

The importance of prior choice in model selection: a density dependence example

James D. Lawrence^{1*}, Robert B. Gramacy², Len Thomas³ and Stephen T. Buckland³

¹Statistical Laboratory, Cambridge University, Wilberforce Road, Cambridge, CB3 9DQ, UK; ²Booth School of Business, University of Chicago, 5807 South Woodlawn Ave, Chicago, IL, 60637, USA; and ³Centre for Research into Ecological and Environmental Modelling, University of St Andrews, St Andrews, KY16 9LZ, UK

Summary

1. It is important to discern the magnitude of density dependence a species exhibits, as well as the time lag over which it operates. Knowledge of a species' likely response to natural as well as synthetic shocks will assist in effective species management. Statistically this is a challenging problem which does not usually admit closed-form mathematical analysis. Consequently, many people have used Bayesian methods to fit state space models of density dependence to many different species, of which we take eleven species of North American duck as our motivating examples. A Bayesian analysis requires a choice of model and parameter prior. The latter is difficult to do without inducing bias in model selection, and we attempt to address this problem.

2. Our priors will be obtained by considering which parameter values are representative of features we expect to see in the data, and which would produce unnatural behaviour. To fit the models, we use a novel sequential Monte Carlo method (particle learning) not previously applied to ecological data sets.

3. We show that existing analyses on the duck data may have been susceptible to a common problem in Bayesian model selection (Lindley's paradox), and suggest methods for prior selection which mitigate this issue. We also discover that although it is possible to detect the existence of density dependence, it is unrealistic to expect to determine the time lag over which it operates without a great deal of data, even if said data are simulated from the model.

4. We demonstrate that prior choices motivated by the above considerations can lead to substantially increased predictive accuracy over surprisingly long time scales whether model selection is of primary concern or not.

5. We conclude from our analysis of real-world data that there is little evidence of density dependence in many duck species, suggesting that such effects, if present, are likely to be small in magnitude.

Key-words: autoregressive, Bayesian, density dependence, Lindley's paradox, model selection, North American ducks, prior selection

Introduction

DENSITY DEPENDENCE

Density dependence within a species is usually the primary means of numerical self-regulation, the mechanism by which a species can maintain a steady population trajectory in an environment that produces unexpected events of both beneficial and harmful natures. Turchin (1995), in a synthesis of several other sources, states that density dependence is necessary for a regulated population. That is, a population without it is almost certain to be numerically unstable, with an undefined carrying capacity.

The debate over the relevance of density dependence has been at times acrimonious, as summarised in Turchin (1995). The quote from that paper (p. 31) which we take as our starting point on this issue is that available evidence 'is entirely consistent with the universal applicability of the density dependence

model'. As such, we seek to make what statistical inferences we can about the magnitude and time period of such effects. There are several biological hypotheses as to the causes of density dependence, both in general and in the specific case of North American ducks, our motivating example. These have differing implications for the likely degree of density dependence to be expected in such species.

We will use as an illustrative example data from US Fish and Wildlife Service (2010) on the abundance of North American ducks, but we stress that this is only one particular application of our methodology. We deal fundamentally with how to perform a conscientious Bayesian model selection process and that problem is found across statistical ecology and beyond.

Using historical count data provided by US Fish and Wildlife Service (2010), we analyse ten species of duck, including both diving and dabbling ducks, between which there is reason to expect a distinction in density dependence profile. The hypothesis tested (and to an extent borne out) by Jamieson & Brooks (2004) was that diving ducks might, in response to a poor year

*Correspondence author. E-mail: j.d.lawrence@statslab.cam.ac.uk

(low habitat and/or food availability), delay breeding for a year. This would imply a delayed density dependence in diving ducks that would not be present in dabbling ducks.

In contrast, Sargeant, Allen, & Eberhardt (1984) looked at red fox (*Vulpes vulpes*) predation on both diving and dabbling ducks, and concluded that dabbling ducks are significantly more vulnerable to predation of this kind. The red fox is only one predator of ducks in North America, but it is one of the primary predators, and, in common with many other duck predators, it is a generalist. A hypothesis of Bjørnstad, Falck & Stenseth (1995), tested in Viljugrein *et al.* (2005), suggests that this would induce more immediate density dependence in the affected species because both ducks and eggs are potential predatory targets. This would imply both first and second order density dependence in dabbling ducks; less so in diving ducks.

BAYESIAN STATE SPACE MODELS

We will take a Bayesian standpoint when analysing these data. This is not because a classical analysis is impossible, but rather because we believe that common sense can be translated into a meaningful, informative parameter prior.

Inference about the degree of density dependence under this framework is a Bayesian model selection problem. Link & Barker (2006) illustrate the principles and some of the issues inherent to this class of problem. Two themes from this study that we will explore in depth are that:

1. An uninformative prior for a particular parameter may induce a great deal of information in the model selection process.
2. Two different, apparently uninformative priors for the same parameter may give very different model selection results.

We demonstrate that choosing an informative prior (using simple rules which we will describe) is both necessary for a balanced model selection procedure and improves the accuracy with which we can predict future population levels.

We will use a state space formulation for our density dependence model. Suppose we make an observation y_t in year t , and the actual quantity of interest is x_t , which we cannot measure directly. A state space model must specify the *Observation Process* $p(y_t|x_t)$ and the *System Process* $p(x_t|x_{t-1}, \dots)$. Both of these processes are usually stochastic, though exceptionally the observation process might not be.

