Dolphins Who's Who: A Statistical Perspective

Teresa Barata and Steve P. Brooks

Statistical Laboratory, Centre for Mathematical Sciences,
Wilberforce Road, Cambridge, CB3 0WB, UK
{T.Barata, S.P.Brooks}@statslab.cam.ac.uk
http://www.statslab.cam.ac.uk/~steve/

Abstract. When studying animal behaviour and ecology the recognition of individuals is very important and in the case of bottlenose dolphins this can be done via photo-identification of their dorsal fins. Here we develop a mathematical model that describes this fin shape which is then fitted to the data by a Bayesian approach implemented using MCMC methods. This project is still at a testing stage and we are currently working with simulated data. Future work includes: extracting the outline of the fin shape from the pictures; fitting the model to real data; and devising a way of using the model to discriminate between individuals.

1 Introduction

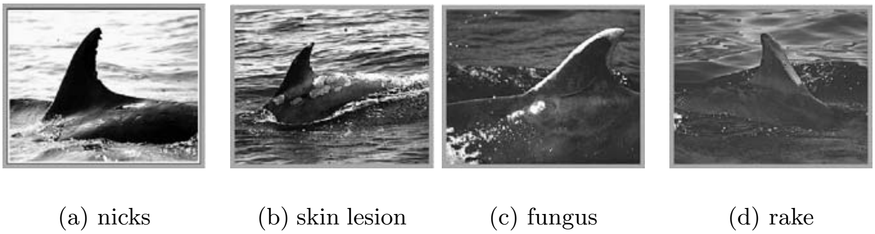
For researchers of animal behaviour and ecology being able to identify individuals plays an important role in their studies and it has been shown that for most large long-living animals this can be done from natural marks. When species live underwater and are only visible for brief moments, individuals are photographed to provide a record of the encounter, which can be compared to a catalogue of pictures for identification purposes, see [9].

For bottlenose dolphins, the dorsal fin (especially its trailing edge) is the most identifying feature and individuals are identified by: the shape of the fin; shading of the fin, scrapes, scratches, nicks and wound marks; and pigment patterns such as fungus. A well-marked dolphin is one that is recognised by more than one single feature.

However there are many factors that complicate the photo-identification procedure: the difficulties in obtaining pictures, the ambiguous nature of the markings on the dolphins and the huge number of matching decisions are some examples. But, probably the biggest problem is the gradual loss of old non-permanent marks and the gain of new ones. As it would be impossible to link the two sets of pictures, the dolphin would then be classified as new a individual. Thus the same animal may appear in the database more than once.

In spite of this, photo-identification has been used on a wide variety of cetaceans, and a few studies now extend for periods of over twenty years. Moreover, the validity of photo-identification by natural markings has been confirmed by studies that combine this technique with tagging.

H. Kalviainen et al. (Eds.): SCIA 2005, LNCS 3540, pp. 429–438, 2005.



(a) nicks (b) skin lesion (c) fungus (d) rake

Fig. 1. Four examples of dolphin fins, courtesy of Paul Thompson, Lighthouse Field Station, Cromarty

In this paper we propose a statistical approach to identify individual dolphins via photographs of their dorsal fins. Section 2 gives a brief overview of the previous approaches to this problem. In section 3 we discuss the mathematical model developed to characterise the dolphin's fin shape. Section 4 is a brief review of Bayesian statistics and MCMC and in section 5 we apply these methodologies to the model in section 3. Some preliminary results using simulated data are presented and analysed in section 6. Finally, we give our conclusions and discuss future work in section 7.

2 Previous Work on Automatic Photo-Identification of Bottlenose Dolphins

The first attempt at developing an automated method for photo-identification of bottlenose dolphins was the Dorsal Ratio method explained in [3]. In this very simple method the top points of the two largest notches on each photograph are labelled A (top) and B (bottom). The Dorsal Ratio (DR) is then defined as being the ratio of the distance between A and B divided by the distance from B to the top of the fin. The catalogue is then examined for fins with similar dorsal ratios, to the one being identified. DR does not depend on the size of the fin and it can also handle moderate cases of parallax. However, it can only be used for dolphins with two or more notches and it may lack consistency as the locations of the tip and notches are defined by the user.

