JOURNAL OF THE ROYAL SOCIETY Interface

CrossMark click for updates

J. R. Soc. Interface (2012) **9**, 3436–3443
doi:10.1098/rsif.2012.0594
Published online 5 September 2012

Prediction of welfare outcomes for broiler chickens using Bayesian regression on continuous optical flow data

Stephen J. Roberts[1,*], Russell Cain[2] and Marian Stamp Dawkins[2]

[1]*Department of Engineering Science, and* [2]*Department of Zoology, University of Oxford, Oxford, UK*

Currently, assessment of broiler (meat) chicken welfare relies largely on labour-intensive or post-mortem measures of welfare. We here describe a method for continuously and robustly monitoring the welfare of living birds while husbandry changes are still possible. We detail the application of Bayesian modelling to motion data derived from the output of cameras placed in commercial broiler houses. We show that the forecasts produced by the model can be used to accurately assess certain key aspects of the future health and welfare of a flock. The difference between healthy flocks and less-healthy ones becomes predictable days or even weeks before clinical symptoms become apparent. Hockburn (damaged leg skin, usually only seen in birds of two weeks or older) can be well predicted in flocks of only 1–2 days of age, using this approach. Our model combines optical flow descriptors of bird motion with robust multivariate forecasting and provides a sparse, efficient model with sparsity-inducing priors to achieve maximum predictive power with the minimum number of key variables.

Keywords: animal welfare; optical flow; Bayesian multivariate modelling; variational Bayes inference

1. INTRODUCTION

Advance warning of undesirable outcomes such as outbreaks of disease, tail-biting in pigs, feather-pecking in hens and even rioting in human crowds would be a major step in controlling and even averting them altogether. In this context, changes in behaviour are increasingly being recognized as key precursors of such events with the possibility of indicating when and where husbandry changes would be most effective [1]. For example, even before overt clinical symptoms appear, a reduction in time spent at the feeder identifies those dairy cows that are most at risk of developing uterine inflammation after calving [2], a decline in exploration behaviour precedes and predicts clinical symptoms of disease in mice used as a model for Huntington's disease [3], pigs that will later show a tendency to biting the tails of other pigs can be identified early by their tendency to bite other objects [4] and behavioural changes are shown by chickens with only sub-clinical levels of *Salmonella* [5]. At a group level, flocks of laying hens that subsequently go on to develop serious feather pecking by 40 weeks can be identified as early as 18–20 weeks, by increased incidence of disturbed behaviour [6].

*Author for correspondence (stephen.roberts@eng.ox.ac.uk).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2012.0594 or via http://rsif.royalsocietypublishing.org.

Using behaviour in this way has a great advantage that it is non-invasive and does not require taking tissue samples for diagnosis. Furthermore, large numbers of animals can be monitored on a continuous basis, using easily available CCTV and video equipment. However, the potential of this approach has not yet been fully used because of a lack of ways of automatically analysing the vast quantities of behavioural data that can be produced. Unless behavioural precursors can be recognized automatically from camera data, their use as widely used diagnostics will be strictly limited.

Here, we show that statistical measures automatically derived from simple optical flow algorithms applied to the movements of large flocks of chickens can be combined to give highly predictive measures of health and welfare, some of them days or even weeks before more conventional measures are taken. The commercial broiler houses we recorded from each contained around 34 000 birds, all similar in appearance and with no possibility of marking or tagging individuals so that our methods could be applied to any situation where there is a need to monitor and analyse information about large numbers of anonymous individuals, such as crowds of people, or flocks and herds of animals. The use of optical flow to analyse behaviour at a group level avoids the heavy computation involved in attempting to track many individuals at once. The computation required by our method is simple enough to be done in real time, giving a continuous minute-by-minute output.

Received 26 July 2012
Accepted 15 August 2012

Three basic statistical measures of optical flow (daily mean, skew and kurtosis) have already been shown to be predictive on their own of key health and welfare measures in broiler chickens [6,7]. Skew and kurtosis, as measures of heterogeneity of flow, were particularly effective at predicting final flock mortality, incidence of hockburn (damaged leg skin) and poor walking in older birds (27–35 days) even when the birds were as young as 15–20 days. In this study, we show the predictive power of this approach by using the optical flow measures as input to a Bayesian multivariate linear regression model, with sparsity-inducing *shrinkage* to highlight which variables are genuinely contributing to the predictive power of the model. Variables that contribute little information are then associated with very small factor weights, while those that contribute a great deal are weighted heavily. The result is a sparse, efficient model that gives the maximum predictive power with the minimum number of key variables.

By using information from different inputs and cumulating this information over the entire preceding life-history of a flock, we show that it is possible to predict end-of-flock welfare measures in birds as young as a week old and over a month before these measures are conventionally taken. The regression model described has potential application to a wide variety of other situations with multiple inputs and high levels of noise.

2. METHODS

2.1. Animals and housing

Data (optical flow, production and welfare) were collected from 24 commercial chicken flocks, all on a single site in the UK and belonging to a major producer company. Four flocks were studied at a time, housed in separate identical broiler houses, each with floor area of 1670 m^2 and with identical numbers and layout of feeders, drinkers and ventilation fans. The same farm staff looked after all the flocks. Six replicates of four flocks were completed over the period between October 2010 and June 2011. Each flock contained approximately 34 000 chickens of mixed sexes and two commercial broiler breeds (sometimes separated and sometimes mixed, according to company needs). Chicks were placed as day-olds and grown to 35 days old with a target final stocking density of 38 kg m^{-2}. Flocks were not thinned (proportion of flock removed before clearance) during the growing period.