Materials and methods

AN AUTOREGRESSIVE MODEL FOR DENSITY DEPENDENCE

We consider a density dependence model from Dennis & Taper (1994). Let x_t be the log-population size in year t . The evolution of x_t over time is modelled by the stochastic update

$$x_t = x_{t-1} + b_0 + \sum_{i=1}^k b_i e^{x_{t-i}} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2). \quad \text{eqn 1}$$

The parameters are interpreted as, k , degree (maximum time lag) of density dependence, in years; b_0 , uninhibited exponen-

tial growth rate; $b_{1,k}$, density dependence effects at different time lags; σ^2 , species (and unmodelled covariate) volatility.

We will refer to the combined vector $(b_0, b_{1,k})^T$ as \mathbf{b} . The number of parameters in \mathbf{b} is $k+1$, so k is a model order parameter. If $k = 0$, then this process simplifies to a random walk with drift. Also, if several different mechanisms induce a density dependence effect at the same time lag, then the appropriate component of \mathbf{b} will in effect be a summary statistic measuring the sum of all effects at that time lag.

Observe that this model is not phenomenological; it does not emphasise demographical parameters. Rather, it seeks to provide a realistic and reasonably flexible framework for assessing the magnitude and duration of density dependence effects.

We do not in general observe a true and accurate count of the species abundance. We observe data y_t that will include noise which may vary in intensity from year to year. We assume that this observation process is Gaussian, that is,

$$y_t \sim N(e^{x_t}, S_t^2) \quad \text{eqn 2}$$

and we assume that S_t is known for each year $t = 1, \dots, T$. The full model as specified by eqns 1 and 2 is thus a state space formulation.

We treat the estimate of the observation error as exact. It is in actuality only an estimate; however, the estimation procedure is independent of the final counts as the data come from double-observer surveys. Thus, any inaccuracy induced by treating the observation errors as exact is likely to have little effect on model selection or parameter estimation.

We will analyse a total of eleven species. Seven of these are dabbling ducks: Mallard (*Anas platyrhynchos*), American Wigeon (*Anas americana*), Gadwall (*Anas strepera*), Green-Winged Teal (*Anas crecca*), Blue-Winged Teal (*Anas discors*), Northern Shoveler (*Anas clypeata*) and Northern Pintail (*Anas acuta*). The remaining four are diving ducks, two of which are amalgamated: Redhead (*Aythya americana*), Canvasback (*Aythya valisineria*) and Greater and Lesser Scaup (*Aythya marila* and *Aythya affinis*). The data, as supplied by US Fish and Wildlife Service (2010), include both an estimated annual count and an estimate of the observation error. An example (the American Wigeon) is given in Fig. 1.

LINDLEY'S PARADOX

It has been known for some time Lindley (1957) that choosing a vague (high-variance) prior for within-model parameters (except for parameters common to all of them, such as σ^2) will bias the model selection routine in favour of simple models. This is discussed in depth in Link & Barker (2006). In the limiting case where an improper flat prior is used for all parameters, the posterior model probabilities will always be degenerate in favour of the model with the fewest parameters. Lindley's paradox therefore implies that we cannot take a diffuse Normal prior for \mathbf{b} , as this would lead to selecting $k = 0$, even if the data produced a likelihood that was higher for other models (hence the paradox).

In the light of this, it is clear that we must choose an informative prior, but the question arises as to how to choose an informative prior when one has, apparently, no information.

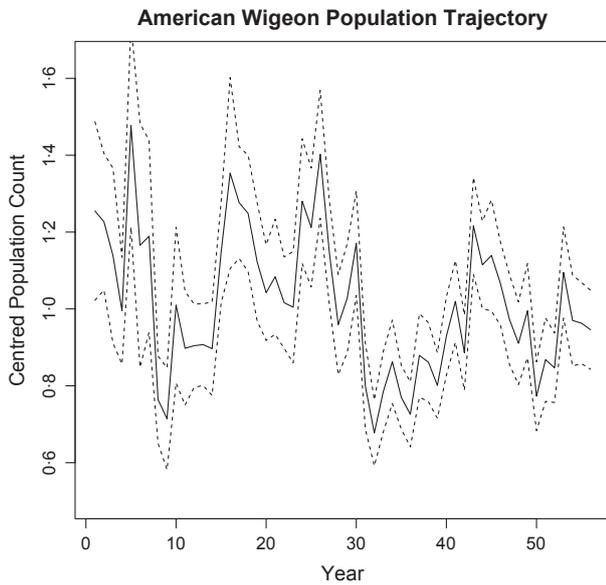


Fig. 1. American Wigeon data, after centring. Dotted lines are mean \pm two standard errors.

We now show that an informative choice can be reached just by excluding certain pathological cases that we would not expect to arise in the biological systems in question.

MODEL STABILITY CONSIDERATIONS

The population evolution model specified in eqn 1 is simple to simulate from. Some examples are given in Fig. 2. One notices that for certain parameter values, the simulated population

