

A simulation-based method for model evaluation

David J. Allcroft and Chris A. Glasbey

Biomathematics & Statistics Scotland

JCMB, King's Buildings, Edinburgh, EH9 3JZ, Scotland

June 21, 2002

Abstract

We wish to evaluate and compare models that are non-nested and fit to data using different fitting criteria. We first estimate parameters in all models by optimising goodness-of-fit to a dataset. Then, to assess a candidate model, we simulate a population of datasets from it and evaluate the goodness-of-fit of all the models, without re-estimating parameter values. Finally, we see whether the vector of goodness-of-fit criteria for the original data is compatible with the multivariate distribution of these criteria for the simulated datasets. By simulating from each model in turn, we determine whether any, or several, models are consistent with the data. We apply the method to compare three models, fit at different temporal resolutions to binary time series of animal behaviour data, concluding that a semi-Markov model gives a better fit than latent Gaussian and hidden Markov models.

Keywords: Binary time series; Hidden Markov model; Latent Gaussian model; Mahalanobis distance; Model comparison; Semi-Markov model.

1 Introduction

Our aim is to identify good models for the short-term behaviour of animals, to aid researchers in formulating plausible biological mechanisms for behavioural choices of individual animals. We have identified three potential models for binary time series of cow feeding behaviour, which we need to evaluate and compare. A latent Gaussian variable model was fitted to data on a discretised time scale, using an *ad hoc* least squares fitting method. A hidden Markov model was fitted by maximum likelihood, again on a discrete time scale. Finally, a semi-Markov model was fitted by maximum likelihood to the original data, recorded on a continuous time scale.

As the models are non-nested and not all fitted by maximum likelihood, the standard approaches to model comparison, using likelihood ratio tests or information criteria such as AIC (Akaike, 1974) or

BIC (Schwarz, 1978), are ruled out. Similarly, modified Neyman-Pearson likelihood ratio tests (Cox, 1961, 1962), exponential combinations of competing models (Atkinson, 1970) and approaches based on Bayes factors (O’Hagan, 1995), are unavailable. Also, as the models have been fitted to data at differing time scales, all parameters cannot be re-estimated from the simulated data, as is required in simulation methods proposed by Williams (1970), Hinde (1992) and Ross (1998), and parametric bootstraps (Efron and Tibshirani, 1993; Davison and Hinkley, 1997). Instead, we propose a new simulation-based method, whereby we first estimate parameters in all models by optimising goodness-of-fit to our data. Then, to assess a candidate model, we simulate a population of datasets from it and evaluate the goodness-of-fit of all the models, without re-estimating parameter values. Finally, we see whether the vector of goodness-of-fit criteria for the original data is compatible with the multivariate distribution of these criteria for the simulated datasets. By simulating from each model in turn, we determine whether any, or several, models are consistent with the data.

In Section 2 we describe the proposed method and in Section 3 we give a simple illustration of its use and compare it with other approaches. In Section 4 we describe the cow feeding data and the three models, and apply our method to compare them. Finally, in Section 5 we discuss the results.

2 Theory

Our proposal is to simulate a population of datasets from each candidate model and see which, if any, the original data resembles. This is shown schematically in Figure 1 for three models: the strategy is to see whether the top four compartments in the graph resemble any of the clusters of four compartments below. To compare r models, M_1, M_2, \dots, M_r , the steps are:

1. Estimate parameters for models M_1, M_2, \dots, M_r from the observed data by optimising goodness-of-fit criteria.
2. Simulate a population of realisations (e.g. 100) from each model.
3. Compare goodness-of-fit criteria of all models to the original data with the multivariate distributions of goodness-of-fit criteria to the simulated data, without re-estimating parameter values.

We note that, if instead, parameters were re-estimated for every simulated dataset, we would be using a form of parametric bootstrap (Efron and Tibshirani, 1993; Davison and Hinkley, 1997). However, because we anticipate situations where models describe data at different time scales, all parameters cannot be re-estimated from the simulated data. In such cases, when data are simulated from a model with less resolution than the model under which the goodness-of-fit is to be evaluated, this should be compared

