Prediction of feather damage in laying hens using optical flows and Markov models

Hyoungh-joo Lee¹, Stephen J. Roberts^{1,*}, Kelly A. Drake²
and Marian Stamp Dawkins²

¹*Department of Engineering Science, and*

²*Department of Zoology, University of Oxford, Oxford OX1 3PS, UK*

Feather pecking in laying hens is a major welfare and production problem for commercial egg producers, resulting in mortality, loss of production as well as welfare issues for the damaged birds. Damaging outbreaks of feather pecking are currently impossible to control, despite a number of proposed interventions. However, the ability to predict feather damage in advance would be a valuable research tool for identifying which management or environmental factors could be the most effective interventions at different ages. This paper proposes a framework for forecasting the damage caused by injurious pecking based on automated image processing and statistical analysis. By frame-by-frame analysis of video recordings of laying hen flocks, optical flow measures are calculated as indicators of the movement of the birds. From the optical flow datasets, measures of disturbance are extracted using hidden Markov models. Based on these disturbance measures and age-related variables, the levels of feather damage in flocks in future weeks is predicted. Applying the proposed method to real-world datasets, it is shown that the disturbance measures offer improved predictive values for feather damage thus enabling an identification of flocks with probable prevalence of damage and injury later in lay.

Keywords: animal welfare; condition monitoring; feather pecking; optical flow; hidden Markov model; Gaussian processes

1. INTRODUCTION

A major welfare problem in commercial egg production is that of injurious feather pecking, where the hens peck each other's feathers and skin, leading to feather damage, injury and even cannibalism [1,2]. The problem is worse in barn and free-range systems than in cages [3,4] and so is likely to become even more widespread with the European Union ban on barren cages in 2012. Feather pecking increases feed consumption owing to heat loss [5], reduces egg production [6], causes pain and suffering of the injured birds [7] and increases bird mortality, including cannibalism [8]. Despite a great deal of research it is currently not possible to prevent feather pecking altogether [1,2]. The main method of controlling it is to blunt the birds' beaks to reduce the damage they can do to each other. This blunting is usually done when they are day-olds by use of hot-blade or infra-red beak treatment [9], but it raises welfare problems in its own right [10]. A major research priority is therefore to eliminate feather pecking without resorting to beak trimming.

An important step forward would be an ability to predict outbreaks of feather pecking in advance and to identify those flocks that were at particular risk of damage [11]. A large number of interventions such as

changing the litter, changing the lighting, changing the methods of providing food and water, changing the times of feeding have been proposed, although none are currently completely effective [1,2]. Some of these would cause considerable expense and so it is important to know those or which combinations are most effective and to target the most at-risk flocks. Being able to analyse the effects of such interventions, or other management practices, at an early stage will help to identify key factors in this problem.

This paper proposes a novel method for predicting feather damage based on visual processing and statistical analysis of flock behaviour taken from video recordings. We show that a disturbance profile can be generated for a flock based on optical flow measures of the movements of the birds [12] and the use of hidden Markov models (HMMs; [13]). We then show that this disturbance profile, together with other measures such as the age of the birds, has predictive power as to which flocks will develop severe feather damage in the future.

2. DATA COLLECTION

The procedure is outlined in figure 1. We used video recordings taken inside commercial free-range laying houses for our basic data.

*Author for correspondence (sjrob@robots.ox.ac.uk).

Figure 1. Outline of the method.

2.1. Video recordings and laying flocks

The recordings were made as part of a larger study into the causes of feather pecking in laying hens [11], in which approximately 335 000 commercially reared laying hens were followed from 12 different rearing farms onto 19 laying farms (44 different houses) between February 2006 and August 2008. Laying flock size varied between 800 and 4000 birds. All flocks were visited on at least four occasions: at end of rear (less than 17 weeks, soon after transfer to the laying house (18–22 weeks), peak lay (23–30 weeks) and towards clearance (approx. 50 weeks). A variety of management factors (e.g. feeder and drinker type, stocking density, litter type, range size, etc) and environmental factors (e.g. levels of light, sound, CO₂, ammonia, litter pH, etc) were recorded, as were key production variables (mortality, per cent flock in lay and food consumption (g/h/d)) and feather scores (see below). Details of methods and results are given in Darke *et al.* [11]. The study included both organic free-range and free-range houses and five different breeds (Hyline, Lohman (Brown and Traditional), Shaver, Goldline and Columbian Black Tail). All but two of the houses contained birds that had been beak-treated as day-olds. Video recordings were made at bird level in each house to record behaviour, but these were found not to be suitable for optical flow analysis, so during the second half of the study, covering 13 separate farms and 18 individual houses, additional overhead video recordings were made.

Digital video camcorders (Canon MV940) were affixed inside houses and used to record 1 h video recordings of hens' behaviour within the houses. The recordings were taken of 18 flocks at various ages on separate days, giving a total of 32 useable video files, with each flock recorded at least once and in most cases twice. The files were also subsequently converted to 8-bit greyscale, with four frames-per-second (FPS) and a frame size of 320 × 240 pixels. Between 20 and 100 birds were visible in most frames.

2.2. Optical flows

Most automated visual processing techniques keep track of individual animals, which is very difficult for

a large number of animals without marking individuals and even then computationally prohibitive. Instead, we extract motion from the hens as a whole group using the optical flow algorithm [14]. An optical flow is an approximation to apparent velocities of image motion and can be used to describe the mass movement of a group of objects.

Consider a video file that consists of $T + 1$ image frames each of 320 × 240 pixels. Each image is divided into 1200 (=40 × 30) 8-by-8 pixel blocks. The optical flow algorithm estimates, for each block, a local velocity vector derived by analysis of the frame-by-frame changes between two consecutive image frames at time t and $t + 1$. The velocity vector contains two elements, horizontal and vertical, i.e. $[H_{t,\ell}, V_{t,\ell}]$, for frames at time $t = 1, \dots, T$, and blocks $\ell = 1, \dots, 1200$. From the velocity vectors, the amount of movement for each block was obtained as the magnitude of the velocity, $M_{t,\ell} = \sqrt{H_{t,\ell}^2 + V_{t,\ell}^2}$. Then, as a snapshot of global movement at time t , the spatial mean, $y_t = \frac{1}{1200} \sum_{\ell} M_{t,\ell}$, was sequentially calculated for each frame in the video file, resulting in a time-series of length T .

2.3. Expert labelling of disturbances

As HMMs are used for identifying disturbance periods in the optical flow datasets there is a need of validation against expert opinions. To evaluate the efficacy of the HMMs to give results that correspond to what a human observer might notice, the visual records used for the optical flow analysis were independently viewed by an expert human observer (K.A.D.) who was familiar with the hen behaviour. While this was not an essential part of validating the method (the method might be more sensitive than any human), correlations between the two methods would serve to link the statistical analysis to an understandable measure in the real world and to give an indication of what the model was detecting. The human observer viewed the tapes and then scored a variety of different types of disturbances, such as running, birds pecking each other or several birds suddenly turning in on the same spot. She viewed the tapes without being given any information about age or identity of the flocks. She identified 19 disturbances in 32 video sequences and these same sequences were then presented to the model.