2.2. Production and welfare data

The company provided the following data for each flock: % mortality (% of all deaths before slaughter), % culls (% of all birds killed before slaughter because of leg problems), daily mortality, daily culls, daily weights (manually weighed), daily growth rate, daily water consumption, daily food consumption, % pododermatitis (% birds scored with foot pad lesions after slaughter at the slaughter plant) and % hockburn (% of birds scored with problematic 'brown hocks' after slaughter at the slaughter plant). To assess the walking ability of living birds, a trained observer used the six-point Bristol Gait Score [8] with a catching pen and gait scored 60 (randomly selected) birds on day 28. The measurements were monitored by independent measurements from a second observer. The results were expressed as the mean gait score for that flock. The observers used a catching pen placed at the same positions in each house. The 24 flocks showed a mean mortality of 3.35 per cent (s.d. = 0.91; range 2.37–6.46%) and an average gait score of 1.92 (s.d. = 0.23; range 1.64–2.38). We note, particularly, the very narrow range of both mortality and gait scores, indicating that these data are well suited to testing the sensitivity of our methods.

2.3. Cameras and hardware

Optical flow was monitored continuously between 08.00 and 20.00 h, using a bespoke recording processing unit[1] designed to operate continuously in the demanding conditions of commercial broiler houses and with enough capacity to store 90 days of data. The lights were continuous during the recordings.

2.4. Optical flow analysis

Most automated visual-processing techniques keep track of individual animals, which is a very difficult task especially with large numbers of animals unless they are marked and even then is computationally prohibitive. Instead, we extract motion from the chickens as a whole group, using an *optical flow* algorithm. This was facilitated by the fact that broiler chickens are white, in contrast to their darker background. 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. We use here the Lucas–Kanade optical flow method [9,10], in which incremental motion in the stack of images, $f(x,y,t,)$, is estimated as follows.

Let $f(x,y,t)$ be the grey-level at pixel x,y at time t. Consider the expansion of partial derivatives

$$\mathrm{d}f = f_x \mathrm{d}x + f_y \mathrm{d}y + f_t \mathrm{d}t,$$

if f remains a constant function (brightness), then $\mathrm{d}f \approx 0$ whence

$$-f_t = f_x v_x + f_y v_y = \nabla f \cdot \mathbf{v},$$

where $\mathbf{v} \stackrel{\mathrm{def}}{=} (v_x, v_y)^{\mathrm{T}}$ is the velocity vector. The most commonly used method to estimate velocities is the *Lucas–Kanade* method which places a small neighbourhood around the pixel of interest, in which it is assumed that the *motion vector* is stationary. If we define the set $\{p_1, p_2, \ldots, p_n\}$ as the pixels in the neighbourhood, then

$$\begin{aligned} f_x(p_1)v_x + f_y(p_1)v_y &= -f_t(p_1), \\ f_x(p_2)v_x + f_y(p_2)v_y &= -f_t(p_2), \\ &\vdots \\ f_x(p_n)v_x + f_y(p_n)v_y &= -f_t(p_n), \end{aligned}$$

which can be written in the form

$$\mathbf{A}\mathbf{v} = \mathbf{b},$$

[1]The Black Box system that is described in www.robots.ox.ac.uk/~parg/projects/welfare/blackBoxSpecfication.pdf and in the electronic supplementary material.

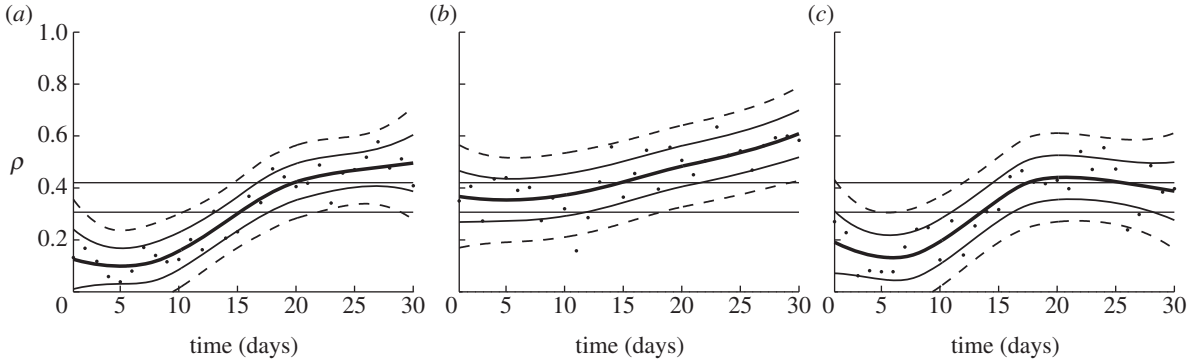

Figure 1. Correlations, $\boldsymbol{\rho}$, of day-by-day regression onto (*a*) final mortality, (*b*) hockburn and (*c*) expert gait score. The solid horizontal lines represent the $p = 0.05$, $p = 0.01$ boundaries and each subplot shows the data (dots) along with best-fit regression (solid thick curve) and associated $\pm 1, 2\boldsymbol{\sigma}$ standard posterior uncertainty (thin curves, dashed curves respectively).

with solution given as

$$\mathbf{v} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{b}, \tag{2.1}$$

i.e. using the pseudo inverse of $\mathbf{A}$. The matrix $\mathbf{A}$ here is the matrix of spatial partial derivatives over the set of pixels in a neighbourhood, and $\mathbf{b}$ is the column vector of (negative) partial derivatives with time

$$\mathbf{A} = \begin{pmatrix} f_x(p_1) & f_y(p_1) \\ f_x(p_2) & f_y(p_2) \\ \vdots & \vdots \\ f_x(p_n) & v_y(p_n) \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} -f_t(p_1) \\ -f_t(p_2) \\ \vdots \\ -f_t(p_n) \end{pmatrix}.$$

