When Ecological Individual Heterogeneity Models and Large Data Collide: An Importance Sampling Approach

Submitted to the Annals of Applied Statistics

WHEN ECOLOGICAL INDIVIDUAL HETEROGENEITY MODELS AND LARGE DATA COLLIDE: AN IMPORTANCE SAMPLING APPROACH

BY RUTH KING[1], BLANCA SARZO[1,2], VÍCTOR ELVIRA[1]

[1]*School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh, Edinburgh, UK.* *Ruth.King@ed.ac.uk; Victor.Elvira@ed.ac.uk*

[2]*Cavanilles Institute of Biodiversity and Evolutive Biology. Department of Microbiology and Ecology, University of Valencia, Valencia, Spain.* *Blanca.Sarzo@uv.es*

We consider the challenges that arise when fitting ecological individual heterogeneity models to "large" data sets. In particular, we focus on (continuous-valued) random effect models commonly used to describe individual heterogeneity present in ecological populations within the context of capture-recapture data, although the approach is more widely applicable to more general latent variable models. Within such models, the associated likelihood is expressible only as an analytically intractable integral. Common techniques for fitting such models to data include, for example, the use of numerical approximations for the integral, or a Bayesian data augmentation approach. However, as the size of the data set increases (i.e. the number of individuals increases), these computational tools may become computationally infeasible. We present an efficient Bayesian model-fitting approach, whereby we initially sample from the posterior distribution of a smaller subsample of the data, before correcting this sample to obtain estimates of the posterior distribution of the full dataset, using an importance sampling approach. We consider several practical issues, including the subsampling mechanism, computational efficiencies (including the ability to parallelise the algorithm) and combining subsampling estimates using multiple subsampled datasets. We initially demonstrate the feasibility (and accuracy) of the approach via simulated data before considering a challenging real dataset of approximately 30,000 guillemots, and, using the proposed algorithm, obtain posterior estimates of the model parameters in substantially reduced computational time compared to the standard Bayesian model-fitting approach.

1. Introduction. The use of continuous random effect models within statistical ecology applications is becoming increasingly common, particularly where individual and/or temporal heterogeneity can be substantial (Gimenez, Cam and Gaillard, 2017). However, the introduction of such random effects often leads to a likelihood that is expressible only in the form of an analytically intractable integral. We focus on the inclusion of individual heterogeneity within the Cormack-Jolly-Seber (CJS) model for capture-recapture data, where the survival probabilities are the primary parameters of interest, and on which we wish to specify individual heterogeneity.

Traditionally, many different approaches have been applied to obtain estimates of the model parameters when the likelihood is analytically intractable. For example, within a classical framework, numerical integration schemes have been applied such as Gaussian-Hermite quadrature for low dimensional problems (Coull and Agresti, 1999; Gimenez and Choquet, 2010); Laplace approximations (Herliansyah, King and King, 2022); Monte Carlo-type estimates for higher dimensional integrals (de Valpine, 2002, 2004); and the reduction to finite mixture models (Pledger, 2000; Pledger, Pollock and Norris, 2003). Alternatively, within a Bayesian framework data augmentation (or complete-data likelihood approach) have been applied (King and Brooks, 2008; Royle, 2008; King et al., 2016).

Large scale capture-recapture-type studies are becoming increasingly common where several thousands of individuals may be ringed/tagged each year. This is particularly true for bird

Keywords and phrases: Capture-recapture, Cormack-Jolly-Seber model, importance sampling, individual heterogeneity, intractable likelihood, random effects.

studies. For example, Hestbeck, Nichols and Malecki (1991) consider data relating to nearly 30,000 Canada Geese; while Francis and Saurola (2009) has data from approximately 20,000 Tawny Owls. However, many traditional model-fitting approaches for heterogeneity models do not scale when the dataset becomes "large" in terms of the number of individuals in the study; and/or when the likelihood increases in complexity due to the given model structure.

More generally within the wider statistical literature, for large dataset two approaches are often used: (i) divide-and-conquer that partitions the data into multiple datasets, analysing each independently and recombining; and (ii) using a subsample of the data to approximate the full posterior. See Bardenet, Doucet and Holmes (2017) for further discussion. Our approach is embedded within the latter idea, but further borrow ideas from the divide-and-conquer approach by combining multiple estimates of the posterior distribution from the different subsamples. In particular, we propose an algorithm that initially analyses a smaller subsample of the data using a Markov chain Monte Carlo (MCMC) sampler (Brooks et al., 2011), and then corrects the sampled parameter values such that we obtain an estimate of the posterior distribution of the full dataset of interest. The subsampled data are such that a Bayesian data augmentation approach can be applied within standard black-box software. The realisations of the Markov chain are then reweighted via an importance sampling algorithm to obtain an estimate of the posterior distribution for the full dataset (for a review of importance sampling, see for example, Tokdar and Kass, 2010; Elvira and Martino, 2021). Multiple sets of subsampled data can be taken and analysed in parallel, independently of each other, and subsequently combined to decrease the mean squared error of the corresponding estimated summary statistics of the posterior distribution. We note that unlike other works that compress the dataset introducing quantified errors, such as the coreset approach (Huggins, Campbell and Broderick, 2016), our proposed approach, in its base form, is asymptotically exact since it targets the posterior distribution of the unknown parameters given the full dataset.

In Section 2, we describe the CJS model and motivating case study relating to common guillemots (*Uria aalge*). In Section 3, we describe the model-fitting algorithm of subsampling the data, and subsequently correcting the output via importance sampling, before discussing associated practical implementation issues in Section 4. We apply the approach to a simulated dataset in Section 5 and the case study in Section 6, for which the traditional Bayesian data augmentation technique becomes computationally very challenging. We conclude with a discussion in Section 7.

2. Model description and case study.

We first introduce the CJS model before presenting the common guillemot case study.

2.1 *Cormack-Jolly-Seber model.*

We consider capture-recapture studies, where data are collected over a series of discrete capture occasions, $t = 1, \ldots, T$. At each occasion, all observed individuals are recorded. The first time an individual is observed, an associated unique identifier is recorded (e.g. natural skin/fur markings) or applied (e.g. a physical ring/tag attached). The capture-recapture data are the associated capture histories of each individual observed within the study, $i = 1, \ldots, I$, indicating whether the given individual was observed or not at each capture occasion. Mathematically, for $i = 1, \ldots, I$ and $t = 1, \ldots, T$, we let,

$$x_{it} = \begin{cases} 0 & \text{if individual } i \text{ is not observed at time } t; \\ 1 & \text{if individual } i \text{ is observed at time } t. \end{cases} \tag{1}$$

We let f_i and l_i denote the first and last time individual $i = 1, \ldots, I$ is observed in the study. The capture history for individual $i = 1, \ldots, I$ is denoted $\boldsymbol{x}_i = \{x_{it} : t = 1, \ldots, T\}$; with the full dataset, $\boldsymbol{x} = \{\boldsymbol{x}_i : i = 1, \ldots, I\}$. We consider only live recaptures but the approach is immediately extendable to include dead recoveries. The CJS model conditions on initial capture

and is defined in terms of (apparent) survival and recapture probabilities. Mathematically for $i = 1, \ldots, I$ we define:

$$\begin{aligned} \phi_{it} &= \mathbb{P}(\text{individual } i \text{ is alive at time } t+1 \mid \text{alive at time } t), && \text{for } t = 1, \ldots, T-1; \\ p_{it} &= \mathbb{P}(\text{individual } i \text{ is observed at time } t \mid \text{alive at time } t), && \text{for } t = 2, \ldots, T. \end{aligned}$$

We let $\boldsymbol{\phi} = \{\phi_{it} : i = 1, \ldots, I; t = 1, \ldots, T-1\}$ and $\boldsymbol{p} = \{p_{it} : i = 1, \ldots, I; t = 2, \ldots, T\}$. More generally, the state of "alive" corresponds to being available for capture, so that ϕ_{it} corresponds to *apparent* survival with emigration and survival confounded. For simplicity, we refer to ϕ_{it} as simply the survival probability. The corresponding likelihood can be expressed in the form,

$$f(\boldsymbol{x}|\boldsymbol{\phi}, \boldsymbol{p}) = \prod_{i=1}^{I} f(\boldsymbol{x}_i|\boldsymbol{\phi}, \boldsymbol{p}).$$

The term $f(\boldsymbol{x}_i|\boldsymbol{\phi}, \boldsymbol{p})$ denotes the probability of the capture history of individual i given by,

$$f(\boldsymbol{x}_i|\boldsymbol{\phi}, \boldsymbol{p}) = \left[\prod_{t=f_i}^{l_i-1} \phi_{it}\, p_{it+1}^{x_{it+1}} (1 - p_{it+1})^{(1-x_{it+1})} \right] \times \chi_{il_i}, \tag{2}$$

where $\prod_{t=f_i}^{f_i-1} \equiv 1$; and χ_{it} denotes the probability individual i is not observed after time t, given they are alive at t. This probability is most often described via the recursion,

$$\chi_{it} = 1 - \phi_{it}(1 - (1 - p_{it+1})\chi_{it+1}), \qquad \text{with} \quad \chi_{iT} = 1.$$

In practice, restrictions are typically specified on the dependence structure of the survival and recapture probabilities. For example, the parameters may be specified as common across individuals, (e.g. $p_{it} = p_t$ for all $i = 1, \ldots, I$); expressed as a function of external environmental covariates and/or observed individual characteristics (such as age, breeding status, condition etc.); or expressed as an (unobserved) random effect at either the temporal and/or individual level. We note that for the applications that we consider, we will assume that the capture probabilities are either constant or a function of the age of an individual at the given capture occasion; while the survival probabilities have an (unobserved) individual random effect component and for the case study are further dependent on the age of the individual and the capture occasion. See King et al., 2010; King, 2014; McCrea and Morgan, 2015; Seber and Schofield, 2019 for further details and a comprehensive review of capture-recapture-type models.