2.4. Scoring feather damage

Levels of feather damage were scored for the flocks using the method described by Bright *et al.* [15], which is summarized in table 1. An observer entered the laying house and visually scored 100 birds inside the laying house and 100 birds on the range outside, using a random number grid to select the birds. The feather scores for the five different body regions (neck, back, rump, tail and wing) were then averaged over all 100 birds to give a single mean feather score for that flock for that day. Further details are given in Darke *et al.* [11]. A total of 198 feather scores were recorded, but where houses were divided into colonies (physical barriers between different parts of the same house), the

Table 1. Description of levels of feather damage (from [15]).

score	description of body	flight feathers
0	well feathered body part with no/little damage	intact
1	slight damage with feathers ruffled, body is completely or almost completely covered	few feathers separated and/or broken/missing
2	severe damage to feathers, but localized naked area ($< 5 \text{ cm}^2$)	all feathers separated and several broken/missing
3	severe damage to feathers and large naked areas ($> 5 \text{ cm}^2$)	most feathers broken/missing
4	severe damage to feathers and large naked areas ($> 5 \text{ cm}^2$)	all feathers missing/broken and/or broken skin

colonies within a house were not considered to be independent, so that the scores for all the colonies within a house were combined, giving a total of $18 \text{ (houses)} \times 4 \text{ (visits)} = 72$ feather scores.

3. EXTRACTION OF DISTURBANCE MEASURES

As we are primarily trying to detect periods of disturbance, which are manifested as state changes in an optical flow time-series, we use a HMM to identify disturbance periods within the dataset. From the state transition identification by the HMM, a set of disturbance measures are obtained for each flock.

3.1. Hidden Markov models

HMMs [13] have been widely used for inferring sequential structures of data in various areas such as speech recognition and bioinformatics. Consider a sequence of observations, $\mathbf{Y} = \{y_1, y_2, \dots, y_T\}$, where $y_t \in \mathcal{R} \forall t$. It is assumed that the probability densities of the observations are determined by the corresponding discrete hidden states, $\mathbf{S} = \{s_1, \dots, s_T\}$, where $s_t \in \{1, \dots, J\}$. Only the observations $\mathbf{Y}$ are available to us, but not the hidden states $\mathbf{S}$. Analogously, in the application of disturbance identification for laying hens, we can observe optical flows, but not the underlying (that is, hidden) states of hens' behaviour, e.g. normal or disturbed. HMMs are hence used to infer a sequence of likely hidden states based on the observed optical flows.

Figure 2 depicts the dependency between the variables as a graphical model. First of all, a hidden state at time $t = 1$, s_1 , is determined by an initial state vector, $\boldsymbol{\pi} = [\pi_1, \dots, \pi_J]^T$, where $\pi_j = p(s_1 = j)$. At time t , given a hidden state at the previous time step s_{t-1} , the current hidden state s_t depends only on s_{t-1} and not on those at time $t - 2$ and before, i.e. the state transitions follow a *Markov* process: $p(s_t | s_{t-1}, \dots, s_1) = p(s_t | s_{t-1})$. By this (first-order) Markov property, the sequential dependency in the data is modelled. The state transition is determined by a transition matrix, $\mathbf{A}$, where $A_{ij} = p(s_t = j | s_{t-1} = i)$. In this paper, the probability density of an observation, given a hidden state, is taken to be a finite mixture of Gaussians (MoG) as this offers the possibility of modelling arbitrarily complex multi-modal distributions (see [16] for example). Taking an MoG emission model introduces another set of hidden variables, $\mathbf{Z} = \{z_1, \dots, z_T\}$, where $z_t \in \{1, \dots, K\}$, an indicator determining which mixture emission component is most responsible for

Figure 2. The graphical model of the HMM with a mixture of Gaussians emission probability. Filled and empty circles represent observations and hidden variables, respectively.

observed y_t . Hence, the emission model has the following probability density for component k :

$$p(y_t | s_t = j, z_t = k) = \mathcal{N}(y_t; \mu_j, \lambda_{jk}^{-1}), \quad (3.1)$$

where μ_j is a mean value for the j th state and λ_{jk} is a precision (or inverse variance) value for the j th state and the k th component. Periods of distinctive behaviour of the laying hens can hence be identified using the HMM, as we interpret the variable s_t as dictating which state the observation y_t is generated from, namely normal or disturbance. Moreover, an MoG emission enables us to model *heteroscedasticity*, i.e. to identify observation sequences with differing variance structures, resulting in a model which is more robust against outliers, thus avoiding spurious state transitions.

A HMM with an MoG emission has the following parameters: an initial state vector $\boldsymbol{\pi}$, a state transition matrix $\mathbf{A}$, a mixture component matrix $\mathbf{C} = [p(z_t = k | s_t = j)]$, a set of mean values $\boldsymbol{\mu} = \{\mu_j\}$ and a set of precision values $\boldsymbol{\Lambda} = \{\lambda_{jk}\}$. Finding a solution for an HMM involves estimation of the parameters, $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \mathbf{A}, \mathbf{C}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$, as well as inference of the posterior distributions of the hidden variables, $p(\mathbf{S}, \mathbf{Z} | \mathbf{Y})$. A two-stage algorithm is usually adopted, known as the Baum–Welch algorithm [17], which itself is a special case of the expectation–maximization (EM) algorithm [18]. However, the conventional EM algorithm has several limitations, most notably its inability to incorporate prior knowledge, its tendency to overfit and its difficulty in model selection.

In order to overcome these limitations, the variational Bayesian (VB) approach [19,20] is adopted in this paper. Given a model order $\mathcal{M} = (J, K)$, the log marginal likelihood, $\log p(\mathbf{Y} | \mathcal{M})$, is approximated by maximizing a lower bound, which is referred to as the *variational free energy* after its original usage in the

Figure 3. Detection of a disturbance period in a section of typical optical flow dataset from a 27-week-old flock. The (blue) solid curve is the observation sequence while the (black) staircase plot is the estimated mean value of the hidden states. One standard deviation error bars are represented by the (pink) shading. The lower panel shows the expert labelling: zero for normal and one for disturbance.

statistical physics literature¹. This quantity is maximized by the VB-EM algorithm where the VB-E and VB-M steps are alternately iterated until convergence. The maximization is equivalent to approximating the joint posterior of the parameters and the hidden variables. In Bayesian model selection, one selects the model order with the highest model posterior, $p(\mathcal{M}|\mathbf{Y}) \propto p(\mathbf{Y}|\mathcal{M})p(\mathcal{M})$. Assuming a flat prior over the model structure, i.e. $p(\mathcal{M}) = \text{const.}$, leads to selecting the model order with the highest (negative) free energy.

Periods of distinctive behaviour can be identified by decoding the most probable path, $\{s_1^*, \dots, s_T^*\}$, resulting in a partition $G_j = \{t | s_t^* = j\}$. The decoding is essentially equivalent to the Viterbi algorithm [13] with the only difference being that the ‘submarginalized’ parameter and likelihood are used. The VB learning algorithm for HMMs is presented in appendix 6. For more details, see Ji *et al.* [21] and Roberts & Penny [22].

3.2. State identification details

The optical flow datasets were log-transformed and a moving median filter was applied over the previous five samples to remove very short-term transients, lasting for one or two time steps. The success of the final predictive model is not sensitive to inclusion of this filter; its inclusion, however, is observed to improve the rate of convergence of the HMM. The data were normalized to have a zero mean and a unit variance, before the application of the HMM. Each frame was classified as normal or disturbance depending on which state it belonged to. In particular, since we know that disturbances are characterized by periods of higher activity, each frame was classified as normal if it was associated with the lowest mean activity hidden state and as a disturbance otherwise.

It is possible, of course, to apply supervised learning, e.g. binary classification, to the disturbance identification. However, this requires a significant amount of labelled data from both ‘classes’, normal and disturbance. For typical datasets, only a small fraction of data belong to the disturbance class, resulting in severe class imbalance.

