Hierarchical Bayesian Survival Analysis of Age-Specific Data From Birds' Nests

Master of Science Thesis in Statistics

Niejing Willgohs

Department of Mathematics

University of Bergen

November, 2010

Foreword

After finishing my thesis and I am thinking that there are some people I have to mention because of their supports.

First I must thank my supervisor Ivar Heuch, who gave me many useful advices that helped me very much during the writing.

Then I need to thank my earlier professor Jostein Paulsen, who taught me a lot of theoretical knowledge and motivated my interest in statistics.

And last I have to thank my family, especially my wife Xi Zeng; The thesis would never be done without their endless supports.

Content

Introduction

In this thesis, I first present the grassland birds ‘data from Wells(2007) which is used by several different methods of estimating the nest survival rates. The hierarchical Bayesian method from Cao(2009) then is introduced as a new model to estimate nest-specific survival rates with double censored, left-truncated data. I compare two methods and during the comparison, cox-proportional model and intrinsic autoregressive prior are studied

In the second half of this thesis, different data analysis methods are introduced, the deviance information criterion is presented and the Bayesian method is compared with the Mayfield method.

The hierarchical Bayesian method is relatively new and is a complicated model indeed for those people who are not familiar with the Bayesian and higher dimension of integration. Nevertheless, it is still a valuable statistical tool. The deviance information criterion is a new method of analyses data; users could choose the different priors in order to get different estimating results, therefore it is very applicable in the statistical world.

1. Information in Wells et al. (2007)

1.1 Practical information in Wells et al.(2007)

In Wells(2007), the main research goal is to explain patterns of survival for two species of grassland birds during the post fledging period in southwestern Missouri. To achieve this goal, Wells(2007) observed the two species of birds, collected the necessary information about them and then used certain statistical model to evaluate the data. Wells(2007) has got the conclusion that the probability for survival for these birds are mainly depended on their body condition (body mass, the heavier the better).

Wells(2007) monitored each nest every 3-4 days until 2-3 days before the bird are fledging and then changed it to daily observations. During this procedures, the birds were also being attached by a band on their legs . Wells(2007) also weighted each bird and attached transmitter on every bird for radio tracking.

There are three outcomes for Wells (2007) in the data collection period :

a. Transmitter was recovered from dead bird.
b. The battery of the transmitter ran out. (50-60 days)
c. The signal of the transmitter could not be located inside the study area.

During the statistical analyses, Wells(2007) have developed some important covariates which were very essential for choosing the right model. There are five different covariates in total, two biological, two temporal and one spatial. The biological covariates are body mass and natal brood size, because according to earlier researches for those two species of birds, the body mass and the number of siblings are the key factor for survival; The seasonal and yearly effect on the study environment are the two temporal covariates, which means the weather and the temperatures have their effects of predation and predator activities in the study area; And Wells(2007) also assume the potential differences in landscape composition may affect the birds' survival, this is the spatial covariate.

Wells(2007) stated that they used Cox proportional hazard models to estimate survival as a function of multiply covariates. The main reason of doing this is that they were able to observe and obtain a continuous measure of time until the birds' deaths at least daily.

Under the statistical analyses, Wells(2007) used days from fledging as the unit of time and determined the number of days of risk for each bird by assuming that every individual was at risk until they observed a certain fate or censored an individual based on some assumptions. To insure their data that only included the birds were successfully fledged, they made some restrictions for their censorship:

1. They removed those individuals who fledged but then without at least one detection after the fledging.
2. They removed those individuals who died within the first few days after fledging because they were accidently stepped under the radio-tracking process.
3. They assumed the individual was at risk at its last shown location if there was a time gap for the bird between the last visual detection and the determination of its fate.
4. They used the last confirmed visual observation as the date of censorship if the individual was missing over 30 days.

Before using Cox proportional hazard model to analysis the data, Wells(2007) tested first on several assumptions related to the statistical analyses of multiple brood members in their observation sample. They wanted to test whether the independence of the survival probabilities were related to the size of brood from the same nest or not, if the hypothesis is false, they assume it would lead to overdispersion to the whole data and underestimates of variance. Wells(2007) used a modified chi-square test to test the assumption of brood independence, X^2/df, where X^2 was the sum of partial chi-square values ($[\text{observed} - \text{expected}]^2/[\text{expected}]$). For example, for a brood size of two, there are three potential outcomes: complete failure, complete success and partial success. They too calculated the expected values for brood loss at each level of brood size as $p^r(1-p)^{n-r}$, where p is the survival rate , r is the number of individuals surviving to independence and n is the brood size. The result from their data showed no evidence of dependence in survival among brood siblings for both species, and therefore they considered that individual survival probabilities were independent.