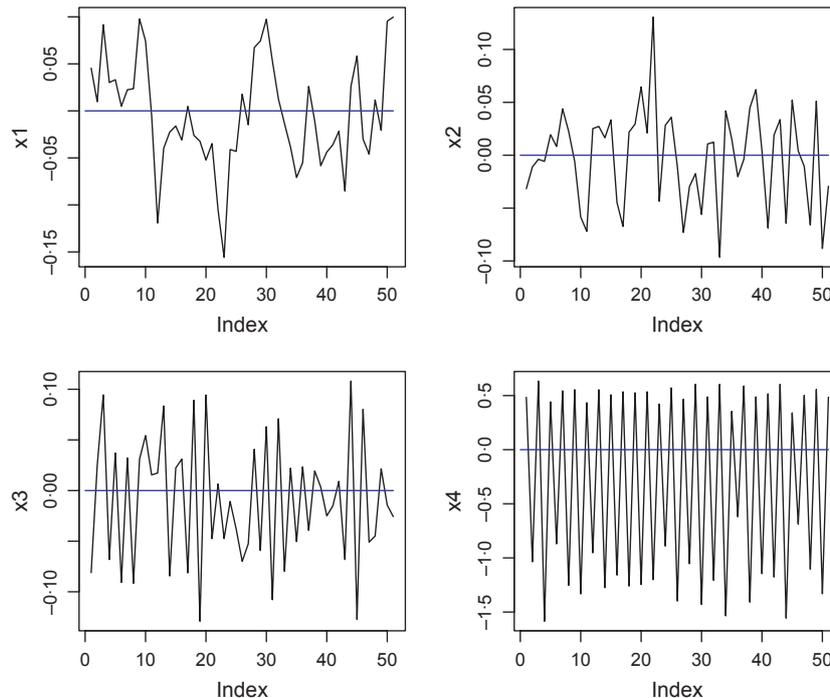


Fig. 2. Simulations from the autoregressive model, with $\mathbf{b} = (1/2, -1/2), (1, -1), (3/2, -3/2)$ and $(5/2, -5/2)$. Note that the last of these is a stable exception to the usually unstable case $b_0 > 2$. $\sigma = 0.05$ for all of these, with the process driving the greatly increased variance for the last simulation. There is no measurement error, and we observe from $t = 100$ to $t = 150$ starting at $(x_1, x_2) = (0, 0)$.

fluctuates wildly or grows very rapidly until the computer suffers numerical overflow. However, for other values, the population reaches a stable threshold after a period of time (regardless of its starting value) and then does not move too far from this. We refer to this level as the model-based carrying capacity, since it is the maximum level for which the expected population trajectory is not downwards. We would like to restrict our parameters to values that produce a (finite) carrying capacity (exempt from this is the null model $k = 0$, as it can never have a carrying capacity). We will demonstrate that a diffuse independent Normal prior does not always lead to the stable scenario, but there are other priors that do (at least much more often).

Consider the deterministic analogue of the model eqn 1 with no measurement error, and suppose that we observe a string of k years where the population is at a constant level $x_{1:k} = x$. Then

$$x_{k+1} = x + b_0 + \sum_{i=1}^k b_i e^x.$$

If b_0 and $\sum_{i=1}^k b_i$ are of different sign (and k is at least 1), then we can solve for when $x_{k+1} = x$, and we find that this corresponds to

$$x = x^* = \log\left(\frac{-b_0}{\sum_{i=1}^k b_i}\right). \tag{eqn 3}$$

This exposes an inherent asymmetry in this model: that b_0 and the sum of the other components of \mathbf{b} need to be of different sign to produce stable populations. This is not captured in an independent Normal prior. In addition, it raises the problem

of estimating the carrying capacity x^* . We are constructing a prior, so ‘peeking’ at the data should be avoided where possible. The approach we suggest (and the one we use to produce our results) is to centre the observed data (on the log scale), so that the carrying capacity should correspond approximately to $x^* = 0$.

Previous work on this problem Jamieson & Brooks (2004); Viljugrein *et al.* (2005) has taken data as specified in millions. The decision to divide the raw data by one million (say) is no less arbitrary than the decision to centre. The model is not invariant to such a transformation; however, we do not move from one model to another (or outside the model space entirely). This is because the family of transformations

$$\mathbf{x} \rightarrow \mathbf{x} + \lambda, b_{1:k} \rightarrow e^{-\lambda} b_{1:k}$$

can be applied to \mathbf{x} and \mathbf{b} without changing the model equation (for any real λ).

To address the problem of the varying scales for different species, Jamieson & Brooks (2004) chose a prior variance for each component of \mathbf{b} that was different from each species, using a data-driven selection method.

We appear to have complicated the problem by introducing a new quantity x^* , which we either have to estimate or make prior assumptions about. This is an illusion because the new parameter relieves the problem that otherwise b_0 and the other components of \mathbf{b} act on different scales (due to the e^{x^*} coefficient of the latter in eqn 1). Correspondingly, we would have to construct a separate prior for b_0 and the rest of \mathbf{b} , and this extra choice is just as difficult or as subjective as choosing a value for x^* .

If we take the carrying capacity to be $x^* = 0$, then, rearranging 3, we get

$$b_0 = - \sum_{i=1}^k b_i. \quad \text{eqn 4}$$

Thus, b_0 is perfectly negatively correlated with each of the other components of b . If we take an independent Normal $(0, \sigma_b^2)$ prior for $b_{1:k}$, then this suggests that the joint prior for $b_{0:k}$ should be the degenerate Normal

$$b_{0:k} \sim N \left(\mathbf{0}, \begin{pmatrix} k\sigma_b^2 & -1 & \dots & -1 \\ -1 & \sigma_b^2 & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ -1 & 0 & 0 & \sigma_b^2 \end{pmatrix} \right). \quad \text{eqn 5}$$

This is degenerate in the sense that the covariance matrix does not have full rank, and only those values of \mathbf{b} for which 4 holds will have nonzero likelihood. In practice, this only applies to the deterministic model, and a small amount h would be added to the variance of b_0 to allow for misestimation of x^* . This is because there will always be probabilistic drift towards the carrying capacity, and by allowing some additional variation in b_0 , we introduce the requisite additional flexibility into the model. The choice of h also dictates the prior under the null $k = 0$ model, so a reasonable value might be obtained by considering the variance of symmetric Gaussian random walks over time. For example, a value of $h = 0.04225$ corresponds to

a process that is as likely as not to at least halve or double in five years. In other words, if $Z \sim N(0, 5 \times 0.04225)$ then $\mathbb{P}(Z \in [-\log(2), \log(2)]) = 1/2$. This is the value we use in all of our priors, which have h as a parameter.

