



# Greater Than the Sum of its Parts: Computationally Flexible Bayesian Hierarchical Modeling

Devin S. JOHNSON , Brian M. BROST, and Mevin B. HOOTEN

We propose a multistage method for making inference at all levels of a Bayesian hierarchical model (BHM) using natural data partitions to increase efficiency by allowing computations to take place in parallel form using software that is most appropriate for each data partition. The full hierarchical model is then approximated by the product of independent normal distributions for the data component of the model. In the second stage, the Bayesian maximum *a posteriori* (MAP) estimator is found by maximizing the approximated posterior density with respect to the parameters. If the parameters of the model can be represented as normally distributed random effects, then the second-stage optimization is equivalent to fitting a multivariate normal linear mixed model. We consider a third stage that updates the estimates of distinct parameters for each data partition based on the results of the second stage. The method is demonstrated with two ecological data sets and models, a generalized linear mixed effects model (GLMM) and an integrated population model (IPM). The multistage results were compared to estimates from models fit in single stages to the entire data set. In both cases, multistage results were very similar to a full MCMC analysis.

Supplementary materials accompanying this paper appear online.

**Key Words:** Approximation; Bayesian MAP estimation; Bayesian hierarchical model; Integrated population model; Linear mixed model; Meta-analysis; Multistage estimation.

## 1. INTRODUCTION

Bayesian hierarchical models (BHMs) have become ubiquitous in many fields of scientific enquiry because of their ability to flexibly model complex natural systems yet remain relatively easy to build and modify for the questions at hand (Hobbs and Hooten 2015). Berliner (1996) clarified conceptualization of a BHM based on three submodels: (1) *data*,

---

D. S. Johnson (✉), Pacific Islands Fisheries Science Center, National Marine Fisheries Service, NOAA, Honolulu, USA (E-mail: [devin.johnson@noaa.gov](mailto:devin.johnson@noaa.gov)). B. M. Brost, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, USA. M. B. Hooten, Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, USA.

2022 This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply

*Journal of Agricultural, Biological, and Environmental Statistics*, Volume 27, Number 2, Pages 382–400  
<https://doi.org/10.1007/s13253-021-00485-9>

(2) *process*, and (3) *parameter* models. The data model accounts for measurement error or other differences between observable data and the true natural process for which the researcher would like to make inference. The process model describes the state of nature using a random process to account for its unknown intricacies. Finally, the parameter model describes our uncertainty about the parameters governing this random process. The full model is built by conditioning each submodel on the one above it. Traditionally, inference for BHMs has been conducted using some form of Markov chain Monte Carlo (MCMC) because the model structure is ideally suited to it (Gelfand and Smith 1990; Gelfand and Ghosh 2015). If not for development of MCMC methods, widespread use of BHMs may never have come to pass (Green et al. 2015). With a large amount of data or complex submodels, however, MCMC can be computationally challenging or infeasible to implement (Hooten et al. 2021; Wikle 2003). Herein, we propose a multistage method for making inference at all levels of a BHM using various software formats in each stage. By breaking the estimation procedure into multiple stages based on a partition of the full data it becomes computationally efficient. By allowing the use of the most suitable software platform in the initial stage, it can be made more efficient yet, not only in computing time, but also in development time saved by using software designed to facilitate each specific initial stage analysis.

To alleviate computational challenges in using MCMC to fit BHMs in “big data” situations, there have been several avenues of approach developing in the literature. One of these is development of the so-called two-stage methods that seek to break the computation into pieces based on partitions of the data (Goudie et al. 2019; Hooten et al. 2016, 2021; Lunn et al. 2013; Mesquita et al. 2020; Scott et al. 2016). The first stage involves fitting separate data models to the partitions using MCMC; then, in the second stage, the separate MCMC samples are combined to produce an MCMC sample as if the partitions had been simultaneously analyzed with the full BHM. A similar approach is the divide-and-conquer approach (Srivastava et al. 2018) that averages posterior distributions from data partitions to approximate an analysis of the full data.

An alternative approach to default MCMC for computationally challenging BHMs is to use optimization methods for exploring the posterior parameter space rather than stochastic sampling (Green et al. 2015). Optimization of the posterior density aims to find the *maximum a posteriori* (MAP) estimate rather than the posterior mean for a point estimate. The main benefit is that optimizing the posterior with respect to the parameters often requires fewer evaluations of the posterior density. That comes at a cost, however, because asymptotic results are usually required to describe parameter uncertainty (Van der Vaart 2000).

In an effort to provide a computationally efficient and flexible method for fitting a general class of BHMs, *conditionally independent hierarchical models* (CIHMs; Kass and Steffey 1989), we consider combining benefits of meta-analysis, two-stage MCMC, and posterior optimization into a multistage method that can be used to quickly fit BHMs in big data situations or when the data-level models are complex themselves. Although meta-analysis and two-stage MCMC are both accomplished in multiple stages, two-stage MCMC was initially designed to analyze a BHM, whereas meta-analysis seeks to summarize results from various analyses by forming them into a BHM, namely a multivariate normal linear mixed model (Gasparrini et al. 2012). We can use the meta-analysis approach to approximate

the second-stage combination of initial stage inference in a two-stage MCMC. This allows us to use multiple approaches to approximate the posterior inference in the initial stage, including optimization methods, deterministic sampling, or MCMC if it is convenient.

In what follows, we provide notation and a working definition of the CIHM as we will use it. We show how the full posterior can be approximated using multivariate normal approximations to first-stage posterior densities. Then, we demonstrate how the first-stage results can be combined using the linear mixed model approach of standard meta-analysis (Gasparri et al. 2012; Higgins et al. 2009) when the first-stage parameters are random effects. An optional third stage is introduced to retroactively revisit the first-stage analysis and can improve estimates after the second-stage results are obtained. The method is then demonstrated for two different types of environmental data, a simple generalized linear mixed model (GLMM) and an integrated population model (IPM) for demographic parameter inference.

## 2. CONDITIONALLY INDEPENDENT HIERARCHICAL MODEL

The description of BHMs by Berliner (1996) is very general, and we only consider a specific, but still very broad, class of hierarchical models (CIHMs; Gelfand and Ghosh 2015; Kass and Steffey 1989). The general form of a CIHM is

$$\begin{aligned}
 \text{Data-level model: } \mathbf{y} &= (\mathbf{y}'_1, \dots, \mathbf{y}'_n)' \sim \prod_{i=1}^n [\mathbf{y}_i | \boldsymbol{\theta}_i, \boldsymbol{\gamma}_i, \boldsymbol{\eta}], \\
 \text{Unit-level effects: } \boldsymbol{\theta} &= (\boldsymbol{\theta}'_1, \dots, \boldsymbol{\theta}'_n)' \sim [\boldsymbol{\theta} | \boldsymbol{\psi}], \\
 \boldsymbol{\gamma} &= (\boldsymbol{\gamma}'_1, \dots, \boldsymbol{\gamma}'_n)' \sim \prod_{i=1}^n [\boldsymbol{\gamma}_i], \\
 \text{Population effects: } (\boldsymbol{\psi}', \boldsymbol{\eta}') &\sim [\boldsymbol{\psi}, \boldsymbol{\eta}],
 \end{aligned} \tag{1}$$

where “[ $A|B$ ]” is used to represent the conditional probability density or distribution function of  $A$  given  $B$  (Gelfand and Smith 1990),  $\mathbf{y}_i$  is the observed data set (or data partition) for the  $i$ th unit ( $i = 1, \dots, n$ ),  $\boldsymbol{\theta}_i$  are random unit-level parameters,  $\boldsymbol{\psi}$  are population-level parameters for  $\boldsymbol{\theta}_i$ ,  $\boldsymbol{\gamma}_i$  are distinct (relative to each  $\mathbf{y}_i$ ) unit-level parameters for modeling the  $i$ th data set, and  $\boldsymbol{\eta}$  are global (fixed) population-level parameters. Typically,  $\boldsymbol{\psi}$  and  $\boldsymbol{\eta}$  are the most scientifically interesting; however, there can be interest in each of the  $\boldsymbol{\theta}_i$  and  $\boldsymbol{\gamma}_i$  depending on the situation. Note that unit-level effects are allowed to be correlated which is slightly more general than Kass and Steffey (1989).