2.2 Individual random effect models. We consider the case where the survival probabilities are expressed in the form of an individual random effect:

$$\text{logit } \phi_{it} = \alpha + \epsilon_i, \qquad \text{where } \epsilon_i \sim N(0, \sigma^2),$$

for $t = 1, \ldots, T-1$ and $i = 1, \ldots, I$. The model parameters are denoted $\boldsymbol{\theta} = \{\alpha, \boldsymbol{p}, \sigma^2\}$, with the random effects, $\boldsymbol{\epsilon} = \{\epsilon_i : i = 1, \ldots, I\}$ integrated out in the observed data likelihood:

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{I} \int_{\epsilon_i} f(\boldsymbol{x}_i|\alpha, \boldsymbol{p}, \epsilon_i) f(\epsilon_i|\sigma^2) d\epsilon_i, \tag{3}$$

where $f(\boldsymbol{x}_i|\alpha, \boldsymbol{p}, \epsilon_i)$ is as in Equation (2); and $f(\epsilon_i|\sigma^2)$ denotes the random effect density, which in our case study, we assume to be Gaussian. The approach immediately generalises to random effects specified on other model parameters and mixed-effects type models, allowing for additional temporal or covariate effects and non-Gaussian random effect distributions.

2.3 Case study: guillemots. We consider capture-recapture data collected on a population of guillemots on the island of Stora Karlsö (Sweden). This is the largest guillemot colony in the Baltic Sea with a recorded breeding population of 15,700 pairs in 2014, corresponding to $\approx 2/3$ of the Baltic Sea population (Olsson and Hentati-Sundberg, 2017). We consider data from 2006-2016 (i.e. $T = 11$), with a total of $I = 28,930$ birds ringed. Recaptures were via resightings during the breeding season (May to July) using long-sighted telescopes. For further details see, for example, Sarzo et al. (2019). Previous work by Sarzo et al. (2021) suggested the presence of individual heterogeneity within the survival process, but due to the computational challenges was not investigated further.

3. Method. The observed data likelihood in Equation (3) is analytically intractable. To fit such models numerical integration techniques may be used to estimate the integral over the individual heterogeneity terms (e.g. Gimenez and Choquet, 2010; Coull and Agresti, 1999) or a Bayesian data augmentation approach applied (Royle, 2008; King et al., 2010). However, as the number or dimension of the random effects increases and/or the model increases in complexity, these approaches become computationally more challenging. We propose a Bayesian model-fitting approach that is scalable to large datasets and more complex models. The idea involves initially fitting the random effects model using a subsample of the data, and then correcting the sampled values using an associated importance weight. In this way, it is possible to approximate posterior summary statistics with consistent importance sampling estimators. We note that the focus of this paper is in relation to the application of the approach to individual heterogeneity capture-recapture models, but the approach described is more generally applicable to (continuous-valued) latent variable models.

The algorithm involves initially subsampling the data, and forming the posterior distribution of the model parameters, given the subsampled data, hereafter referred to as the *subposterior*; with the posterior distribution of the parameters given the full dataset is referred to as the *full posterior* for clarity. In our case, the subsampling is at the individual capture history level. The subsampled data are designed such that it is computationally feasible, using a standard Bayesian data augmentation technique, to obtain a set of sampled parameter values from the subposterior (see for example, Royle (2008); King et al. (2010)). We correct this set of sampled parameter values by taking into account the remaining (unsampled) data via importance sampling, i.e. by assigning each sampled value with an importance weight to estimate the full posterior distribution. The algorithm can be summarised as follows:

Step 1: Draw a (random) subsample of the data by sampling without replacement a set of individuals from the set of observed individuals.

Step 2: Using the set of subsampled individuals, implement a standard Bayesian MCMC data augmentation approach to obtain a set of sampled parameter values from the given subposterior.

Step 3: Apply an importance sampling algorithm to correct the sampled parameter values from the subposterior (by assigning an importance weight to each of them) to obtain a weighted sample from the full posterior.

Steps 1-3 provide a set of weighted sample parameter values that can be used to obtain Monte Carlo estimates of the associated summary statistics or moments of interest for the full posterior distribution. However, the steps can be repeated multiple times to obtain multiple estimates of this posterior distribution. Thus we advocate for an additional step to improve the estimation procedure:

Step 4: Repeat Steps 1-3 a total of M times and combine the posterior estimates of the parameters to obtain an improved estimate of the full posterior distribution.

Steps 1-3 can be undertaken in parallel across each of the subsamples $m = 1, \ldots, M$ as they are independent of each other. Thus, these steps are embarrassingly parallisable so that using multiple cores will significantly improve the computational efficiency of the algorithm. Although Step 4 is not strictly necessary, as each posterior obtained for a given subsample is an estimate of the posterior distribution of the parameters given the full data set, combining multiple posterior estimates improves the robustness and precision of the estimated full posterior distribution. We now describe in further detail each individual steps.

Step 1 - Subsampling the data. Recall that the dataset is denoted by $\boldsymbol{x} = \{\boldsymbol{x}_i : i = 1, \ldots, I\}$. We define a subsampled dataset by $\boldsymbol{x}^1 = \{\boldsymbol{x}_j : j \in \mathcal{J}\}$, where $\mathcal{J} \subset \{1, \ldots, I\}$ denotes the elements of the data that are contained in the given subsample. The associated individual random effects are denoted by $\boldsymbol{\epsilon}^1 = \{\epsilon_j : j \in \mathcal{J}\}$. Further, we let $\mathcal{J}^c = \{1, \ldots, I\} \backslash \mathcal{J}$ denote the complement of $\mathcal{J}$ corresponding to the set of non-subsampled individuals, with associated capture histories $\boldsymbol{x}^2$ (so that $\boldsymbol{x}^2 = \boldsymbol{x} \backslash \boldsymbol{x}^1$). We refer to $\boldsymbol{x}^2$ as the *remaining* data.

The simplest sampling scheme is to sample (without replacement) each individual with equal probability. However, this scheme can lead to poor precision due to large sampling variability. Alternatively, stratified sampling may be applied, for suitably defined strata (such as via cohort or other characteristics) to reduce this variability. We discuss different subsampling schemes in Section 4.2.

Step 2 - Sampling from the subposterior. For a given subsample of the data, $\boldsymbol{x}^1$, we form the corresponding subposterior of the model parameters given by,

$$\pi^1(\boldsymbol{\theta}|\boldsymbol{x}^1) \propto f(\boldsymbol{x}^1|\boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

The likelihood $f(\boldsymbol{x}^1|\boldsymbol{\theta})$ is analytically intractable. We implement a data augmentation scheme, with auxiliary variables $\boldsymbol{\epsilon}^1$, to obtain a set of sampled values from $\pi^1(\boldsymbol{\theta}|\boldsymbol{x}^1)$. Mathematically, we form the joint subposterior distribution of the parameters and auxiliary variables:

$$\pi^1(\boldsymbol{\theta}, \boldsymbol{\epsilon}^1|\boldsymbol{x}^1) \propto f(\boldsymbol{x}^1|\boldsymbol{\theta}, \boldsymbol{\epsilon}^1) f(\boldsymbol{\epsilon}^1|\boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

We assume that we can use a standard MCMC algorithm to obtain a set of K sampled values $\{\boldsymbol{\theta}_k, \boldsymbol{\epsilon}_k^1 : k = 1, \ldots, K\}$ following a suitable burn-in period (Gelman et al., 2014; Robert et al., 2018; van de Schoot et al., 2021). For example, black-box software, such as BUGS (Lunn et al., 2000), JAGS (Plummer, 2003), NIMBLE (de Valpine et al., 2017) or Stan (Carpenter et al., 2017), may be used to obtain posterior samples from $\pi^1(\boldsymbol{\theta}, \boldsymbol{\epsilon}^1|\boldsymbol{x}^1)$. For specific code for capture-recapture models, see for example, Gimenez et al. (2009); King et al. (2010); Kéry and Schaub (2011). Note that there is control over the size of the subsample, so we can ensure a feasible computational time for obtaining the set of subposterior sampled values. We discuss the practical considerations regarding subsample size in Section 4.1.

The sampled simulated parameter values from the MCMC algorithm can be used to approximate moments of the subposterior, $\pi^1(\boldsymbol{\theta}|\boldsymbol{x}^1)$, i.e. the posterior distribution of the model parameters given the subset of data $\boldsymbol{x}^1$. However, we are interested in the full posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{x})$. We would expect that the subposterior would be similar to the full posterior distribution but not identical. More precisely, we would expect the subposterior density to be wider compared to the full posterior distribution, due to a reduction of information in the subsample. In order to account for the full dataset, we apply a correction to the parameter values simulated from the subposterior using an importance sampling algorithm, to obtain estimates of the moments of the full posterior distribution using a set of weighted sample values from the subposterior distribution.