We now consider the effect of small perturbations about carrying capacity. We will see that this restricts even further the set of parameter values that yield a dynamical system we might expect to see in a natural population.

Suppose that $x_{1:k} = (0, \dots, 0, \delta)$. Then we may be in one of several scenarios (eqn 3 is assumed to hold):

1. $\sum_{i=1}^k b_i$ is positive. In this case, regardless of the sign of δ , the population is unstable and will diverge from 0. Carrying capacity is undefined.
2. $-1 < \sum_{i=1}^k b_i < 0$. The population returns monotonically towards capacity.
3. $-2 < \sum_{i=1}^k b_i < -1$. The population oscillates around capacity, with decreasing magnitude.
4. $\sum_{i=1}^k b_i < -2$. The population oscillates around zero, but usually with much greater magnitude than (1)–(3). If all of $b_{1:k}$ are negative, then the oscillations will quickly reach a consistent (perhaps large) magnitude, but if any of $b_{1:k}$ are positive, then the population is probabilistically unbounded. That is, with probability 1, as $t \rightarrow \infty$, $x_t \rightarrow \infty$. In the latter case, capacity is again undefined.

Plots of simulated population trajectories for all four cases are given in Fig. 2. We contend that the second of these is most likely to be characteristic of a natural population, but that perhaps some allowance might be made for the third. The first and fourth are considered unlikely to arise in the natural world.

This means that had we chosen a prior of the form 5 then we would unintentionally be making a strong prior assumption about the model order. For example, if $k = 1$, then $\sum_{i=1}^k b_i$ is *a priori* a $N(0, \sigma_b^2)$ random variable, with a corresponding probability of lying in $[-2, 0]$. If $k = 2$, then $\sum_{i=1}^k b_i$ has a $N(0, 2\sigma_b^2)$ distribution, with correspondingly reduced probability of lying in this interval. This could be thought of as a manifestation of Lindley’s paradox. If for example $\sigma_b^2 = 1$, then the chance of being in the prior-plausible region under $k = 1$ would be 48%. Under $k = 5$, that chance shrinks to 31%. The difference is even more pronounced if σ_b^2 is higher. Thus, we would be accidentally favouring simple models.

A logical refinement of 5 is to keep the distribution of $\sum_{i=1}^k b_i$ constant, and to restrict to cases where $-2 < \sum_{i=1}^k b_i < 0$. This is

$$b_{0:k} \sim N \left(\mathbf{0}, \begin{pmatrix} \sigma_b^2 + h & -1 & \dots & -1 \\ -1 & \sigma_b^2/k & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ -1 & 0 & 0 & \sigma_b^2/k \end{pmatrix} \right) \quad \text{eqn 6}$$

restricted to the aforementioned set. This is easy and quick to sample from by rejection sampling. We do not do this ourselves in the following analysis because we have already constructed a prior that has a high chance of producing parameters in this region. However, restricting could be an alternative approach if constructing such a prior were infeasible otherwise.

The restricted prior also has the attractive property that the marginal distribution of b_0 is the same under all models except $k = 0$, so we are equally willing to entertain density dependence effects at different time lags, and we have not unintentionally biased our model prior towards small k , since the prior probability of a model where carrying capacity is defined is the same for all $k > 0$.

Prior selection

To perform a full Bayesian analysis and fit of this model, we need to specify a prior for each parameter that is not directly specified by the model itself.

We give a uniform prior to k , over $\{0, \dots, 5\}$. We believe it is implausible that density dependence effects could operate on a longer timescale than this. In particular, the hypotheses that we wish to assess are only concerned with density dependence up to second order. Our prior gives no preference to one time lag over another in this range, so that we can assess the evidence provided by the data in favour of each model. This is similar to a Bayes Factor, which can be used to compare the fit of different models Kass & Raftery (1995).

An improper inverse gamma (0,0) prior is assigned to σ^2 . This is mostly for reasons of Bayesian conjugacy – the rate of learning is high for this parameter, and the prior shape makes little difference.

The distribution of $x_{1:5}$ might not be specified by the model (depending on k – if $k = 2$ for example, then we need to specify the distribution of x_1 and x_2 , and the model gives us the distribution of $x_{3:5}$). To have a consistent likelihood across all models, we consider the observed likelihood function $p(y_{1:5}|x_{1:5})$ as a (density) function of $x_{1:5}$ and treat it as our prior. Naturally, we do not count it twice, so it is removed from the likelihood, as well as those systemic terms relating to the evolution of $x_{1:5}$. Thus, for all models, the first model-driven term in the likelihood is $p(x_6|x_{1:5}, \sigma^2, \mathbf{b})$. The final parameter that requires a prior is \mathbf{b} , but we must remember to take account of Lindley's paradox before we make our choice.

PARTICLE LEARNING

We use a Particle learning method Carvalho *et al.* (2010) combined with Reversible Jump MCMC Green (1995) to produce a sample from the posterior for each simulation. The algorithm can be summarised thus (full R code available as an online appendix to this paper):

1. Produce an initial sample of N 'particles' $P_{1:N}$, each of which consists of a value for k , \mathbf{b} , σ and $x_{1:k_{\max}}$ drawn from their respective priors.
2. To update from time t to time $t+1$, first calculate for each particle its predictive accuracy

$$w^{(t+1)} = \int L(y_{t+1}|x_{t+1}, S_{t+1})p(x_{t+1}|x_{1:t}, k, \mathbf{b})dx_{t+1}.$$

This behaves as the incremental weight for that particle.