The optical flow algorithm was executed *in situ* by the Fit-2 computer inside each Black Box unit, using 15 min blocks of recorded video footage from each camera and alternating between its two cameras. Consider a video file that consists of T image frames each of 320×240 pixels. Each image is divided into 1200 ($=40 \times 30$) eight-by-eight 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. $\mathbf{v}[t,\ell] = (v_x[t,\ell], v_y[t,\ell])^{\mathrm{T}}$, for frames at time $t = 1, \ldots, T$, and blocks $\ell = 1, \ldots, 1200$. From the velocity vectors, the amount of movement for each block was obtained as the magnitude of the velocity, $V[t,\ell] = |\mathbf{v}|$. As a snapshot of global movement statistics at time t, the spatial mean, variance, skewness and kurtosis of $V[t,\ell]$ were sequentially calculated for each frame in the video file, resulting in a multivariate (four-dimensional) time-series of length $t = 1 \ldots T$ samples. Each camera recorded data in a series of 15 min blocks throughout the day, and in the analysis presented in this study, we average over all daily blocks to obtain an aggregate measure of daily mean flow, variance, skewness and kurtosis, giving rise to a daily four-dimensional vector characterizing the birds' motion[2].

[2]The processed data used in this study may be downloaded from www.robots.ox.ac.uk/~parg/projects/welfare/OFwelfaredata.mat.

3. RESULTS

The dataset was analysed using a Bayesian multivariate linear model with *shrinkage*, such that sparsity-inducing priors are placed over the factor weights in the model, leading to weights associated with inputs that have little information regarding the target regression to shrink towards zero. There are two main reasons for this choice. Firstly, given a relatively small, independent dataset size (remember that the number of trials is 24), we should naturally bias our analysis towards simpler models, thus mitigating against potentially poor generalization due to overfitting. Secondly, we may entertain a simple, yet effective, shrinkage mechanism that we can relate directly to factor weights associated with the optical flow parameters over time. The mathematical details of the approach are given in full in the appendix of this study, with particular details of the inputs to the model being covered in the final subsection. All the results we present in the data are strictly causal; i.e. we regress onto future values of welfare and include in our models only observations from the past.

The model predictions based on combined (mean, variance, skew and kurtosis) optical flow daily aggregate measures for individual days are shown in figure 1. Figure 1*a* shows that the model predicts total flock % mortality when the chickens are only 15 days old. Figure 1*b* shows similar predictions for hockburn, 20 days or so *before* hockburn is assessed (at the slaughterhouse). The predictions for gait score become significant by day 13, even though gait scoring was not carried out until day 28. The solid horizontal lines represent the $p = 0.05$, $p = 0.01$ boundaries, and each subplot shows the data (dots) along with best-fit regression (solid thick curve) and associated $\pm 1, 2\boldsymbol{\sigma}$ standard posterior uncertainty (thin curves and dashed curves, respectively). The latter is not used in any of our analysis and is offered solely to highlight the underlying trend. This regression was performed using a Bayesian model with spline basis functions, based upon the model presented in the appendix. Details of the approach may be found in the final subsection of the appendix. Figure 2 depicts the factor weights associated with this set of predictions, for each of the three welfare measures, broken down by optical flow statistic. We note that mean optical flow

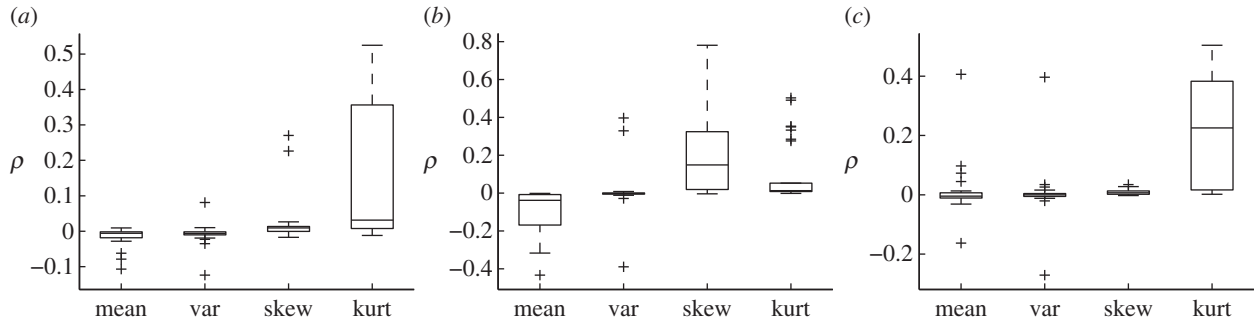

Figure 2. Boxplots of optical flow factor weights aggregated from the predictions shown in figure 1 for (*a*) final mortality, (*b*) hockburn and (*c*) expert gait score.

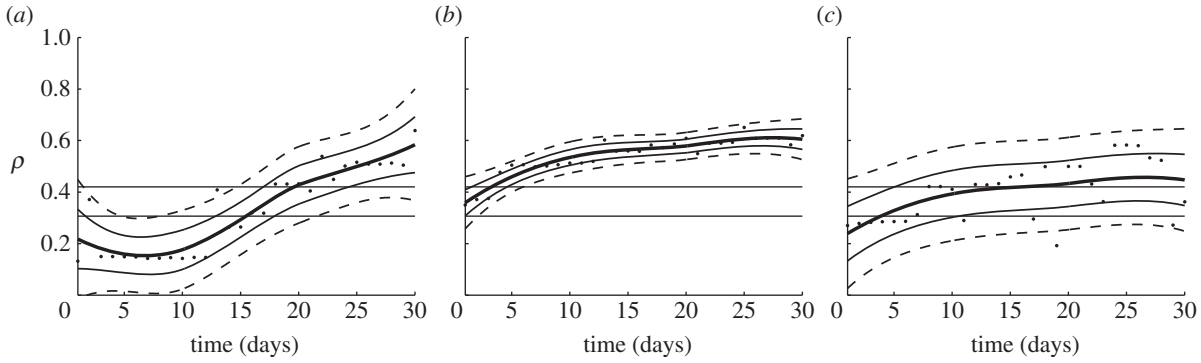

Figure 3. Correlations, ρ, of running regression onto (*a*) final mortality, (*b*) hockburn and (*c*) expert gait score. The solid horizontal lines represent the $p = 0.05$, $p = 0.01$ boundaries, and each subplot shows the data (dots) along with best-fit regression (solid thick curve) and associated $\pm 1, 2\sigma$ standard posterior uncertainty (thin curves and dashed curves, respectively). We note the improved prediction performance of hockburn and gait score.

