant). London: Blackwell.
Jones, G. and Teeling, E. (2006) The evolution of echolocation in bats. *Trends Ecol. Evoln*, **21**, 149–156.
Jung, K., Molinari, J. and Kalko, E. K. V. (2014) Driving factors for the evolution of species-specific echolocation call design in new world free-tailed bats (molossidae). *PLOS ONE*, **9**, article e85279.
Kim, H.-C. and Ghahramani, Z. (2006) Bayesian Gaussian process classification with the em-ep algorithm. *IEEE Trans. Pattn Anal. Mach. Intell.*, **28**, 1948–1959.
Kuss, M. and Rasmussen, C. E. (2005) Assessing approximate inference for binary Gaussian process classification. *J. Mach. Learn. Res.*, **6**, 1679–1704.
Lundy, M., Teeling, E., Boston, E. S. M., Scott, D. D., Buckley, D. J., Prodöhl, P. A., Marnell, F. and Montgomery, W. I. (2011) The shape of sound: elliptic Fourier descriptors (efd) discriminate the echolocation calls of myotis bats (m. daubentonii, m. nattereri and m. mystacinus). *Bioacoustics*, **20**, 101–115.
MacLeod, N., Krieger, J. and Jones, K. E. (2013) Geometric morphometric approaches to acoustic signal analysis in mammalian biology. *Hystrix*, **24**, 110–125.
Magurran, A. E., Baillie, S. R., Buckland, S. T., Dick, J. M., Elston, D. A., Scott, E. M., Smith, R. I., Somerfield, P. J. and Watt, A. D. (2010) Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends Ecol. Evoln*, **25**, 574–582.
Maltby, A., Jones, K. and Jones, G. (2009) Understanding the origination and diversification of bat echolocation calls. In *Handbook of Mammalian Vocalization. an Integrative Neuroscience Approach* (ed. S. Brudzynski), pp. 37–48. London: Academic Press.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Medellín, R. A., Arita, H. and Sánchez, O. (2008) *Identificación de los Murciélagos de México: Clave de Campo*. Mexico City: Asociación Mexicana de Mastozoología.
Minka, T. (2001) Expectation propagation for approximate bayesian inference. In *Proc. 17th A. Conf. Uncertainty in Artificial Intelligence* (eds J. S. Breese and D. Koller), pp. 362–369. San Francisco: Morgan Kaufmann.
Mira, A., Solgi, R. and Imparato, D. (2013) Zero variance Markov chain Monte Carlo for bayesian estimators. *Statist. Comput.*, **23**, 653–662.
Murray, I. and Adams, R. P. (2010) Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems*, vol. 23 (eds J. Lafferty, C. K. I. Williams, R. Zemel, J. Shawe-Taylor and A. Culotta), pp. 1723–1731.
Murray, K., Britzke, E. and Robbins, L. (2001) Variation in search phase calls of bats. *J. Mammlgy*, **82**, 728–737.
Newson, S. E., Evans, H. E. and Gillings, S. (2015) A novel citizen science approach for large-scale standardised monitoring of bat activity and distribution, evaluated in eastern England. *Biol. Conservn*, **191**, 38–49.
Ng, A. Y. and Jordan, M. I. (2002) On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems*, vol. 14 (eds T. Dietterich, S. Becker and Z. Ghahramani), pp. 841–848. Cambridge: MIT Press.
Obrist, M., Boesch, R. and Flückiger, P. (2007) Variability in echolocation call design of 26 Swiss bat species: consequences, limits and options for automated field identification with a synergetic pattern recognition approach. *Mammalia*, **68**, 307–322.
Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2007) A general framework for the parametrization of hierarchical models. *Statist. Sci.*, **22**, 59–73.
Parsons, S. and Jones, G. (2000) Acoustic identification of twelve species of echolocating bat by discriminant function analysis and artifact neural networks. *J. Exptl Biol.*, **203**, 2641–2656.
Rasmussen, C. E. and Williams, C. K. I. (2005) *Gaussian Processes for Machine Learning*. Cambridge: MIT Press.