To place the development of the multi-stage methodology in context, we will demonstrate its use on the analysis of two ecological data sets. The first one uses a generalized linear mixed model (GLMM) to assess the effect of wolf (*Canis lupus*) presence on the occurrence of moose (*Alces alces*) browse in willow stands. Although these data are not large in the modern sense, we demonstrate a two-stage analysis of them because the model can easily be fit in its full hierarchical form for comparison. Another common group of models in the ecological literature are integrated population models (IPMs). Initially proposed by

Besbeas et al. (2002), IPMs are routinely used to collect information in multiple sparse data sets for improved inference on shared parameters. To demonstrate the two-stage meta-analysis approach for IPM model fitting, we analyze the classic northern lapwing (*Vanellus vanellus*) data from Besbeas et al. (2002).

**2.1. MOOSE BROWSE IN THE PRESENCE WOLVES (GLMM)**

This example involves  $n = 2,914$  binary responses,  $y_{ij}$ , which indicate the presence of moose browsing on tree  $j = 1, \dots, n_i$  of plantation  $i = 1, \dots, 24$ . The original study assessed whether or not tree browsing is reduced in plantations exposed to high wolf activity (van Beeck Calkoen et al. 2018). In addition, the study accounted for inherent moose preferences for tree height. To perform a similar analysis, we use a random effects model to treat each plantation as a unit to learn about the fixed effect of high wolf presence on moose browsing while accounting for natural variation in moose response across the 24 plantations. The CIHM is given by

$$\begin{aligned}
 y_{ij} &\sim \text{Bernoulli}(p_{ij}), \\
 \text{logit}(p_{ij}) &= \delta_0 + \delta_{\text{height}}h_{ij} + \delta_{\text{wolf}}w_i + u_{0i} + u_{\text{height},i}h_{ij}, \\
 \mathbf{u}_i &= (u_{0i}, u_{\text{height},i}) \sim \text{N}(\mathbf{0}, \text{diag}(\sigma_1^2, \sigma_2^2)),
 \end{aligned}$$

where  $h_{ij}$  equals the height of the  $j$ th tree in the  $i$ th plantation and  $w_i$  is a binary indicator of high wolf presence in the  $i$ th plantation.

To use the two-stage approach, we must hierarchically center the model because  $\delta_0$  and  $\delta_{\text{wolf}}w_j$  are not identifiable within plantation units. Using hierarchical centering (Gelfand et al. 1996), the model can be rewritten as

$$\begin{aligned}
 \text{logit}(p_{ij}) &= \theta_{0i} + \theta_{\text{height},i}h_{ij}, \\
 \boldsymbol{\theta}_i &= (\theta_{0i}, \theta_{\text{height},i}) \sim \text{N}(\mathbf{L}_i\boldsymbol{\delta}, \boldsymbol{\Sigma}),
 \end{aligned}$$

where

$$\mathbf{L}_i = \begin{bmatrix} 1 & 0 & w_i \\ 0 & 1 & 0 \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2).$$

In terms of the CIHM specification (1),  $\boldsymbol{\psi} = (\boldsymbol{\delta}, \sigma_1^2, \sigma_2^2)$ . The parameters  $\boldsymbol{\gamma}$  and  $\boldsymbol{\eta}$  are not used.

**2.2. LAPWING DEMOGRAPHIC MODELING (IPM)**

The typical IPM (Schaub and Abadi 2011) has the CIHM form

$$[\mathbf{y}] = \prod_{i=1}^n [y_i | \boldsymbol{\gamma}_i, \boldsymbol{\eta}], \text{ and } [\boldsymbol{\gamma}] = \prod_{i=1}^n [\boldsymbol{\gamma}_i],$$

where  $\boldsymbol{\gamma}_i$  are data-type distinct parameters and  $\boldsymbol{\eta}$  are common parameters. Traditionally, there are no random unit-level effects ( $\boldsymbol{\theta}_i$ ) in IPM models. Multiple data sets are used to obtain better estimates of the common parameters,  $\boldsymbol{\eta}$ , and then can be obtained from any single  $\boldsymbol{y}_i$ .

The lapwing data consist of  $n = 2$  data sets, a ring-recovery data set,  $\boldsymbol{y}_1$ , and an index measure of the abundance of adult female birds obtained from a generalized additive model of counts,  $\boldsymbol{y}_2$ . These data have been analyzed previously to demonstrate IPM methodology for ecological data (Besbeas et al. 2002; Besbeas and Morgan 2019; Brooks et al. 2004; Goudie et al. 2019).

Ring-recovery models are based on marking animals and releasing them back into the wild. In subsequent years, the marked animals are then recovered after they have died. The likelihood for this type of model is a product-multinomial, where numbers of ring-recoveries in years following release are multinomial distributed with cell probabilities that are functions of  $\phi_t$  (for  $t = 1963, \dots, 1997$ ), the probability of survival from time  $t$  to  $t + 1$  and  $\lambda_t$ , the probability that an animal that died between  $t - 1$  and  $t$  is recovered. To assess the influence of weather on survival and trends in recovery, the ring-recovery parameters are modeled with

$$\begin{aligned}\text{logit}(\phi_{y_t}) &= \eta_{y0} + \eta_{y1}x_t, \\ \text{logit}(\phi_{a_t}) &= \eta_{a0} + \eta_{a1}x_t, \\ \text{logit}(\lambda_t) &= \gamma_{\lambda 0} + \gamma_{\lambda 1}t,\end{aligned}\tag{2}$$

where  $\phi_{y_t}$  is the survival of first year birds,  $\phi_{a_t}$  is the adult female survival, and  $x_t$  is the number of days below freezing.

The census index data consist of noisy measures of female adult lapwing abundance in the study area. To make inference about population dynamics, we use the state-space model of Besbeas et al. (2002),

$$\begin{aligned}\mathbf{N}_{t+1} &= \mathbf{T}_t \mathbf{N}_t + \boldsymbol{\epsilon}_t \\ y_t &\sim \mathbf{N}(N_{a_t}, \sigma^2),\end{aligned}\tag{3}$$

where  $\mathbf{N}_t = (N_{y_t}, N_{a_t})'$  is the number of yearling and adult females, respectively, in year  $t$ , and  $y_t$  are the noisy count of adult females. The transition matrix,  $\mathbf{T}_t$ , is parameterized with  $\phi_t$  and the rate of offspring production,  $\rho_t$ . The survival parameters are the same as the ring-recovery model, but the production is modeled on the log scale using

$$\log(\rho_t) = \gamma_{\rho 0} + \gamma_{\rho 1}t.$$

Full model details are provided in online appendix S2.

The IPM melds the information in the two data sets to allow inference for all parameters of the joint model,

$$\begin{aligned}\boldsymbol{\eta} &= (\eta_{y0}, \eta_{y1}, \eta_{a0}, \eta_{a1})', \\ \boldsymbol{\gamma}_1 &= (\gamma_{\lambda 0}, \gamma_{\lambda 1})', \text{ and} \\ \boldsymbol{\gamma}_2 &= (\gamma_{\rho 0}, \gamma_{\rho 1}, \log \sigma, \log N_{y0}, \log N_{a0})'.\end{aligned}$$

Random unit-level  $\theta$  parameters are not traditionally used, but see the analysis in Sect. 4.2 for an alternate version. In each of the previous studies that analyze these data, bespoke code was written in software such as MATLAB, JAGS (Plummer 2003), NIMBLE (de Valpine et al. 2017) or some other language to optimize the joint likelihood or execute an MCMC. For a rather simple IPM model like this, it is straightforward to code. But for more complicated capture–recapture models or abundance models, this can become an impediment for practitioners. Many of the individual submodels in IPMs have software available for fitting them separately. For the ring-recovery data in this analysis, there has been production-level software to fit the appropriate capture–recapture model for over 20 years. Thus, we should be able to utilize this substantial development in capture–recapture data analysis.

### 3. METHODS

We propose a two-stage method (extended to an optional third stage) for analyzing CIHM models (1). In the first stage, each data set,  $y_i$ , is analyzed individually according to its individual data-level model in the CIHM using the best method and software available for that particular data type. Then, in the second stage, the results are combined (melded) using the unit-level and population-level models to approximate the inference as if the entire CIHM was fit in one step (e.g., via MCMC). In an optional third stage, we can revisit estimates of the distinct unit parameters,  $\gamma_i$  to see if further updating is possible.