has dominantly negative factor weights in all cases, variance is not significantly related, and both optical flow skewness and kurtosis are positively related to the welfare measures. We note that increases in these measures are indicative of welfare problems and hence should be associated with lower optical mean flows, indicative of birds that are less able to move. A positive skewness is associated with a longer tail to the right and a left-dominant peak, again compatible with our hypothesis of populations of less-able moving birds. The positive factor weighting of kurtosis indicates a potential population spread with the emergence of a low-mobility group of birds.

The previous model used optical flow observations from each day in turn to forecast welfare measures. Figure 3 shows the effect on the models predictions of including optical flow data not just from one day at a time (as in figure 1) but from a given day and *all preceding days* for that flock. The vector of information hence grows as we proceed on a day-by-day basis. We see that the predictions for mortality are not much changed (figure 3*a*), but both % hockburn and gait score are accurately predicted from a much earlier age. Particularly, we see that there is a very significant early correlation with final % hockburn (figure 3*b*) and with expert gait score (figure 3*c*).

4. DISCUSSION

These results clearly show the increased predictive power that can be obtained from combining individual optical flow measures into a single Bayesian regression model and using the cumulative information that can be obtained from data collected continuously over time. The narrow range of recorded mortalities and gait scores in the flocks used for this analysis showed that the method is sensitive enough to discriminate between similar flocks and therefore has potential as a commercial management tool. It appears to be particularly effective at predicting which flocks are at risk of hockburn or walking defects. This is particularly clear for predicting which flocks will develop high levels of hockburn and which ones will have significant numbers of birds with walking problems.

Hockburn is a form of contact dermatitis caused by birds sitting for long periods on wet, poor-quality litter [11,12]. Ammonia from faecal matter in the litter damages the skin and leaves black or brown lesions on the legs. It is used as a measure of welfare in its own right because, although usually assessed post mortem, it gives an indication of the conditions that have been experienced by a bird during its lifetime [11]. The first signs of hockburn may appear as early as two weeks of age [12] and an indication that a flock that will end up with a high % of hockburn in birds after slaughter include the weight and density of the birds at two weeks [13] and high skew and kurtosis of optical flow at the same age [7]. By predicting which flocks will have later hockburn problems from optical flow information in the first few days of life, the model thus seems to be able to give a warning at a very early stage, before overt symptoms appear. Simply using

the amount of movement is not enough, because the mean level of optical flow is not, on its own, sufficient to predict later hockburn at least until 15 days [7]. It is only with the full predictive power of all the available optical flow measures that the model provides that flocks at risk of hockburn can be identified.

A similar increase in predictive power is shown by the model's ability to predict gait score. The skew and kurtosis of optical flow, taken separately, allow prediction of which flocks are likely to be scored as having a high % of poor walkers over a week before the gait scoring is actually carried out, i.e. on day 18–20 when gait scoring is done on day 28 [7]. The model, however, allows such predictions to be made much earlier in the birds lives when they are only 5 or 6 days old—over three weeks before actual gait scoring was carried out. As young birds with severe walking difficulties would be culled well before 28 days (and would therefore not be present for the formal gait scoring), the model must be using information about the optical flow caused by the movement of the remaining birds that may be walking healthily in their first week but will, nevertheless, by the age of 28 days, be part of a flock that that has a significant number of birds with gait deficiencies.

The ability to predict health and welfare problems before they become serious is an important way in which producers can improve both the production and welfare of their animals through early diagnosis and husbandry changes, such as increasing house ventilation. The Bayesian model described here automatically extracts key information from the optical flow patterns made by flocks of broiler chickens. It is, however, equally suitable for a wide range of other applications wherever there is a need for 'early warnings' in continuous real-time data.

This work was partly funded via the UK BBSRC, to whom we are grateful. The authors thank Andy Morris for his generous help in collecting the data and Sibo (Spark) Lu for his help with developing the recording devices.

APPENDIX A. BAYESIAN MULTIVARIATE LINEAR MODEL WITH SHRINKAGE

We start by recasting the multivariate linear model discussed in this study such that the observed $y(x)$ is modelled as a noise-corrupted linear combination of a set of B-dimensional observations, $\boldsymbol{\phi}_i$, which are indexed via an independent variable x (this represents an index into the data array, for example and equates to a timing index in our application).

$$\left.\begin{aligned} y(x) &= \hat{y}(x) + \boldsymbol{\eta} \\ \text{and} \quad y(x) &= \sum\nolimits_{i=1}^{B} w_i \boldsymbol{\phi}_i(x) + w_0 + \boldsymbol{\eta}, \end{aligned}\right\} \tag{A 1}$$

in which w_i are the factor weights, w_0 a bias, or offset, term and $\boldsymbol{\phi}_i(x)$ is the observable set indexed to variable x. The noise term, $\boldsymbol{\eta}$, is taken to be drawn from a normal distribution, $\mathcal{N}(0, \boldsymbol{\beta}^{-1})$, in which $\boldsymbol{\beta}$ is the precision (inverse variance). Without loss of generality, we may augment the observation set with a column of ones, and so fold w_0 and $\{w_i\}$ into a single vector $\mathbf{w}$ and rewrite the above as

$$y(x) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(x) + \boldsymbol{\eta}, \tag{A 2}$$

where $\boldsymbol{\phi}$ is the $(B+1)$-dimensional vector composed of all B observed variables associated with x and a column of ones.