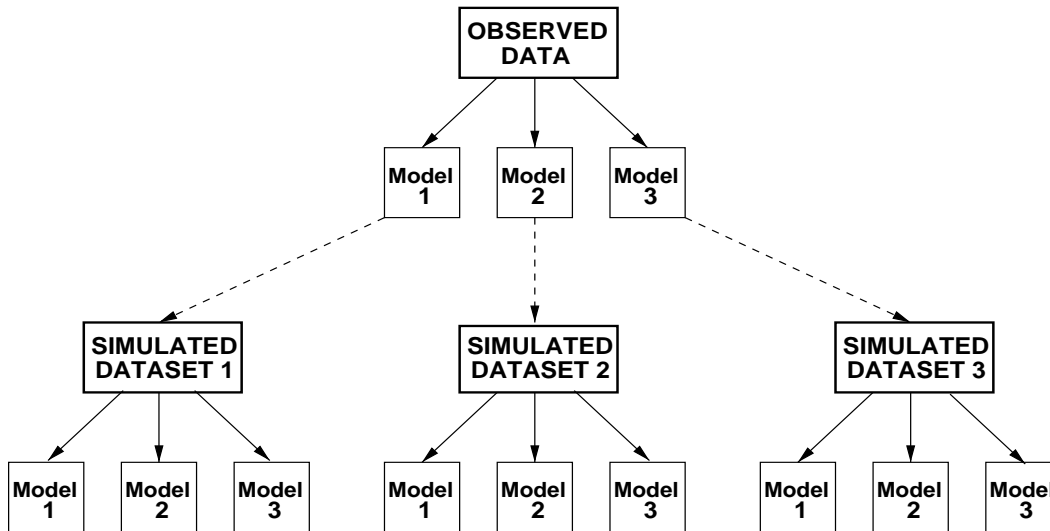


Figure 1: Diagrammatic representation of simulation-based method.

with the goodness-of-fit of other datasets converted to the same resolution, even if available at a finer resolution, in order to retain a fair comparison. This will be demonstrated in Section 4.

If $r = 2$, a two-dimensional scatterplot, of the criterion evaluated for the first model plotted against the criterion for the second model, can be used to assess visually whether the point representing the observed data is consistent with the bivariate distribution of points for either population of simulated datasets. However for three or more competing models, projections onto pairs of axes may fail to reveal lack of fit in higher dimensions. If the joint distributions of the goodness-of-fit criteria are approximately multivariate normal, the Mahalanobis squared distance (Mardia et al., 1979) is sufficient for making model comparisons. This is the standardised squared distance between the point obtained from the observed data and the mean of the i th model, given by

$$D_i^2 = (\mathbf{u} - \bar{\mathbf{u}}_i)' \mathbf{V}_i^{-1} (\mathbf{u} - \bar{\mathbf{u}}_i).$$

Here, \mathbf{u} is the r -dimensional array of goodness-of-fit criteria of the observed data, $\bar{\mathbf{u}}_i$ is the r -dimensional mean of the criteria for the n datasets simulated from the i th model and \mathbf{V}_i is the sample variance matrix for these criteria. If the i th model is correct, then D_i^2 is distributed as $r \times F_{r,n-1}$ distribution. Hence p -values can be obtained to assess evidence against the observed data coming from each model.

3 Illustration

To illustrate and assess the proposed method we compare it with Cox statistics (Cox, 1961, 1962), using a worked example from Cox (1961), which compares the fit of exponential and log-normal distributions

<i>Correct model</i>	<i>Log Normal</i>		<i>Expl</i>	$-\mathcal{L}_1$	$-\mathcal{L}_2$	T_1^*	T_2^*
	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\beta}$				
(a) $\log X \sim N(4, 0.75^2)$	4.00	0.76	72.94	5147	5290	-0.30	7.96
(b) $X \sim \text{Exp}(1/70)$	3.64	1.32	68.58	5333	5228	-7.35	-0.49
(c) $X \sim \text{Gamma}(2, 40)$	4.02	0.79	72.45	5201	5283	-6.81	5.88

Table 1: Maximum likelihood estimates, maximised log-likelihood values, \mathcal{L}_1 and \mathcal{L}_2 , and standardised Cox statistics, T_1^* and T_2^* , for the log-normal, H_1 , and exponential, H_2 , models, for data simulated under each of three models: log-normal, exponential and gamma.

to a sample of data. The hypotheses to be tested are

$$H_1 : \log X \sim N(\mu, \sigma^2)$$

and

$$H_2 : X \sim \text{Exp}(1/\beta).$$

We consider 1000 observations simulated from each of three distributions: (a) log-normal with $\mu = 4$ and $\sigma = 0.75$, (b) exponential with $\beta = 70$, and (c) gamma with shape parameter 2 and scale parameter 40. Parameter values were chosen to give distributions with similar means.

Cox statistics compare the expected value of the likelihood ratio statistic under each of two non-nested models, and conclude data to be consistent with one, both or neither of the two models. The forms of the statistics for this example are given in Cox (1961, equations 60, 71). We denote by T_1^* and T_2^* the standardised values of the statistics, each of which is asymptotically distributed as a standard normal deviate under the correct model. Figure 2 shows histograms of the three datasets along with the maximum likelihood fits and Table 1 gives summary statistics. It can be seen from both Figure 2 and Table 1 that the correct distribution gives a good fit to datasets (a) and (b), and that for dataset (c) neither incorrect distribution gives a good fit, although graphically the log-normal is seen to fit more closely than the exponential.