#### 3.1. TWO-STAGE APPROXIMATE BAYESIAN INFERENCE

If we were to fit the CIHM model in the traditional way, we would use summaries from the full posterior distribution

$$[\psi, \eta, \gamma, \theta | y] \propto [y | \theta, \gamma, \eta][\theta | \psi][\gamma | \psi, \eta].$$

Suppose, however, we only possess a single  $y_i$ . Then, we might estimate  $\vartheta_i = (\theta'_i, \gamma'_i, \eta'_i)'$  using the posterior distribution

$$[\vartheta_i | y_i] \propto [y_i | \vartheta_i][\vartheta_i],$$

where  $[\vartheta_i]$  is a prior distribution that would be used in the absence of additional individuals. Using these individual posterior distributions, we can rewrite the full hierarchical model posterior as

$$[\psi, \eta, \gamma, \theta | y] \propto \left( \prod_{i=1}^n \frac{[\vartheta_i | y_i]}{[\vartheta_i]} \right) [\theta | \psi][\gamma | \psi, \eta]. \tag{4}$$

Stage I of the method we propose begins by first finding estimates  $\hat{\vartheta}_i$  and covariance matrices  $\hat{S}_i$  such that we can approximate

$$\frac{[\vartheta_i | y_i]}{[\vartheta_i]} \propto [y_i | \vartheta_i] \approx N(\vartheta_i | \hat{\vartheta}_i, \hat{S}_i) \tag{5}$$

for each data set or partition. This approximation is motivated by standard large-sample asymptotic results. For large-sample theory to motivate the approximation,  $[y_i|\vartheta_i]$  should satisfy the standard regularity conditions on the likelihood function, namely the “true”  $\vartheta_i$  is not on the edge of the parameter space and  $[y_i|\vartheta_i]$  is continuously twice differentiable (Kass and Steffey 1989; Le Cam and Yang 2000). However, if the log-likelihood is approximately quadratic in  $\vartheta_i$ , then the approximation can work well even for modest sample sizes (Geyer 2005).

An estimate and covariance matrix for use in (5) is the maximum likelihood estimate (MLE) of  $\vartheta_i$  and  $\hat{S}_i$  is the negative inverse Hessian. For now, we assume that the size of  $y_i$  is large enough and the parameters are identifiable enough such that  $\hat{\vartheta}_i$  is well defined and  $\hat{S}_i^{-1}$  is nonnegative definite. The main benefit of using the MLE is that there is virtually always software available for finding the MLE and large-sample covariance matrix for common data-level models.

If  $\vartheta_i$  is only weakly identified in  $[y_i|\vartheta_i]$ , then the MLE may be difficult to find. Thus, calculating an estimate from the posterior distribution  $[\vartheta_i|y_i] = [y_i|\vartheta_i][\vartheta_i]$  might be desirable, where the temporary prior  $[\vartheta_i]$  is chosen to make all parameters identifiable. In this case, one might choose to use the maximum *a posteriori* (MAP) by optimizing  $[\vartheta_i|y_i]$  and using the inverse Hessian for uncertainty as with the MLE. Or one could use a posterior sample (e.g., MCMC or importance sampling) to find the posterior mean and covariance matrix. Similar general regularity conditions apply to the MLE, MAP, or MCMC estimates with respect to asymptotic approximation of well-identified parameters. However, for nearly unidentifiable parameters in  $[\vartheta_i|y_i]$ , asymptotic approximations can be poor for even large samples because no matter how large the data set is, the posterior will be nearly (or exactly) equal to the prior distribution. Therefore, because we are free to choose the temporary prior to aid stage I model fitting, we suggest using a normal prior

$$[\vartheta_i] = N(\vartheta_i|\vartheta_0, S_0).$$

A reparameterization of  $\vartheta_i$  might be necessary such that a normal density is appropriate as a prior, but this will help the accuracy of the normal approximation in stage I even for poorly identified parameters. Note that the support of the temporary prior needs to be the same as the desired parameter model  $[\theta_i|\psi][y_i][\eta]$ . With the use of the temporary prior, the approximation becomes

$$\frac{[\vartheta_i|y_i]}{[\vartheta_i]} \approx \frac{N(\vartheta_i|\check{\vartheta}_i, \check{S}_i)}{N(\vartheta_i|\vartheta_0, S_0)} \propto N(\vartheta_i|\hat{\vartheta}_i, \hat{S}_i), \tag{6}$$

where  $\check{\vartheta}_i$  and  $\check{S}_i$  are the posterior mode (or mean) and negative inverse Hessian (or covariance matrix) of  $[\vartheta_i|y_i]$  and

$$\hat{S}_i = \left(\check{S}_i^{-1} - S_0^{-1}\right)^{-1} \quad \text{and} \quad \hat{\vartheta}_i = \hat{S}_i \left(\check{S}_i^{-1}\check{\vartheta}_i - S_0^{-1}\vartheta_0\right). \tag{7}$$

This ratio was also used by Goudie et al. (2019) in an approximate two-stage MCMC method.

Using (4) and (5) for  $[\vartheta_i | y_i] / [\vartheta_i]$ , the desired full posterior density function for all the parameters can be approximated (up to a proportional constant) by

$$[\boldsymbol{\psi}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\theta} | \mathbf{y}] \approx \left( \prod_{i=1}^n N(\hat{\boldsymbol{\vartheta}}_i | \boldsymbol{\vartheta}_i, \hat{\boldsymbol{S}}_i) \right) [\boldsymbol{\theta} | \boldsymbol{\psi}] [\boldsymbol{\gamma}] [\boldsymbol{\psi}, \boldsymbol{\eta}]. \tag{8}$$

Note that, in (8), we have put  $\boldsymbol{\vartheta}_i$  in the mean position of the normal specification. Due to the symmetry of the normal density function, it is equivalent, but this way  $\hat{\boldsymbol{\vartheta}}_i$  can be interpreted as an observation. In stage II, we can make full inference for all model parameters by maximizing the approximate full posterior (8) to obtain the posterior mode  $(\tilde{\boldsymbol{\psi}}', \tilde{\boldsymbol{\eta}}', \tilde{\boldsymbol{\gamma}}', \tilde{\boldsymbol{\theta}})'$ , or *maximum a posteriori* (MAP) estimate and the negative inverse Hessian of the log posterior density estimates the posterior covariance matrix. Hats are used to represent stage I estimates and tildes to represent stage II estimates. For the remainder of the paper, we will assume that some measure of location,  $\hat{\boldsymbol{\vartheta}}_i$ , and scale  $\hat{\boldsymbol{S}}_i$  are available where  $\hat{\boldsymbol{S}}_i^{-1}$  is nonnegative definite. These can come from any combination of MLEs, MAPs, or MCMC samples, whatever software is most appropriate or available for the data model can be used. Different methods can be used between stage I units as well.

### 3.2. NORMAL RANDOM UNIT-LEVEL EFFECTS

While there is no requirement that the unit-level random effects,  $\boldsymbol{\theta}$ , be normally distributed in (8), it is the case in the vast majority of applications, or  $\boldsymbol{\theta}_i$  can often be parameterized such that a normal distribution is appropriate. Thus, for the remainder of the paper, we assume that

$$[\boldsymbol{\theta} | \boldsymbol{\psi}] = N(\boldsymbol{\theta} | \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where  $\boldsymbol{\psi} = (\boldsymbol{\mu}', \boldsymbol{\sigma}')'$  and  $\boldsymbol{\Sigma}$  is a covariance matrix parameterized by  $\boldsymbol{\sigma}$ . When this is the case, stage II becomes equivalent to fitting a multivariate normal random effects linear model.