¹The variational free energy provides a bound to the true (negative) log evidence of the data under the model. The quantity refers to the statistical modelling properties of the model and is not in any way a measure of physical movement or energy in this context.

Table 2. Contingency table of counts (percentages) for state identification results.

true label	estimated label		total
	disturbance	normal	
disturbance	1797 (94.93)	96 (5.07)	1893
normal	366 (2.14)	16 741 (97.86)	17 107
total	2163	16 837	19 000

Moreover, labelled data are, in practice, often difficult, expensive and time-consuming to obtain, as they require the intensive efforts of human experts. We, therefore, focus on the development of unsupervised methodologies in this paper.

Figure 3 shows a typical example of an identification result on a section of data from a 27-week-old flock. The optical flows are close to zero (the global mean over all data) on the whole, but there is a clear increase, around $t = 600$. Bayesian model selection led to a model with two hidden states, $J = 2$, which are interpreted as normal and disturbance, and three components for the mixture model, $K = 3$. The HMM identified the disturbance period at the same location as the expert labelling. Although there are other very visible spikes in the optical flow, e.g. at $t \approx 530$ and $t \approx 900$, the HMM raised no false alarms, as transients in the optical flow were assigned to a high-variance non-disturbance state.

The identification results on 19 such sections of data comprising 19 000 normal and disturbance events were compared with the expert labelling. Table 2 presents the resultant contingency table, where both the true positive and true negative rates were around 95 per cent. A χ^2 -test was carried out, using bootstrap under-sampling so as to alleviate the severe class imbalance, with the null hypothesis that the cells in the table are evenly distributed. It was resoundingly rejected with a p -value of almost zero. We can conclude therefore that disturbance periods can be accurately identified by the methodology we present in this paper. As full state probabilities are returned by the model it is also possible to adjust the trade-offs between false positives and false negatives if there are unequal costs associated with them. In this paper, however, we take the cost associated with all misclassifications to be the same.

Figure 4. A schematic of an identification result. Of T frames, three distinctive states were identified, one for normal and two for disturbance. The j th state lasted for τ_j frames with the mean value of μ_j .

3.3. Disturbance measures

The HMM algorithm detailed in previous subsections was applied to each full-length recording with the model order $(J, K) = (3, 3)$, which was determined by Bayesian model selection using the variational free energy over all datasets. Using the same model structure for all the datasets allows for easier collation and comparison of the results and avoids the computation required to evaluate a full set of models for each and every dataset. An example result is illustrated in figure 4. State 1 is designated as normal, while the other two as disturbances. For the j th state, for example, the mean value and its total duration are denoted by μ_j and τ_j , respectively. We derived three simple measures representing the disturbance profile as shown in table 3.

4. PREDICTION OF FUTURE FEATHER DAMAGE

Our goal is to predict which flocks will develop severe feather damage. As described in §3, we derived a set of disturbance measures, which we will denote by the three-dimensional vector $\mathbf{d}$. Another relevant and readily available piece of information is the age of each flock, a , when the video was taken. We define the interval, Δ , as being the time interval between the dates of video recording and future feather scoring. All variables, $\mathbf{d}$, a , Δ are used for predicting future feather scores. In other words, given the disturbance measures and the ‘video age’, a feather score will be predicted Δ days ahead. This section describes the prediction models employed and then presents the prediction results.

4.1. Prediction models

Consider a dataset of input–output instances, $(\mathbf{X}, \mathbf{f}) = \{(\mathbf{x}_n, f_n)\}$, $n = 1, \dots, N$. An input vector consists of the disturbance measure $\mathbf{d}$, the video age a , and the time interval Δ , i.e. $\mathbf{x}_n = [\mathbf{d}_n, a_n, \Delta_n]^T \in R^5$. An output value $f_n \in R$ corresponds to a future feather score. The number of instances available is 198, i.e. $N = 198$. We here consider two prediction models, one linear and the other nonlinear.

Linear regression (e.g. [23]) is one of the most popular regression models. Given an input vector $\mathbf{x}_*$, the prediction is given by

$$\hat{f}_* = b_0 + \mathbf{x}_*^T \mathbf{b}, \quad (4.1)$$

Table 3. Disturbance measures.

measure of disturbance	formula
frequency	$(\# \text{ disturbances})/T$
proportion	$(\tau_2 + \tau_3)/T$
magnitude	$(\mu_2 - \mu_1) \cdot \tau_2/T + (\mu_3 - \mu_1) \cdot \tau_3/T$

where b_0 and $\mathbf{b} = [b_1, \dots, b_d]^T$ are $d + 1$ parameters to be estimated from the training data.

Gaussian processes (GPs; [24]) perform Bayesian inference by assuming a Gaussian distribution over the infinite numbers of possible outputs of the function to be estimated. A GP is specified by a mean vector, $\mathbf{m}$, and a covariance matrix, $\mathbf{K}$, over the function outputs:

$$\begin{aligned} p(\mathbf{f}|\mathbf{x}) &= \mathcal{N}(\mathbf{f}; \mathbf{m}, \mathbf{K}) \\ &= (2\pi)^{-1/2} |\mathbf{K}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{f} - \mathbf{m})^T \mathbf{K}^{-1}(\mathbf{f} - \mathbf{m})\right), \end{aligned} \quad (4.2)$$

where the covariance matrix is defined using a covariance function, i.e. $K_{nn'} = k(\mathbf{x}_n, \mathbf{x}_{n'})$. In this paper, we adopt the squared exponential covariance function with additive independent output noise. The squared exponential covariance function offers a canonical model which is ideally suited to interpolating smooth functions with sparse data. The reader is pointed to, (e.g. [24]) for full discussion regarding the choice of covariance function.

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \sigma_{\text{out}}^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\sigma_{\text{in}}^2}\right) \\ &\quad + \sigma_{\text{noise}}^2 \delta(\mathbf{x}, \mathbf{x}'), \end{aligned} \quad (4.3)$$

where the hyperparameters σ_{in} and σ_{out} are the input and output length-scales, respectively. In the second term, σ_{noise}^2 is the output noise variance and $\delta(\mathbf{x}, \mathbf{x}')$ is a Kronecker delta, which is one for $\mathbf{x} = \mathbf{x}'$ and zero otherwise.

When a new (test) input vector, $\mathbf{x}_*$, is given, the joint prior distribution of the training outputs, $\mathbf{f}$, and the test output, f_* , is

$$\begin{bmatrix} \mathbf{f} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \mathbf{K} & \mathbf{k}_*^T \\ \mathbf{k}_* & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right), \quad (4.4)$$

where $\mathbf{k}_* \equiv [k(\mathbf{x}_*, \mathbf{x}_1), \dots, k(\mathbf{x}_*, \mathbf{x}_N)]^T$. We can hence easily obtain the predictive distribution as a conditional Gaussian distribution:

$$p(f_*|\mathbf{X}, \mathbf{f}, \mathbf{x}_*) = \mathcal{N}(m_*, \text{Var}[f_*]), \quad (4.5)$$

where the mean and variance are expressed as follows:

$$m_* = \mathbf{k}_*^T (\mathbf{K} + \sigma_{\text{noise}}^2 \mathbf{I})^{-1} \mathbf{f} \quad (4.6)$$

and

$$\text{Var}[f_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T (\mathbf{K} + \sigma_{\text{noise}}^2 \mathbf{I})^{-1} \mathbf{k}_*. \quad (4.7)$$

The mean value, m_* , is usually taken as the point prediction, i.e. $\hat{f}_* = m_*$.