A computer-assisted approach to determine the Dorsal Ratio was implemented in [8]. In this case a digitised image constitutes the input to the system, followed by an edge detection module to generate the edge of the dolphin's dorsal fin of which only the trailing section is considered. The user selects the start and end points of the fin's trailing edge and is also required to prune false notches. This edge can be uniquely represented by its curvature function and is used to automatically extract the Dorsal Ratio.

A more sophisticated approach is the string matching method described in [1]. Edge detection and curvature description are done in exactly the same way as in the previous method, but this time the curvature function, or string, is used to identify each dolphin. As in the previous method only the trailing section is considered. To identify an individual, its string representation is aligned pairwise with the ones in the database and a distance measure between the two is used to assess the degree of matching. This has the disadvantage of only taking into account the trailing edge of the fin shape, and relying on notches to identify dolphins. This method has been coded in a user-friendly graphical interface software called CEDR (Computer Extracted Dorsal Ratio) that can be downloaded from: http://ee.tamu.edu/∼mckinney/Dolf.html.

Another option is DARWIN, a computer vision application developed to assist researchers with photo-identification of dolphins. It starts by using an edge detection algorithm followed by a preprocessing step which allows for the manipulation of the curve in three dimensions (to account for the angle at which the dolphin might have been photographed). First, the centroid of the outline (i.e., the centre of the fin shape, see [4]) is calculated. Next, the distance from the centroid to the points along the outline is found and plotted, in what is known as a signature. The new fin is matched against each database fin at relative rotations to allow for dolphins that were surfacing or diving. Also the position of the centroid is affected by the user designation of the beginning and ending of the outline, so a process that corrects the positions of the centroids was also implemented. When matching is complete, the database fins are presented in rank order of their mean squared error. DARWIN deals with the fact that the dolphins might have been photographed at an angle, but it does so by letting the user estimate this angle in a preprocessing step. This has the disadvantage of being user dependent and the estimated angle can not be changed further during the analysis. DARWIN can be downloaded from: http://darwin.eckerd.edu/

Recently the mismatch area method was introduced, see [7], which is an affine invariant curve matching technique for photo-identification of marine mammals based on silhouettes. The process starts by carrying out edge detection and the output curve is smoothed using cubic B-splines. Subsequently, this curve is matched to the ones in the database by using an affine transformation to force both curves to overlap as closely as possible. Having done so, the mismatch area is computed and the database curves are ranked accordingly.

3 A Model for the Fin Shape

Although automated photo-identification of dolphins usually relies on the position of nicks and notches on their dorsal fins, we propose the use of the overall fin shape instead. The advantages are the following: the overall fin shape does not change even when new marks are acquired, but most importantly, it is a very good way of identifying poorly marked dolphins (which is crucial in areas where dolphins lack predators). As it is quite difficult to distinguish different fin shapes "by eye", a parametric curve can be used to characterise them.

In this section we introduce a parametric model for the dorsal fin shape of bottlenose dolphins. As there is not a single parametric line that describes this

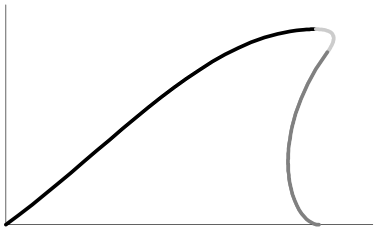


Fig. 2. Example of a fin shape curve

shape accurately, we use segments from three curves, matching their start and end points, as well as, their first derivatives to get a smooth curve.

The three curves chosen were: a Gaussian shaped exponential, which models the back of the fin; a parabola, to model the tip of the fin and a logarithmic spiral, that models the trailing edge of the fin. The model has a total of eight parameters that must be estimated and seven fixed parameters, to do the matching between the curves, as follows:

Exponential:
Parameters to be estimated: E_1 and E_2.

$$\begin{cases} x_e(t) = t \\ y_e(t) = \exp(-(t-E_1)^2\ E_2) - \exp(-E_1^2\ E_2) \end{cases} \quad t \in [0, E_1] \tag{1}$$

Parabola:
Parameters to be estimated: P_1 and P_2.