A.1. Maximum likelihood

The maximum likelihood (ML) solution for the weights is given by the standard pseudo-inverse equation, namely

$$\mathbf{w}_{ML} = (\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{Y}, \tag{A 3}$$

where $\mathbf{Y}$ is the array of all observed $\{y_1, y_2, \ldots, y_N\}$ associated with $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ and $\boldsymbol{\Phi}$ is the matrix of all observables over all $x_i \in \mathbf{X}$ (and so is of size $B \times N$). The problem with the ML approach, however, is model-overfitting. Some relief from the inherent problems associated with ML solutions may be obtained by introducing a prior over the weights and obtaining the *maximum a posteriori* solution. This approach still relies on *point values* and fails to take into account the intrinsic uncertainty by marginalizing over the posterior distributions of the variables, thus providing a full Bayesian solution. While the latter could be achieved using, for example, sample-based approaches. In the sampling framework, integration is performed via a stochastic sampling procedure such as Markov chain Monte Carlo. The latter, however, can be computationally intensive, scales poorly and assessment of convergence is often problematic. In this paper, we advocate an alternative solution based on the *variational Bayes* (VB) framework. In recent years, VB has been extensively used as a method of choice for approximate Bayesian inference as it offers computational tractability even on very large datasets. A full tutorial on VB is given in [14,15]. In what follows, we describe the key features and concentrate on the derivations of update equations for the linear basis model at the core of this study.

A.2. The probabilistic model

To start with, let us consider our Bayesian model such that the weights w_i are all drawn from a zero-mean Gaussian distribution with an isotropic covariance having precision (inverse variance) $\boldsymbol{\alpha}$

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \left(\frac{\boldsymbol{\alpha}}{2\boldsymbol{\pi}}\right)^{k/2} \exp\left(-\frac{\boldsymbol{\alpha}}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right). \tag{A 4}$$

As $\boldsymbol{\alpha} \geq 0$, the precision hyperparameter has a gamma distribution with hyper-hyperparameters b_α, c_α:

$$p(\boldsymbol{\alpha}) = \mathcal{G}(\boldsymbol{\alpha}; b_\alpha, c_\alpha). \tag{A 5}$$

The noise term, $\boldsymbol{\eta}$, is taken as zero-mean with precision hyperparameter $\boldsymbol{\beta}$ over which we place another Gamma distribution

$$p(\boldsymbol{\beta}) = \mathcal{G}(\boldsymbol{\beta}; b_\beta, c_\beta). \tag{A 6}$$

We concatenate all the parameters and hyperparameters of the model into the vector $\boldsymbol{\theta} = [\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}]$. As the weights depend on the scale $\boldsymbol{\alpha}$ (but not on the

noise precision β), the joint distribution over $\boldsymbol{\theta}$ factorizes as

$$p(\boldsymbol{\theta}) = p(\mathbf{w}|\alpha)p(\alpha)p(\beta). \tag{A 7}$$

APPENDIX B. VARIATIONAL BAYESIAN INFERENCE

The central quantity of interest in Bayesian learning is the posterior distribution $p(\boldsymbol{\theta}|\,\text{data})$, which fully describes our knowledge regarding the parameters of the model, $\boldsymbol{\theta}$. Given a probabilistic model of the data with parameters $\boldsymbol{\theta}$, the 'evidence' or 'marginal likelihood' of the data under the model is given by

$$p(\mathbf{D}) = \int p(\mathbf{D}, \boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{B 1}$$

The log evidence can be written as

$$\log p(\mathbf{D}) = \log \int q(\boldsymbol{\theta}|\mathbf{D})\frac{p(\mathbf{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}|\mathbf{D})}\mathrm{d}\boldsymbol{\theta}, \tag{B 2}$$

where $q(\boldsymbol{\theta}|\mathbf{D})$ is a *tractable* posterior proposal density. This has been introduced in both denominator and numerator in equation (B 2). Hence,

$$\log p(\mathbf{D}) = \int q(\boldsymbol{\theta}|\mathbf{D})\log\frac{p(\mathbf{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}|\mathbf{D})}\mathrm{d}\boldsymbol{\theta} + \int q(\boldsymbol{\theta}|\mathbf{D})\log\frac{q(\boldsymbol{\theta}|\mathbf{D})}{p(\boldsymbol{\theta}|\mathbf{D})}\mathrm{d}\boldsymbol{\theta}. \tag{B 3}$$

We may write the latter equation as

$$\log p(\mathbf{D}) = F(p, q) + KL(p, q), \tag{B 4}$$

where

$$F(p, q) \stackrel{\text{def}}{=} \int q(\boldsymbol{\theta}|\mathbf{D})\log\frac{p(\mathbf{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}|\mathbf{D})}\mathrm{d}\boldsymbol{\theta}, \tag{B 5}$$

is known as the (negative) variational free energy and

$$KL(p, q) \stackrel{\text{def}}{=} \int q(\boldsymbol{\theta}|\mathbf{D})\log\frac{q(\boldsymbol{\theta}|\mathbf{D})}{p(\boldsymbol{\theta}|\mathbf{D})}\mathrm{d}\boldsymbol{\theta}, \tag{B 6}$$

is the Kullback–Leibler (KL) divergence between the approximate posterior $q()$ and the true posterior $p()$.