First, we reparameterize the model using  $\boldsymbol{\theta}_i = \boldsymbol{\mu} + \mathbf{u}_i$  where  $\mathbf{u} = (\mathbf{u}'_1, \dots, \mathbf{u}'_n) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ ,  $\boldsymbol{\alpha} = (\boldsymbol{\eta}', \boldsymbol{\gamma}'_1, \dots, \boldsymbol{\gamma}'_n)'$ ,  $\boldsymbol{\alpha}_i = (\boldsymbol{\eta}', \boldsymbol{\gamma}'_i)'$ , and  $\boldsymbol{\beta} = (\boldsymbol{\mu}', \boldsymbol{\alpha}')'$ . Now, the approximation (8) becomes

$$[\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \mathbf{u} | \mathbf{y}] \approx \prod_{i=1}^n \left( N(\hat{\boldsymbol{\vartheta}}_i | \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i, \hat{\boldsymbol{S}}_i) \right) N(\mathbf{u} | \mathbf{0}, \boldsymbol{\Sigma}) [\boldsymbol{\gamma}] [\boldsymbol{\mu}, \boldsymbol{\sigma}] [\boldsymbol{\eta}], \tag{9}$$

where  $\mathbf{X}_i = \text{diag}(\mathbf{I}, \mathbf{K}_i)$ ,  $\mathbf{I}$  is a  $|\boldsymbol{\mu}|$  size identity matrix,  $\mathbf{K}_i$  is a  $|\boldsymbol{\alpha}_i| \times |\boldsymbol{\alpha}|$  indicator matrix that selects the appropriate parameter, and  $\mathbf{Z}_i$  is a  $|\boldsymbol{\vartheta}_i| \times |\boldsymbol{\theta}_i|$  indicator matrix that selects  $\mathbf{u}_i$  to align with the  $\boldsymbol{\theta}_i$  portion of  $\boldsymbol{\vartheta}_i$ . If  $\boldsymbol{\mu}$  is represented as a linear model as in the moose browse example, then  $\mathbf{I}$  is replaced with the appropriate design matrix (e.g.,  $\mathbf{L}_i$  in Sect. 2.1). After concatenation,  $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}'_1, \dots, \boldsymbol{\vartheta}'_n)'$ ,  $\mathbf{X} = (\mathbf{X}'_1, \dots, \mathbf{X}'_n)'$ ,  $\mathbf{Z} = (\mathbf{Z}'_1, \dots, \mathbf{Z}'_n)'$  and  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \text{diag}(\hat{\boldsymbol{S}}_1, \dots, \hat{\boldsymbol{S}}_n))$ , finding the MAP of  $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \mathbf{u})$  is equivalent to fitting the multivariate linear mixed model

$$\hat{\boldsymbol{\vartheta}} = \mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \mathbf{u} + \boldsymbol{\epsilon}, \tag{10}$$



subject to the prior distribution  $[\boldsymbol{\gamma}][\boldsymbol{\mu}, \boldsymbol{\sigma}][\boldsymbol{\eta}]$ .

In the analysis of the two ecological data sets, we use the R package TMB to fit (9), but it is just a multivariate linear mixed model, so any software that can account for the parameter prior distributions can be used. TMB uses Laplace's method to find the MAP of  $\boldsymbol{\psi} = (\boldsymbol{\mu}, \boldsymbol{\sigma}), \boldsymbol{\eta}$ , and  $\boldsymbol{\gamma}$  using the marginal posterior  $[\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\eta}, \boldsymbol{\gamma}|\mathbf{y}]$ , and then a generalized delta-method (Kass and Steffey 1989) is employed to estimate  $\mathbf{u}$  accounting for uncertainty in the other parameters (Kristensen et al. 2016; Skaug and Fournier 2006). At first, it might seem as if we are adding another layer of approximation by using the Laplace method; however, given the normal approximation already made in (8) and the normality of the unit-level parameters, the Laplace method of approximating the marginal posterior density of  $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\eta}, \boldsymbol{\gamma})|\mathbf{y}$  obtained from (9) is actually exact because the log-posterior is quadratic in  $\mathbf{u}$  (Goutis and Casella 1999). After finding the MAP, traditionally, the inverse Hessian of (9) is used to estimate the covariance matrix of the full model (9), which is what we use in the examples that follow.

### 3.3. STAGE III: REVISITING STAGE I

Using the description to this point, we are set to perform two-stage inference for a CIHM. However, we noticed, in our initial analysis of the lapwing data in Section 4.2, when  $\mathbf{S}_i^{-1} \approx \mathbf{0}$  (relative to  $\boldsymbol{\vartheta}_i$ ) for a particular unit in stage I, then updating  $\hat{\boldsymbol{\gamma}}_i$  might not be as efficient as desired. Therefore, we propose revisiting stage I estimation to update  $\boldsymbol{\gamma}_i$  given the results of the second stage using the generalized delta method initially proposed by Kass and Steffey (1989). Note that we are not proposing that stage III is always necessary, but, that it is an option, especially if  $\boldsymbol{\gamma}_i$  is not well identified in the data model alone.

For a heuristic example of lack of updating for  $\hat{\boldsymbol{\gamma}}_i$ , consider the simple two-parameter model structure:

$$\mathbf{y}_1 \sim [\mathbf{y}_1|\boldsymbol{\eta}] \text{ and } \mathbf{y}_2 \sim [\mathbf{y}_2|\boldsymbol{\eta}, \boldsymbol{\gamma}_2]. \quad (11)$$

Further, suppose that  $\boldsymbol{\eta}$  and  $\boldsymbol{\gamma}_2$  jointly are only vaguely identifiable in  $[\mathbf{y}_2|\boldsymbol{\eta}, \boldsymbol{\gamma}_2]$ , whereas it is fully identifiable in  $[\mathbf{y}_1|\boldsymbol{\eta}]$ . In stage I, we can obtain well-informed estimates  $\hat{\boldsymbol{\eta}}_1$  and  $\hat{\mathbf{S}}_1$ . However, for  $\mathbf{y}_2$ ,  $\hat{\mathbf{S}}_2^{-1}$  will be small. In the extreme case that  $\hat{\mathbf{S}}_2^{-1} \rightarrow \mathbf{0}$ , one can show (through the generalized least-square estimate of  $\boldsymbol{\beta} = (\boldsymbol{\eta}, \boldsymbol{\gamma}_2)$  in (10)) that  $\tilde{\boldsymbol{\eta}} \rightarrow \hat{\boldsymbol{\eta}}_1$  for any  $\hat{\boldsymbol{\eta}}_2$ . However,  $\tilde{\boldsymbol{\gamma}}_2 \rightarrow \hat{\boldsymbol{\gamma}}_2$ , remaining unchanged from stage I. So, we see satisfactory updating of  $\hat{\boldsymbol{\eta}}_2$ , but no updating of  $\hat{\boldsymbol{\gamma}}_2$  in stage II. This case is admittedly pathological, but it illustrates that the Hessian matrix of  $[\mathbf{y}_2|\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}_2, \hat{\boldsymbol{\gamma}}_2]$  may not be the best representation of the likelihood shape at the full MAP.

Stage III proceeds by first marginalizing over  $\boldsymbol{\gamma}_i$ . This can be done analytically if possible. However, based on the initial normal approximation after stage I, after  $\hat{\boldsymbol{\vartheta}}_i$  and  $\hat{\mathbf{S}}_i$  are calculated, the rows and columns associated with  $\boldsymbol{\gamma}_i$  can be ignored when moving to stage II. Let  $\boldsymbol{\xi}$  represent all parameters except  $\boldsymbol{\gamma}_i$ ; after stage II, we obtain the marginal posterior approximation:

$$[\boldsymbol{\xi}|\mathbf{Y}] \approx \mathbf{N}(\boldsymbol{\xi} | \tilde{\boldsymbol{\xi}}, \tilde{\mathbf{V}}_{\boldsymbol{\xi}}),$$

where  $\tilde{\mathbf{V}}_{\xi_i}$  is the negative inverse of the Hessian of the log-posterior distribution represented by the stage II model (9) and the associated prior distributions. If  $\gamma_i$  is a nuisance parameter and of little interest, we can stop here and there is no need for stage III. However, if an estimate of  $\gamma_i$  is desired, we can obtain an updated mean and covariance matrix for the full posterior approximation such that

$$[\gamma_i, \xi | \mathbf{Y}] \approx \mathbf{N}((\gamma_i, \xi) | \bar{\xi}, \bar{\mathbf{V}}),$$

where  $\bar{\xi} = (\bar{\gamma}_i, \bar{\xi})$ ,  $\bar{\gamma}_i$  is the value that maximizes the conditional posterior  $[\gamma_i | \tilde{\theta}_i, \tilde{\eta}, y_i]$  and  $\bar{\mathbf{V}}$  is a updated posterior covariance matrix. To calculate  $\bar{\mathbf{V}}$ , we need to define the matrices:

$$\mathbf{H}_{\gamma\gamma} = \frac{\partial^2}{\partial \gamma_i \partial \gamma_i'} \left\{ \log[y_i | \tilde{\theta}_i, \tilde{\eta}, \bar{\gamma}_i] + \log[\bar{\gamma}_i] \right\} \text{ and}$$

$$\mathbf{H}_{\gamma\xi} = \frac{\partial^2}{\partial \gamma_i \partial \xi_i'} \log[y_i | \tilde{\theta}_i, \tilde{\eta}, \bar{\gamma}_i].$$