$$\begin{cases} x_p(t) = -\frac{\cos(\theta)\ (E_1+1-t)^2}{P_1} - \frac{\sin(\theta)\ (E_1+1-t)}{P_1 P_2} + a \\ y_p(t) = -\frac{\sin(\theta)\ (E_1+1-t)^2}{P_1} + \frac{\cos(\theta)\ (E_1+1-t)}{P_1 P_2} + b \end{cases} \quad t \in [E_1, E_1+2] \tag{2}$$

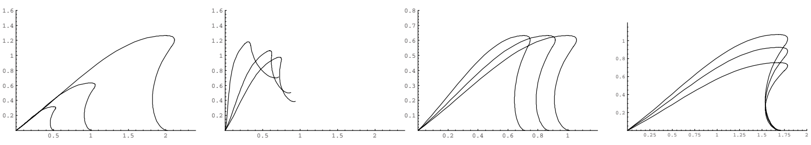
Logarithmic spiral:
Parameters to be estimated: S_1, S_2, S_3, and S_4.

$$\begin{cases} x_s(t) = g\ \sin(S_1\ (E_1+2+2\pi-t))\ \exp(S_2\ (E_1+2+2\pi-t)) + d \\ y_s(t) = g\ c\ \cos(S_3\ (E_1+2+2\pi-t))\ \exp(S_4\ (E_1+2+2\pi-t)) + f \end{cases}$$
$$t \in \left[E_1+2, E_1+2+\frac{3\pi}{4}\right] \tag{3}$$

This gives rise to the shape in Fig. 2, where the exponential curve is given in black, the parabola in light grey and finally the logarithmic spiral in dark grey.

3.1 Angles and Size

The above model must then be fitted to the outline of the fin extracted from the photographs. Hence four more parameters need to be included. These are *size*



(a) *size* (b) *pitch* (c) *yaw* (d) *roll*

Fig. 3. The model curve for different values of the parameters *size*, *pitch*, *yaw* and *roll*

and the angles: *pitch*, if the dolphin is diving or emerging; *yaw*, that takes into account the angle between the camera and the dolphin; and *roll* if the dolphin is rolling on its side. The updated model is given in equation 4,

$$\begin{cases} x(t) = size \;\; [x_{\cdot}(t)\cos(pitch) - y_{\cdot}(t)\sin(pitch)] \;\; \cos(yaw) \\ y(t) = size \;\; [x_{\cdot}(t)\sin(pitch) + y_{\cdot}(t)\cos(pitch)] \;\; \cos(roll) \end{cases} \tag{4}$$

where the subscript is e, p or s according to the value of t and equations 1, 2 and 3. Fig. 3 shows how these parameters affect the curve given in Fig 2.

The inclusion of these additional parameters has the advantage of dealing with the angles in the photograph, while the shape parameters are still being estimated. However, it is now possible for two different fin shapes (i.e. with different shape parameters) to look very similar when one of them has been rotated, making it impossible for the model to distinguish between these two cases. In any case, our goal is not to find a perfect match for the dolphin in the photograph, but to give a ranking of the most likely individuals. Hence, the two fin shapes being discussed would both be considered.

4 Bayesian Statistics and MCMC

As we use Bayesian statistics and MCMC to fit the model discussed in the last section, next we give a brief overview of these methods.

Suppose we are fitting a model with parameters θ to a dataset x. The Bayesian statistics approach allows us to combine information from this dataset with expert prior information on the parameters, see for example [5] for a full account of the Bayesian approach. Both sources of information have to be summarised as probability distributions, these are the likelihood function $L(\theta; x)$ for the data and the prior distribution $p(\theta)$ for the expert information. Bayes' theorem then combines these to obtain the posterior distribution,

$$p(\theta|x) \propto L(\theta; x)p(\theta) \tag{5}$$

Inference on θ usually involves integrating $p(\theta|x)$ to get estimates of its expected value and variance, for example. Often these integrals are either too complicated to be done explicitly or $p(\theta|x)$ is only known up to proportionality. In this case, Markov Chain Monte Carlo (MCMC) methods can be used to sample from $p(\theta|x)$ and obtain sample estimates of the quantities of interest. An overview of the MCMC methods is given in [2].