Both models contain parameters or hyperparameters to be optimized. Their performance should be evaluated

Figure 5. GP fitting of feather scores based only on ages. The (blue) circles represent the feather scores against ages, while the (black) curve and the (pink) shading are the GP prediction and its 1 s.d. error bar.

on a sample independent of the training data to avoid overfitting. As only 198 instances were available, leave-one-out validation was adopted to estimate performance on out-of-sample data. Specifically, a single instance from the dataset was assigned as the test data, and the remaining instances as the training data. A model is then trained using the training data and evaluated on the test data. This procedure is repeated for all possible choices of the test data.

In addition, since a feather score is a strictly positive quantity, each output value, f_n , was transformed to $\tilde{f}_n = \log(f_n + 1)$. A prediction $\tilde{m}_*$ was made for $\tilde{f}_*$ and then transformed to $m_* = \exp(\tilde{m}_*) - 1$. For this reason, the error bars on the figures in the next subsection are asymmetric.

4.2. Prediction results

Feather scores depend on the birds' ages when the scores were recorded, since they tend to increase monotonically [11,25]. Figure 5 shows the GP fitting result based only on the age ($=a + \Delta$). The relationship is slightly nonlinear although the prediction by linear regression (in a log scale) was not significantly different. Note that uncertainty increases, as expected, as the hens grow older. While this result shows an obvious relationship between ages and feather scores, better performance could be obtained using the disturbance measures in conjunction with age information.

Figure 6 shows boxplots of absolute errors made by linear regression and Gaussian process regressors. For each model, two sets of input variables were compared: with age variables only (a and Δ) and with both the age variables and the disturbance measures. Linear regression performed slightly better with the disturbance measures, although the improvement was statistically insignificant according to a Brown–Forsythe test [26]. On the other hand, the GP was considerably better with the disturbance measures included. The variances of the errors were significantly different between the two sets of input variables with a p -value of less than 1 per cent (using the Brown–Forsythe test).

Table 4 lists the correlation coefficients between the true and the predicted feather scores with respect to the

Figure 6. Boxplots of absolute errors by the two prediction models with respect to the input variables used. Each box has three lines at the lower quartile, median and upper quartile values, with the extended lines showing the range of the rest of the data. (a) Linear regression, (b) GP.

Table 4. Correlation coefficients between true and predicted feather scores. All the values are statistically significant with a p -value of 0.01 or lower.

variables	linear regression	GP
age	0.7197	0.7136
disturbance and age	0.7497	0.8710

input variables used. Respectable performance was observed when the age variables were used. However, the best results were obtained using both sets of variables. The correlation coefficient of 0.8710 from the GP was higher than any other values with a p -value of less than 1 per cent according to a test based on Fisher's Z -transformation (e.g. [27]). Linear regression did not perform much better with the disturbance measures included, implying that the relationship between the disturbance measures and the (log) feather scores is nonlinear. Figure 7a shows an example plot of GP out-of-sample prediction for a 10-week-old flock as a function of prediction time. Figure 7b shows, for comparison, the GP output using age only as a predictive measure for future feather score.

5. DISCUSSION

Our results suggest that flocks at greatest risk of later developing severe feather damage can be recognized

Figure 7. GP prediction results on a 10-week-old flock. The (blue) circles represent feather scores against the days ahead of the prediction. The (black) curves and the (pink) shadings are the predicted feather scores and their 1 s.d. error bars, respectively.

several weeks earlier using optical flow analysis of flock behaviour. Such prediction would be most valuable if it could be used to trigger the use of measures to prevent or control feather pecking, which is currently very difficult to do, especially if birds are not beak trimmed. However, being able to predict which flocks are most at risk might at least allow control measures such as lowered light levels or additional environmental enrichment to be targeted at these at-risk flocks, rather than used indiscriminately on all flocks—owing to cost or because some, such as straw bales or vegetation, are a biosecurity risk. Furthermore, detecting the early warning signs of severe feather pecking will be a valuable research tool in the search for more ethically acceptable ways of preventing and controlling feather pecking. As a caveat, it should be noted that the proposed method was used for predicting feather scores, not feather pecking itself. Feather pecking and feather scores, though closely related to each other, are not necessarily equivalent.

Optical flow measurements are now increasingly used for a range of applications such as monitoring traffic flow [28], the movement of glaciers [29], cell and sperm motility [30], the study of human crowds [31] and for studying groups of non-human animals [32]. Optical flow can be used both to track individuals and to recognize specific behaviour patterns [33], but its great advantage is that it can also be used to analyse the mass movements of whole groups such as herds, flocks or schools, where tracking individual animals is difficult or impossible. Whole frames or sections of frames, containing tens or hundreds of individuals can

be assessed together and the basic statistical properties of these flow patterns can be derived automatically and continuously [34] using algorithms simple enough to deliver results continuously in real time. The flow patterns can of course potentially be analysed in many different ways and the key issue then becomes to find the appropriate statistical analysis that detects patterns of biological significance. In this particular case, we have shown that statistical anomalies detected by using HMM appear to be of significance to the future welfare of laying hens in that they appear to predict future feather damage before it has actually occurred. At risk flocks show signs of behavioural disturbance that this method is able to detect, but the method described here is only one of a range of possible models which have yet to be explored such as an HMM with an autoregressive (AR; [35]), or a generalized AR emission model [22] or even, a HMM with an infinite number of states [36]. We suggest that similar approaches might also be useful for early detection of other signs of poor welfare such as tail-biting in pigs or even subclinical signs of disease. Current research is investigating whether the spatial distribution of optical flow, on the image frames, improves the analysis. It also remains to be seen whether, on short timescales, these measures offer predictive information regarding the occurrence or prevalence of disturbance events.

All the methods considered in this paper were implemented using the MATLAB programming language. Code for performing generic modelling using Bayesian dynamical models may be found under <http://www.robots.ox.ac.uk/~parg/software.html>.

6. CONCLUSIONS

This paper addresses the problem of predicting feather damage of laying hen flocks based on visual processing and statistical analysis. From video recordings, optical flows were extracted as indicators of the hens' movements. HMM were then applied to the optical flow datasets to identify disturbance periods and to derive the disturbance measures. Using these disturbance measures and age-related variables, feather scores were predicted via linear regression and nonlinear regression using Gaussian processes. It was shown that the Gaussian process model performed well and that the disturbance measures provided predictive information regarding future feather scores.

The work was carried out as part of the BBSRC Animal Welfare Initiative (grant no. BBC5189491) and we would like to thank the BBSRC for supporting H.j.L. The data were collected with support from Defra (contract no. AW1134) and we would like to thank them for supporting K.A.D.

APPENDIX A. VARIATIONAL BAYESIAN HMM

This section describes the VB learning algorithm for a HMM. First of all, the priors are defined as $p(\theta) = p(\pi)p(\mathbf{A})p(\mathbf{C})p(\mu)p(\Lambda)$, each of which is to be conjugate, that is, Dirichlet for π , $\mathbf{A}$ and $\mathbf{C}$, Gaussian for

μ , and Gamma for Λ as follows:

$$p(\pi) = \mathcal{D}(\pi; \mathbf{u}^\pi), \quad (\text{A } 1)$$

$$p(\mathbf{A}) = \prod_{i=1}^J \mathcal{D}(\mathbf{A}_i; \mathbf{u}_i^A), \quad (\text{A } 2)$$

$$p(\mathbf{C}) = \prod_{j=1}^J \mathcal{D}(\mathbf{C}_j; \mathbf{u}_j^C), \quad (\text{A } 3)$$

$$p(\mu) = \prod_{j=1}^J \mathcal{N}(\mu_j; w_j, r_j^{-1}) \quad (\text{A } 4)$$

$$\text{and } p(\Lambda) = \prod_{j=1}^J \prod_{k=1}^K \mathcal{G}(\lambda_{jk}; \nu_{jk}, v_{jk}), \quad (\text{A } 5)$$

where $\mathbf{u}^\pi$, $\mathbf{u}^A$, $\mathbf{u}^C$, $\{w_j\}$, $\{r_j\}$, $\{\nu_{jk}\}$ and $\{v_{jk}\}$ are hyperparameters, and $\mathcal{D}(\cdot)$, $\mathcal{N}(\cdot)$, $\mathcal{G}(\cdot)$ denote the Dirichlet, Gaussian and Gamma distributions, respectively. The density functions and other relevant quantities of the distributions are given in appendix A.5.