3. Resample the particles with replacement, sampling each with probability proportional to $w^{(t+1)}$ for that particle.

4. For each particle, sample a value of x_{t+1} from $p(x_{t+1}|x_t, k, \mathbf{b})$.
5. Update the values of k and \mathbf{b} for each particle by sampling from the equilibrium distribution of the update detailed in Troughton & Godsill (1997).
6. Repeat steps 2–5 until $t = T$.

This produces a weighted sample from the posterior distribution of models, parameters and hidden states. We are also able to chart the posterior as it evolves over time, as more data are added.

SIMULATION SETUP

Before we look at observed abundance data, we analyse some simulations of populations which follow the specified dynamics. We have twenty simulated data sets. Ten have the parameters $k = 1$, $\mathbf{b} = (0.5, -0.5)$. These will be referred to as set (1). The remaining ten have parameters $k = 2$, $\mathbf{b} = (0.5, -0.1, -0.4)$ and comprise simulation set (2). All simulations share the parameter $\sigma = 0.05$, and S_t were chosen to have mean 0.1 after centring. Both series have 500 years of data (this is considerably longer than the real survey, so we can see how much we can expect to learn about the model parameters in the future). We consider five prior choices for \mathbf{b} :

1. Independent Normal, variance 5 (primarily as an illustration of Lindley's Paradox).
2. Independent Normal, variance 1 (a baseline for comparison).
3. Multivariate Normal with covariance matrix from the modified version of 5, and $\sigma_b^2 = 1, h = 0.04225$.
4. A shrinkage-inspired prior. Shrinkage is a recurring theme when trying to predict normally distributed data. The central idea, as first exhibited in Stein (1955), is to reduce the prior variance of parameters corresponding to the mean of said distribution (i.e. to 'shrink' towards a certain value). Shrinkage offers improved (in mean squared error terms) prediction at the expense of parameter bias. In the spirit of this, our fourth prior is Normal with covariance matrix based on eqn 6:

$$b_{0:k} \sim N \left(\mathbf{0}, \begin{pmatrix} \left(\begin{array}{ccccc} \sigma_b^2 + h & -\sigma_b^2/k & \dots & \dots & -\sigma_b^2/k \\ -\sigma_b^2/k & \sigma_b^2/k & 0 & 0 & 0 \\ \vdots & 0 & \sigma_b^2/k & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ -\sigma_b^2/k & 0 & 0 & 0 & \sigma_b^2/k \end{array} \right) \end{pmatrix} \right) \quad \text{eqn 7}$$

and again $\sigma_b^2 = 1, h = 0.04225$.

5. As (4), but with smaller variance ascribed to later components of \mathbf{b} :

$$b_{0:k} \sim N \left(\mathbf{0}, \begin{pmatrix} \left(\begin{array}{ccccc} \sigma_b^2 + h & -\sigma_b^2 * d & \dots & \dots & -\sigma_b^2 * d/k \\ -\sigma_b^2 * d & \sigma_b^2 * d & 0 & 0 & 0 \\ \vdots & 0 & \sigma_b^2 * d/2 & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ -\sigma_b^2 * d/k & 0 & 0 & 0 & \sigma_b^2 * d/k \end{array} \right) \end{pmatrix} \right) \quad \text{eqn 8}$$

d is suitably defined so that the sum of variances of $b_{1:k}$ is σ_b^2 . In fact, under this restriction

$$d = \frac{1}{\sum_{j=1}^k 1/j}$$

in eqn 8 for $k \geq 1$. Notice that both priors (4) and (5) have the same total variance for \mathbf{b} , as long as $k > 0$. This is deliberate, as discussed earlier.

The choice between the last two priors largely depends on whether one considers the assumption that longer lags tend to be smaller in size to be suitable *a priori*. We will see that they do not provide substantially different estimates or predictions, but then we only consider simulations for low values of k .

MEASURES OF PREDICTIVE PERFORMANCE

We use three measures to assess the predictive accuracy of the priors proposed in this study.

1. The expected mean square error (MSE) is the expected squared difference between the predicted hidden state $\exp(\hat{x}_{t+1})$ at time $t+1$ (based on all the data up to time t) and the datum y_{t+1} . This is calculated for each particle in the particle set, and a weighted average taken. A lower value of MSE indicates better predictive accuracy. The MSE can be calculated for a particular time t or averaged across the whole time series.

2. The Mahalanobis distance Mahalanobis (1936) is based on taking a Gaussian approximation to the predictive distribution and calculating the expected total squared error over the whole time series. It is given by

$$D_M = (\mathbf{y} - \exp(\hat{\mathbf{x}}))^T S^{-1} (\mathbf{y} - \exp(\hat{\mathbf{x}})). \quad \text{eqn 9}$$

S is the covariance matrix of $\exp(\hat{\mathbf{x}})$. The Mahalanobis distance is not a function of time, it measures performance from start to finish. As with the MSE, a low Mahalanobis distance means good overall predictive accuracy. It is a measure of performance across the whole time series.

3. For our simulated data, we know the true value of \mathbf{x} , so we can use $\exp(\mathbf{x})$ instead of \mathbf{y} and calculate the Mahalanobis distance this way. Of course, we cannot do this for the real-world data.