Now, the full covariance matrix can be approximated by (Kass and Steffey 1989; Skaug and Fournier 2006):

$$\bar{\mathbf{V}} = \begin{bmatrix} -\mathbf{H}_{\gamma\gamma}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \mathbf{J} \tilde{\mathbf{V}}_{\xi} \mathbf{J}', \text{ where}$$

$$\mathbf{J} = \begin{bmatrix} -\mathbf{H}_{\gamma\gamma}^{-1} \mathbf{H}_{\gamma\xi} \\ \mathbf{I}_{\xi} \end{bmatrix}.$$

This may not be applicable in all situations, because one must be able to evaluate  $[\gamma_i | \tilde{\theta}_i, \tilde{\eta}, y_i]$  to a proportional constant as well as its gradient and Hessian for various parameter values. This may not be practical for most production software used in stage I.

One could question whether or not  $\theta_i$  might suffer the same lack of sufficient updating in stage II as well as  $\gamma_i$  if the likelihood is generally flat. However, this is much less likely due to the fact that there is a unit-level model,  $[\theta | \psi]$ , which is being explicitly fit in the second stage. This allows information in other data sets to directly influence and update  $\theta_i$  at stage II. However,  $y_j, j \neq i$ , can only contribute to updates of  $\hat{\gamma}_i$  via the off-diagonal elements of  $\hat{\mathbf{S}}_i^{-1}$ .

## 4. MODEL FITTING AND RESULTS

### 4.1. MOOSE BROWSE GLMM

To effectively use the two-stage approach in this example, it was necessary to use a stage I temporary prior (6) and (7) because browsing is rare (or absent) in some of the plantations. We used the normal  $g$ -prior (Hanson et al. 2014) where  $\theta_0 = \mathbf{0}$ ,  $\mathbf{S}_0 = \pi^2 n_i (\mathbf{H}'_i \mathbf{H}_i)^{-1} / 6$ , and  $\mathbf{H}_i$  is the usual design matrix with a column for intercept and the other containing tree height ( $h_{ij}$ ). This prior produces  $p_{ij}$  that are approximately uniform(0,1) distributed for a randomly selected  $h_{ij}$ . The effect of this prior was removed prior to stage II using (7).

For stage I model fitting, we used two approaches. In the first version (A), we fitted each individual model using the R package `mgcv` (Wood 2011) to obtain the posterior mode and Hessian covariance matrix from  $[\theta_i | y_i] \propto [y_i | \theta_i][\theta_i]$ . In the second version (B), we used the approximate posterior mean and covariance matrix obtained from a deterministic posterior sampling procedure detailed in Johnson et al. (2011) (see online appendix S1). This deterministic procedure estimates the posterior mean and variance without asymptotic results. For version B, the likelihood must be calculated with bespoke code. We are not aware of any production-level software that returns only the value of the posterior for logistic regression, which makes this version more challenging in practice than the first version, but we explored here nonetheless. In stage II, we used `TMB` (Kristensen et al. 2016) to fit the linear mixed model in (10). Code for this example as well as the IPM example is provided in an online supplement.

We examined predictive estimates  $\tilde{\theta}_i = X_i \tilde{\beta} + \tilde{u}_i$ , fixed effects  $\tilde{\beta}$ , and variance parameters ( $\log \tilde{\sigma}_1$ ,  $\log \tilde{\sigma}_2$ ) with respect to a full MCMC analysis. Both versions of the two-stage approach produced results similar to a gold standard MCMC analysis using all the data simultaneously (Figs. 1 and 2). Heuristically, for both  $\theta_i$  and  $\Sigma$ , the mean and covariance matrix obtained from stage I, version B (deterministic posterior sample), produced estimates more similar to the full-model MCMC estimates than those derived from the Bayesian MAP fitting with `mgcv`, but only slightly so. Scientific inference for any of the three methods would be identical, especially for the fixed parameters  $\beta$ .

#### 4.2. LAPWING IPM

To demonstrate how we might utilize production software, in the first stage, we fitted the ring-recovery model using `MARK` (White and Burnham 1999) via the user interface R package `RMark` (Laake 2013). `MARK` provides the MLE and the large-sample covariance matrix for  $\hat{\alpha}_1 = (\hat{\eta}_1, \hat{\gamma}_1)'$  and  $\hat{S}_1$ . For the state-space model of the female count data, we used bespoke model code written for analysis with the R package `TMB` (Kristensen et al. 2016). However, the state-space model is the simpler of the two data models in terms of programming complexity. From `TMB`, we obtain the MAP estimate and large-sample covariance matrix,  $\hat{\alpha}_2 = (\hat{\eta}_2, \hat{\gamma}_2)'$  and  $\hat{S}_2$ . Because the survival and production parameters are not well identified in the state-space model, we used vaguely informative temporary stage I prior distributions. For  $(\eta_{y0}, \eta_{y1})$  and  $(\eta_{a0}, \eta_{a1})$ , we used independent normal  $g$ -priors to induce approximately uniform priors on  $\phi_{yt}$  and  $\phi_{at}$  (Hanson et al. 2014). For the production parameters,  $(\gamma_{0\rho}, \gamma_{1\rho})$ , we used a normal  $g$ -prior such that  $\rho_t$  lies within [0.01, 2] with prior probability  $\approx 0.95$ . Based on other analyses, this prior covers a fairly broad range of values for the average number of females produced per female. For the remaining parameters, temporary priors were not used. In stage II, we used the linear mixed model in (10) along with the population prior distributions:

$$\begin{aligned} [\eta_{y0}, \eta_{y1}, \eta_{a0}, \eta_{a1}] &= N(\mathbf{0}, 10^2 \mathbf{I}), \\ [\gamma_{\lambda 0}, \gamma_{\lambda 1}] &= N(\mathbf{0}, 10^2 \mathbf{I}), \quad [\gamma_{\rho 0}, \gamma_{\rho 1}] = N(\mathbf{0}, 10^2 \mathbf{I}) \\ [\log N_{y0}, \log N_{a0}] &= N(0, 1000^2 \mathbf{I}), \quad \text{and } [\log \sigma] \propto \text{constant (improper)}. \end{aligned}$$