Equation (B 4) is the fundamental equation of the VB-framework. Importantly, because the KL-divergence is always positive, $F(p,q)$ provides a strict *lower bound* on the model evidence. Moreover, because the KL-divergence is zero when the two densities are the same, $F(p,q)$ will become equal to the model evidence when the approximating posterior is equal to the true posterior, i.e. if $q(\boldsymbol{\theta}|\mathbf{D}) = p(\boldsymbol{\theta}|\mathbf{D})$.

The aim of VB-learning is therefore to maximize $F(p,q)$ and so make the approximate posterior as close as possible to the true posterior. This requires the extremization of an integral with respect to a functional, which is typically achieved using the *calculus of variations*. However, to obtain a *practical* inference algorithm, we can ensure that extremization with respect to the *function* $q(\boldsymbol{\theta})$ can be replaced exactly by extremization with respect to the *parameters* $\boldsymbol{\theta}$. This holds so long as the distributions are in the exponential family. As this includes all the commonly used probability distributions, this constraint is not normally problematic. What we obtain is then a set of coupled update equations over the parameters that are cycled until a convergence criterion is met. This approach is a generalization of the *expectation–maximization* (EM) algorithm, which is obtained as a special case of variational Bayes in the limit of the $q()$ distributions being replaced by delta functions at their ML values [14].

We consider parameter $\boldsymbol{\theta}_i$, and define $f(\boldsymbol{\theta}_i)$ as the marginal integral over all other $\boldsymbol{\theta}_{\neg i}$, where we use the nomenclature $\neg i$ to mean 'all variables except for i',

$$f(\boldsymbol{\theta}_i) \stackrel{\text{def}}{=} \int q(\boldsymbol{\theta}_{\neg i}|\mathbf{D})\log[p(\mathbf{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})]\mathrm{d}\boldsymbol{\theta}_{\neg i}. \tag{B 7}$$

The update form can be shown to be such that the updated distribution for parameter $\boldsymbol{\theta}_i$ is

$$q(\boldsymbol{\theta}_i^{\text{new}}) = \frac{\exp[f(\boldsymbol{\theta}_i^{\text{old}})]}{\int \exp[f(\boldsymbol{\theta}_i^{\text{old}})]\mathrm{d}\boldsymbol{\theta}_i}. \tag{B 8}$$

In the next subsections, we cycle through the parameters of the model, detailing the update equations associated with each $q()$ distribution in turn. In the following sections, we use $\hat{\theta}$ to denote the 'new', updated variable and θ^0 to be the prior value (i.e. our initialization). We note that, such as the EM algorithm, each iteration is *guaranteed* to improve the marginal likelihood of the data under the model.

B.1. Updating the weight parameters, w

We begin by defining

$$f(\mathbf{w}) = \iint q(\beta|\mathbf{D})q(\alpha|\mathbf{D})log[p(\mathbf{D}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)]\mathrm{d}\alpha\mathrm{d}\beta. \tag{B 9}$$

Our update equations, which maximize the functional, are hence given via

$$q(\mathbf{w}|\mathbf{D}) \propto \exp[f(\mathbf{w})]. \tag{B 10}$$

Substituting for the terms in equation (B 9) gives

$$\begin{aligned} f(\mathbf{w}) &= -\int q(\beta|\mathbf{D})\frac{\hat{\beta}}{2}(\mathbf{Y} - \boldsymbol{\Phi}\mathbf{w})^{\mathrm{T}}(\mathbf{Y} - \boldsymbol{\Phi}\mathbf{w})\mathrm{d}\beta \\ &\quad - \int q(\alpha|\mathbf{D})\frac{\hat{\alpha}}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\mathrm{d}\alpha \\ &= -\frac{\hat{\beta}}{2}(\mathbf{Y} - \boldsymbol{\Phi}\mathbf{w})^{\mathrm{T}}(\mathbf{Y} - \boldsymbol{\Phi}\mathbf{w}) - \frac{\hat{\alpha}}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}, \end{aligned} \tag{B 11}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the mean weight and noise precisions from the posterior distributions over α and β (see next two sections). The weight posterior is therefore a normal density $q(\mathbf{w}|\mathbf{D}) = \mathcal{N}(\mathbf{w}; \hat{\mathbf{w}}, \hat{\boldsymbol{\Sigma}})$, where

$$\left.\begin{aligned} \hat{\mathbf{w}} &= \hat{\boldsymbol{\Sigma}}\hat{\beta}\,\boldsymbol{\Phi}^{\mathrm{T}}\,\mathbf{Y} \\ \text{and}\quad \hat{\boldsymbol{\Sigma}} &= (\hat{\beta}\,\boldsymbol{\Phi}^{\mathrm{T}}\,\boldsymbol{\Phi} + \hat{\alpha}\mathbf{I})^{-1}, \end{aligned}\right\} \tag{B 12}$$

where $\mathbf{I}$ denotes the identity matrix. Thus, the posterior precision matrix, $\hat{\boldsymbol{\Sigma}}^{-1}$, takes the usual Bayesian form of being the sum of the data precision, plus the prior

precision, $\hat{\alpha}\mathbf{I}$. With $\hat{\alpha} = 0$, i.e. no prior on the weights, we recover the ML solution of equation (A 3).