We applied the method of Section 2 to the three datasets, to assess the two hypotheses. The left side of Figure 3 shows scatterplots of the goodness-of-fit criterion, here a negative log-likelihood, evaluated under H_1 plotted against the criterion under H_2 . For (a), the log-normal data, the point corresponding to the original data falls within the range covered by the simulated data from the log-normal distribution, and for (b), the exponential data, it falls within the range covered by the exponential distribution simulations. For (c), the gamma data, the point for the original data falls outside both regions, although much closer to the log-normal simulations, reflecting the histogram of Figure 2(c). In this application it is also possible to compute a parametric bootstrap, by re-fitting the candidate models to each simulated dataset before evaluating the goodness-of-fit criteria. The results, given on the right of Figure 3, are very similar to those on the left though showing slightly less spread, suggesting that, when available, the parametric bootstrap is a more powerful test.

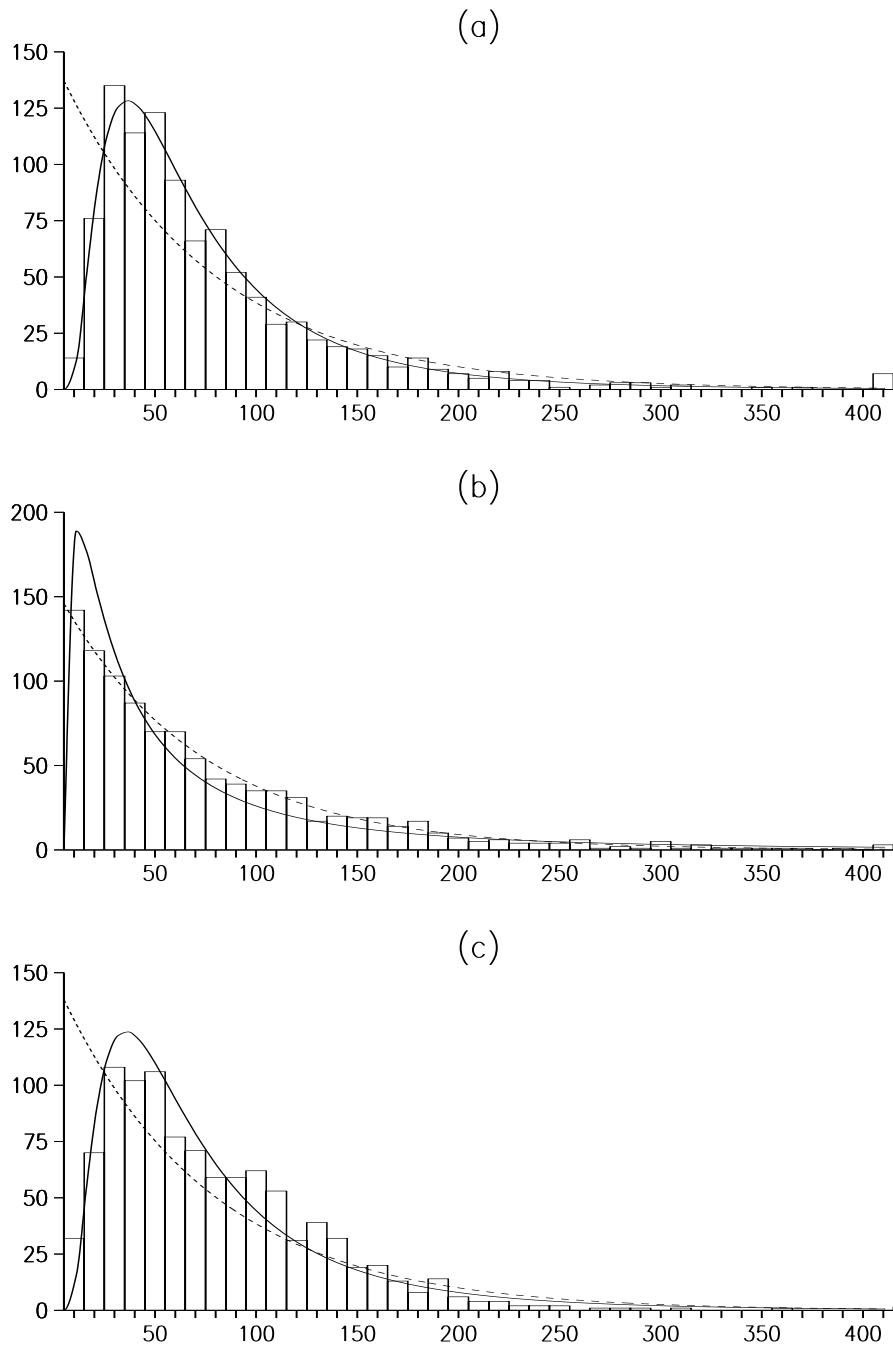


Figure 2: Histograms of 1000 observations from (a) log-normal model, (b) exponential model, (c) gamma model; maximum likelihood fit of (—) log-normal model, (- - -) exponential model.