Results

SIMULATED DATA

Model selection results

There is a fair amount of variation on model posteriors across simulations. The evolving model posterior for a typical $k = 2$ simulation is shown in Fig. 3. Average model posteriors at various time points are given in Tables 1 and 2. Even with data simulated from the model, model selection is far from conclusive. After 500 years, models of greater order than the truth cannot be ruled out. It is pleasing to see that models that are too small (in this case, $k = 1$ and $k = 0$) disappear from the posterior relatively quickly and have no noticeable posterior mass by the end of the simula-

tion. However, they still have a small presence after 50 years, which would normally be considered quite long for an ecological data set.

The posterior at $t = 6$ is of some interest, as it provides an indication of the prior model bias. The independent priors can be seen to initially favour small models, before the likelihood has much of an effect. Notice that prior (3) has a very close to uniform posterior at $t = 6$, which is pleasing as it indicates a lack of model bias.

We can see that in this particular simulation, the vaguest prior (1) has the highest support for the low-order models. This is in keeping with Lindley's paradox. Priors (4) and (5) have the least support for small models, which we expect because we know that shrinkage procedures of this kind will favour a more complex model if it can predict future data well. Priors (2) and (3) fall somewhere in between. This pattern is shared amongst all the simulations, despite the considerable difference in the posteriors themselves.

The posteriors from the $k = 1$ simulations (not pictured) in general place more weight on the correct model than for the $k = 2$ simulations. This suggests that it is easier to fit to an instance of a simpler model.

Predictive accuracy

Tables 3–5 summarise the predictive accuracy of all the priors. The tables can each be summarised in the same way: The vague prior (1) shows very poor predictive accuracy. The standard Normal prior (2) is an improvement, but the accuracy is best for our suggested priors (3), (4) and (5). Of these, (4) and (5) perform a little better than (3). This is to be expected because shrinkage is designed to maximise predictive accuracy, but this comes at the expense of model selection.

ANALYSIS OF OBSERVED DATA

The posterior model probabilities for each species, using the shrinkage prior (4), are summarised in Table 6.

None of the posteriors are conclusive as to the order of density dependence. We expect this from the simulation study; even with data that we *know* follows a particular instance of the model, we can only expect perhaps a 60% posterior probability for that model after this length of time. It would be optimistic to expect the same level of agreement with real data, though we do see that for some species.

The Mahalanobis distances for each combination of species and prior are shown in Table 7. These mirror the results for the simulated data sets, with our proposed priors making more accurate predictions than the independent priors.

Discussion

We can see from Table 6 that for all but the Canvasback and Scaup, the posterior has a reasonable proportion ($\geq 10\%$) of its mass on $k = 0$. This might be indicative that these species are not in fact at carrying capacity, but it could also be the case that any density-dependent effects are swamped by the random

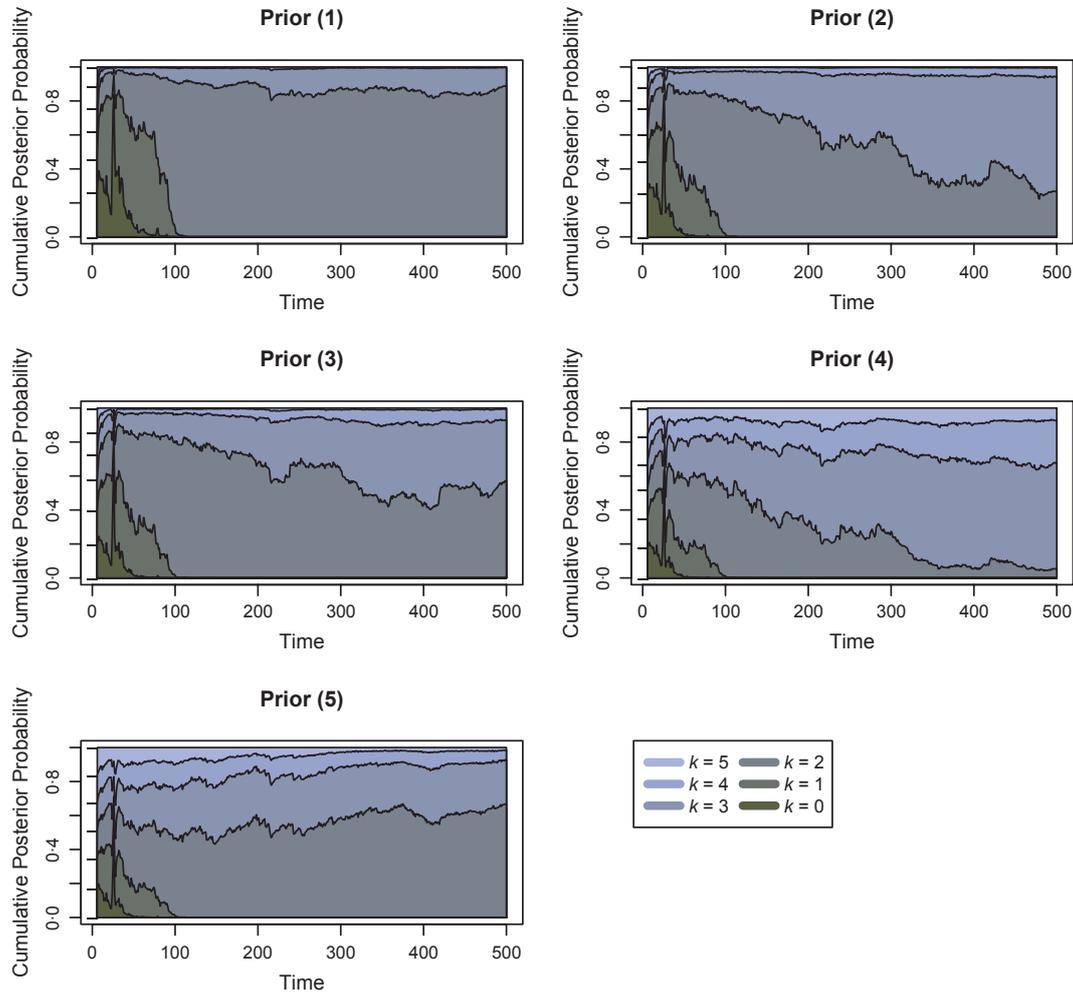


Fig. 3. Evolution of the posterior model distribution over a long time span when $k = 2$, for the five different prior choices. The tick marks to the left of the plot show the model posterior at $t = 6$.