A lower bound of a log marginal likelihood can be obtained using Jensen's inequality:

$$\begin{aligned} \log p(\mathbf{Y}) &= \log \int \sum_{\mathbf{S}, \mathbf{Z}} p(\mathbf{Y}, \mathbf{S}, \mathbf{Z}, \theta) d\theta \\ &= \log \int \sum_{\mathbf{S}, \mathbf{Z}} q(\mathbf{S}, \mathbf{Z}, \theta) \frac{p(\mathbf{Y}, \mathbf{S}, \mathbf{Z}, \theta)}{q(\mathbf{S}, \mathbf{Z}, \theta)} d\theta \\ &\geq \int \sum_{\mathbf{S}, \mathbf{Z}} q(\mathbf{S}, \mathbf{Z}, \theta) \log \frac{p(\mathbf{Y}, \mathbf{S}, \mathbf{Z}, \theta)}{q(\mathbf{S}, \mathbf{Z}, \theta)} d\theta \\ &= F(q(\mathbf{S}, \mathbf{Z}, \theta)). \end{aligned} \quad (\text{A } 6)$$

The last term, $F(q(\mathbf{S}, \mathbf{Z}, \theta))$, is called the variational free energy. It is easy to show that $\log p(\mathbf{Y}) = F(q(\mathbf{S}, \mathbf{Z}, \theta)) + \text{KL}(q(\mathbf{S}, \mathbf{Z}, \theta) \| p(\mathbf{S}, \mathbf{Z}, \theta | \mathbf{Y}))$, where the $\text{KL}(\cdot \| \cdot)$ is the Kullback–Leibler (KL) divergence between two distributions. Hence, maximizing $F(q(\mathbf{S}, \mathbf{Z}, \theta))$ is equivalent to approximating $p(\mathbf{S}, \mathbf{Z}, \theta | \mathbf{Y})$ by $q(\mathbf{S}, \mathbf{Z}, \theta)$. For brevity, the model order, $\mathcal{M} = (J, K)$, is omitted from all the conditional terms and assumed implicit.

Since approximation of the joint posterior is analytically intractable, the *mean-field* approximation is adopted, i.e. the approximate posterior can be factorized as

$$\begin{aligned} q(\mathbf{S}, \mathbf{Z}, \theta) &= q(\mathbf{S}, \mathbf{Z}) q(\theta) \\ &= q(\mathbf{S}, \mathbf{Z}) q(\pi) q(\mathbf{A}) q(\mathbf{C}) q(\mu) q(\Lambda). \end{aligned} \quad (\text{A } 7)$$

Consequently, the free energy is decomposed as

$$\begin{aligned} F(q(\mathbf{S}, \mathbf{Z}, \theta)) &= \int \sum_{\mathbf{S}, \mathbf{Z}} q(\mathbf{S}, \mathbf{Z}, \theta) \log \frac{p(\mathbf{Y}, \mathbf{S}, \mathbf{Z} | \theta) p(\theta)}{q(\mathbf{S}, \mathbf{Z}, \theta)} d\theta \\ &= \int \sum_{\mathbf{S}, \mathbf{Z}} q(\mathbf{S}, \mathbf{Z}) q(\theta) \log \frac{p(\mathbf{Y}, \mathbf{S}, \mathbf{Z} | \theta)}{q(\mathbf{S}, \mathbf{Z})} d\theta \\ &\quad + \int q(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta \\ &= F(q(\mathbf{S}, \mathbf{Z})) + F(q(\theta)). \end{aligned} \quad (\text{A } 8)$$

The approximate posteriors are found using the VB-EM algorithm. The hidden variables part, $F(q(\mathbf{S}, \mathbf{Z}))$, is maximized in the VB-E step while the parameters part, $F(q(\theta))$, is maximized in the VB-M step.

In the VB-E step, we have

$$\begin{aligned} F(q(\mathbf{S}, \mathbf{Z})) &\stackrel{c}{=} \int \sum_{\mathbf{S}, \mathbf{Z}} q(\mathbf{S}, \mathbf{Z}) q(\theta) \log \frac{p(\mathbf{Y}, \mathbf{S}, \mathbf{Z} | \theta)}{q(\mathbf{S}, \mathbf{Z})} d\theta \\ &\stackrel{c}{=} \sum_{\mathbf{S}, \mathbf{Z}} q(\mathbf{S}, \mathbf{Z}) \log \frac{q^\star(\mathbf{S}, \mathbf{Z})}{q(\mathbf{S}, \mathbf{Z})} \\ &= -KL(q(\mathbf{S}, \mathbf{Z}) \| q^\star(\mathbf{S}, \mathbf{Z})), \end{aligned} \quad (\text{A } 9)$$

where $q^\star(\mathbf{S}, \mathbf{Z}) \propto \exp[\langle \log p(\mathbf{Y}, \mathbf{S}, \mathbf{Z} | \theta) \rangle_{q(\theta)}]$. The relation $\stackrel{c}{=}$ means that the equality holds up to a constant, and $\langle \cdot \rangle_{q(\cdot)}$ indicates an expectation with respect to the specified distribution. From equation (A 9), maximizing $F(q(\mathbf{S}, \mathbf{Z}))$ is equivalent to approximating $q(\mathbf{S}, \mathbf{Z})$ to $q^\star(\mathbf{S}, \mathbf{Z})$. Likewise, in the VB-M step, notating $q(\theta) = \prod_{\ell=1}^L q(\theta_\ell)$, we have

$$\begin{aligned} F(q(\theta_\ell)) &\stackrel{c}{=} \int q(\theta_\ell) \log \frac{q^\star(\theta_\ell)}{q(\theta_\ell)} d\theta_\ell \\ &= -KL(q(\theta_\ell) \| q^\star(\theta_\ell)). \end{aligned} \quad (\text{A } 10)$$

The posterior of a parameter is approximated as

$$q^\star(\theta_\ell) \propto p(\theta_\ell) \exp[\langle \log p(\mathbf{Y}, \mathbf{S}, \mathbf{Z} | \theta) \rangle_{q(\mathbf{S}, \mathbf{Z}), q(\theta_{-\ell})}], \quad (\text{A } 11)$$

where $\theta_{-\ell}$ denotes all the parameters excluding θ_ℓ . Since the conjugate priors are defined, $q^\star(\theta_\ell)$ can be specified without actually computing the normalization terms. These two steps are iterated alternately until some terminating conditions. After finding the optimal model order, the most probable path of the hidden states is decoded. Algorithm 1 outlines the VB learning algorithm. The following subsections provide more detailed descriptions of the algorithm.