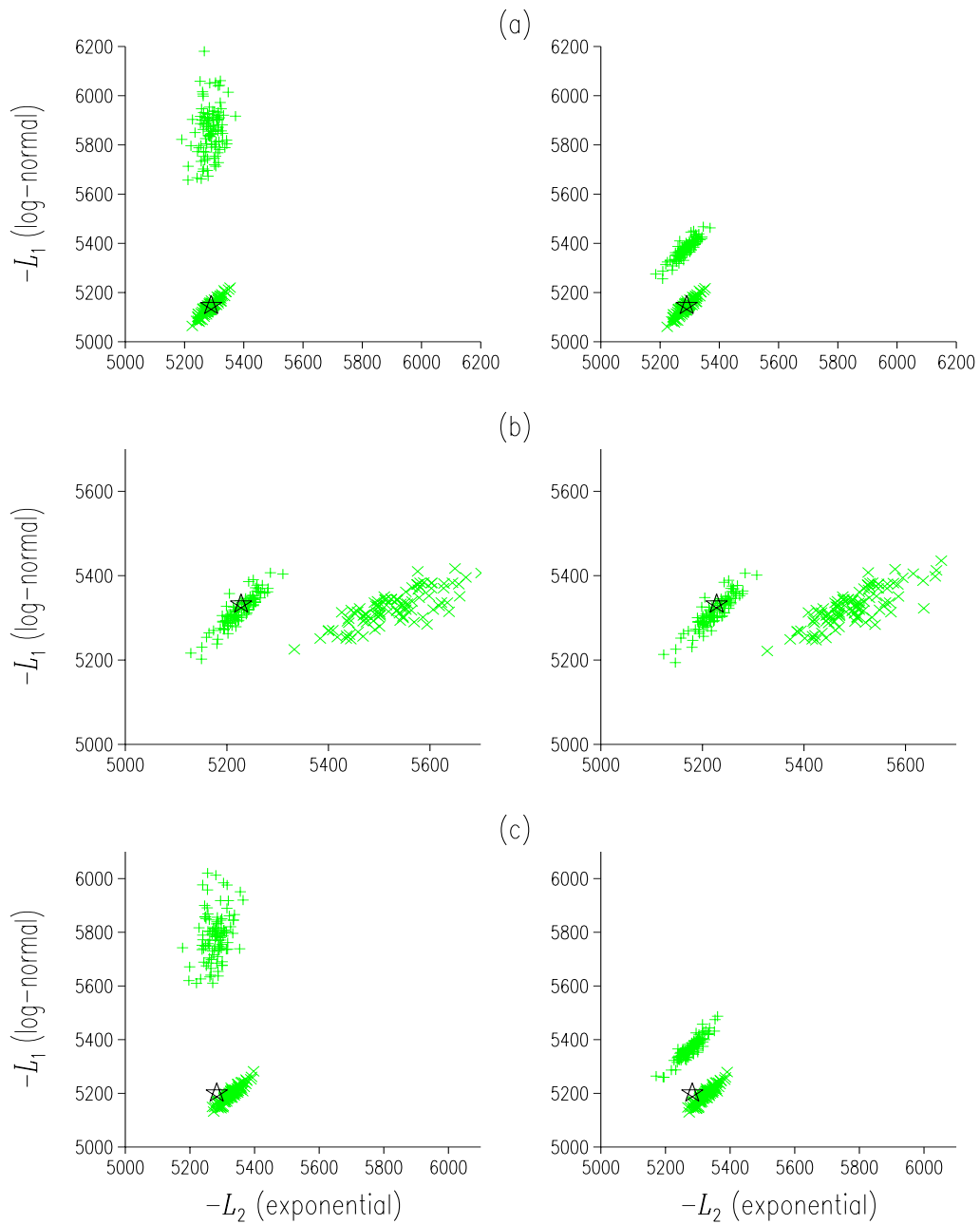


Figure 3: Negative log-likelihoods under log-normal and exponential models; (x) data simulated from the fitted log-normal model, (+) data simulated from the fitted exponential model, (★) original data. Original data is (a) log-normal, (b) exponential or (c) gamma; (left) likelihoods, (right) maximised likelihoods.

<i>Correct Model</i>	<i>Fitted Model</i>	<i>Based on likelihoods</i>		<i>Based on maximised likelihoods</i>	
		D^2	p	D^2	p
		(a) Log-normal	Log-normal	0.06	0.97
	Exponential	52.92	<0.001	179.95	<0.001
(b) Exponential	Log-normal	36.76	<0.001	48.48	<0.001
	Exponential	0.57	0.75	0.70	0.71
(c) Gamma	Log-normal	25.18	<0.001	26.37	<0.001
	Exponential	47.31	<0.001	110.12	<0.001

Table 2: Mahalanobis squared distances, D^2 , and p -values for the original data being consistent with either of the two populations of simulated datasets, summarising Figure 3.