Step 3 - Importance sampling. We implement an importance sampling step on the sampled parameter and random effect values, $(\boldsymbol{\theta}_1, \boldsymbol{\epsilon}_1^1), \ldots, (\boldsymbol{\theta}_K, \boldsymbol{\epsilon}_K^1)$ where the corresponding proposal distribution is the subposterior, $\pi^1(\boldsymbol{\theta}, \boldsymbol{\epsilon}^1 | \boldsymbol{x}^1)$, with target distribution, $\pi(\boldsymbol{\theta}, \boldsymbol{\epsilon}^1 | \boldsymbol{x})$. For $k = 1, \ldots, K$, the corresponding importance sampling weight, $\{w_k\}_{k=1}^K$, is given by,

$$\begin{aligned} w_k &= \frac{\pi(\boldsymbol{\theta}_k, \boldsymbol{\epsilon}_k^1 | \boldsymbol{x})}{\pi^1(\boldsymbol{\theta}_k, \boldsymbol{\epsilon}_k^1 | \boldsymbol{x}^1)} \propto \frac{f(\boldsymbol{x} | \boldsymbol{\theta}_k, \boldsymbol{\epsilon}_k^1) p(\boldsymbol{\epsilon}_k^1 | \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k)}{f(\boldsymbol{x}^1 | \boldsymbol{\theta}_k, \boldsymbol{\epsilon}_k^1) p(\boldsymbol{\epsilon}_k^1 | \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k)} \\ &= f(\boldsymbol{x}^2 | \boldsymbol{\theta}_k), \end{aligned}$$

where $f(\boldsymbol{x}^2 | \boldsymbol{\theta}_k) = \prod_{i \in \mathcal{J}^c} f(\boldsymbol{x}_i | \boldsymbol{\theta}_k)$. In other words, the associated importance weight, w_k, is the observed data likelihood for $\boldsymbol{x}^2$ evaluated at $\boldsymbol{\theta}_k$. However, this weight is again analytically intractable. We extend the importance sampling approach and replace the likelihood expression with an estimate of this function denoted $\widehat{f}(\boldsymbol{x}^2 | \boldsymbol{\theta}_k)$, and estimate the weight as,

$$\widehat{w}_k \propto \widehat{f}(\boldsymbol{x}^2 | \boldsymbol{\theta}_k).$$

Tran et al. (2016) show that if the estimate is unbiased i.e. $\mathbb{E}(\widehat{f}(\boldsymbol{x}^2 | \boldsymbol{\theta}_k)) = f(\boldsymbol{x}^2 | \boldsymbol{\theta}_k)$ the corresponding importance sampling estimate converges almost surely to the distribution of interest (and termed this approach IS^2). The result is akin to that of Andrieu and Roberts (2009); Andrieu, Doucet and Holenstein (2010) for particle MCMC, where replacing the likelihood with an unbiased estimate within an MCMC algorithm leads to the desired posterior distribution.

We propose a Monte Carlo (MC) approach to obtain $\widehat{w}_k$. In the simplest case, for each $i \in \mathcal{J}^c$ we simulate N values of the random effects, $\boldsymbol{\epsilon}_i = \{\epsilon_i(1), \ldots, \epsilon_i(N)\}$ such that $\epsilon_i(j) \sim N(0, \sigma_k^2)$ for $j = 1, \ldots, N$. The unnormalised importance sampling weight is estimated as,

$$\widehat{w}_k^* = \prod_{i \in \mathcal{J}^c} \left[\frac{1}{N} \sum_{j=1}^{N} f(\boldsymbol{x}_i | \boldsymbol{\theta}_k, \epsilon_i(j)) \right], \tag{4}$$

where $f(\boldsymbol{x}_i | \boldsymbol{\theta}_k, \epsilon_i(j))$ denotes the closed form conditional likelihood contribution for capture history $\boldsymbol{x}_i$, given the model parameters, $\boldsymbol{\theta}_k$, and associated individual random effect, $\epsilon_i(j)$.

We subsequently estimate the normalised sampling weights $\{\widehat{w}_k : k = 1, \ldots, K\}$ using,

$$\widehat{w}_k = \frac{\widehat{w}_k^*}{\sum_{j=1}^K \widehat{w}_j^*}, \tag{5}$$

through the self-normalized importance sampling (SNIS) estimator (Elvira and Martino, 2021).

Using the normalised weights $\widehat{w}_k$ converges almost surely to the quantities of interest (such as the moments of the full posterior distribution) as both K and N go to infinity. First, when K goes to infinity, any error associated with sampled values not being from the stationary distribution (i.e. burn-in phase) vanishes. Second, when N goes to infinity, $\widehat{w}_k^*$ converges almost surely to the marginal observed data likelihood analogous to Equation (3) for the subsampled dataset (i.e. $\widehat{w}_k^*$ is a consistent estimator). This is apparent from the fact that $\widehat{w}_k^*$ in Equation (4) is a product of (a finite number of) terms, each of them converging to the true quantity almost surely due to the strong law of large numbers. Thus, the product also converges almost surely to the true quantity (this latter argument is often used in proof of convergence of the SNIS estimator, see for instance (Owen, 2013, Theorem 9.2)).

Subsequently, the importance weights, $\widehat{w}_k$, for $k = 1 \ldots, K$, can be used to obtain summary statistics/distributions of interest. For example, to obtain the posterior mean of some

parameter, ψ, say, we use,

$$\mathbb{E}_\pi(\psi) = \sum_{k=1}^{K} \widehat{w}_k \, \psi_k,$$

where ψ_k denotes the sampled value of the parameter ψ for iteration $k = 1, \ldots, K$ from the subposterior distribution. Further, a sampling importance resampling (SIR) approach can be used to obtain a set of parameter values which can be used, for example, to obtain posterior density estimates and/or 95% credible intervals (CIs).

The MC estimate of the likelihood may become computationally expensive as I, K and Nincrease. Further, the MC estimates are required for each posterior subsample, $m = 1, \ldots, M$, although these computations are parallelisable across subsamples, $m = 1, \ldots, M$ and individuals $i \in \mathcal{J}^c$. We discuss further computational considerations in Section 4 and suggest approaches to decrease the computational component, including a stratified MC estimate; two-step algorithm and alternative approximate (biased but consistent) weight estimates.

Step 4 - Combined posterior estimate. Steps 1-3 are embarrassingly parallelisable over $m = 1, \ldots, M$; each subsampled dataset is independently drawn and separate MCMC algorithms applied. (Step 3 is also parallelisable over MC samples). This means that for no extra computational cost we can obtain multiple estimates of the full posterior distribution (at least up to the number of processors available). These posterior distributions can be combined to obtain a more reliable and robust estimate of the full posterior. Thus, this final step is similar to the divide-and-conquer concept of combining multiple estimates. However of substantial note is that within the standard divide-and-conquer approach, there is a fixed "dimension" in that the number of subsamples is determined by the number of data points within each subsample (and vice versa), as the data are partitioned. However, within our approach, we do not partition the data into the different subsamples to be considered but instead each subsample is drawn independently from the full dataset (we discuss how this may be done in Section 4.1). This means that for our approach we are not limited in the number of subsamples that may be drawn, and the subsamples are, in general, not independent, since the same individuals (i.e. data points) may be included within multiple subsamples.

The combined estimate of the posterior distribution of the full data over all the different subsampled datasets is defined to be a (weighted) average of the corresponding M subposterior distributions. For example, to obtain the posterior mean of the parameter, ψ we use the weighted average,

$$\mathbb{E}_\pi(\psi) = \sum_{m=1}^{M} z_m \mathbb{E}_{\pi(m)}(\psi), \tag{6}$$

where $\mathbb{E}_{\pi(m)}(\psi)$ denotes the full posterior mean of ψ estimated using subsampled data $m = 1, \ldots, M$; and $z_1, \ldots, z_M$ are corresponding weights such that $\sum_{m=1}^{M} z_m = 1$ and $0 \leq z_m \leq 1$. We discuss different possible weights in Section 4.4.

4. Practical considerations.

We now discuss some practical considerations relating to the proposed algorithm.

4.1 Subsample size.

A decision within the algorithm relates to the proportion of the data to subsample (i.e. $|\boldsymbol{x}^1|$). The larger the subsample, the closer the subposterior should be to the full posterior, so that the importance sampling algorithm increases in efficiency; however also the larger the computational cost in sampling from the subposterior. This computational cost

is in terms of (1) time per each iteration (due to the number of auxiliary variables and cost to evaluate the likelihood function), and (2) length of MCMC simulations required since poorer mixing is often observed due to increased correlation between the parameters (notably for the random effects, $\boldsymbol{\epsilon}^1$, and σ^2). Alternatively, smaller subsamples provide subposteriors for which it is (relatively) computationally fast to obtain a sample from but where the following importance sampling algorithm may suffer from increased particle depletion due to differences between the subposterior and full posterior (see, Elvira and Martino, 2021 for further discussion). Further, in this case, there is an increased computational cost in the calculation of the importance sampling weight, as this is a function of the remaining data, though this is minimised when using an alternative (biased) weight calculation or deterministic approximation (see Section 4.3, considerations (iii) and (iv)) which each reduce consideration to only unique capture histories. In practice, the proportion of the data to sample will be dependent on the computational resources available, with the general advice to take as large a sample as possible that is computationally reasonable. For both the simulation and case studies, a subsample size of 20% appeared to be a good choice since (a) the subposterior is similar to the full posterior and (b) a relatively low computational cost can be achieved.

4.2 Sampling schemes. We focus on stochastic schemes to subsample datasets. Ideally, the subposterior should be as similar as possible to the full posterior, to maximise the efficiency of the importance sampling approach. Random subsampling, selecting each capture history with equal probability, ignores any structure within the data, and thus typically leads to relatively non-similar distributions and poor performance (this is easily seen via simulation). Thus we consider a stratified sampling approach, where we initially stratify the individual capture histories, and then perform proportional random sampling within each strata. For instance, consider the simplified scenario where, for the I capture histories, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_I$, we stratify the histories into $I/10$ different strata, with 10 individuals contained within each strata (in practice the different strata will not necessarily be of the same size). Then, if the subsampled dataset is of size $|\mathcal{J}| = I/5$, we may either sample 2 capture histories from each strata (for fixed strata sampling); or include an additional stochastic step to determine the number of individuals to sample from each strata, using a multinomial distribution with probabilities proportional to the number of individuals within each strata before randomly sampling within each strata (for stochastic strata sampling).

Such a stratified subsampling approach is designed to replicate data structures in the subsample that are present within the full dataset. For example, strata may be defined via observable covariate information (such as age/gender); cohort (i.e. year of first capture); unique capture histories; or capture histories with defined characteristics, such as the number of times observed alive; or initial and final capture times. In practice, it may also be desirable to pool several strata when frequency sizes are small. An 'optimal' scheme will typically depend on the dependence of the model parameters, as for standard sampling techniques (Hankin, Mohr and Newman, 2019), and model being fitted to the data. For example, if the model parameters are assumed to be age dependent, then this suggests that including age within the subsampling stratification may be useful.

4.3 Estimation of importance sampling weights. We initially consider a MC approach for the estimation of the importance sampling weight, w_k, since these provide an unbiased estimate of the importance sampling weights, and hence the associated theoretical guarantees discussed in Section 3, before considering additional efficient, but biased, estimation approaches. We discuss in further detail computational efficiency relating to: (i) stratified MC; (ii) 2-step MC; (iii) repeated histories; and (iv) deterministic approximations.