MCMC is a generalisation of the Monte Carlo simulation methods and is used whenever it is impossible to sample from the posterior distribution directly. In this case a Markov chain with stationary distribution $p(\theta|x)$ is used as an indirect sampling method. Thus after a large enough number of iterations has been simulated, and under certain regularity conditions, these values can be treated as a sample from $p(\theta|x)$. In practice, long simulations are run and iterations within an initial transient phase or burn-in period are discarded.

There are many important implementational issues associated with Bayesian statistics and MCMC. Some of these are technical (choice of priors and convergence diagnosis, for example) but most of them are problem dependent. A good reference is [6] which has a chapter on Markov Chain Monte Carlo and Image Analysis that gives an overview of a variety of image models and the use of MCMC methods to deal with them. It also reviews some of the methodological innovations in MCMC stimulated by the needs of image analysis.

5 Fitting the Model to the Data

In order to use the statistical framework discussed in the last section a likelihood function and priors for the parameters are needed. We assume that the data, i.e. the coordinates for each pixel on the edge of the fin, follows a bivariate normal distribution centred on the model values for these coordinates which were defined in equation 4. That is,

$$(x_i,\ y_i) \sim \mathrm{N}\left((x(t_i),\ y(t_i)), \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right),\ i = 1,\ \cdots,\ k \tag{6}$$

where k is the number of pixels on the edge of the fin and σ^2 models the edge detection and pixelisation errors and is a parameter to be estimated.

As for the prior densities, for the shape parameters we chose the following uniform priors to make sure the resulting curve would look like a fin shape:

$$\begin{aligned}
&p(E_1) = \mathrm{Unif}\left(\left[0,\ 3/\sqrt{2\ E_2}\right]\right),\ p(E_2) = \mathrm{Unif}([0,\ 10]) \\
&p(P_1) = \mathrm{Unif}([10, 70]),\ p(P_2) = \mathrm{Unif}([0.5,\ 2]),\ p(S_1) = \mathrm{Unif}([0.7,\ 1.3]) \\
&p(S_2) = \mathrm{Unif}([0,\ 2.5]),\ p(S_3) = \mathrm{Unif}([0.7,\ 1.3]),\ p(S_4) = \mathrm{Unif}([0,\ 3])
\end{aligned} \tag{7}$$

With respect to the nuisance parameters, the prior for *size* was chosen to depend upon the width and hight of the image (respectively n and m) as the outline of the fin should not be too small and the entire outline has to be within the image. As for the angles, *pitch* does not change the shape of the dolphins' fin and its prior takes that into account by allowing quite big variations of this parameter. The same is not true for *yaw* and *roll*, and only images that, "by eye", have both angles equal to zero, are entered in the database. If the dolphin's right side has been photographed, and not the left as in Fig. 2, *yaw* will be close to π, and not zero. In our case, this problem is dealt with in a preprocessing step where the user selects the side of the dolphin being considered. This is common practice in marine biology, and usually two separate databases are kept for each side. To summarise, we have chosen the following priors for these parameters,

$$\begin{aligned}
&p(size) = \text{Unif}\left[1/64\sqrt{n \times m},\ \sqrt{n \times m}\right],\ p(pitch) = \text{Unif}\left[-\pi/3,\ \pi/3\right] \\
&p(roll) = \text{Unif}\left[0,\ \pi/6\right] \\
&\begin{cases} p(yaw) = \text{Unif}\left[0,\ \pi/6\right], \text{ if side} = \text{left} \\ p(yaw) = \text{Unif}\left[5\pi/6,\ \pi\right], \text{ if side} = \text{right} \end{cases}
\end{aligned} \tag{8}$$

As for σ^2, as there is no prior information, we will assume a non-informative positive prior Gamma(ϵ, ϵ), with ϵ small. Putting this together with equations 6, 7 and 8 we get the following posterior density function (θ represents the shape parameters; μ, *size* and the angles; and $(x,\ y)$ the data),

$$\begin{aligned}
&p\left(\sigma^2, \theta, \mu | (x,\ y)\right) \propto \frac{1}{\sigma^{2k}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{k} \left((x_i - x(t_i))^2 + (y_i - y(t_i))^2\right)\right\} \\
&p(\sigma^2)\ p(\theta)\ p(\mu)
\end{aligned} \tag{9}$$