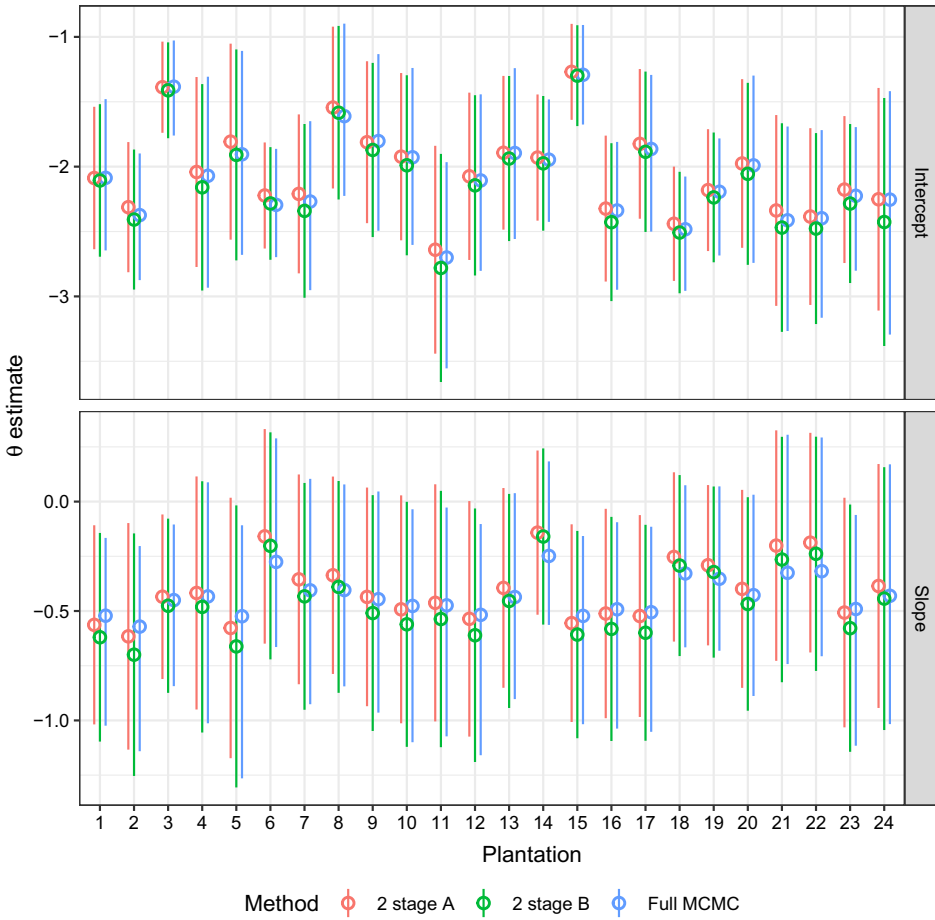


Figure 1. Random effect inference for Example I using two-stage and full MCMC methods. In the two-stage analyses, this was estimated using  $\tilde{\theta}_i = X_i \tilde{\beta} + \tilde{u}_i$ , where the tilde represents the second-stage estimates .

We also consider an additional complexity beyond the traditional IPM specification. One of the tenets of the IPM is that the multiple data models share common parameters of scientific interest, in our case the survival parameters,  $\eta$ . It seems reasonable, however, that differences in data collection procedures or model specification might imply separate “true” values of the survival parameters, that is,  $\hat{\eta}_i$  is actually estimating  $\eta_i \neq \eta$ . So, one might consider each  $[y_i | \eta, \mathcal{Y}_i]$  to be an imperfect source of information about  $\eta$ . This is how meta-analysts would traditionally envision combining these separate studies. Therefore, we also fit a model with random effects  $\theta_i \sim N(\eta, \sigma_\theta^2 I)$ . If the available data suggest common survival parameters between the two data sets, then  $\tilde{\sigma}_\theta$  will be small and all the stage II estimates of each  $\tilde{\theta}_i \approx \tilde{\eta}$ . However, if  $\tilde{\sigma}_\theta \gg 0$ , then a fixed  $\eta$  model is probably inappropriate and there are possible biases in one or both of the different data sets. Thus,  $\sigma_\theta$  serves as a regulator to assess the degree to which the data sets support a fixed  $\eta$  value. For this model, we used  $[\sigma_\theta] = \text{Exponential}(1)$  in addition to the previous prior specifications.

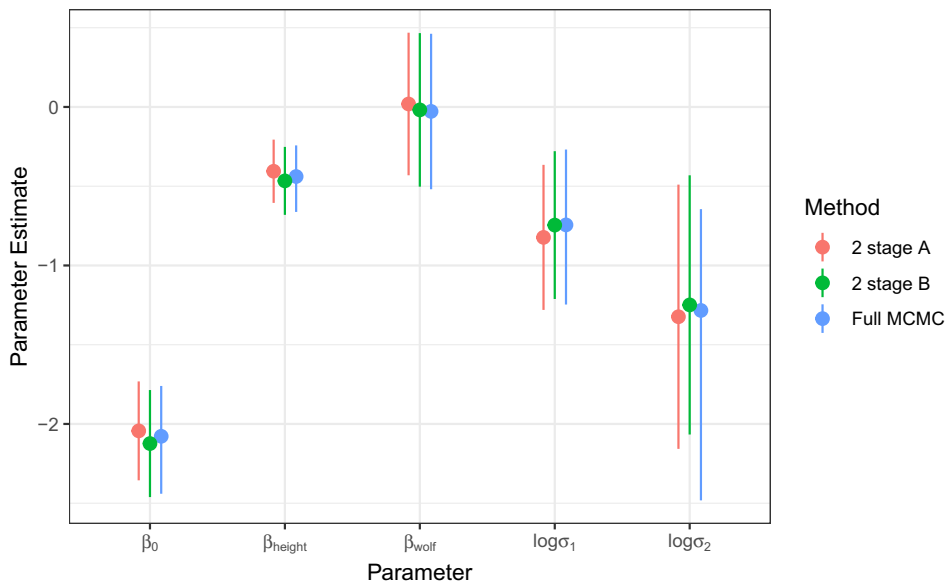


Figure 2. Fixed effect and variance parameter inference for Example I using two-stage and full MCMC methods

To judge the quality of the posterior approximation, we compared our results to those presented by [Besbeas and Morgan \(2019\)](#), hereafter, B&M. Table 2 (pg. 483) of that publication provides the results of a full joint analysis for several different models and inference methods. Our specification is most similar to their “MLE KF” model, that is a model where  $\log N_{y0}$  and  $\log N_{a0}$  are explicitly estimated. We only compare our results to that model (Fig. 3), but the other results are very similar. All of the multistage approximate fitting methods produced results very similar to the B&M results (Fig. 3). A full table of the multistage results is available in Appendix B. For the two- and three-stage fixed  $\eta$  models, estimates of  $\eta$  were virtually indistinguishable from those given by B&M, both point estimates and standard deviations. Also, the band reporting parameters,  $\gamma_1$ , are nearly identical. From stage II to stage III,  $\bar{\gamma}_{\rho 1}$  became more similar to the B&M results, while  $\bar{\gamma}_{\rho 0}$  became less similar. Although they are both close to the B&M results overall, a marked improvement in estimates from stage II to III is evident in the  $\log \sigma$ ,  $\log N_{y0}$ , and  $\log N_{a0}$  parameters, with all of them becoming nearly equivalent to the B&M estimates after stage III.

We now turn our attention to the random effect version of the model. The estimate of  $\tilde{\sigma}_\eta = 0.017$  (0.006–0.049), which is small, but not trivial. Thus, there is some evidence that each data set contains different information about the true survival process. The population-level effect of frost days on survival becomes slightly steeper and more uncertain (Fig. 3) for young animals. For adult animals, frost days have less effect on survival, but this is again, more uncertain. The increase in uncertainty and change in frost effect are plainly evident when examining survival on the natural scale (Fig. 4). Although they did not fit separate survival processes, [Goudie et al. \(2019\)](#) noted this discrepancy as well in their two-stage MCMC analysis of these same data.

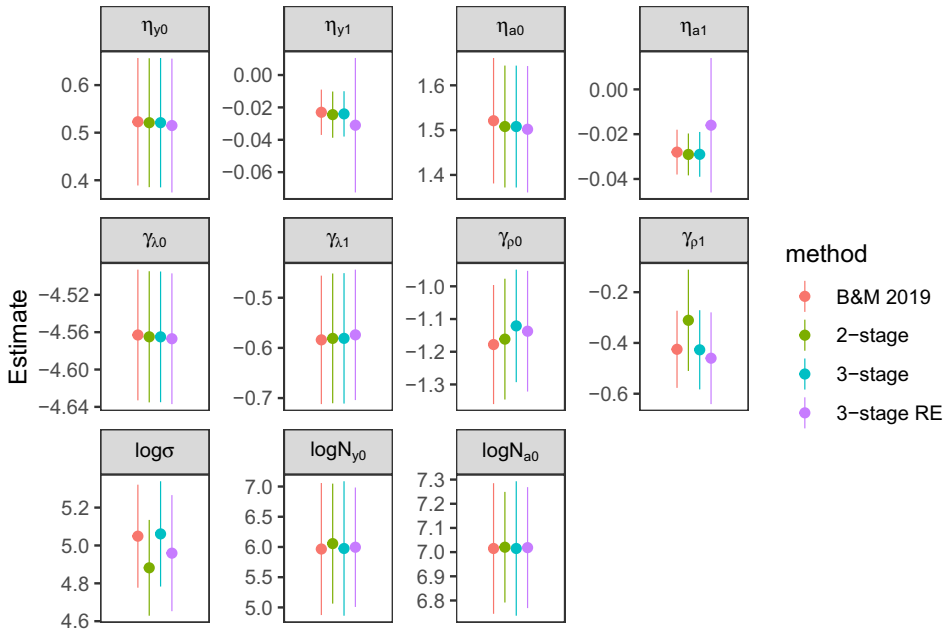


Figure 3. Parameter estimates for the lapwing IPM model. The B&M 2019 results were taken from Table 2 of Besbeas and Morgan (2019) and represent estimates from single-stage estimation of the full model. In the three-stage RE model,  $\theta_i \sim N(\eta, \sigma_\theta^2 \mathbf{I})$ . Parameter estimates are also provided in online supplement Table S1 .