(i) Stratified MC approach. For increased computational efficiency, we apply a stratified MC approach (Owen, 2013, Chapter 8). In particular, we partition $\mathbb{R}$ into N strata, separated by the $N-1$ quantiles of the given $N(0,\sigma^2)$ distribution, and simulate a single particle in each strata. This leads to strata of varying length but, by definition, have equal probability. Consequently, the estimate for the unnormalised weight is as in Equation (4), due to the equal probability of each stratum, but reduces the associated variability of the estimate.

(ii) Two-step MC approach. Many of the sampled values from the subposterior will typically have a negligible importance sampling weight (leading to "particle depletion" when using a subsequent resampling approach), with this issue increasing as the number of parameters increases. The proportion of sampled values with a non-negligible weight will depend on how close the subposterior is to the full posterior distribution. To improve computational efficiency we thus distribute the computational effort to focus on the sampled values that dominate the Monte Carlo estimate using a two-step approach. In particular, we initially undertake a fast "screening-type" process to identify the dominant sampled values (i.e. those with potential non-negligible weight). Thus in the first step, we use a coarse (stratified) MC approach, using a relatively small number of MC particles, to obtain an estimate of the (unnormalised) weight. In the second step, we obtain a refined, more accurate, estimate for those sample parameters values identified as dominant in the first step using a substantially larger number of MC particles. For example, this may be defined to be the top-ranked sampled values, such as those values corresponding to the largest estimated 10-20% weights, or with a non-negligible (normalised) weight > 0.001. In practice, obtaining a fast and reliable "ball-park" value in the first step is generally straightforward. For example, for the real data application, using as few as $N = 25$ MC values where we simulate a single value within each 4% quantile range of the random effect distribution (or even use the mid-point of the quantile ranges) led to stable estimates in terms of the ranking of the sampled values to be retained for obtaining a more accurate MC estimate of the weight (the top 10% were retained for Step 2). We note that, in general, the approach works when the variability within the MC estimates for given sampled parameter values is smaller than the variability of the weights across parameter values.

(iii) Repeated histories. The weight in Equation (4) is a product over the number of individuals in $\boldsymbol{x}^2$, and thus scales linearly with the number of histories. However many individuals will have the same capture history, and hence marginal likelihood contribution. In other words, for individuals i and j that have the same history, $f(\boldsymbol{x}_i|\boldsymbol{\theta}) = f(\boldsymbol{x}_j|\boldsymbol{\theta})$. In the MC scheme described above we obtain an estimate of the marginal likelihood for each individual, independently, leading to an unbiased estimate of the marginal likelihood, $f(\boldsymbol{x}^2|\boldsymbol{\theta})$. However, we can consider an alternative (biased) estimate of the weight that is computationally faster by only estimating the marginal likelihood for *unique* histories.

Let $\mathcal{J}_U^c$ denote the set of unique capture histories in $\boldsymbol{x}^2$ and $n(\boldsymbol{\omega})$ the number of individuals in $\boldsymbol{x}^2$ with capture history $\boldsymbol{\omega} \in \mathcal{J}_U^c$. For each history $\boldsymbol{\omega} \in \mathcal{J}_U^c$ and sampled parameter value $\boldsymbol{\theta}_k$, for $k = 1, \ldots, K$, we simulate N values of the random effects $\boldsymbol{\epsilon}_{\boldsymbol{\omega}} = \{\epsilon_{\boldsymbol{\omega}}(1), \ldots, \epsilon_{\boldsymbol{\omega}}(N)\}$, such that $\epsilon^2_{\boldsymbol{\omega}}(i) \sim N(0, \sigma_k^2)$. We estimate the importance sampling weight using,

$$\widehat{w}_k^* = \prod_{\boldsymbol{\omega} \in \mathcal{J}_U^c} \left[\frac{1}{N} \sum_{j=1}^{N} f(\boldsymbol{\omega}|\boldsymbol{\theta}_k, \epsilon_{\boldsymbol{\omega}}(j)) \right]^{n(\boldsymbol{\omega})},$$

where $f(\boldsymbol{\omega}|\boldsymbol{\theta}_k, \epsilon_{\boldsymbol{\omega}}(j))$ denotes the conditional likelihood contribution for capture history $\boldsymbol{\omega}$, given $\boldsymbol{\theta}_k$ and $\epsilon_{\boldsymbol{\omega}}(j)$. This estimate (though biased) is a consistent estimator of the unnormalised weight, since it converges almost surely as N goes to infinity, as discussed in Section

3 (Step 3). For improved efficiency a stratified MC approach can be used as described in (i) above.

The number of unique capture histories is, in general, significantly smaller than the number of individuals observed, and hence scales significantly slower as the number of individuals increases. More precisely, the maximum number of unique capture histories is 2^T and hence limited by the number of capture occasions, and in most cases not all histories will be observed within the dataset. Thus, this estimate is significantly faster computationally in general, at the expense of the property of unbiasedness for finite sample size. Further, we note that given the substantially reduced required number of MC estimates at the capture history level, we can use a significantly larger value for N. In practice, for the case study in Section 6, where the number of individuals observed with the same capture history is of the order of 1000s, we use a hybrid approach. In this approach we essentially specify the data using multiple replicates of the same capture history, such that the number of individuals with each of these (repeated) capture histories is limited to be at most some specified maximum value (for the case study a value of 200 was used). Within the MC estimate of the unnormalised weight we then consider each of these histories as unique. This hybrid approach led to improved convergence of the MC estimate of the weight.

(iv) Deterministic approximations. Similarly to the previous repeated histories approach, if we are willing to consider an accurate (but biased) estimate of the importance sampling weight we can consider alternative (deterministic) estimation schemes. For example, quadrature and Laplace approximations have both been applied to estimate the marginal likelihood in the presence of individual random effects for capture-recapture models (Gimenez and Choquet, 2010; Herliansyah, King and King, 2022). In particular, we consider Gauss-Hermite quadrature (GHQ) to estimate the integral within the importance sampling weight, as this is known to be an accurate and computationally efficient estimate for low dimension integrals with a Gaussian random effect (Butler and Moffit, 1982; Hedeker and Gibbons, 1994; Liu and Pierce, 1994; Elvira, Martino and Closas, 2020). Since GHQ is a deterministic algorithm, this means that the estimate of the weight can again be calculated at the unique capture history level (as the integral of the associated likelihood contribution is identical for all individuals with the same capture history). Further, an analogous two-step algorithm can again be applied to GHQ as for the MC approach (as described in (ii) above), for additional computational efficiency, if required. However, for the applications that we consider, using 20 nodes appeared to be provide fast and accurate estimates so that a two-step approach was not required.

4.4 Combining multiple importance sampling estimates.

The importance sampling algorithm is naively parallelisable for the subsampled datasets. Thus, given sufficient computer cores, we can obtain multiple posterior estimates for no additional computational time. Further, the estimates across different subsampled data can be combined to obtain an improved estimate of the full posterior, as in Equation (6). The function is a linear combination of the posterior estimates for each subsampled dataset, for any set of positive weights that sum to unity. For example, in the simplest case, $z_m = \frac{1}{M}$, for $m = 1, \ldots, M$. However, this implicitly assumes that all the subsampled posterior estimates are equally informative, which in general will not be the case. To address this, we may consider, for example, setting z_m as proportional to the inverse of the variance of the weights (Douc et al., 2007; Luengo et al., 2018), effective sample size or unique number of non-negligible weights (Nguyen et al., 2014). The ideas extend immediately to using the analogous SIR argument for obtaining additional posterior quantities of interest.

5. Simulated data. We conduct a simulation with $I = 10,450$ individuals and $T = 11$ capture occasions. We consider a constant capture probability and specify the survival probabilities to be a function of individual heterogeneity:

$$p_{it+1} = p; \quad \text{and} \quad \text{logit}\,(\phi_{it}) = \alpha + \epsilon_i,$$

for $i = 1, \ldots, I$; $t = 1, \ldots, T-1$, where $\epsilon_i \sim N(0, \sigma^2)$. We set $p = 0.13$, $\alpha = 0.62$, and $\sigma = 0.5$, corresponding to a realistic capture probability for many species, a median survival probability of 0.65 with lower and upper 2.5% quantiles (0.41, 0.83). This is the same length of study as for the case study but for a reduced number of individuals and simpler model, so that we are able to analyse the full dataset using a standard Bayesian data augmentation approach for comparison.

We used a stratified sampling approach, with strata defined to be the set of individuals released at time $t = 1, \ldots, T-1$ and observed for the final time at occasion $\tau = t, t+1, \ldots, T$ (a total of 54 strata). The number of individuals sampled from each strata was set equal to its observed proportion (rounded up to an integer). Within each strata, we uniformly selected the individual histories without replacement. To determine subsample size, we implemented a pilot-tuning stage using subsample sizes between 5%-30%. Sample sizes $\geq 20\%$ had consistently similar subposterior distributions; whereas the subposterior distribution of subsamples $\leq 10\%$, displayed much greater variability and level of particle depletion within the importance sampling step. Thus, we used a subsample size of 20% (2,090 individuals) as a compromise between consistently similar subposterior distributions and reasonable computational cost. We simulated $M = 100$ subsampled datasets. Finally, we specified the prior distributions: $p \sim \text{U}(0, 1)$, $\alpha \sim \text{N}(0, 10)$, and $\sigma \sim \text{U}(0, 10)$.