Table 2 shows Mahalanobis squared distances between the original data and the simulated datasets, corresponding to Figure 3, together with p -values for $2 \times F_{2,99}$ distributions. Results can be seen to agree with those obtained graphically and from Cox statistics. For the gamma data, the log-normal model is preferred over the exponential one, though there is still evidence that the data are not from this model.

Anticipating the comparison of models fit using different fitting criteria, the above exercise was repeated, but observations were grouped into bins of various widths (e.g. as in the histograms of Figure 2). Letting O_i be the observed number of observations in group i , and E_i the expected number under the fitted model, the two criteria $\sum_i (O_i - E_i)^2 / E_i$ and $\sum_i (O_i - E_i)^2$ were minimised as the goodness-of-fit criteria. In both cases, similar pictures were obtained to those in Figure 3. By increasing the bin width, the direct effect of using a less efficient estimator can be investigated. Results showed that as the bin width increased, the variability within each cloud of points increased, making the overall position of the clouds relatively closer together. This indicates a decrease in power, nevertheless it was still quite clear from which model the data had been simulated. This demonstrates that the method is robust to comparison of models fit by different criteria.

4 Application

Our data are the records of feeding behaviour for each of a group of eight cows over a 30 day period in April–May 1995 at the Langhill Dairy Cattle Research Centre, Roslin, Midlothian. The cows had continuous access to a high-protein silage/concentrate feed in computerised feeders, transponders worn around their necks enabling the feeders to automatically record the time the cow entered and exited the feeder (Tolkamp and Kyriazakis, 1997; Tolkamp et al., 1998). The data are binary variables, $Z(t)$, recorded on a continuous time scale, t , taking the value 0 when a cow is not feeding and 1 when feeding occurs. To illustrate, Figure 4 shows two days of data for one cow. Let D_j denote the duration of the j th non-feeding interval and E_j the duration of the following feeding event. Thus, if we assume that the

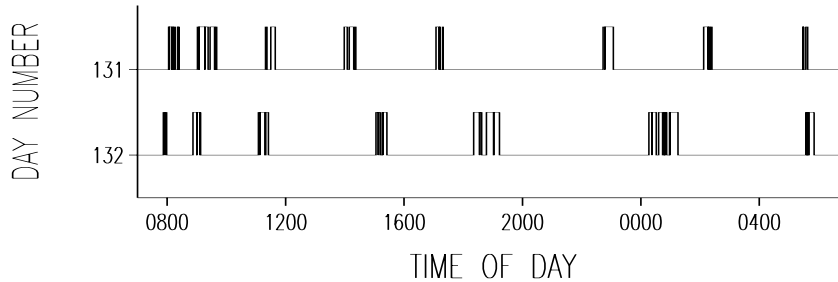


Figure 4: Two days of data for one cow. Raised values of the signal denote periods of feeding.

animal is not feeding at time zero, $Z(t) = 1$ if and only if there exists a J such that

$$\sum_{j=1}^{J-1} (D_j + E_j) + D_J \leq t < \sum_{j=1}^J (D_j + E_j).$$

The semi-Markov model is fit using these data directly. The latent Gaussian and hidden Markov models are formulated in discrete time, and we discretise the data at a one-minute timescale, using X_t , defined by

$$X_t = \begin{cases} 0 & \text{if } \int_{t-1}^t Z(u)du \leq 0.5 \\ 1 & \text{otherwise} \end{cases} \quad \text{for } t = 1, \dots, T.$$

The one-minute discretisation scale is arbitrary, and is a compromise between retaining enough detail and having data series of manageable length, as 30 days of data produces a series of 43200 data points. We now describe our three candidate models and how they were fitted to the data.

4.1 Latent Gaussian model

For the latent Gaussian model, let Y_t denote the value of a latent variable at time t , related to X_t by

$$X_t = \begin{cases} 0 & \text{if } Y_t \leq \Phi^{-1}(1-p) \\ 1 & \text{otherwise.} \end{cases}$$

Here, Φ^{-1} is the inverse normal distribution function and p denotes the overall feeding rate. We model Y as a stationary ARMA(2,1) process (see Allcroft and Glasbey, 2002, for details), given by

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t - \theta \epsilon_{t-1} \quad \text{where } \epsilon_t \sim i.i.d. N(0, \sigma^2),$$

and σ^2 is chosen such that Y has unit variance. Latent variable models are considered by Cox and Snell (1989, pages 101-102), Kedem (1980) and Glasbey and Nevison (1997). As the full data are censored,

maximum likelihood techniques are not available. Various *ad hoc* methods for estimating parameters have been investigated by Allcroft and Glasbey (2002). They found that a method based on maximisation of a spectral quasi-likelihood was computationally quick and sometimes more efficient than other methods based on least squares and pairwise likelihoods. However for simplicity here we have used parameter estimates obtained using least squares, based on matching observed and expected autocorrelations up to a sufficiently large lag (720 minutes).