Table 1. Mean posterior probabilities under the $k=1$ simulation scenario, using prior (3)

Time	Posterior					
	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
$t=6$	0.196	0.189	0.172	0.155	0.147	0.141
$t=25$	0.102	0.455	0.278	0.105	0.042	0.018
$t=50$	0.034	0.488	0.315	0.108	0.041	0.013
$t=500$	0.000	0.381	0.454	0.142	0.018	0.004

Table 2. Mean posterior probabilities under the $k=2$ simulation scenario, using prior (3)

Time	Posterior					
	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
$t=6$	0.186	0.185	0.169	0.159	0.153	0.148
$t=25$	0.194	0.304	0.222	0.132	0.086	0.061
$t=50$	0.036	0.332	0.272	0.209	0.113	0.038
$t=500$	0.000	0.000	0.176	0.484	0.280	0.059

Table 3. Average (SD) MSE for each prior and simulation scenario

Prior	Simulation	
	$k=1$	$k=2$
(1)	0.76 (0.17)	0.71 (0.20)
(2)	0.041 (0.0088)	0.041 (0.0080)
(3)	0.0024 (0.00039)	0.0030 (0.00045)
(4)	0.0023 (0.00033)	0.0027 (0.00033)
(5)	0.0022 (0.00027)	0.0027 (0.00035)

Table 4. Average (SD) Mahalanobis distance (compared with the observations \mathbf{y})

Prior	Simulation	
	$k=1$	$k=2$
(1)	35 000 (9700)	34 000 (13 700)
(2)	1900 (550)	1900 (630)
(3)	120 (26)	150 (23)
(4)	110 (21)	130 (18)
(5)	110(20)	140(21)

Table 5. Average (SD) Mahalanobis distance (compared with the truth \mathbf{x})

Prior	Simulation	
	$k=1$	$k=2$
(1)	35 000 (9800)	34 000 (13 900)
(2)	2200 (570)	2300 (650)
(3)	370 (25)	540 (38)
(4)	370 (22)	530 (38)
(5)	360 (26)	530 (37)

Table 6. Posterior model probabilities for each duck species, using a shrinkage prior

Species	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Mallard	0.172	0.188	0.208	0.078	0.240	0.114
A.Wigeon	0.024	0.419	0.264	0.134	0.060	0.010
Gadwall	0.681	0.166	0.055	0.019	0.045	0.034
G.W.Teal	0.682	0.098	0.052	0.053	0.059	0.056
B.W.Teal	0.442	0.320	0.135	0.058	0.028	0.018
N.Shoveler	0.491	0.076	0.249	0.094	0.042	0.048
N.Pintail	0.273	0.232	0.271	0.106	0.032	0.086
Redhead	0.324	0.067	0.201	0.149	0.136	0.124
Canvasback	0.030	0.528	0.192	0.094	0.092	0.064
Scaup	0.576	0.145	0.135	0.068	0.042	0.033

Table 7. Mahalanobis error for different prior choices, USFWS data

Species	$N(0, 5)$	$N(0, 1)$	Corr.	Shrink. 1	Shrink.2
Mallard	20 726	5416	1767	1622	1575
A.Wigeon	4803	1616	949	818	807
Gadwall	2884	1498	1302	1044	1029
GW.Teal	4266	2171	1717	1326	1396
BW.Teal	7553	2286	1088	1035	992
N.Shoveler	4384	1789	1268	1122	1130
N.Pintail	19,561	5228	2367	1798	1763
Redhead	3361	1636	1117	1005	981
Canvasback	3118	1073	529	479	481
Scaup	16 153	3702	1037	972	879

year-to-year effects ϵ_t . In any case, there appears to be little distinction between diving and dabbling ducks in terms of the degree of density dependence present (if at all).

This lack of density dependence is not consistent with the hypothesis that diving ducks will delay breeding in a year of poor resources. If that were the case we would expect to see a qualitative difference between the posteriors of the two categories, which we do not, using any of our suggested unbiased priors. This was the hypothesis investigated in Jamieson & Brooks (2004). The authors of that study concluded that there was evidence of increased density dependence in the Redhead and Canvasback (both diving ducks). These two species are also those with the lowest numbers out of all species analysed (of the order of half a million individuals, the next lowest being the Gadwall at around 1.5 million). The difference between

their conclusions and ours arises from the fact that they did not use centred data. This combined with a prior in which b_0 had the same marginal as $b_{1:k}$ led to much greater posterior concentration on $k = 0$ than our results for all species with more than 1 million individuals, and the reverse for those species with fewer.

On the issue of red fox predation, given that the red fox is a generalist, it is consistent with our findings that when there are many ducks, the red fox will prey on them. However, if there are sufficient alternative food sources in a year with few ducks, the causal link of lower duck numbers leading to lower fox numbers will be greatly diminished, which might explain why this fails to induce any lagged density dependence in dabbling ducks over diving ducks.