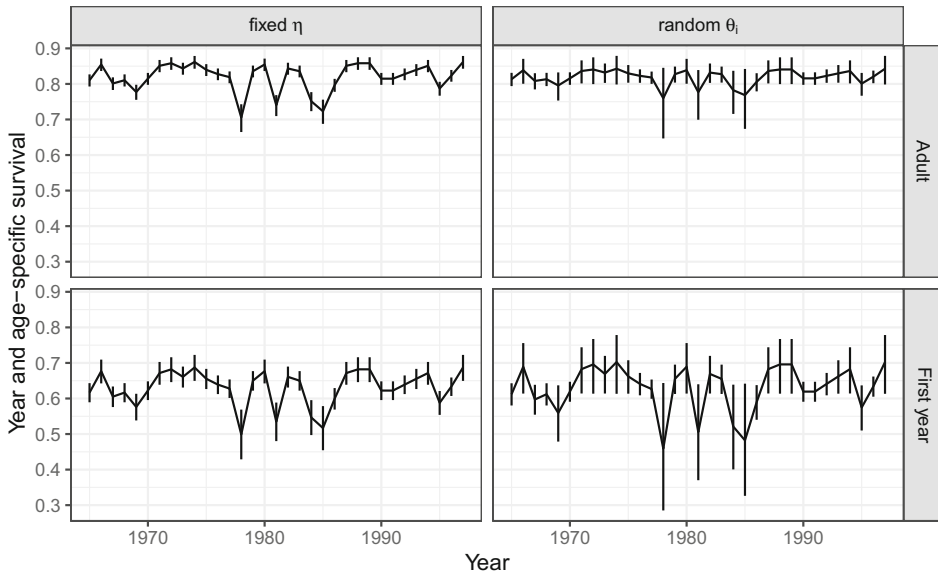


Figure 4. Estimates of year and age-specific survival for lapwings. The different age classes are represented in the rows; adult and first year. The columns illustrate estimates from the two different models. The first column is the traditional IPM where the survival parameters are common between the data sets (i.e.,  $\eta$  is fixed). In the second column, we modeled  $\theta_i \sim N(\eta, \sigma_\theta^2 \mathbf{I})$ , where  $\sigma_\theta$  is estimated in the second stage .

## 5. DISCUSSION

We have proposed a multistage method for estimating parameters in a broad class of BHMs. Our approach is computationally efficient and flexible and draws inspiration from traditional meta-analysis where individual studies using differing methods contribute unit-level parameter estimates which are then combined using a CIHM format. In stage I, practitioners can use software expressly designed for a particular data model. We use traditional GLM/GAM software in the first example, and a mixture of production software (MARK/RMark) and pseudo-coding software (TMB) in the IPM example. This increases the efficiency of the analysis in development time because software expressly designed for the data at-hand is being used. Analysts do not have to create bespoke MCMC that will often be substandard to production-level software in terms of ease-of-use and computational speed. In addition, as with other two-stage methods, the stage I analyses are “embarrassingly parallel,” so they can be executed in parallel if necessary, further increasing efficiency. In stage II, the multivariate linear mixed model is very quick to analyze because the dimension of the problem is substantially reduced from the data models. The linear mixed model represents a compression of the information in the various data sets about the parameters. Finally, the generalized delta method allows one to revisit the estimates of distinct parameters that is usually not possible in traditional meta-analysis because the raw data are not available.

This efficiency comes at a cost, however, because it relies on approximations in two places during the development, the normal approximation in stage I and the normal approximation of the stage II posterior distribution. We do not view this as a detriment because this method is designed for computationally challenging situations. These usually occur in settings with a large amount of data and hence a substantial amount of information about the parameters. That is, situations for which the large sample approximations were developed. Both of the models in our examples could be fitted readily in a single stage BHM analysis. Even so, the multistage method produced estimates very close to the “gold standard” single-stage MCMC estimates. The two-stage approach may not be optimal in all settings, thus if one is capable of fitting the full CIHM in one step that would be the recommended course of action. We suggest our two-stage approach for instances where computation or development of software to fit the full model at once is impractical.

Even though the approximations produced good results in our examples, we should examine the seriousness of these approximations and what options practitioners have for checking the degree of error introduced by these approximations. The first approximation is the Gaussian approximation of the data-level likelihood shape at the first stage. This approximation is justified by the standard large sample results that coincide for MLE and posterior modes (MAP) and means. There are some structural situations with stage I models where the assumptions underlying those results would be questionable. Gelman et al. (2013, Section 4.2) presents several scenarios where the assumptions of large sample theory fail to be met. Most of these situations are able to be fixed with appropriate reparameterization or prior selection. For example, parameters should be specified on the log or logit scale, as well as the use of the Gaussian temporary prior in (6).

If the data-level models are sufficiently regular, one may still want to assess whether the sample size is sufficient to warrant a Gaussian approximation. We note that, in a variety

of previous studies, this assumption has often worked well with little technical justification (e.g., Geyer 2005; Goudie et al. 2019; Scott et al. 2016). If the validity of the MLE or MAP estimates and Hessian derived covariance are in question, one can also use bootstrap bias and variance corrections to improve the results (Geyer 2005; Scott et al. 2016). One can also increase data size by combining the original data partitions where appropriate. To ascertain whether sample size is too small for the shape of the normal density function to be appropriate, one could also perform a MCMC analysis of some of the smallest units to see how posterior density appears. If the approximating normal posterior is used as a proposal distribution in a simple independent Metropolis–Hastings algorithm, then the acceptance rate can be used to judge the closeness of the approximation. For example, an acceptance rate of 0.9 (i.e., 90% of proposals accepted) would imply that the stage I approximation is very close to the actual stage I posterior shape. Michelot et al. (2019) use this type of diagnostic for assessing adequacy of Euler approximations for Langevin diffusion models. Because the partitions were designed such that individual data models are relatively easy to evaluate this should pose little challenge for at least a few low sample size partitions. After diagnostic checking, if a normal approximation appears inappropriate, Gelman et al. (2013) suggests a normal mixture,

$$[y_i | \theta_i] \approx \sum_{k=1}^{K_i} \omega_{ik} N(\vartheta_i | \vartheta_{ik}^*, S_{ik}^*),$$

where  $\sum_k \omega_{ik} = 1$ , and  $\{\vartheta_{ik}^*, S_{ik}^*\}$  are selected in an advantageous way in stage I. If  $\theta_i \sim N(\mu, \Sigma)$ , then we can follow the same development in Section 3.2 to obtain the stage II model that is a mixture of linear mixed models. That is, the approximate posterior would be

$$[\mu, \sigma, \eta, \gamma, u | y] \propto \prod_{i=1}^n \left( \sum_{k=1}^{K_i} \omega_{ik} N(\vartheta_{ik}^* | X_i \beta + Z_i u_{ik}, S_{ik}^*) \right) N(u | \mathbf{0}, \Sigma) [\gamma | [\mu, \sigma] | \eta].$$

Although it is no longer a traditional linear mixed model, it is still relatively easy to fit because the weights are known.

In stage II, when using the generalized delta method for estimating  $\theta$  and the stage III method for reassessing  $\gamma$ , we make the additional assumption that  $[\mu, \sigma, \eta, \gamma | y]$  is approximately normal. Again, given we have already made the assumption that the stage I posteriors are approximately normal. This new approximation is not adding a substantial new element of error to the analysis because, for a given value of the random effect variance parameters,  $\sigma$ , normal random effects, and the original normal assumption, the stage II posterior is multivariate normal (assuming normal priors on  $\beta$ ). However, if needed, then bootstrap, MCMC, or other method could be used for the stage II inference instead of the MAP and Hessian.