For each subsampled dataset, we fitted the model using NIMBLE, specifying three independent MCMC chains, running each for 15,000 iterations, following a burn-in of 5,000 iterations. The simulations took approximately 5 minutes on an IntelXeon CPU E5-2683 v4 at 2.10 GHz and 64-bit Scientific Linux Mint 18.2 Sonya. For each subposterior, we thinned the sampled parameter values by 15 (i.e. retaining 1000 sampled values) and calculated their associated IS weights using (i) a stratified MC approach with $N = 100$ particles (this took approximately 4 minutes); and (ii) a GHQ approach using 20 nodes (approximately 2 minutes). Essentially identical results were obtained from both the MC and GHQ approaches (with only negligible differences). Across the subsamples, the mean number of particles with non-negligible weight (> 0.001) was 203, and ranged from 78-260. We used an SIR approach to obtain the associated 95% symmetric credible intervals (CIs). For comparison, we also fitted the model to the full database directly using a Bayesian data augmentation approach. Due to the increased level of auto-correlation of the parameters (and posterior correlation between the random effect terms and associated variance), the simulations were run for 1 million iterations, with the first 100,000 discarded as burn-in (approximately 14 hours to run). Table 1 provides a summary comparison of the computational times for the different approaches.

The subposterior distributions where over-dispersed compared to the full posterior distribution, as expected. This can be seen in Table 2 where we provide summary statistics of the lower and upper 2.5% quantiles for the subposterior compared to (corrected) full posteriors across subsamples. The corresponding results for each (corrected) posterior for each subsampled dataset and associated estimate obtained from directly fitting the model to the full data are provided in Figure 1. These posterior estimates are generally very similar to those obtained using the full data. (The corresponding subposteriors are provided in Web Appendix A in the Supplementary Material for each subsample). We combine the corrected posteriors across each subsample into a single estimate of the posterior. For simplicity, we assume an equal weight over subsamples, although using alternative weights gave essentially identical estimates. Table 3 provides a summary of the associated posterior means, standard deviations

TABLE 1

Computational times (to nearest minute) for fitting the individual heterogeneity model to the simulated data and case study. For the simulated data and subsampling approach, the MCMC was run for a total of 60,000 iterations (with 15,000 sampled values discarded as burn-in), with 1000 (thinned) sampled parameter values used in the importance sampling step. For the MC approach, $N = 100$ particles are used; and for GHQ approach 20 nodes. For comparison a further MCMC simulation was run using the standard Bayesian data augmentation approach on the full dataset using 1 million MCMC iterations (to ensure convergence). For the case study 40,000 MCMC iterations were run (5,000 sampled values were discarded as burn-in). For the importance sampling step, a two-step approach was applied for the MC approach; while for the GHQ approach a single step was used. For the MC approach, $N = 25$ MC particles were used in Step 1 for 5000 (thinned) sampled parameter values and $N = 250$ particles in Step 2 retaining the top 500 ranked particles. For the GHQ approach 20 nodes was used. The MCMC algorithm was implemented in NIMBLE

.

(*) Computational times are stated per subsample.

	MCMC iterations	Importance sampling weights	Total
Simulated data: full data approach	14 hours	–	14 hours
Simulated data: subsampling using MC approach (*)	5 minutes	4 minutes	9 minutes
Simulated data: subsampling using GHQ approach (*)	5 minutes	2 minutes	7 minutes
Case study: subsampling using 2-step MC approach (*)	21 minutes	29 minutes	50 minutes
Case study: subsampling using GHQ approach (*)	21 minutes	9 minutes	30 minutes

and 95% CIs. The combined estimate of the model parameters are very similar to those obtained from directly fitting the model to the full data; for all quantities displayed in Table 3, the estimates all differ by less than 1%. However, the estimates obtained by our proposed approach are at a substantially reduced computationally cost.

TABLE 2

Simulation study: Mean lower and upper 2.5% quantiles for the model parameters across the 100 subsamples for the subposterior distribution and full posterior distribution.

	Subposterior distributions		Full posterior distribution	
	Mean lower 2.5% quantile	Mean upper 2.5% quantile	Mean lower 2.5% quantile	Mean upper 2.5% quantile
α	0.1740	0.8643	0.3935	0.7318
p	0.1134	0.1658	0.1248	0.1499
σ	0.2159	1.1677	0.3272	0.8565

TABLE 3

Simulation study: Posterior mean, standard deviation and 95% symmetric CIs for the model parameters using the proposed importance sampling approach and combined across the 100 subsampled datasets and using a Bayesian data augmentation approach for the full simulated dataset.

	Combined approach			Full simulated dataset		
	Mean	Sd	95% CI	Mean	Sd	95% CI
α	0.5670	0.0887	[0.3915, 0.7361]	0.5679	0.0879	[0.3934, 0.7384]
p	0.1369	0.0065	[0.1245, 0.1500]	0.1369	0.0065	[0.1246, 0.1499]
σ	0.6128	0.1379	[0.3179, 0.8613]	0.6118	0.1380	[0.3149, 0.8621]

Following this "proof-of-concept" simulated dataset we apply the approach to the more challenging case study, where the model is more complex in terms of age and temporal dependencies in addition to the individual heterogeneity on the survival component.

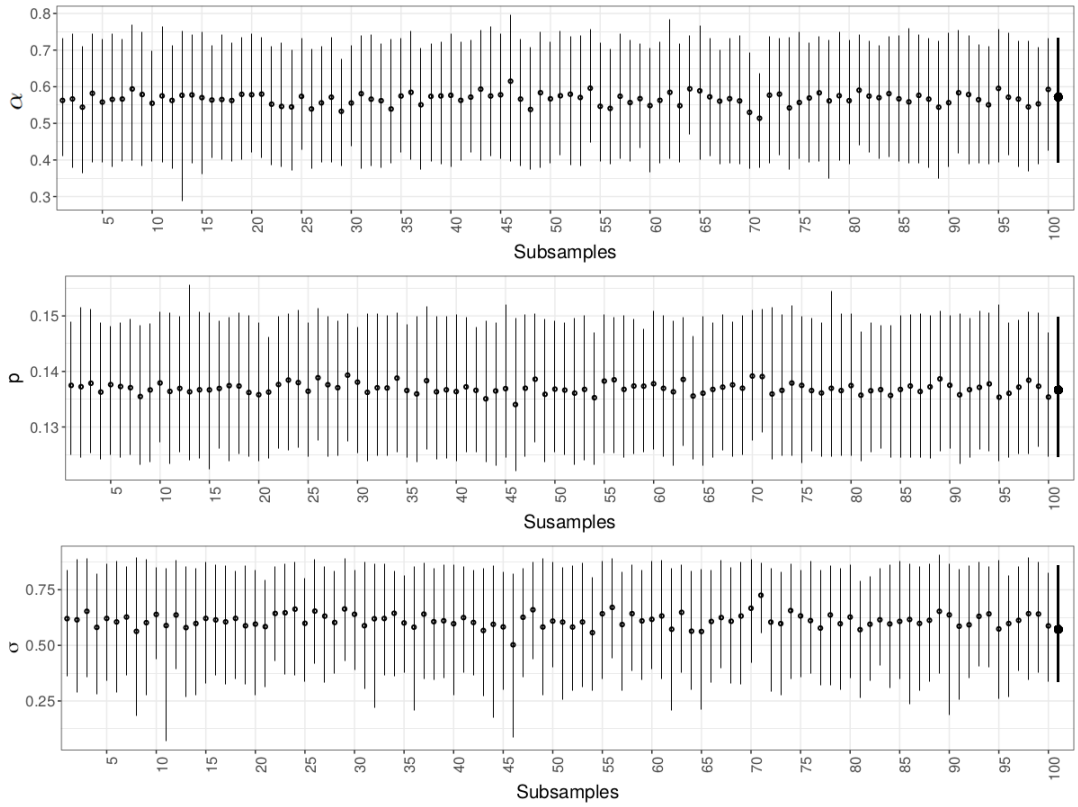

Fig 1: Simulation study: Corrected posterior means and 95% symmetric CIs for model parameters (α, p, and σ) and for the full simulated database (thick solid error bar).

6. Case study: guillemots. We consider the case study described in Section 2.3. Given the number of ringed birds (28,930), the inclusion of individual heterogeneity on the survival probabilities is computationally challenging using the standard Bayesian data augmentation approach, even for relatively simple parameter dependence models. Incorporating additional biologically sensible parameter dependencies leads to added computational challenges. Motivated by Sarzo et al. (2021), and incorporating the known life cycle of guillemots, we consider an age-dependent model, where the survival and recapture probabilities have age structures: 1, 2, 3 and 4+, and 2, 3, 4, 5+, respectively. The survival probabilities are assumed to have additional temporal effects to reflect (unobserved) environmental heterogeneity over time, such as food availability, environmental conditions, etc. Mathematically, we let $a(i,t)$ denote the age of individual $i = 1,\dots,I$ at time $t = 1,\dots,T-1$, such that the parameters are of the form:

$$p_{it+1} = p_{a(i,t+1)}; \qquad \text{and} \qquad \text{logit}\,\phi_{it} = \alpha_{a(i,t)} + \beta_t + \epsilon_i, \qquad \text{where } \epsilon_i \sim N(0,\sigma^2),$$

for $t = 1,\dots,T-1$ and $i = 1,\dots,I$. We specify vague prior distributions. For the temporal survival effects, we use a hierarchical distribution, such that $\beta_t \sim \text{N}(\mu,\kappa^2)$, where $\mu \sim \text{N}(0,10)$ and $\kappa \sim \text{U}(0,10)$. For the age effect survival terms we set $\alpha_1 = 0$ (for identifiability) and $\alpha_a \sim \text{N}(0,4)$, for $a = 2,\dots,4+$. For the resighting probabilities, we specify $p_a \sim \text{U}(0,1)$, for $a = 2,\dots,5+$. Finally for the individual effects variance term, we set $\sigma \sim \text{U}(0,2)$.