B.2. Updating the weight precision, α

We let

$$f(\alpha) = \iint q(\boldsymbol{\beta}|\mathbf{D})q(\mathbf{w}|\mathbf{D})\log[\,p(\mathbf{w}|\alpha)p(\alpha)]\mathrm{d}\mathbf{w}\mathrm{d}\boldsymbol{\beta}$$
$$= \int q(\mathbf{w}|\mathbf{D})\log[p(\mathbf{w}|\alpha)p(\alpha)]\mathrm{d}\mathbf{w}. \tag{B 13}$$

As before, the negative free energy is maximized when

$$q(\boldsymbol{\alpha}|\mathbf{D}) \propto \exp[f(\alpha)]. \tag{B 14}$$

By substituting for the terms in equation (B 13), we find that the updated weight precision posterior density is a gamma density of the form $q(\boldsymbol{\alpha}|\mathbf{D}) = \mathcal{G}(\alpha; \hat{b}_\alpha, \hat{c}_\alpha)$ where the updated hyper-hyperparameters, $\hat{b}_\alpha$ and $\hat{c}_\alpha$, are given by

$$1/\,\hat{b}_\alpha = \hat{\mathbf{w}}^{\mathrm{T}}\,\hat{\mathbf{w}} + \frac{1}{2}\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}) + \frac{1}{b_\alpha^0}, \tag{B 15}$$

$$\hat{c}_\alpha = \frac{B}{2} + c_\alpha^0,$$
$$\hat{\alpha} = \hat{b}_\alpha \hat{c}_\alpha, \tag{B 16}$$

where B is the number of basis functions.

B.3. Updating the noise precision, β

Again, we start by writing the function

$$f(\boldsymbol{\beta}) = \iint q(\alpha|\mathbf{D})q(\mathbf{w}|\mathbf{D})\log[\,p(\mathbf{D}|\mathbf{w},\alpha)p(\alpha)]\mathrm{d}\mathbf{w}\mathrm{d}\alpha,$$
$$= \int q(\mathbf{w}|\mathbf{D})\log[p(\mathbf{D}|\mathbf{w},\alpha)p(\alpha)]\mathrm{d}\mathbf{w}. \tag{B 17}$$

The negative free energy is then maximized when

$$q(\boldsymbol{\beta}|\mathbf{D}) \propto \exp[f(\boldsymbol{\beta})]. \tag{B 18}$$

By substituting for the terms in equation (B 17) we find, as with $\boldsymbol{\alpha}$, that the posterior distribution over $\boldsymbol{\beta}$ is of gamma form, $q(\boldsymbol{\beta}|\mathbf{D}) = \mathcal{G}(\boldsymbol{\beta}; \hat{b}_\beta, \hat{c}_\beta)$, with updates to the hyper-hyperparameters $\hat{b}_\beta$ and $\hat{c}_\beta$ given by

$$\left.\begin{aligned} \frac{1}{\hat{b}_\beta} &= \frac{1}{2}(\mathbf{Y} - \boldsymbol{\Phi}\hat{\mathbf{w}})^{\mathrm{T}}(\mathbf{Y} - \boldsymbol{\Phi}\hat{\mathbf{w}}) + \frac{1}{2}\mathrm{Tr}(\boldsymbol{\Sigma}\,\boldsymbol{\Phi}^{\mathrm{T}}\,\boldsymbol{\Phi}) + \frac{1}{b_\beta^0} \\ \hat{c}_\beta &= \frac{N}{2} + c_\beta^0 \\ \text{and} \quad \hat{\boldsymbol{\beta}} &= \hat{b}_\beta \hat{c}_\beta, \end{aligned}\right\} \tag{B 19}$$

where N is the number of data points.

B.4. Structured priors

Instead of using the isotropic Gaussian of equation (A 4), where the distribution over all the weights has a common scale (defined by the single hyperparameter $\boldsymbol{\alpha}$), we can split the weights into groups and allow different groups to have different scales in their distributions; each weight w_i can indeed have its own scale hyperparameter. This approach is often referred to as *automatic relevance determination* [14,16,17], so-called because by inspecting the inferred scales associated with the weights, we can see which (groups of) weights are relevant to the problem at hand; those that are not helpful to our problem will evolve with distributions with vanishingly small variance, i.e. there is strong evidence that the weight lies close to zero. Conversely, those weights that are well supported by the data will entertain larger variances in their pdfs. The importance of this lies with each weight acting as a scaling associated with an observed variable (denoted here via the $\boldsymbol{\phi}_i$). Hence, weights that are unsupported by the data shrink to close to zero because their *associated variable is not explanatory of the data*. This means we may operate with a rich set of observed variables and allow the Bayesian model to select only those that have explanatory power in the data.

For our linear models, we want to allow for different w_i to have such different scales, and this property can be captured with 'structured priors'. For example, if we have strong domain knowledge that certain parameters should have a similar posterior scale, then we can group them. Our structured priors are hence of the form:

$$p(\mathbf{w}|\{\alpha_j\}) = \prod_{j=1}^{G}\left(\frac{\alpha_j}{2\pi}\right)^{B_j/2}\exp(-\alpha_j E_j(\mathbf{w})), \tag{B 20}$$

where the weights have been split into $j = 1 \ldots G$ groups with B_j weights in the jth group and where we define

$$E_j(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{I}_j\mathbf{w}, \tag{B 21}$$

with $\mathbf{I}_j$ being a matrix with 1s along the diagonal that pick off coefficients in the jth group, and zeros elsewhere. Use of structured priors results in VB updates for the posterior weight covariance and weight precision as follows:

$$\left.\begin{aligned} \hat{\boldsymbol{\Sigma}} &= \left(\hat{\boldsymbol{\beta}}\,\boldsymbol{\Phi}^{\mathrm{T}}\,\boldsymbol{\Phi} + \sum_{j=1}^{G}\hat{\alpha}_j\,\mathbf{I}_j\right)^{-1}, \\ \frac{1}{\hat{b}_\alpha(j)} &= E_j(\hat{\mathbf{w}}) + \frac{1}{2}\mathrm{Tr}(\mathbf{I}_j\hat{\boldsymbol{\Sigma}}\mathbf{I}_j) + \frac{1}{b_\alpha^0}, \\ \hat{c}_\alpha(j) &= \frac{B_j}{2} + c_\alpha^0 \\ \text{and} \quad \hat{\alpha}_j &= \hat{b}_\alpha(j)\hat{c}_\alpha(j). \end{aligned}\right\} \tag{B 22}$$

The other updates are exactly the same as for the global variance scale over the parameters.