Riihimaki, J., Jylanki, P. and Vehtari, A. (2013) Nested expectation propagation for Gaussian process classification with a multinomial probit likelihood. *J. Mach. Learn. Res.*, **14**, 75–109.
Robert, C. P. and Casella, G. (2005) *Monte Carlo Statistical Methods*. New York: Springer.

Sakoe, H. and Chiba, S. (1978) Dynamic programming algorithm optimization for spoken word recognition. *IEE Trans. Acoust. Spch Signal Process.*, **26**, 43–49.
Schnitzler, H. U., Moss, C. and Denzinger, A. (2003) From spatial orientation to food acquisition in echolocating bats. *Trends Ecol. Evoln*, **18**, 386–394.
Seeger, M. and Jordan, M. I. (2004) Sparse Gaussian process classification with multiple classes. *Technical Report*. University of California at Berkeley, Berkeley.
Simmons, N. B. (2005) Order Chiroptera. In *Mammal Species of the World: a Taxonomic and Geographic Reference*, 3rd edn (eds D. E. Wilson and D. M. Reeder), pp. 312–529. Baltimore: Johns Hopkins University Press.
Skowronski, M. D. and Harris, J. G. (2006) Acoustic detection and classification of microchiroptera using machine learning: lessons learned from automatic speech recognition. *J. Acoust. Soc. Am.*, **119**, 1817–1833.
Snoek, J., Larochelle, H. and Adams, R. P. (2012) Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, vol. 25 (eds F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger), pp. 2951–2959. Red Hook: Curran Associates.
Szewczak, J. M. (2010) Sonobat v.3. SonoBat, Arcata. (Available from `http://www.sonobat.com`.)
Tittensor, D. P., Walpole, M., Hill, S. L. L., Boyce, D. G., Britten, G. L., Burgess, N. D., Butchart, S. H. M., Leadley, P. W., Regan, E. C., Alkemade, R., Baumung, R., Bellard, C., Bouwman, L., Bowles-Newark, N. J., Chenery, A. M., Cheung, W. W. L., Christensen, V., Cooper, D., Crowther, A. R., Dixon, M. J. R., Galli, A., Gaveau, V., Gregory, R. D., Gutierrez, N. L., Hirsch, T. L., Höft, R., Januchowski-Hartley, R., Karmann, M., Krug, C. B., Leverington, F. J., Loh, J., Kutsch Lojenga, R., Malsch, K., Marques, A., Morgan, D. H. W., Mumby, P. J., Newbold, T., Noonan-Mooney, K., Pagad, S. N., Parks, B. C., Pereira, H. M., Robertson, T., Rondinini, C., Santini, L., Scharlemann, J. P. W., Schindler, S., Sumaila, U. R., Teh, L. S. L., van Kolck, J., Visconti, P. and Ye, Y. (2014) A mid-term analysis of progress toward international biodiversity targets. *Science*, **346**, 241–244.
Walters, C. L., Collen, A., Lucas, T., Mroz, K., Sayev, C. A. and Jones, K. E. (2013) Challenges of using bio-acoustics to globally monitor bats. In *Bat Evolution, Ecology, and Conservation* (eds R. A. Adams and S. C. Pedersen), pp. 479–499. New York: Springer.
Walters, C. L., Freeman, R., Collen, A., Dietz, C., Brock Fenton, M., Jones, G., Obrist, M. K., Puechmaille, S. J., Sattler, T., Siemers, B. M., Parsons, S. and Jones, K. E. (2012) A continental-scale tool for acoustic identification of European bats. *J. Appl. Ecol.*, **49**, 1064–1074.
Williams, C. and Barber, D. (1998) Bayesian classification with Gaussian processes. *IEEE Trans. Pattn Anal. Mach. Intell.*, **20**, 1342–1351.
Yu, Y. and Meng, X. L. (2011) To center or not to center: that is not the question—an ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *J. Computnl Graph. Statist.*, **20**, 531–570.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material for: Bat echolocation call identification for biodiversity monitoring: a probabilistic approach'.