We apply the same subsampling scheme as in Section 5, stratifying the histories based on initial and final capture times (54 possible strata). We subsampled $M = 100$ datasets of

sample size corresponding to 20% of the database (i.e. 5,789 individuals). For each dataset, the model was fitted via NIMBLE, using 35,000 MCMC iterations, following a burn-in of 5,000 iterations (consideration of selected subsamples suggested that this was sufficient for convergence). Each MCMC simulation took approximately 21 minutes on an IntelXeon CPU E5-2683 v4 at 2.10 GHz and 64-bit Scientific Linux Mint 18.2 Sonya. We again considered both a stratified MC approach and a GHQ approach to estimate the importance sampling weights, using 5000 sampled values from the MCMC sampled values (i.e. we thinned the sampled values by 70). For the MC approach we implemented a two-step approach. For the first (coarser) step, we used $N = 25$ MC particles, retaining the top 10% (i.e. 500) sampled values; and for the second (finer resolution) step, we used $N = 250$ MC particles. To assess for convergence of the two-step MC approach we repeated the analysis multiple times for a number of the subsampled datasets (i.e. estimated the importance sampling weights for given subsampled datasets). In all cases, we consistently retained all the particles with non-negligible weight following the first step, and obtained consistent weights for the second step. For the GHQ approach as we were able to obtain consistent estimates using only 20 nodes, withing a single-step approach. Increasing the number of nodes led to essentially identical results. Table 1 provides a summary of the computational times. The mean number of particles with a minimum weight of 0.0001 was 42 (range 5-97). The increased level of particle depletion (compared to the simulated data) is unsurprising given the increased dimension of the parameter space (18 parameters).

Table 4 provides the (corrected) full posterior mean and standard deviation (SD) for each parameter combined over the subsamples; while Figures 2 to 5 provide the estimated (corrected) full posterior mean and 95% CIs for each subsample, and combined across all subsamples. There is some variability of the posterior distribution per subsample (though generally overlapping), which is unexpected given the reduced effective sample sizes. However, we are able to obtain an estimate of the posterior by combining the subsample estimates, immediately increasing the sample size and providing increased accuracy. To investigate the robustness of this approach, we randomly selected 25 and 50 samples (without replacement) of the estimated posterior distributions obtained from the full set of subsamples and calculated the associated posterior mean and SD. We repeated this a total of 100 times and calculated the corresponding root mean square error of the given posterior summary statistics, compared to the estimate obtained using all subsamples. The results are given in Table 4, which suggests that the estimates of the posterior summary statistics are fairly robust when combining across subsamples, even when some individual subsamples lead to low effective sample sizes. As expected there is smaller variability when using 50 subsamples compared to 25.

From Figure 5, there appears to be a substantial random effect variance component (on the logit scale), with the posterior mean of σ equal to 0.96, with 95% CI $[0.65, 1.24]$. This suggests a reasonable amount of unobserved individual heterogeneity present, unexplained by the individual age effects. This may, for example, be representative of inherent differences in individual quality or condition. We compare the posterior estimates of the parameters with the model omitting the individual heterogeneity component in Web Appendix B in the Supplementary Material. We note that the inclusion of the individual heterogeneity leads to similar parameter estimates for the survival temporal effects and capture probabilities, but with substantially larger credible intervals for the survival probabilities across ages.

7. Discussion. Advances in computational resources and readily available computer packages have permitted the fitting of more complex models to real data across the breadth of the scientific community. However, computational limitations remain for many real applications, particularly as increasing amounts of data become available. In such circumstances, applying standard computational algorithms may become prohibitive. In this paper, we were

TABLE 4

Case study: Posterior mean and standard deviation (SD) of the model parameters for the combined full posterior distribution (using 100 subsampled datasets from the full dataset) and associated root mean square error (RMSE) for the posterior mean and SD using 50 and 25 randomly sampled posterior distribution (without replacement), repeated 100 times.

	100 subsamples		50 subsamples		25 subsamples	
	Posterior		RMSE		RMSE	
	Mean	SD	Mean	SD	Mean	SD
β_1	0.842	0.156	0.010	0.008	0.017	0.012
β_2	0.571	0.161	0.009	0.008	0.016	0.011
β_3	0.177	0.158	0.013	0.010	0.019	0.015
β_4	-0.788	0.103	0.005	0.006	0.010	0.009
β_5	-0.242	0.101	0.009	0.005	0.011	0.007
β_6	-0.729	0.105	0.006	0.005	0.011	0.008
β_7	-0.518	0.106	0.006	0.007	0.009	0.009
β_8	-0.097	0.107	0.006	0.006	0.011	0.008
β_9	-0.081	0.104	0.009	0.005	0.012	0.009
β_{10}	-0.303	0.147	0.012	0.007	0.019	0.012
α_2	3.871	0.930	0.089	0.133	0.143	0.149
α_3	0.472	0.176	0.011	0.010	0.020	0.013
α_{4+}	-0.248	0.223	0.018	0.015	0.025	0.026
p_2	0.072	0.003	0.015	0.012	0.019	0.015
p_3	0.251	0.009	0.023	0.019	0.028	0.025
p_4	0.330	0.014	0.033	0.024	0.037	0.030
p_{5+}	0.429	0.015	0.029	0.025	0.038	0.029
σ	0.957	0.130	0.017	0.021	0.018	0.022
μ	-0.121	0.216	0.004	0.009	0.018	0.017
κ	0.638	0.197	0.013	0.014	0.019	0.022

motivated by fitting (continuous-valued) individual heterogeneity models to a large capture-recapture dataset, for which using the standard Bayesian data augmentation approach is impractical, although the approach is more generally applicable to other latent variable models.

Previous approaches for dealing with large datasets leading to computational challenges typically consider either a divide-and-conquer approach or via consideration of only a suitable subsample of the data. We proposed a new efficient approach that essentially borrows concepts from both of the previous approaches by considering how to obtain "good" subsamples of the data, and subsequently combining the estimated posterior distribution obtained from each subsample to obtain an improved estimate of the full posterior distribution of interest. For each subsample, the corresponding subposterior distribution is corrected via importance sampling to obtain an estimate of the full posterior distribution. The number of subsamples that may be drawn (of a given size) is not limited, as within the standard divide-and-conquer approach. The approach is embarrassingly parallelisable in two aspects: in terms of the multiple subsamples, and calculating the importance sampling (unnormalised) weights of each subsample. Thus, the proposed mechanism is particularly well suited for architectures that allow a high level of parallelisation (e.g., GPUs). Further, the algorithm can be easily implemented requiring essentially a black-box MCMC sampler (such as in JAGS/NIMBLE) and one additional bespoke function corresponding to the numerical estimate of the probability of a given capture history (expressed as an analytically intractable integral). For an efficient application of the algorithm, consistent (but biased) numerical estimate of the integral are considered, at the expense of the associated theoretical guarantees associated with unbiased estimates. In particular, we consider both a deterministic Gaussian-Hermite quadrature approach and a stochastic (stratified) Monte Carlo approach at the unique capture history level. Reliable and consistent estimates were obtained using both approaches, although, as expected, Gaussian-Hermite quadrature was computationally more efficient.

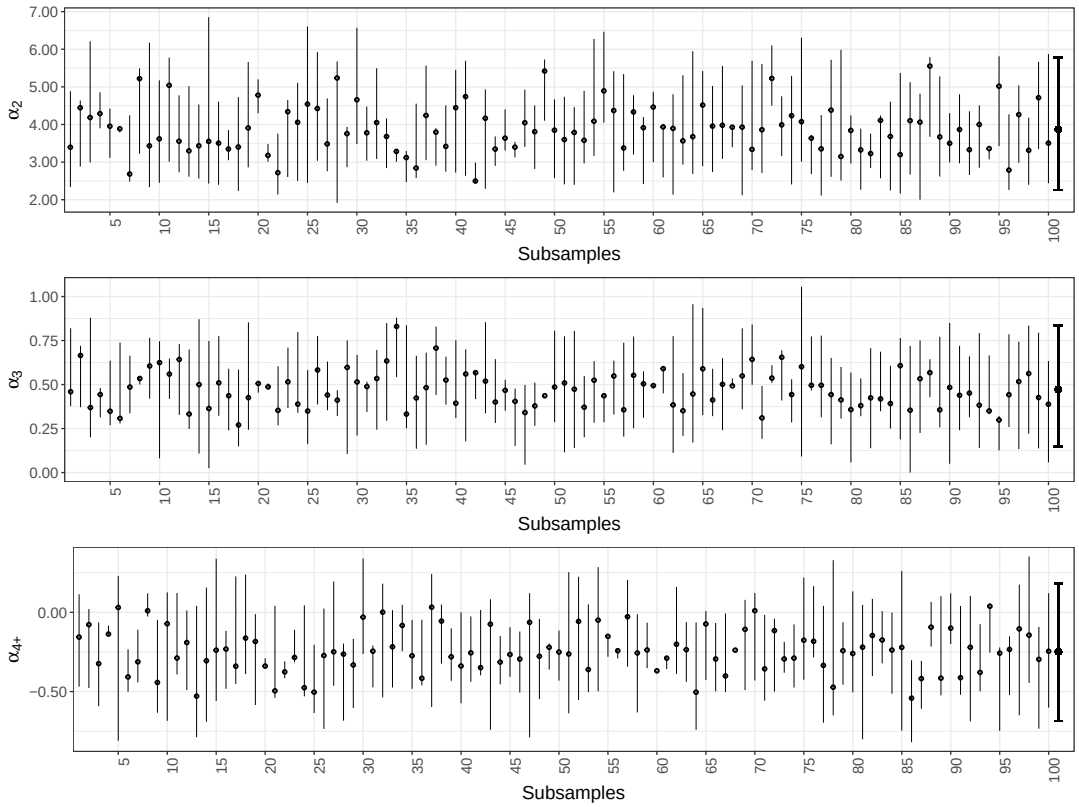

Fig 2: Case study: Corrected posterior means and 95% symmetric CIs for α_a parameters for each subsample and combined across all subsamples (thick solid error bar) by age $a = 2, \ldots, 4+$.

For the guillemot case study, we consider an individual heterogeneity effect on the survival probability, for which using the standard Bayesian data augmentation approach becomes infeasibly slow. However, using our proposed approach, we were able to obtain an estimate of the full posterior distribution using NIMBLE combined with a single bespoke function written in `R` in less than one hour, considering 20% of the capture histories within the subsampled datasets. Moreover, multiple subsamples can be run simultaneously, with the limiting factor simply the number of computing cores available, and combined to obtain more robust and reliable results. The corresponding results estimated the posterior mean of the random effect standard deviation to be equal to 0.96 (where the random effect is on the logistic scale), suggesting a reasonably high level of heterogeneity present in the (*apparent*) survival probabilities of individuals.