Causes of mortality were part of the observations in Wells(2007). The main cause was predation in the study area, and rest four minor causes were deaths related to general equipment(farm and management), death related to research accident, death related to weather and temperatures and death from unknown causes(natural death). To estimate daily mortality rates, they used the number of mortalities from each cause and the total number of exposure days for the birds were at risk during the study period. And at last , they combined the mortality rates with brood size,

to estimate if the fate of multiple individuals from the same brood were not independent, or caused by the same predator (practically for simultaneously predation or within the short interval).

1.2 Survival function and Cox proportional hazard model:

In order to understand the data analysis from Wells (2007), it is necessary to explain briefly about the Survival Function and the Cox proportional hazard model. Let T represent survival time. We regard T as a random variable with cumulative distribution function $P(t) = Pr(T \leq t)$ and probability density function $p(t) = \frac{(dP(t))}{dt}$. Then the survival function $S(t)$ is the complement of the distribution function, $S(t) = Pr(T > t) = 1 - P(t)$. A fourth representation of the distribution of survival times is the hazard function, which assesses the instantaneous risk of demise at time t, conditional on survival to that time:

$$h(t) = \lim_{\Delta t \to 0} Pr\left[\left((t \leq T \leq t + \Delta t)\middle| T \geq t\right)\right]$$
$$= \frac{f(t)}{S(t)}$$

Modeling of survival data usually employs the hazard function or the log hazard. For example, assuming a constant hazard, $h(t) = v$, implies an exponential distribution of survival times, with density function $p(t) = v \cdot exp(-vt)$.

Normally the survival analysis examines the relations between the survival distribution and its corresponding covariates. This examination commonly uses a linear-like model for the log hazard. For example, a parametric model based on the exponential distribution could be written like this:
$log\ h_i(t) = \alpha + \beta_1 x_{i1} + \beta_2 x_{ik} + \cdot\ \cdot\ \cdot + \beta_k x_{ik}$ or equivalently be written as: $h_i(t) = exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{ik} + \cdot\ \cdot\ \cdot + \beta_k x_{ik})$

that is, as a linear model for the log-hazard or as a multiplicative model for the hazard. Here, i is a subscript for observation, and the x's are the covariates. The constant α in this model represents a kind of log-baseline hazard, since $\log h_i(t) = \alpha$ [or $h_i(t) = \exp(\alpha)$] when all of the x's are zero.

The Cox model, in contrast, leaves the baseline hazard function $\alpha(t) = \log h_0(t)$ unspecified:

$log\ h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{ik} + \cdot\ \cdot\ \cdot + \beta_k x_{ik}$ or equivalently be written as: $h_i(t) = h_0(t)\ exp(\beta_1 x_{i1} + \beta_2 x_{ik} + \cdot\ \cdot\ \cdot + \beta_k x_{ik})$

Consider, now, two observations i and i_ that differ in their x-values, with the corresponding linear predictors

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{ik} + \cdot\ \cdot\ \cdot + \beta_k x_{ik}$$

and

$$\eta_{i\_} = \beta_1 x_{i\_1} + \beta_2 x_{i\_k} + \cdot\ \cdot\ \cdot + \beta_k x_{i\_k}$$

The hazard ratio for these two observations,

$$h_i(t) \,/\, h_{i\_}(t) = h_0(t)exp(\eta i) \,/\, h_0(t)exp(\eta i\_) = exp(\eta i\,)/exp(\eta i\_)$$

is independent of time t. Consequently, the Cox model is a proportional-hazards model.

1.3 Results and brief analyses from Wells (2007)

The final sample size in Wells (2007) for survival analysis was:

Dickcissels(from 69 broods)		**Meadowlarks(from 30 broods)**	
Year 2002	40	Year 2002	17
Year 2003	42	Year 2003	43
Year 2004	73	Year 2004	48
Total	155	Total	107

The confirmed or estimated mortality rate for individuals:

44% of individual Dickcissels (n=69), 60 out of 69 died within the first week of fledging.
28% of individual Meadowlarks (n=30), 27 out of 30 died within the first week of fledging.