A.1. VB-E step

Suppose that the posteriors of the parameters are available from the previous iteration, that is, equations (A 19)–(A 23). Then, the ‘submarginalized’ likelihood for each observation, y_t , given a hidden state and a mixture component variables, is evaluated as

$$\begin{aligned} \log \tilde{p}(y_t | s_t = j, z_t = k) &\stackrel{c}{=} \langle \log \mathcal{N}(y_t; \mu_j, \lambda_{jk}^{-1}) \rangle_{q^{(\text{old})}(\mu_{jk}), q^{(\text{old})}(\lambda_{jk})} \\ &= -\frac{1}{2} \log 2\pi + \frac{1}{2} \langle \log \lambda_{jk} \rangle \\ &\quad - \frac{1}{2} \tilde{v}_{jk}^{(\text{old})} \tilde{v}_{jk}^{(\text{old})} \left\{ (y_t - \tilde{w}_j^{(\text{old})})^2 + \frac{1}{\tilde{r}_{jk}^{(\text{old})}} \right\}, \end{aligned} \quad (\text{A } 12)$$

where $\langle \log \lambda_{jk} \rangle$ is given in equation (B 7). A superscript of ‘(old)’ means that the posterior or the hyperparameter has been obtained in the previous iteration. A tilde ($\tilde{\cdot}$) indicates an updated posterior hyperparameter, probability, or parameter obtained by marginalizing nuisance variables.

Algorithm 1. The pseudo code for the VB learning algorithm for a HMM.

```

1: INITIALIZATION
2:   Specify the model order,  $\mathcal{M} = (J, K)$ .
3:   Define the prior hyperparameters. //Equations (A 1)–(A 5)
4: REPEAT
5:   VB-E step to infer the hidden variables. //Section A.1
6:   VB-M step to estimate the posteriors of the parameters. //Section A.2
7:   Calculate the variational free energy,  $F$ . //Section A.3
8: UNTIL  $F$  is converged or the maximum number of iterations is reached.
9: MODEL SELECTION
10:  Find the most probable model order. //Section A.4
11: STATE IDENTIFICATION
12:  Find the most probable path of states. //Section A.5
    
```

Consider the following relation:

$$\begin{aligned}\tilde{p}(y_t|s_t) &= \sum_{z_t} \tilde{p}(y_t, z_t|s_t) \\ &= \sum_{z_t} \tilde{p}(y_t|s_t, z_t) \tilde{p}(z_t|s_t),\end{aligned}\quad (\text{A } 13)$$

where $\tilde{p}(z_t|s_t)$ corresponds to $\tilde{\mathbf{C}} = [\tilde{C}_{jk}] \propto [\exp(\langle \log C_{jk} \rangle)]$, given in equation (B 2). The forward and backward recursions are, respectively, as follows:

$$\alpha_t \equiv \tilde{p}(s_t|y_{1:t}) \propto \tilde{p}(y_t|s_t) \sum_{s_{t-1}} \tilde{p}(s_t|s_{t-1}) \alpha_{t-1} \quad (\text{A } 14)$$

and

$$\begin{aligned}\beta_t &\equiv \tilde{p}(y_{t+1:T}|s_t) \\ &= \sum_{s_{t+1}} \tilde{p}(y_{t+1}|s_{t+1}) \tilde{p}(s_{t+1}|s_t) \beta_{t+1},\end{aligned}\quad (\text{A } 15)$$

where $\alpha_1 \propto \tilde{p}(y_1|s_1) \tilde{p}(s_1)$ and $\beta_T = [1, \dots, 1]^T$. The initial state probability, $\tilde{p}(s_1)$, and the transition probability, $\tilde{p}(s_t|s_{t-1})$, correspond to $\tilde{\pi} = [\tilde{\pi}_j] \propto [\exp(\langle \log \pi_j \rangle)]$ and $\tilde{\mathbf{A}} = [\tilde{A}_{ij}] \propto [\exp(\langle \log A_{ij} \rangle)]$, respectively. Then, the posteriors of the hidden variables are now written as

$$\begin{aligned}\gamma_{tj} &= q^\star(s_t = j) = \frac{1}{\tilde{p}(y_{t+1:T}|y_{1:t})} \\ &\tilde{p}(s_t = j|y_{1:t}) \tilde{p}(y_{t+1:T}|s_t = j) \propto \alpha_{tj} \beta_{tj},\end{aligned}\quad (\text{A } 16)$$

$$\begin{aligned}\xi_{tij} &= q^\star(s_{t-1} = i, s_t = j) \propto \alpha_{t-1,i} \tilde{p}(s_t = j|s_{t-1} = i) \\ &\tilde{p}(y_t|s_t = j) \beta_{tj}\end{aligned}\quad (\text{A } 17)$$

$$\begin{aligned}\text{and } \eta_{tjk} &= q^\star(s_t = j, z_t = k) \\ &= \frac{\tilde{p}(y_t|s_t = j, z_t = k) \tilde{p}(z_t = k|s_t = j)}{\tilde{p}(y_t|s_t = j)} q^\star(s_t = j).\end{aligned}\quad (\text{A } 18)$$

A.2. VB-M step

The Dirichlet posteriors for the initial state, the transition, and the mixture component parameters are

$$q^\star(\pi) = \mathcal{D}(\pi; \tilde{\mathbf{u}}^\pi) \quad \text{where } \tilde{u}_j^\pi = u_j^\pi + \gamma_{1j}, \quad (\text{A } 19)$$

$$q^\star(\mathbf{A}) = \prod_{i=1}^J \mathcal{D}(\mathbf{A}_i; \tilde{\mathbf{u}}_i^A) \quad \text{where } \tilde{u}_{ij}^A = u_{ij}^A + \sum_{t=2}^T \xi_{tij} \quad (\text{A } 20)$$

$$\begin{aligned}\text{and } q^\star(\mathbf{C}) &= \prod_{j=1}^J \mathcal{D}(\mathbf{C}_j; \tilde{\mathbf{u}}_j^C) \quad \text{where } \tilde{u}_{jk}^C \\ &= u_{jk}^C + \sum_{t=1}^T \eta_{tjk}.\end{aligned}\quad (\text{A } 21)$$

The Gaussian posteriors for the mean values are

$$q^\star(\mu_j) = \mathcal{N}(\mu_j; \tilde{w}_j, \tilde{r}_j^{-1}), \quad (\text{A } 22)$$

$$\text{where } \tilde{r}_j = r_j + \sum_{t=1}^T \sum_{k=1}^K \eta_{tjk} \tilde{v}_{jk}^{(\text{old})} \tilde{v}_{jk}^{(\text{old})}$$

and

$$\tilde{w}_j = \tilde{r}_j^{-1} \left(r_j w_j + \sum_{t=1}^T \sum_{k=1}^K \eta_{tjk} \tilde{v}_{jk}^{(\text{old})} \tilde{v}_{jk}^{(\text{old})} y_t \right).$$

The Gamma posteriors for the precisions are

$$q^\star(\lambda_{jk}) = \mathcal{G}(\lambda_{jk}; \tilde{\mathbf{v}}_{jk}, \tilde{v}_{jk}), \quad (\text{A } 23)$$

$$\text{where } \tilde{\mathbf{v}}_{jk} = \mathbf{v}_{jk} + \frac{1}{2} \sum_{t=1}^T \eta_{tjk}$$

and

$$\tilde{v}_{jk}^{-1} = v_{jk}^{-1} + \frac{1}{2} \sum_{t=1}^T \eta_{tjk} \left\{ (y_t - \tilde{w}_{jk}^{(\text{old})})^2 + \frac{1}{\tilde{r}_{jk}^{(\text{old})}} \right\}.$$