The proposed algorithm is more generally applicable to intractable likelihood problems of large datasets. There are a number of practical implementation issues to be considered for such problems, including, for example, the "optimal" sub-sampling size and/or subsampling strata to be used (in order to minimize the mismatch between subposterior and the full posterior). The efficiency of the approach relies on the subposterior being similar to the full posterior, to minimise particle depletion and reduce the effective sample size. Thus an additional step that may be considered is the inclusion of an accept/reject step following the simulation of a subsampled dataset, retaining the subsample only if it has similar enough "properties" to the full data (with the aim that this increases the probability that the posteriors are similar). For example, such properties could be a function of (scaled) sufficient statistics of the given

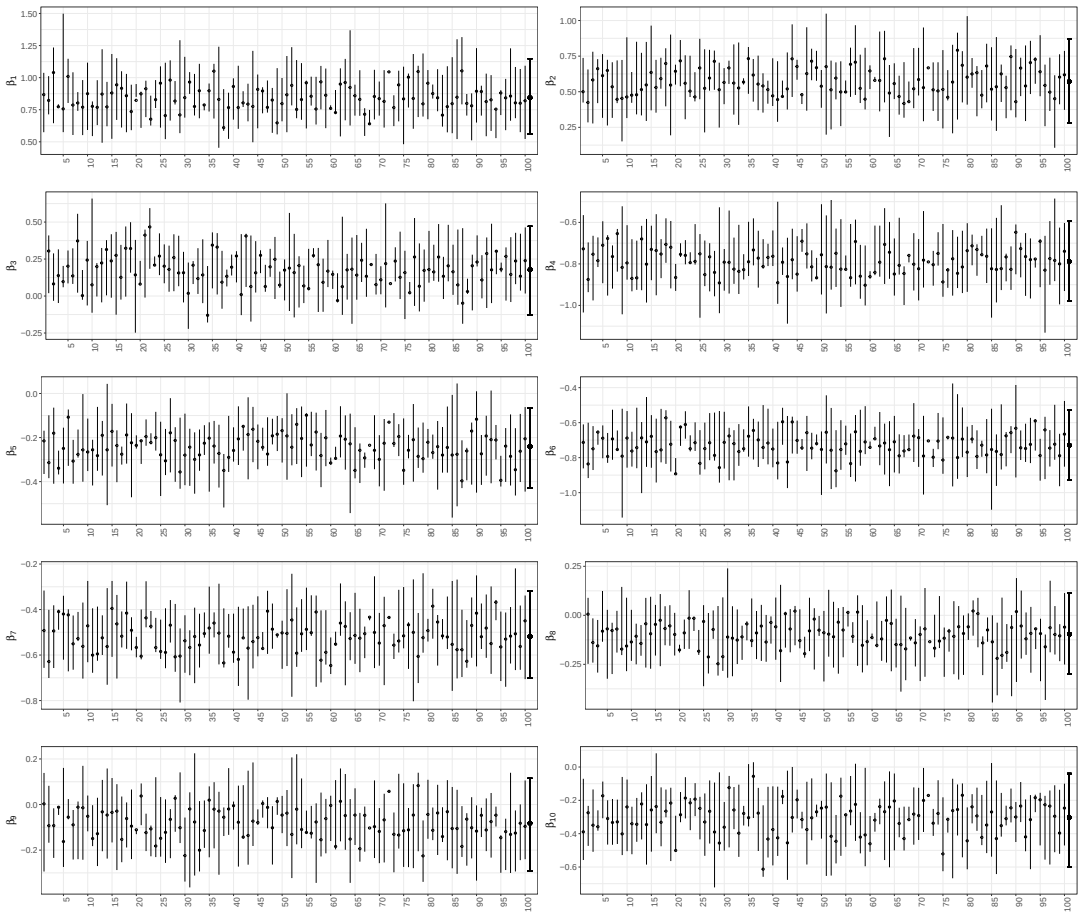

Fig 3: Case study: Corrected posterior means and 95% symmetric CIs for β_t parameters for each subsample and combined across all subsamples (thick solid error bar) for occasions $t = 1, \ldots, T-1$.

dataset. Alternatively to decrease the particle depletion, the selection of sampled MCMC parameter values to be used may be considered further, considering the autocorrelation of the parameter values and/or using a multi-step algorithm for selecting the set of parameter values, following the calculation of the weights of a given set of parameter values in an initial step. Finally, other potential extensions may be explored within importance weight calculation step for increased efficiency. For example, within an efficient two-step approach, further investigation of the threshold used to determine the samples to retain for the second step for computational efficiency whilst retaining high precision may be useful. In particularly, the threshold may be specified to depend on the variability of the (coarse) weights, the use of a nonlinear transformation of the importance weights in order to reduce particle depletion (for example, as in Ionides (2008); Vehtari et al. (2015)), or the variance estimation of the whole scheme. The latter extension is readily possible due to the fact that we have many weighted approximations of the full posterior before performing the combination step. For a sufficiently large M, there is also the potential for bootstrapping the M estimators to obtain an estimate of the variance of the combined estimator (and even to improve the combination strategy). These areas are the focus of current research.

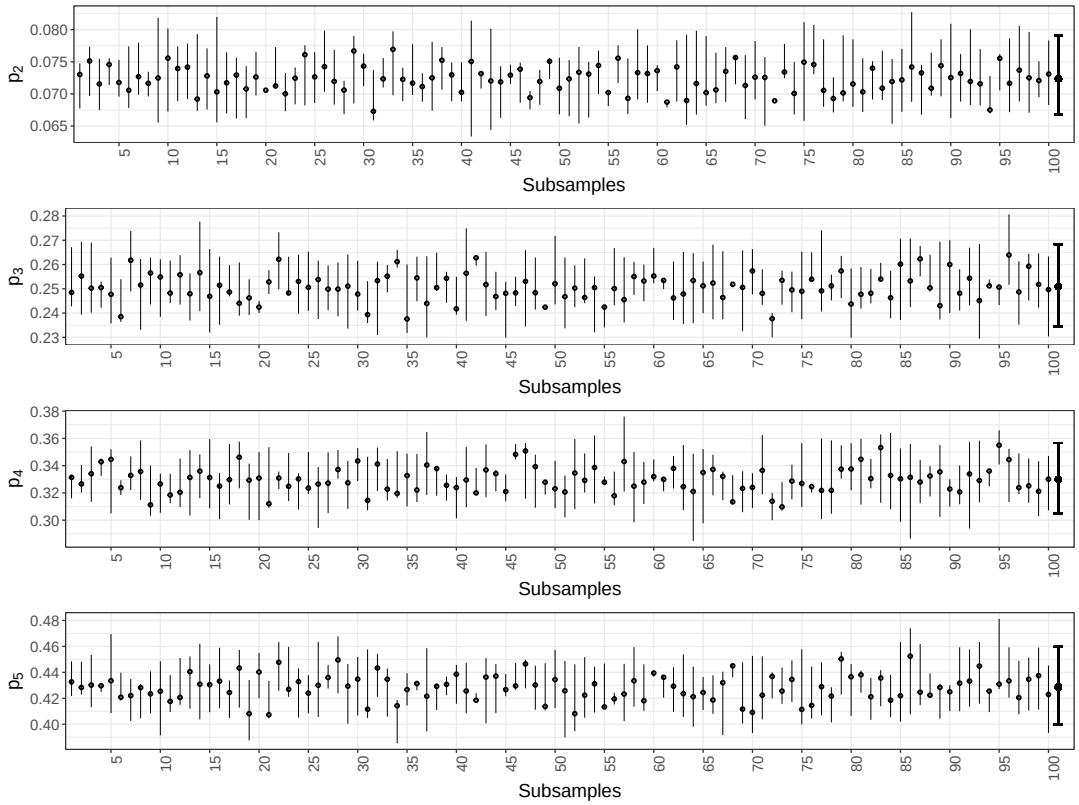

Fig 4: Case study: Corrected posterior means and 95% symmetric CIs for recapture probabilities for each subsample and combined across all subsamples (thick solid error bar) by age ($a = 2, \ldots, 5+$).

Acknowledgments. We thank the Baltic Seabird Project for making the data available and the large number of field workers and volunteers at Stora Karlsö. Field work on Stora Karlsö has been made possible through a long-term engagement in the Baltic Seabird project by WWF Sweden. We would also like to thank the two reviewers and Associate Editor for their helpful and insightful feedback in relation to the initial submission of the paper, leading to an improved manuscript. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

Funding. RK was supported by the Leverhulme research fellowship RF-2019-299. BS was supported by Margarita Salas fellowship from Ministry of Universities-University of Valencia (MS21-013). VE was supported by the *Agence Nationale de la Recherche* of France (ANR-17-CE40-0031-01), the Leverhulme research fellowship (RF-2021-593), and by ARL/ARO (grants W911NF-20-1-0126 and W911NF-22-1-0235).

SUPPLEMENTARY MATERIAL

Appendices

Web appendices A and B referenced in Sections 5 and 6.