IDENTIFIABILITY, CLASSICAL INFERENCE AND DATA CLONING

Data cloning Lele, Dennis & Lutscher (2007) is a tool that can produce a classical analysis (MLEs, confidence intervals and so on) from a Bayesian one. This is achieved by replicating the data and treating it as a larger data set (i.e. literally 'cloning the data'). In our state space context, this implies having multiple observations for each species in each year, each of which have the same mean value and standard deviation. Thus, data cloning in our case amounts to investigating the posterior when we shrink the observational variance. In the limit where this variance tends to zero, the posterior mean would be the maximum likelihood estimate (including over the model space). A confidence interval can also be obtained by analysing how the posterior shrinks with the observational standard error.

In addition, data cloning can be used to assess parameter identifiability: if the posterior for a certain parameter or combination of parameters does not become more precise as we clone the data more times, then that is an indication that the parameter cannot be learned from the data.

We choose not to take a data cloning approach for two reasons. Firstly, we believe that a Bayesian approach adds value over a classical approach. MLEs are in general biased in a non-linear system with measurement error if the error is ignored Carroll *et al.* (2006). In addition, we have chosen our priors to contain some information which we believe reflects reality, and this would be lost if we were to use classical tools. Secondly, and of a more technical nature, data cloning is currently unable to perform model averaging. We have seen that there is considerable posterior uncertainty across models, and we would rather not constrain ourselves to one choice. (Data cloning is capable of selecting a model, but not of averaging across them.) It is also worth noting that as the posterior becomes more and more peaked, the effective sample size from the particle method will drop (if one were to use an alternative Monte Carlo Method, then one would experience alternative problems such as low MCMC acceptance rates, and in any case, the Monte Carlo error would be increased).

We hope that we have demonstrated the importance of a considered choice of prior. A default choice is rarely safe in model selection problems, and we have shown how, by

considering whether the carrying capacity is well defined and trying to exclude cases where it is not, we can arrive at an informative prior without peeking at the data.

A more general principle is that of excluding the so-called unphysical possibilities from the prior, that is, not allowing parameters to take values which would produce behaviour we know does not happen. We excluded models which did not give rise to a well-defined carrying capacity; the precise nature of the prior restrictions will vary from problem to problem.

It is important to consider how a parameter's prior varies between models: a parameter with a different interpretation in different models may well require a different prior in each case. In our example, b_0 typically had a prior that was different under the null model $k = 0$ than in more complex cases. This mirrored the fact that in the null model, b_0 was interpreted as an overall drift, whereas otherwise it was the counterbalance to the density-dependent effects.

When we exercise such caution in choosing our parameter priors, we are in a position to judge much more effectively whether the data provide evidence in favour of our hypotheses or not.

Acknowledgements

J.D.L. was funded by an Engineering and Physical Sciences Research Council grant. This work was also partially funded by EPSRC grant EP/D06570 4/1 to R.B.G. Most of this research was conducted whilst R.B.G. was a Lecturer at the Statistical Laboratory, University of Cambridge.

References

- Bjørnstad, O., Falck, W. & Stenseth, N. (1995) A geographic gradient in small rodent density fluctuations: a statistical modelling approach. *Proceedings of Biological Sciences*, **262**, 127–133.
- Carroll, R., Ruppert, D., Stefanski, L. & Crainiceanu, C. (2006) *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Monographs on Statistics and Applied Probability. Taylor & Francis, Boca Raton, FL.
- Carvalho, C., Johannes, M., Lopes, H. & Polson, N. (2010) Particle learning and smoothing. *Statistical Science*, **25**, 88–106.
- Dennis, B. & Taper, M. (1994) Density dependence in time series observations of natural populations: estimation and testing. *Ecological Monographs*, **64**, 205–224.
- Green, P. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.
- Jamieson, L.E. & Brooks, S.P. (2004) Density dependence in North American ducks. *Animal Biodiversity and Conservation*, **27**, 113–128.
- Kass, R. & Raftery, A. (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.
- Lele, S., Dennis, B. & Lutscher, F. (2007) Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters*, **10**, 551–563.
- Lindley, D. (1957) A statistical paradox. *Biometrika*, **44**, 187–192.
- Link, W. & Barker, R. (2006) Model weights and the foundations of multimodel inference. *Ecology*, **87**, 2626–2635.
- Mahalanobis, P. (1936) On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, **12**, 49–55.
- Sargeant, A. B., Allen, S. H. & Eberhardt, R. T. (1984) Red fox predation on breeding ducks in midcontinent North America. *Wildlife Monographs*, 3–41.
- Stein, C. (1955) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium* (ed. J. Neyman), pp. 197–206. University of California Press, Berkeley, CA.
- Troughton, P. & Godsill, S. A. (1997) Reversible Jump Sampler for Autoregressive Time Series, Employing Full Conditionals to Achieve Efficient Model Space Moves. Technical Report CUED/F-INFENG/TR.305, Signal Processing and Communications Laboratory, Dept. of Engineering, University of Cambridge.
- Turchin, P. (1995) Population regulation: old arguments and a new synthesis. *Population dynamics: new approaches and synthesis* (eds N. Cappuccino & P. Price), pp. 19–40. Academic Press, San Diego, CA.
- US Fish and Wildlife Service (2010) Waterfowl Population Status 2010.
- Viljugrein, H., Stenseth, N.C., Smith, G.W. & Steinbakk, G.H. (2005) Density dependence in North American ducks. *Ecology*, **86**, 245–254.

Received 7 August 2012; accepted 2 September 2012

Handling Editor: Robert Freckleton

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. R code to fit density-dependent models to ducks data.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be reorganized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.