We developed the proposed multistage method under the assumptions of the CIHM (1); however, we can relax some of the independence assumptions in the CIHM at the data level and the method is still valid in its current form. First, the data do not need to be strictly independent at the unit level. That is, the multistage method can still be applied when the



data models can be partitioned as

$$\mathbf{y} \sim \prod_i [\mathbf{y}_i | \mathbf{y}_{-i}, \boldsymbol{\theta}_i, \boldsymbol{\gamma}_i, \boldsymbol{\eta}],$$

where  $\mathbf{y}_{-i} = \{\mathbf{y}_1, \dots, \mathbf{y}_{i-1}, \mathbf{y}_{i+1}, \dots, \mathbf{y}_n\}$ . Having a *meaningful* partition of the data is not the important part of any multistage analysis, it is the ability to partition the data-level model (Hooten et al. 2021). In stage I, these dependent data models can be individually fit in the same way.

Hierarchical Bayesian models have brought a wealth of analytical capability for extracting scientific signal out of complex and noisy data. Moreover, they provide an honest assessment of uncertainty by accounting for many different forms of stochasticity and relationships in the observable data. By assembling relatively simple components, a complex model can emerge. With the advent of modern data collection capabilities, default methods for inference have started to become obsolete. Gelfand and Ghosh (2015) note that BHMs have been a great benefit to science, but the current rate of data acquisition and model complexity are “beginning to stretch our computational capabilities” and future inference will have to be the result of “an artful mixture of model specification and approximation.” Hierarchical conditioning provides a beneficial way to build large complex models; thus, it is sensible that inferential computing for these models may also be best accomplished in a hierarchical fashion (McCaslin et al. 2021).

## ACKNOWLEDGEMENTS

The findings and conclusions of the NOAA authors in the paper are their own and do not necessarily represent the views of the National Marine Fisheries Service, NOAA. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. Funding was provided by National Science Foundation (US) (Grant No: 1927177).

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

[Received November 2020. Revised December 2021. Accepted December 2021. Published Online March 2022.]

## REFERENCES

- Berliner LM (1996) Hierarchical Bayesian time series models. In *Maximum entropy and bayesian methods*. Springer, pp 15–22
- Besbeas P, Freeman SN, Morgan BJ, Catchpole EA (2002) Integrating mark-recapture-recovery and census data to estimate animal abundance and demographic parameters. *Biometrics* 58:540–547

- Besbeas P, Morgan BJ (2019) Exact inference for integrated population modelling. *Biometrics* 75(2):475–484
- Brooks SP, King R, Morgan BJT (2004) A Bayesian approach to combining animal abundance and demographic data. *Anim Biodivers Conserv* 27(1):515–529
- de Valpine P, Turek D, Paciorek CJ, Anderson-Bergman C, Lang DT, Bodik R (2017) Programming with models: writing statistical algorithms for general model structures with NIMBLE. *J Comput Graph Stat* 26(2):403–413
- Gasparrini A, Armstrong B, Kenward M (2012) Multivariate meta-analysis for non-linear and other multi-parameter associations. *Stat Med* 31(29):3821–3839
- Gelfand A, Ghosh S (2015) Hierarchical modeling. In: Damien P, Dellaportas P, Polson NG, Stephens DA (eds) *Bayesian theory and applications*. Oxford University Press, Oxford, pp 33–49
- Gelfand AE, Sahu SK, Carlin BP (1996) Efficient parameterizations for generalized linear mixed models. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds) *Bayesian Statistics 6*. Oxford University Press, Oxford
- Gelfand AE, Smith AF (1990) Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc* 85:398–409
- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) *Bayesian data analysis*. CRC Press, Boca Raton
- Geyer C. J (2005). *Le Cam made simple: Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity*. Technical report, Technical Report 643, School of Statistics, University of Minnesota
- Goudie RJ, Presanis AM, Lunn D, De Angelis D, Wernisch L (2019) Joining and splitting models with Markov melding. *Bayesian Anal* 14:81–109
- Goutis C, Casella G (1999) Explaining the saddlepoint approximation. *Am Stat* 53(3):216–224
- Green PJ, Łatuszyński K, Pereyra M, Robert CP (2015) Bayesian computation: a summary of the current state, and samples backwards and forwards. *Stat Comput* 25(4):835–862
- Hanson TE, Branscum AJ, Johnson WO et al (2014) Informative *g*-priors for logistic regression. *Bayesian Anal* 9:597–612
- Higgins JP, Thompson SG, Spiegelhalter DJ (2009) A re-evaluation of random-effects meta-analysis. *J R Stat Soc A Stat Soc* 172(1):137–159
- Hobbs N. T, Hooten M. B (2015) *Bayesian models: a statistical primer for ecologists*. Princeton University Press, Princeton
- Hooten MB, Buderman FE, Brost BM, Hanks EM, Ivan JS (2016) Hierarchical animal movement models for population-level inference. *Environmetrics* 27:322–333
- Hooten MB, Johnson DS, Brost BM (2021) Making recursive Bayesian inference accessible. *Am Stat* 75:185–194
- Johnson DS, London JM, Kuhn CE (2011) Bayesian inference for animal space use and other movement metrics. *J Agric Biol Environ Stat* 16(3):357–370
- Kass RE, Steffey D (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J Am Stat Assoc* 84(407):717–726
- Kristensen K, Nielsen A, Berg CW, Skaug H, Bell B (2016) TMB: automatic differentiation and Laplace approximation. *J Stat Softw* 70(1):1–21. <https://doi.org/10.18637/jss.v070.i05>
- Laake J (2013) RMark: An R interface for analysis of capture-recapture data with MARK. AFSC Processed Rep. 2013-01, Alaska Fish. Sci. Cent., NOAA, Natl. Mar. Fish. Serv., Seattle, WA
- Le Cam L, Yang GL (2000) *Asymptotics in statistics: some basic concepts*. Springer-Verlag, New York, 2nd edition
- Lunn D, Barrett J, Sweeting M, Thompson S (2013) Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. *J R Stat Soc Ser C (Appl Stat)* 62(4):551–572
- McCaslin HM, Feuka AB, Hooten MB (2021) Hierarchical computing for hierarchical models in ecology. *Methods Ecol Evol* 12(2):245–254
- Mesquita D, Blomstedt P, Kaski S (2020) Embarrassingly parallel MCMC using deep invertible transformations. In *Uncertainty in artificial intelligence*. PMLR, pp 1244–1252
- Michelot T, Etienne M-P, Glogaen P (2019) The Langevin diffusion as a continuous-time model of animal movement and habitat selection. *Methods in ecology and evolution (In press)*

- 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. Vienna, Austria
- Schaub M, Abadi F (2011) Integrated population models: a novel analysis framework for deeper insights into population dynamics. *J Ornithol* 152(1):227–237
- Scott SL, Blocker AW, Bonassi FV, Chipman HA, George EI, McCulloch RE (2016) Bayes and big data: the consensus Monte Carlo algorithm. *Int J Manag Sci Eng Manag* 11(2):78–88
- Skaug HJ, Fournier DA (2006) Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Comput Stat Data Anal* 51(2):699–709
- Srivastava S, Li C, Dunson DB (2018) Scalable Bayes via barycenter in Wasserstein space. *J Mach Learn Res* 19(1):312–346
- van Beeck Calkoen ST, Kuijper DP, Sand H, Singh NJ, van Wieren SE, Cromsigt JP (2018) Does wolf presence reduce moose browsing intensity in young forest plantations? *Ecography* 41(11):1776–1787
- Van der Vaart AW (2000) *Asymptotic statistics*, vol 3. Cambridge University Press, Cambridge
- White GC, Burnham KP (1999) Program MARK: survival estimation from populations of marked animals. *Bird Study* 46(sup1):S120–S139
- Wikle CK (2003) Hierarchical models in environmental science. *Int Stat Rev* 71(2):181–199
- Wood SN (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J R Stat Soc Ser B (Stat Methodol)* 73(1):3–36

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.