4.2 Hidden Markov model

For the hidden Markov model (MacDonald and Zucchini, 1997), at each discrete time, t , the observed behaviour X_t of the animal is assumed to depend on an underlying state C_t , i.e.

$$(X_t | C_t = k) = \begin{cases} 0 & \text{with probability } (1 - p_k) \\ 1 & \text{otherwise} \end{cases} \quad \text{for } k = 1, \dots, K.$$

This model has K states, and p_k is the Bernoulli probability of feeding in state k . C_t forms a stationary Markov process with

$$P(C_t = l | C_{t-1} = k) = p_{kl} \quad \text{for } k, l = 1, \dots, K,$$

representing the underlying motivational state of the animal. Hidden Markov models have previously been used extensively for speech processing (see, for example, Rabiner, 1989), and were introduced in the behaviour literature by MacDonald and Raubenheimer (1995). The likelihood l_K for a K -state hidden Markov model can be calculated as

$$l_K = \sum_{k=1}^K \alpha_{Tk},$$

where α_{tk} are *forward probabilities*, calculated recursively for $t = 1, \dots, T$ as $\alpha_{tk} = P(X_1 = x_1, \dots, X_t = x_t, C_t = k)$, see, for example, Le et al. (1992). The likelihood can then be maximised numerically to estimate the parameters. Models of orders $K = 2$ and $K = 3$ were considered (Allcroft, 2001), AIC (Akaike, 1974) and BIC (Schwarz, 1978) showing that the addition of a third state was generally beneficial.

4.3 Semi-Markov model

We formulate a semi-Markov model in continuous time. The duration of feeding events are taken to be independent, with exponential marginal distribution, i.e. $E_j \sim \text{Exp}(\lambda)$. Non-feeding events have a marginal distribution that is a mixture of L log-normal distributions, representing different types of event,

e.g. within-meal or between-meal, hence

$$(\log D_j \mid S_j = l) \sim N(\mu_l, \tau_l^2) \quad \text{for } l = 1, \dots, L,$$

where L is the number of non-feeding states. The model also incorporates first-order dependency in these types of non-feeding period, as S is a stationary Markov process with

$$P(S_j = m \mid S_{j-1} = l) = q_{lm} \quad \text{for } l, m = 1, \dots, L.$$

The states S_j are unobserved so estimation is analogous to that for a hidden Markov model, parameter estimates being obtained by numerical maximisation of the likelihood, again calculated *via* forward probabilities. The parameter for the exponential marginal distribution of feeding events can be directly estimated separately. Models of order $L = 2$ and $L = 3$ were considered. Although in some cases the third non-feeding state was found to be beneficial (again using AIC and BIC, see Allcroft, 2001), we have used $L = 2$ for the comparisons here. This type of model are described more fully by Cox and Isham (1980) and, in the context of animal behaviour data, by Haccou and Meelis (1994).

4.4 Model comparison

We applied the method of Section 2 to assess which, if any, of the three models are consistent with the cow feeding data. As data from the latent Gaussian and hidden Markov models are simulated at a 1-minute timescale, a potential problem arises when the likelihoods of these simulated datasets under the semi-Markov model are to be compared with those of the observed data or semi-Markov simulated data, as both of these are on the continuous scale. To make the comparison fair, we therefore must compare all these likelihoods evaluated from data at a 1-minute scale, i.e. even for the observed and semi-Markov simulated data we use discretised versions of the datasets. In practice this causes no problem, and in fact if, as here, the observed data were recorded in continuous time, the discretisation scale could be smaller, the only disadvantage then being the obvious increase in volume of data and some parameter estimates getting closer to boundaries. The scale should be chosen at least small enough such that most of the detail in the dataset is preserved — in the case here, if several short feeding events were occurring within a minute, a 1-minute discretisation would not be sufficient.

Figure 5 shows bivariate scatterplots of goodness-of-fit criteria for each pair of models fit to data from a single cow. Note that these fitting criteria are negative log-likelihoods for the hidden Markov and semi-Markov models, and sums of squares for the latent Gaussian model. In order to improve the normality of the sums of squares, they have been log-transformed. The plots show the observed data to be clearly inconsistent with the latent Gaussian model. The hidden Markov and semi-Markov models appear much more similar to each other as the regions of the plots containing these simulated data are overlapping.