B.5. Application details

The above methodology is used twice within this paper. Once for forecasting future welfare variables and secondly for fitting a trend curve to the results. The latter usage does not form part of the analysis, and the trend curve is solely used for making presentation of the data clearer. These two cases are detailed as follows.

Forecasts: In our forecasts, the target variables are welfare measures and the observations vector, $\boldsymbol{\phi}(x)$ are

the optical flow statistics either observed on day x or from all days 1 to x. In the former case, $\boldsymbol{\phi}$ is a four-dimensional vector of optical flow mean, variance, skewness and kurtosis. In the latter case, $\boldsymbol{\phi}$ is a vector of increasing length, as it represents a vector of optical flow statistics from each day prior to and including day x.

Trend curves: To form smooth trend curves through the day-by-day correlations of our forecasts with the welfare targets, we exploit the above model with $\boldsymbol{\phi}$ being a set of *thin-plate spline* basis functions, one centred on each day of our results. The spline functions are given by

$$\boldsymbol{\phi}_i(x) = |x - x_i|^2 \ln|x - x_i|,$$

where x_i are locations of the spline basis functions, here located at each day of the study.

REFERENCES

1 Koolhaas, J. M. 2008 Coping style and immunity in animals: making sense of individual variation. *Brain Behav. Immun.* **22**, 662–667. (doi:10.1016/j.bbi.2007.11.006)

2 Urton, G., von Keyserlingk, M. A. G. & Weary, D. M. 2005 Feeding behavior identifies dairy cows at risk for metritis. *J. Dairy Sci.* **88**, 2843–2849. (doi:10.3168/jds.S0022-0302(05)72965-9)

3 Littin, K., Acevedo, A., Browne, W., Edgar, J., Mendl, M., Owen, D., Sherwin, C., Wurbel, H. & Nicol, C. 2008 Towards humane end points: behavioural changes precede clinical signs of disease in a Huntingtons disease model. *Proc. R. Soc. B* **275**, 1865–1874. (doi:10.1098/rspb.2008.0388)

4 Beattie, V. E., Breuer, K., O'Connell, N. E., Sneddon, I. A., Mercer, J. T., Rance, K. A., Sutcliffe, M. E. M. & Edwards, S. A. 2005 Factors identifying pigs predisposed to tail biting. *Anim. Sci.* **80**, 307–312. (doi:10.1079/ASC40040307)

5 Toscano, M. J., Sait, L., Jorgensen, F., Nicol, C. J., Powers, C., Smith, A. L., Bailey, M. & Humphrey, T. J. 2010 Sub-clinical infection with salmonella in chickens differentially affects behaviour and welfare in three inbred strains. *Br. Poult. Sci.* **51**, 703–713. (doi:10.1080/00071668.2010.528748)

6 Lee, H.-J., Roberts, S. J., Drake, K. A. & Dawkins, M. S. 2011 Prediction of feather damage in laying hens using optical flow and Markov models. *J. R. Soc. Interface* **8**, 489–499. (doi:10.1098/rsif.2010.0268)

7 Dawkins, M. S., Lee, H.-J., 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)

8 Kestin, S. C., Knowles, T. G., Tinch, A. E. & Gregory, N. 1992 Prevalance of leg weakness in broiler chickens and its relationship with genotype. *Vet. Rec.* **131**, 190–194. (doi:10.1136/vr.131.9.190)

9 Beauchemin, S. & Barron, J. 1995 The computation of optical flow. *ACM Comput. Surv.* **27**, 433–467. (doi:10.1145/212094.212141)

10 Fleet, D. J. & Weiss, Y. 2005 Optical flow estimation. In *Handbook of mathematical models for computer vision* (eds N. Paragios, Y. Chen & O. Faugeras), pp. 239–258. New York, NY: Springer.

11 Haslam, S. M., Brown, S. N., Wilkins, L. J., Kestin, S. C., Warriss, P. D. & Nicol, C. J. 2006 Preliminary study to examine the possibility of using foot burn or hockburn to assess aspects of housing conditions in broiler chickens. *Br. Poult. Sci.* **47**, 13–18. (doi:10.1080/00071660500475046)

12 Kjaer, J. B., Su, G., Nielsen, B. L. & Särenson, P. 2006 Foot pad dermatitis and hock burn in broiler chickens and degree of inheritance. *Poult. Sci.* **85**, 1342–1348.

13 Hepworth, P. J., Nefedov, A. V., Muchnik, I. B. & Morgan, K. L. 2010 Early warning indicators for hock burn in broiler flocks. *Avian Pathol.* **39**, 405–409. (doi:10.1080/03079457.2010.510500)

14 Bishop, C. M. 2006 *Pattern recognition and machine learning*. New York, NY: Springer.

15 Fox, C. W. & Roberts, S. J. 2012 A tutorial on variational Bayesian inference. *Artif. Intell. Rev.* **38**, 1–11. (doi:10.1007/s10462-011-9236-8)

16 Neal, R. 1998 Assessing relevance determination methods using DELVE. In *Neural networks and machine learning* (ed. C. Bishop), pp. 97–129. Berlin, Germany: Springer-Verlag.

17 Penny, W. & Roberts, S. 2002 Bayesian multivariate autoregressive models with structured priors. *IEE Proc. Vision, Image Signal Proc.* **149**, 33–41. (doi:10.1049/ip-vis:20020149)