A.3. Variational free energy

The variational free energy is evaluated after each round of the two steps. From equation (A 8), the

variational free energy can be expanded as

$$\begin{aligned} F(q(\mathbf{S}, \mathbf{Z}, \theta)) &= \langle \log p(\mathbf{Y}, \mathbf{S}, \mathbf{Z} | \theta) \rangle_{q(\mathbf{S}, \mathbf{Z}, \theta)} + H(\mathbf{S}, \mathbf{Z}) \\ &\quad - \text{KL}(q(\boldsymbol{\pi}) \parallel p(\boldsymbol{\pi})) - \text{KL}(q(\mathbf{A}) \parallel p(f \mathbf{A})) \\ &\quad - \text{KL}(q(\mathbf{C}) \parallel p(\mathbf{C})) - \text{KL}(q(\boldsymbol{\mu}) \parallel p(\boldsymbol{\mu})) \\ &\quad - \text{KL}(q(\Lambda) \parallel p(\Lambda)), \end{aligned} \quad (\text{A } 24)$$

where the first term can be evaluated as

$$\begin{aligned} &\sum_j \gamma_{1j} \langle \log \pi_j \rangle + \sum_{t=2}^T \sum_{i,j} \xi_{tij} \langle \log A_{ij} \rangle \\ &\quad + \sum_{t=1}^T \sum_{j,k} \eta_{tjk} \langle \log C_{jk} \rangle + \frac{1}{2} \sum_{t=1}^T \sum_{j,k} \eta_{tjk} \\ &\quad \left[-\log 2\pi + \langle \log \lambda_{jk} \rangle - \tilde{\mathbf{v}}_{jk} \tilde{\mathbf{v}}_{jk} \left\{ (y_t - \tilde{w}_j)^2 + \frac{1}{\tilde{r}_j} \right\} \right]. \end{aligned} \quad (\text{A } 25)$$

The second term, the entropy of the hidden variables, is

$$\begin{aligned} H(\mathbf{S}, \mathbf{Z}) &= - \sum_{\mathbf{S}, \mathbf{Z}} q(\mathbf{S}, \mathbf{Z}) \log q(\mathbf{S}, \mathbf{Z}) \\ &= - \sum_{t=1}^T \sum_{j,k} \eta_{tjk} \log \eta_{tjk}. \end{aligned} \quad (\text{A } 26)$$

The remaining terms associated to the KL divergences can be evaluated as given in equations (B 3), (B 5) and (B 8).

A.4. Model order selection

In Bayesian model selection, one selects a model order with the highest model posterior,

$$p(J, K | \mathbf{Y}) = \frac{p(\mathbf{Y} | J, K) p(J, K)}{p(\mathbf{Y})}, \quad (\text{A } 27)$$

where $p(\mathbf{Y} | J, K)$, $p(J, K)$ and $p(\mathbf{Y})$ are the marginal likelihood, the model prior and the total evidence, respectively. Since the variational free energy is an approximation to the log-marginal likelihood, if we assume a flat model prior, that is, $p(J, K) = \text{const.}$, $\forall J, K$, the model posterior is approximated as

$$p(J, K | \mathbf{Y}) \approx \frac{\exp\{F(q(\mathbf{S}, \mathbf{Z}, \theta | J, K))\}}{\sum_{J, K} \exp\{F(q(\mathbf{S}, \mathbf{Z}, \theta | J, K))\}}. \quad (\text{A } 28)$$

In essence, we select a model order that leads to the highest free energy.

A.5. Decoding of the most probable path

The decoding of the most probable path of hidden states is essentially equivalent to the Viterbi algorithm described in [13]. The only difference is that we use the submarginalized parameter, $\tilde{\mathbf{A}}$, and likelihood $\tilde{p}(y_t | s_t) = \exp[\langle \log p(y_t | s_t, z_t) \rangle_{q^*(z_t, \boldsymbol{\mu}, \Lambda)}]$.

APPENDIX B. RELEVANT DISTRIBUTIONS

This section presents the density functions, KL divergences and some relevant statistics for the Dirichlet, Gaussian and Gamma distributions.

B.1. Dirichlet distribution

For a variable $\boldsymbol{\pi} \in R_+^J$, $\sum_{j=1}^J \pi_j = 1$, the Dirichlet density function with a parameter $\mathbf{u} \in R_+^J$ is

$$\mathcal{D}(\boldsymbol{\pi}; \mathbf{u}) = \frac{(\sum_{j'} u_{j'})}{\prod_{j'} \Gamma(u_{j'})} \prod_{j=1}^J \pi_j^{u_j-1}. \quad (\text{B } 1)$$

The expectation of the logarithm of the Dirichlet variable is

$$\langle \log \pi_j \rangle = \psi(u_j) - \psi\left(\sum_{j'} u_{j'}\right), \quad (\text{B } 2)$$

where $\psi(\cdot)$ is the digamma function. For two Dirichlet distributions, $q(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}; \tilde{\mathbf{u}})$ and $p(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}; \mathbf{u})$, the KL divergence is

$$\begin{aligned} \text{KL}(q \parallel p) &= \log \frac{\Gamma(\sum_j \tilde{u}_j)}{\Gamma(\sum_j u_j)} + \sum_j \frac{\Gamma(u_j)}{\Gamma(\tilde{u}_j)} + \sum_j (\tilde{u}_j - u_j) \\ &\quad \left\{ \psi(\tilde{u}_j) - \psi\left(\sum_{j'} \tilde{u}_{j'}\right) \right\}. \end{aligned} \quad (\text{B } 3)$$

B.2. Gaussian distribution

The Gaussian density function of $y \in R$ is specified by a mean $\boldsymbol{\mu}$ and a precision λ :

$$\mathcal{N}(y; \boldsymbol{\mu}, \lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} \exp\left\{-\frac{\lambda}{2}(y - \boldsymbol{\mu})^2\right\}. \quad (\text{B } 4)$$

For two Gaussian distributions, $q(y) = \mathcal{N}(y; \tilde{\boldsymbol{\mu}}, \tilde{\lambda}^{-1})$ and $p(y) = \mathcal{N}(y; \boldsymbol{\mu}, \lambda^{-1})$, the KL divergence is

$$\text{KL}(q \parallel p) = \frac{1}{2} \left\{ \log \frac{\tilde{\lambda}}{\lambda} + \lambda(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})^2 + \frac{\lambda}{\tilde{\lambda}} - 1 \right\}. \quad (\text{B } 5)$$

B.3. Gamma distribution

For a variable $\lambda \in R_+$, the Gamma density function with a shape parameter $\nu \in R_+$ and a scale parameter $v \in R_+$ is

$$\mathcal{G}(\lambda; \nu, v) = \frac{v^{-\nu} \lambda^{\nu-1}}{\Gamma(\nu)} \exp\left(-\frac{\lambda}{v}\right). \quad (\text{B } 6)$$

The expectation of the logarithm of the Gamma variable is written as

$$\langle \log \lambda \rangle = d(\psi(\nu) + \log v). \quad (\text{B } 7)$$

For two Gamma distributions, $q(\lambda) = \mathcal{G}(\lambda; \tilde{\nu}, \tilde{v})$ and $p(\lambda) = \mathcal{G}(\lambda; \nu, v)$, the KL divergence is

$$\begin{aligned} \text{KL}(q \parallel p) &= -\nu \log \frac{\tilde{v}}{v} + (\tilde{\nu} - \nu) \psi(\tilde{\nu}) \\ &\quad - \log \frac{\Gamma(\tilde{\nu})}{\Gamma(\nu)} + \frac{\tilde{\nu} \tilde{v}}{v} - \tilde{\nu}. \end{aligned} \quad (\text{B } 8)$$