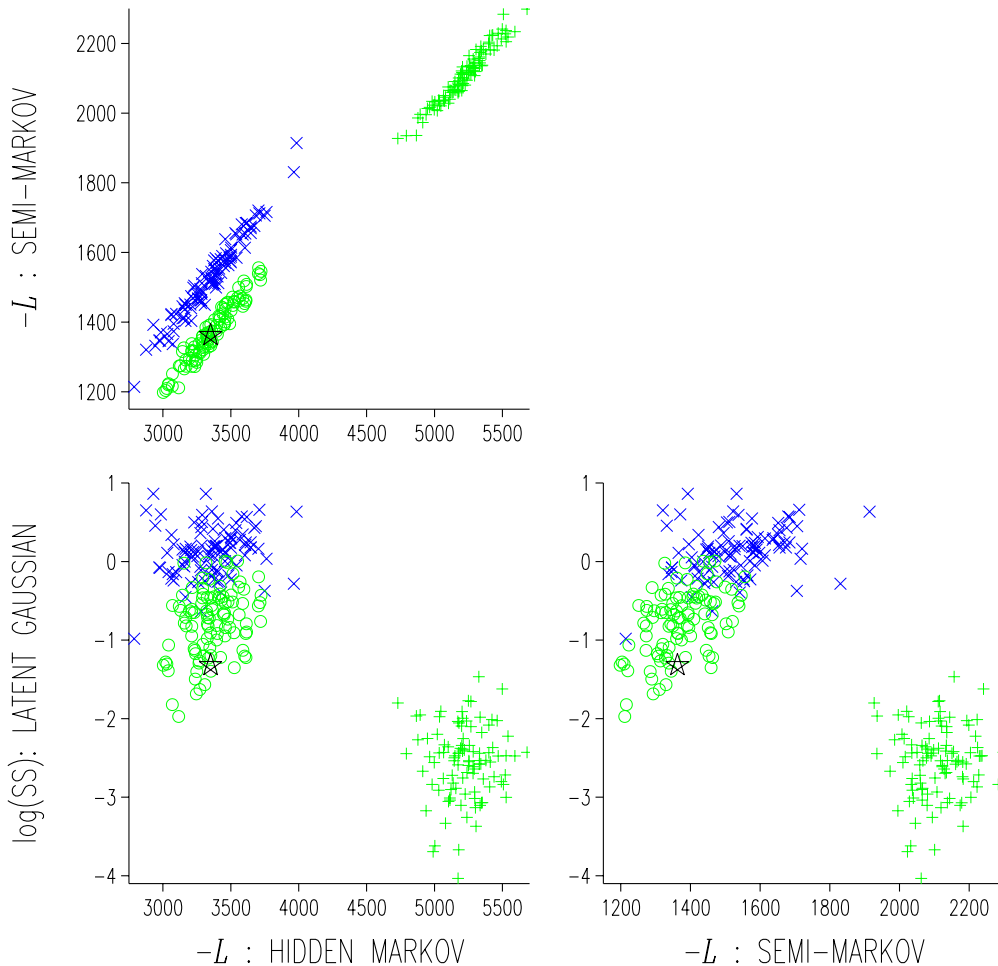


Figure 5: Bivariate scatterplots of pairs of goodness-of-fit criteria for data simulated under (+) latent Gaussian model, (x) hidden Markov model, (o) semi-Markov model; and (★) for the original data.

Nevertheless, the top plot shows the data to be inconsistent with the hidden Markov model. Table 3 gives Mahalanobis squared distances for all eight cows, together with p -values based on $3 \times F_{3,99}$ distributions. Cow 1 in the table is that illustrated in Figure 5. Although there is evidence of lack of fit of the semi-Markov model for some of the animals, in all cases it gives the smallest Mahalanobis distance, confirming that of the models considered, the semi-Markov is the only reasonable choice.

From a biological point of view, the semi-Markov model has the advantage of incorporating the amount of time in the current state into the probability of moving into the next state. In contrast, for the hidden Markov model, the state at the next time-step (minute) is only allowed to depend on what is happening at the current minute and any amount of time already spent in the same state in the immediately preceding minutes is ignored. Both these models are based on moving between discrete states, whereas the latent Gaussian model has a continuous underlying variable. Although simulations from this model can produce

<i>Cow No.</i>	<i>Latent Gaussian</i>		<i>Hidden Markov</i>		<i>Semi-Markov</i>	
	D^2	p	D^2	p	D^2	p
1	117.5	<0.001	41.1	<0.001	2.7	0.45
2	26.3	<0.001	39.4	<0.001	4.3	0.24
3	69.8	<0.001	124.4	<0.001	32.0	<0.001
4	121.4	<0.001	14.9	0.003	0.4	0.94
5	71.7	<0.001	38.1	<0.001	5.2	0.16
6	85.7	<0.001	34.2	<0.001	1.2	0.25
7	105.2	<0.001	58.3	<0.001	8.9	0.04
8	49.5	<0.001	120.3	<0.001	38.4	<0.001

Table 3: Mahalanobis squared distances, D^2 , and p -values for the observed data belonging to each of the three sets of simulated data, with Cow 1 corresponding to Figure 5.

data very similar to that observed, biologically this may not be as intuitive a model. A full discussion of the suitability of these types of models for animal behaviour data can be found in Allcroft et al. (2002), concluding that the extra memory incorporated by the semi-Markov model is important for many types of animal behaviour.

5 Discussion

We have seen that a simulation-based approach can be used to compare models which are not only non-nested, but fit in different ways to data at different temporal resolutions. For the animal behaviour application, we have compared models fit according to different criteria, i.e. maximum likelihood and least squares, and that use the dataset in different forms.