GitHub code

Simulated data and `R` code used for the simulation study implemented in the paper in Section

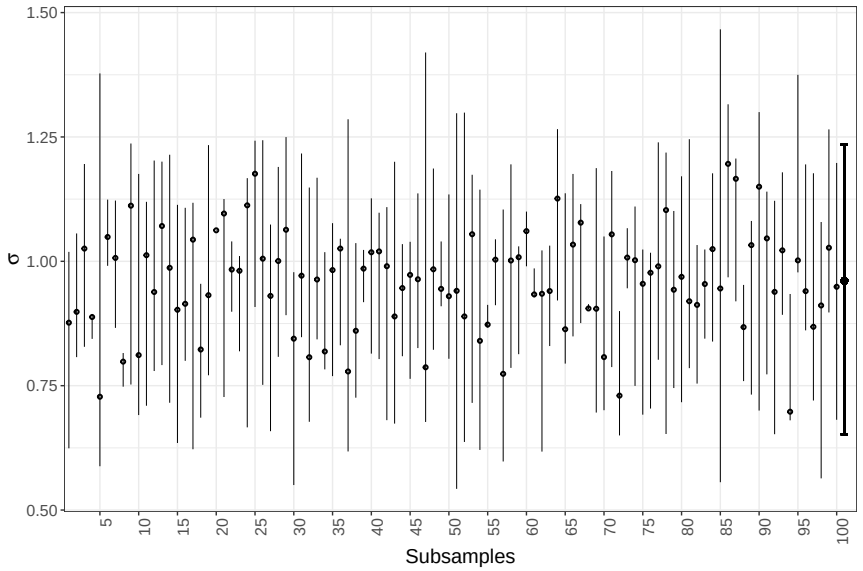

Fig 5: Case study: Corrected posterior means and 95% symmetric CIs for the variance of the individual effects (σ) for each subsample and combined across all subsamples (thick solid error bar).

5. This material is also available at: https://github.com/sarzoblanca/King-Sarzo-and-Elvira.-2022.-When-Worlds-Collide.

References.

ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B* **72** 269–342.

ANDRIEU, C. and ROBERTS, G. O. (2009). The Pseudo-Marginal Approach for Efficient Monte Carlo Computations. *The Annals of Statistics* **37** 697–725.

BARDENET, R., DOUCET, A. and HOLMES, C. (2017). On Markov Chain Monte Carlo Methods for Tall Data. *Journal of Machine Learning Research* **18** 1515–1557.

BROOKS, S. P., GELMAN, A., JONES, G. and MENG, X., eds. (2011). *Handbook of Markov Chain Monte Carlo; Methods and Applications*. CRC Press.

BUTLER, J. and MOFFIT, R. (1982). A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica* 761-764.

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BENTANCOURT, M., BRUBAKER, M., GUO, J. and RIDDELL, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software* **76**.

COULL, B. A. and AGRESTI, A. (1999). The Use of Mixed Logit Models to Reflect Heterogeneity in Capture-Recapture Studies. *Biometrics* **55** 294–301.

DE VALPINE, P. (2002). Review of methods for fitting time-series models with process and observation error and likelihood calculations for nonlinear, non-Gaussian state-space models. *Bulletin of Marine Science* **70** 455–471.

DE VALPINE, P. (2004). Monte Carlo state-space likelihoods by weighted posterior kernel density estimation. *Journal of the American Statistical Association* **99** 523–534.

DE VALPINE, P., TUREK, D., PACIOREK, C., ANDERSON-BERGMAN, C., LANG, D. and BODIK, R. (2017). Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE. *Journal of Computational and Graphical Statistics* **26** 403–413.

DOUC, R., GUILLIN, A., MARIN, J. M. and ROBERT, C. P. (2007). Minimum variance importance sampling via Population Monte Carlo. *ESAIM: Probability and Statistics* **11** 427–447.

ELVIRA, V., MARTINO, L. and CLOSAS, P. (2020). Importance Gaussian Quadrature. *IEEE Transactions on Signal Processing* **69** 474–488.

ELVIRA, V. and MARTINO, L. (2021). Advances in Importance Sampling. *Wiley StatsRef: Statistics Reference Online* 1–14.

FRANCIS, C. M. and SAUROLA, P. (2009). Estimating Demographic Parameters from Complex Data Sets: A Comparison of Bayesian Hierarchical and Maximum-Likelihood Methods for Estimating Survival Probabilities of Tawny Owls, *Strix aluco* in Finland. In *Modeling Demographic Processes In Marked Populations* (D. L. Thomson, E. G. Cooch and M. J. Conroy, eds.) 617–637. Springer, Boston, MA.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2014). *Bayesian Data Analysis* **2**. Chapman & Hall/CRC Boca Raton, FL, USA.

GIMENEZ, O., CAM, E. and GAILLARD, J.-M. (2017). Individual heterogeneity and capture–recapture models: what, why and how? *Oikos* **127** 664–686.

GIMENEZ, O. and CHOQUET, R. (2010). Individual heterogeneity in studies on marked animals using numerical integration: capture-recapture mixed models. *Ecology* **91** 951–957.

GIMENEZ, O., BONNER, S., KING, R., PARKER, R. A., BROOKS, S. P., JAMIESON, L. E., GROSBOIS, V., MORGAN, B. J. T. and THOMAS, L. (2009). WinBUGS for Population Ecologists: Bayesian Modelling using Markov chain Monte Carlo (MCMC) Methods. In *Modeling Demographic Processes In Marked Populations* (D. L. Thomson, E. G. Cooch and M. J. Conroy, eds.) 885–918. Springer, Boston, MA.

HANKIN, D., MOHR, M. and NEWMAN, K. (2019). *Sampling Theory*. Oxford University Press.

HEDEKER, D. and GIBBONS, R. (1994). A random effects ordinal regression model for multilevel analysis. **50** 933-944.

HERLIANSYAH, R., KING, R. and KING, S. E. (2022). Laplace Approximations for Individual Heterogeneity Capture-Recapture Models. *Journal of Agricultural, Biological, and Environmental Statistics* **22** 401-418.

HESTBECK, J. B., NICHOLS, J. D. and MALECKI, R. A. (1991). Estimates of Movement and Site Fidelity Using Mark-Resight Data of Wintering Canada Geese. *Ecology* **72** 523-533.

HUGGINS, J., CAMPBELL, T. and BRODERICK, T. (2016). Coresets for scalable Bayesian logistic regression. *Advances in Neural Information Processing Systems* **29**.

IONIDES, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics* **17** 295–311.

KÉRY, M. and SCHAUB, M. (2011). *Bayesian Population Analysis using WinBUGS: A hierarchical perspective*. Academic Press.

KING, R. (2014). Statistical Ecology. *Annual Review of Statistics and its Application* **1** 401–426.

KING, R. and BROOKS, S. P. (2008). On the Bayesian estimation of a closed population size in the presence of heterogeneity and model uncertainty. *Biometrics* **64** 816–824.

KING, R., SARZO, B. and ELVIRA, V. (2023). Supplement to "When Ecological Individual Heterogeneity Models and Large Data Collide: An Importance Sampling Approach".

KING, R., MORGAN, B. J. T., GIMÉNEZ, O. and BROOKS, S. P. (2010). *Bayesian Analysis for Population Ecology*. CRC Press.

KING, R., MCCLINTOCK, B. T., KIDNEY, D. and BORCHERS, D. (2016). Capture-recapture abundance estimation using a semi-complete data likelihood approach. *The Annals of Applied Statistics* **10** 264–285.

LIU, Q. and PIERCE, D. A. (1994). A Note on Gauss-Hermite Quadrature. *Biometrika* **81** 624–629.

LUENGO, D., MARTINO, L., ELVIRA, V. and BUGALLO, M. (2018). Efficient linear fusion of partial estimators. *Digital Signal Processing* **78** 265–283.

LUNN, D. J., THOMAS, A., BEST, N. and SPIEGELHALTER, D. (2000). WinBUGS: a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10** 325–337.

MCCREA, R. S. and MORGAN, B. J. T. (2015). *Analysis of Capture-Recapture Data*. CRC Press.

NGUYEN, T. L. T., SEPTIER, F., PETERS, G. W. and DELIGNON, Y. (2014). Improving SMC sampler estimate by recycling all past simulated particles. In *Statistical Signal Processing (SSP), 2014 IEEE Workshop on* 117–120. IEEE.

OLSSON, O. and HENTATI-SUNDBERG, J. (2017). Population trends and status of four seabird species (*Uria aalge*, *Alca torda*, *Larus fuscus*, *Larus argentatus*) at Stora Karlsö in the Baltic Sea. *Ornys Svecica* **27** 64–93.

OWEN, A. (2013). *Monte Carlo theory, methods and examples*. http://statweb.stanford.edu/∼owen/mc/.

PLEDGER, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics* **56** 434–442.

PLEDGER, S., POLLOCK, K. H. and NORRIS, J. L. (2003). Open capture-recapture models with heterogeneity. I Cormack-Jolly-Seber. *Biometrics* **59** 786–794.

PLUMMER, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* **124** 1–9.

ROBERT, C. P., ELVIRA, V., TAWN, N. and WU, C. (2018). Accelerating MCMC algorithms. *Wiley Interdisciplinary Reviews: Computational Statistics* **10** 1–14.

ROYLE, J. A. (2008). Modeling individual effects in the Cormack–Jolly–Seber model: A state–space formulation. *Biometrics* **64** 364–370.

SARZO, B., ARMERO, C., CONESA, D., HENTATI-SUNDBERG, J. and OLSSON, O. (2019). Bayesian immature survival analysis of the largest colony of Common murre *Uria aalge* in the Baltic sea. *Waterbirds* **42** 304–313.

SARZO, B., KING, R., CONESA, D. and HENTATI-SUNDBERG, J. (2021). Correcting bias in survival probabilities for partially monitored populations via integrated models. *Journal of Agricultural, Biological and Environmental Statistics* **26** 200–219.

SEBER, G. A. F. and SCHOFIELD, M. R. (2019). *Capture-Recapture: Parameter Estimation for Open Animal Populations*. Springer.

TOKDAR, S. T. and KASS, R. E. (2010). Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics* **2** 54–60.

TRAN, M.-N., SCHARTH, M., PITT, M. K. and KOHN, R. (2016). Importance sampling squared for Bayesian inference in latent variable models. *arXiv:1309.3339*.

VAN DE SCHOOT, R., DEPAOLI, S., KING, R., KRAMER, B., MÄRTENS, K., TADESSE, M. G., VANNUCCI, M., GELMAN, A., VEEN, D., WILLEMSEN, J. and YAU, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers* **1** 1–26.

VEHTARI, A., SIMPSON, D., GELMAN, A., YAO, Y. and GABRY, J. (2015). Pareto smoothed importance sampling. *arXiv:1507.02646*.