REFERENCES

- 1 Rodenburg, T. B., van Hierden, Y. M., Buitenhuis, A. J., Riedstra, B., Koene, P., Korte, S. M., van der Poel, J. J., Groothuisd, T. G. G. & Blokhuis, H. J. 2004 Feather pecking in laying hens: new insights and directions for research? *Appl. Anim. Behav. Sci.* **86**, 291–298.
- 2 Dixon, L. M. 2008 Feather pecking behaviour and associated welfare issues in laying hens. *Avian Biol. Res.* **1**, 73–87. (doi:10.3184/175815508X363251)
- 3 Blokhuis, H. J. et al. 2007 The laywel project: welfare implications of changes in production systems for laying hens. *Worlds Poult. Sci. J.* **63**, 101–114. (doi:10.1017/S0043933907001328)
- 4 Fossum, O., Jansson, D. S., Etterlin, P. E. & Vågsholm, I. 2009 Causes of mortality in laying hens in different housing systems in 2001 to 2004. *Acta Vet. Scand.* **51**, 3. (doi:10.1186/1751-0147-51-3)
- 5 Tauson, R. & Svensson, S. A. 1980 Influence of plumage condition on the hen's feed requirement. *Swedish J. Agric. Res.* **10**, 35–39.
- 6 El-Lethey, H., Aerni, V., Jungi, T. W. & Wechsler, B. 2000 Stress and feather pecking in laying hens in relation to housing conditions. *Br. Poult. Sci.* **41**, 22–28. (doi:10.1080/00071660086358)
- 7 Gentle, M. J. & Hunter, L. N. 1991 Physiological and behavioural responses associated with feather removal in *Gallus gallus* var. *domesticus*. *Res. Vet. Sci.* **50**, 95–101.
- 8 Huber-Eicher, B. & Sebö, F. 2001 Reducing feather pecking when raising laying hen chicks in aviary systems. *Appl. Anim. Behav. Sci.* **73**, 59–68. (doi:10.1016/S0168-1591(01)00121-6)
- 9 Dennis, R. L., Fahey, A. G. & Cheng, H. W. 2009 Infrared beak treatment method compared with conventional hot-blade trimming in laying hens. *Poult. Sci.* **88**, 38–43. (doi:10.3382/ps.2008-00227)
- 10 Gentle, M. J., Waddington, D., Hunter, L. N. & Jones, R. B. 1990 Behavioural evidence for persistent pain following partial beak amputation in chickens. *Appl. Anim. Behav. Sci.* **27**, 149–157. (doi:10.1016/0168-1591(90)90014-5)
- 11 Drake, K. A., Donnelly, C. A. & Dawkins, M. S. In press. The influence of rearing and lay risk on propensity for feather damage in laying hens. *Br. Poult. Sci.*
- 12 Dawkins, M. S., Lee, H., Waitt, C. D. & Roberts, S. J. 2009 Optical flow patterns in broiler chicken flocks as automated measures of behaviour and gait. *Appl. Anim. Behav. Sci.* **119**, 203–209. (doi:10.1016/j.applanim.2009.04.009)
- 13 Rabiner, L. R. 1989 A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286. (doi:10.1109/5.18626)
- 14 Beauchemin, S. S. & Barron, J. L. 1995 The computation of optical flow. *ACM Comput. Surveys* **27**, 433–467. (doi:10.1145/212094.212141)
- 15 Bright, A., Jones, T. A. & Dawkins, M. S. 2006 A non-intrusive method of assessing plumage condition in commercial flocks of laying hens. *Anim. Welf.* **15**, 113–118.
- 16 Bishop, C. M. 2006 *Pattern recognition and machine learning*. New York, NY: Springer.
- 17 Baum, L. E., Petrie, T., Soules, G. & Weiss, N. 1970 A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41**, 164–171. (doi:10.1214/aoms/1177697196)
- 18 Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977 Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.* **39**, 1–38.
- 19 Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. 1999 An introduction to variational methods for graphical models. *Mach. Learn.* **37**, 183–233. (doi:10.1023/A:1007665907178)
- 20 Beal, M. J. & Ghahramani, Z. 2003 The variational Bayesian EM algorithm for incomplete data. In *Bayesian statistics*, vol. 7 (eds J. M. Bernardo, M. J. Bayarri, A. P. Dawid, J. O. Berger, D. Heckerman, A. F. M. Smith & M. West), pp. 453–464. Oxford, UK: Oxford University Press.
- 21 Ji, S., Krishnapuram, B. & Carin, L. 2006 Variational Bayes for continuous hidden Markov models and its application to active learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 522–532. (doi:10.1109/TPAMI.2006.85)
- 22 Roberts, S. J. & Penny, W. D. 2002 Variational Bayes for generalized autoregressive models. *IEEE Trans. Signal Process.* **50**, 2245–2257. (doi:10.1109/TSP.2002.801921)
- 23 Johnson, R. A. & Wichern, D. W. 1998 *Applied multivariate statistical analysis*, 4th edn. Upper Saddle River NJ: Prentice-Hall.
- 24 Rasmussen, C. E. & Williams, C. K. I. 2006 *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.
- 25 Bright, A. 2009 Time course of plumage damage in commercial layers. *Vet. Rec.* **164**, 334–335.
- 26 Brown, M. B. & Forsythe, A. B. 1974 Robust tests for the equality of variances. *J. Am. Statist. Assoc.* **69**, 364–367. (doi:10.2307/2285659)
- 27 Meng, X., Rosenthal, R. & Rubin, D. B. 1992 Comparing correlated correlation coefficients. *Psychol. Bull.* **111**, 172–175. (doi:10.1037/0033-2909.111.1.172)
- 28 Bellomo, N., Bianca, C. & Delitala, M. 2009 Complexity analysis and mathematical tools towards the modelling of living systems. *Phys. Life Rev.* **6**, 144–175. (doi:10.1016/j.plev.2009.06.002)
- 29 Giles, A., Massom, R. A. & Warner, R. 2009 A method for sub-pixel scale feature-tracking using Radarsat images applied to the Mertz Glacier Tongue, East Antarctica. *Remote Sensing Environ.* **113**, 1691–1699. (doi:10.1016/j.rse.2009.03.015)
- 30 Shi, L., Nascimento, J., Chandsawangbhuwana, C., Botvinick, C. & Berns, M. 2008 An automatic system to study sperm motility and energetics. *Biomed. Microdevices* **10**, 573–583. (doi:10.1007/s10544-008-9169-4)
- 31 Courty, N. & Corpetti, T. 2007 Crowd motion capture. *Comput. Animat. Virtual Worlds* **18**, 361–370. (doi:10.1002/cav.199)
- 32 Bremond, F., Thonnat, M. & Zuniga, M. 2006 Video-understanding framework for automatic behavior recognition. *Behav. Res. Methods* **38**, 416–426.
- 33 Johansson, A. & Helbing, D. 2008 From crowd dynamics to crowd safety: a video-based analysis. *Adv. Complex Systems* **11**, 497–527. (doi:10.1142/S0219525908001854)
- 34 Sonka, M., Hlavac, V. & Boyle, R. 1999 *Image processing, analysis, and machine vision*. London, UK: PWS publishing.
- 35 Penny, W. D. & Roberts, S. J. 1999 Dynamic models for nonstationary signal segmentation. *Comput. Biomed. Res.* **32**, 483–502. (doi:10.1006/cbmr.1999.1511)
- 36 Beal, M. J., Ghahramani, Z. & Rasmussen, C. E. 2002 The infinite hidden Markov model. In *Advances in neural information processing systems (NIPS 2001)*, vol. 14 (eds T. Diettrich, S. Becker & Z. Ghahramani), pp. 577–584. Cambridge, MA: MIT Press.