We have used both bivariate scatterplots and the Mahalanobis squared distance to see whether the observed data are consistent with the fitted models. Mahalanobis distances assume the goodness-of-fit criteria to be multivariate normal, i.e. to be elliptical shapes, and assuming this, approximate p -values can be obtained. An alternative would have been to use bootstrap-style p -values, i.e. to order Mahalanobis distances for 99 simulated series and then calculate the p -value for the data as $(100 - i)/100$, where i is the rank of the observed data when included in the ranking of the simulated data. This would be a more satisfactory approach if the spread of the data did not look elliptical; in our case the elliptical assumption looked reasonable and so we chose to use the parametric p -values. Other ideas include the construction of a convex hull around the set of points in r -space and the stripping down of these in order to obtain a p -value for the observed data. A discussion of this, and other ideas of rectangular peeling and elliptical peeling, is given in Green (1981). Projection along principal components or use of other multivariate techniques may also be useful. In Step 2 of the method in Section 2, an alternative, Bayesian, option would have been to simulate using parameters drawn from their posterior distributions instead of point estimates. Lastly, we note that if a model is being used for a specific purpose, there may be more relevant quantities to use as a basis for comparison, other than the fitting criteria.

Acknowledgements

We thank Ilias Kyriazakis, Bert Tolcamp, Colin Aitken and Elizabeth Austin for advice with this work. We are grateful to Langhill Dairy Cattle Research Centre, Edinburgh, for the data and the Scottish Executive Environment and Rural Affairs Department for financial support. We also thank the referees for their comments.

References

- Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, 19:716–723.
- Allcroft, D. J. (2001). *Statistical models for short-term animal behaviour*. PhD thesis, University of Edinburgh.
- Allcroft, D. J. and Glasbey, C. A. (2002). A spectral estimator of ARMA parameters from thresholded data. *Statistics and Computing*. (in press).
- Allcroft, D. J., Tolcamp, B. J., Glasbey, C. A., and Kyriazakis, I. (2002). The importance of ‘memory’ in statistical models for animal feeding behaviour. (submitted).
- Atkinson, A. C. (1970). A method for discriminating between models (with discussion). *Journal of the Royal Statistical Society, Series B*, 32:323–353.
- Cox, D. R. (1961). Tests of separate families of hypotheses. *Proceedings of the Fourth Berkeley Symposium*, 1:105–123.
- Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society, Series B*, 24:406–424.
- Cox, D. R. and Isham, V. (1980). *Point Processes*. Chapman & Hall, London.
- Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*. Chapman & Hall, London, second edition.
- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.
- Glasbey, C. A. and Nevison, I. M. (1997). Rainfall modelling using a latent Gaussian variable. In T. G. Gregoire et. al., editor, *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions*, number 122 in Lecture Notes in Statistics, pages 233–242. Springer, New York.

- Green, P. J. (1981). Peeling bivariate data. In Barnett, V., editor, *Interpreting Multivariate Data*, pages 3–19. John Wiley, Chichester.
- Haccou, P. and Meelis, E. (1994). *Statistical Analysis of Behavioural Data: An Approach Based on Time-structured Models*. Oxford University Press, Oxford.
- Hinde, J. P. (1992). Choosing between non-nested models: a simulation approach. In Fahrmeir, L., Francis, B., Gilchrist, R., and Tutz, G., editors, *Advances in GLIM and Statistical Modelling: Proceedings of the GLIM92 Conference and the 7th International Workshop on Statistical Modelling*, New York. Springer-Verlag.
- Kedem, B. (1980). *Binary Time Series*. Lecture Notes in Pure and Applied Mathematics; v.52. Dekker, New York.
- Le, N. D., Leroux, B. G., and Puterman, M. L. (1992). Exact likelihood evaluation in a Markov mixture model for time series of seizure counts. *Biometrics*, 48:317–323.
- MacDonald, I. L. and Raubenheimer, D. (1995). Hidden Markov models and animal behaviour. *Biometrical Journal*, 37:701–711.
- MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman & Hall, London.
- Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.
- O’Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B*, 57:99–138.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286.
- Ross, G. J. S. (1998). Comparing the fit of non-nested non-linear models. In Payne, R. and Green, P., editors, *COMPSTAT98 Proceedings in Computational Statistics*, pages 431–436, Heidelberg. Physica-Verlag.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.
- Tolkamp, B. J., Dewhurst, R. J., Friggens, N. C., Kyriazakis, I., Veerkamp, R. F., and Oldham, J. D. (1998). Diet choice by dairy cows. 1. Selection of feed protein content during the first half of lactation. *Journal of Dairy Science*, 81:2657–2669.
- Tolkamp, B. J. and Kyriazakis, I. (1997). Measuring diet selection in dairy cows: effect of training on selection of dietary protein level. *Animal Science*, 64:197–207.
- Williams, D. A. (1970). Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. *Biometrics*, 28:23–32.