This is an extremely complicated posterior distribution as the model points $(x(t_i),\ y(t_i))$, defined in equation 4, depend on the parameters in quite a non-linear way. We used MCMC, namely the Gibb's sampler algorithm, in order to do the estimation. The technical details for this algorithm can be found in [2], but the idea is as follows: as the posterior distribution is too hard to work with, a random sample is simulated, assuming at each step that only one of the parameters is a random variable and the others are fixed, that is, we use their posterior conditional distributions. Unfortunately, even in this case, for most parameters (*size* is the exception) we need to sample from a non-standard distribution which is only known up to proportionality. Thus we must use the Metropolis-Hastings algorithm as an indirect sampling method, see [2] for a full account on this methodology. All the parameters were simulated in a similar way, hence details are given for E_1, as an example.

The conditional posterior distribution for E_1 is

$$\begin{aligned}
&p\left(E_1 \mid \sigma^2,\ \theta \setminus \{E_1\},\ \mu,\ (x,\ y)\right) \propto \\
&\exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{k} \left((x_i - x(t_i))^2 + (y_i - y(t_i))^2\right)\right\},\ E_1 \in \left[0,\ 3/\sqrt{2\ E_2}\right]
\end{aligned} \tag{10}$$

And the Metropolis-Hastings algorithm works as follows. Suppose that at a given step the current value of the Markov chain is E_1. A proposed new value E_1' is simulated from the auxiliary distribution defined below.

$$E_1' \sim q(E_1, E_1') = N\left(E_1,\ \sigma^2_{E_1}\right) \tag{11}$$

If $E_1' \notin \left[0,\ 3/\sqrt{2\ E_2}\right]$ it is promptly rejected and the old value E_1 is kept. Otherwise, the probability of E_1' being accepted is given by,

$$\alpha(E_1, E_1') = \min\left\{1, \frac{p(E_1'|.)q(E_1', E_1)}{p(E_1|.)q(E_1, E_1')}\right\}, \text{ where } \frac{p(E_1'|.)q(E_1', E_1)}{p(E_1|.)q(E_1, E_1')} = \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\left((x_i - x'(t_i))^2 - (x_i - x(t_i))^2 + (y_i - y'(t_i))^2 - (y_i - y(t_i))^2\right)\right\} \tag{12}$$

Hence E_1' is more likely to be accepted if its conditional posterior is higher than that of E_1 under the current values of the other parameters. The above steps are repeated until a large enough number of values has been simulated.

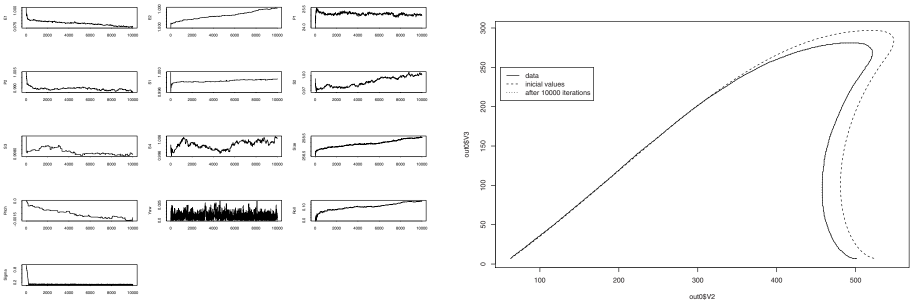
The convergence rate of the chain depends heavily on $\sigma^2_{E_1}$ and this must be small enough so that there is a fair chance E_1' will be accepted but big enough so that different values of E_1 are explored. In this very preliminary stage of our work, while still working with simulated data, we have chosen $\sigma^2_{E_1}$ to be 0.005 on the basis of pilot tuning.

6 Preliminary Results Using Simulated Data

The use of simulated data has the advantage of the true parameter values being known and thus it is a good way to verify the estimation methods described above.

We used the black and white version of Fig. 2 as a simulated curve and transformed it into a JPEG file. This file is scanned into the program that extracts the coordinates of the black pixels and fits the model to them. Fig. 4 provides plots of some preliminary results.

In Fig. 4 (b), it seems the model is fitting the data extremely well, as after 10,000 iterations the model curve is indistinguishable from the data. However, these simulations were based upon starting points chosen to be the true values of the shape parameters. Future work will include running simulations with different starting values. On the other hand, the traceplots given in Fig. 4 (a) show that some of the parameters are highly correlated. For example, the exponential parameters E_1 (the left most graph on the first line) and E_2 (the graph to its right) have inverse behaviours, when E_1 decreases, E_2 increases. These high correlations imply that if the estimated value for one of the parameters is wrong the other parameters are able to compensate so that the overall shape will not be affected. This is an extremely undesirable behaviour, and we are currently looking at ways of making the parameters more independent.



(a) traceplots for the simulated values of the parameters

(b) model curves

Fig. 4. Preliminary results

7 Discussion and Future Work

In this paper we have presented a mathematical model for the fin shape of bottlenose dolphins which also takes into account the angles at which this shape is viewed from. We have also shown preliminary results of fitting the model to simulated data, however this is still very much work on progress and although our results are promising much is still to be done. Future developments include the following. (a) Working with real data. (b) An edge detection algorithm, which ideally would take into account the uncertainty in extracting the outline of the fin. Hence instead of the output being a single curve, for each point we would have the probability of it being on the edge. An alternative would be to deal with this uncertainty while fitting the model, which is the approach we followed in this paper. (c) Devising a way of using the model parameters to compare fin shapes and hence identify individuals. This can be achieved by using Reversible Jump MCMC. The idea is very similar to the Metropolis-Hastings method explained earlier. Suppose at a given iteration it is assumed that the new dolphin is dolphin A in the database. We then propose this new dolphin to be dolphin B, say. This move is accepted with a given probability calculated in a similar way to equation 12. The database dolphins can then be ranked with respect to the proportion of iterations where it was assumed they were the new dolphin. (d) Finally we also wish to compare our method with other available alternatives.

Acknowledgements

The first author is sponsored by Fundação para a Ciência e Tecnologia with the scholarship SFRH/BD/8184/2002.

References

1. B. Araabi, N. Kehtarnavaz, T. McKinney, G. Hillman and B. Würsig, A String Matching Computer-Assisted System for Dolphin Photoidentification , *Annals of Biomedical Engineering*, 2000, vol. 28, pp. 1269-1279.
2. S. P. Brooks, Markov Chain Monte Carlo method and its applications, In: *The Statistician*, 1998, vol.47, pp. 69-100.
3. R. H. Defran, G. M. Shultz, and D. W. Weller, A Technique for the Photographic Identification and cataloguing of dorsal Fins of the Bottlenose Dolphin *Tursiops truncatus*, In: *Individual Recognition of Cetaceans: Use of Photo Identification and other Techniques to Estimate Population Parameters*, Report of the International Whaling Commission, edited by P.S. Hammond, S.A.Mizroch and G.P. Donovan. Cambridge: Cambridge University Press, 1990, vol. 12, pp. 53-55.
4. I. L. Dryden and K. V. Mardia, *Statistical Shape Analysis*, John Wiley, 1998.
5. A. Gelman, J. B. Carlin, H. S. Stern and D. R. Rubin *Bayesian Data Analysis*, Chapman & Hall, 1995. .
6. P. J. Green, MCMC in Image Analysis, In: *Markov Chain Monte Carlo in practice*, W. Gilks, S. Richardson and D. J. Spiegelhalter (eds.), Chapman & Hall,1996, pp. 381-399.
7. C. Gope, N. Kehtarnavaz, G. Hillman and B. Würsig, An affine invariant curve matching method for photo-identification of marine mammals, In: *Pattern Recognition*, 2005, vol.38, pp. 125-132.
8. A. Kreko, N. Kehtarnavaz, B. Araabi, G. Hillman, B. Würsig and D. Weller, Assisting Manual Dolphin Identification by Computer Extraction of Dorsal Ratio", In: *Annals of Biomedical Engineering*, 1999, vol. 27, pp. 830-838.
9. B. Würsig and T. Jefferson, Methods of Photo-Identification for Small Cetaceans, In: *Individual Recognition of Cetaceans: Use of Photo Identification and other Techniques to Estimate Population Parameters*, Report of the International Whaling Commission, edited by P.S. Hammond, S.A.Mizroch and G.P. Donovan. Cambridge: Cambridge University Press, 1990, vol. 12, pp. 43-51.