Other important observing results:

Average body mass at the time of transmitter attachment	Age associated with individuals confirmed or assumed dead	Average body mass at the time of transmitter attachement	Age associated with censored individuals
14.3g-15.1g (range: 9g – 27g) Dickcissels	2.9days-4.1days (range: 0day-29days) Dickcissels	14.9g-15.5g (range:7g – 22g) Dickcissels	29.5days-32.1days (range: 3days-58days) Dickcissels
42.4g-45.0g (range: 43.7g – 46.3g) Meadowlarks	4.2days-6.4days (range: 5days-7days) Meadowlarks	44.9g-46.5g (range: 29g – 59g) Meadowlarks	38.5days-41.5days (range: 12days-72days) Meadowlarks

These results above in Wells (2007) had proved the importance of body condition on the probability of individual survival. In other word, Wells' study had estimated or assumed that the heavier individuals had an advantage over the lighter individuals.

We know that by using the Cox model, there might be a lot of more covariates that do not affect the hazard rate. Therefore it is desirable to work with as less as covariates as possible. However, the results from Wells(2007) show that there are still some covariates included. The reason for that spurious variables are often included(especially AIC) is that because as far as model performance is concerned, it is a lot worse to exclude an important variable than to include a spurious variable. As a result, in the evolutionary process, a model that contains all the important variables will have a higher fitness score than a model that does not

contain all the variables, because all other spurious variable in the model will be regarded as important.

Wells(2007) used AIC_c to test the pattern of survival. For AIC, the lower fitness score means the better. Generally, the AIC is:

$$AIC = 2p - 2\ln(L)$$

Where p is numbers of parameters and L is the maximized likelihood function for the estimated model (pattern in Wells). And as we mentioned above, AIC_c is used here, which is AIC with second order correction for small sample size:

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$$

AIC_c will converge to AIC when n becomes large. (Results are shown in TABLE 1 and TABLE 2 in Wells(2007))

1.4 Difference between Wells et al. (2007) and Cao et al.(2009)

Wells(2007) used each individual bird as the observation unit and Cao (2009) used each nest as the observation unit.

Cao(2009) used the same data that Wells(2007) had collected, but analyses by using another statistic method: the Hierarchical Bayesian approach.
The main idea in Cao(2009) is they proposed a Bayesian hierarchical model, and this model is easier than the Cox model in Wells(2007) to be applied in the nest survival study with unknown nest ages, double interval censored and left-truncated data, and some other nest-specific covariates.

Cao(2009) pointed that the Bayesian model does not need continuous time measure as the Cox model, it only needed the number of days that a nest is required to survive and the following information of every observed nest:

a. The outcome of the nest, either success or failure
b. The date of the first encounter
c. The date of the second-to-last revisit
d. The date of the last revisit
e. The specific value of covariates

Cao(2009) assumed that let J is total days required for a nest to survive successfully, n is the total observed nests, and k is the kth observed nest. Then they defined U_k and T_k to be the nest age at the first discovery and the nest age at the outcome of the kth nest respectively, and they are both positive discrete variables, measured in days. Let $T_k = J + 1$ if the nest is a success. Under an irregular visiting schedule, it is obviously helpful that we let $[U_{Lk}\ U_{Rk}]$ to be the lower and upper interval for U_k, and let $[T_{Lk}\ T_{Rk}]$ to be the lower and upper interval for T_k . This is so called double interval censored data.
The nests would be categorized into three different groups:

1. Undiscovered nests, which means U > J .
2. Truncated nests that failed before thery were even discovered: T < U <= J .
3. Observed nests, U <= T

Because only observed nests were recorded in the data, therefore the nest survival data were truncated.

And Cao(2009) let:

$$\delta_i = P(U = i | U \leq J) \text{ for } i = 1, 2, \ldots\ldots, J,$$

$$q_{jk} = P(T_k = j) \text{ for } j = 1, 2, \ldots\ldots, J + 1;\ k = 1, 2, \ldots\ldots, n.$$

δ_i is the conditional probability that nest age at first encounter is i given that the nest is discoverable. $q_{jk}(j \leq J)$ is the kth nest's failure probability at age j, and the nest success probability for the kth nest is $q_{(J+1)k}$.(Because for the failure probabilities, both the age effect and the nest-specific covariates are different for each nest, so we have to consider this fact and therefore each q is different, that's the reason we mark a second lower index for each q.) Then the following equations are:

$$\delta_1 + \delta_2 + \ldots\ldots + \delta_J = 1$$

$$q_{1k} + q_{2k} + \dots \dots + q_{Jk} + q_{(J+1)k} = 1$$

Consider the nests are discovered at age i and would either be failed or be succeeded at age j, then the probability for these nests are $\delta_i q_j$. If Cao(2009) set the exact discover age to be U_k and set the exact outcome age to be T_k for the kth nest, then the kth nest has probability

$$\delta_{U_k} q_{T_k k} \Big/ \sum_{j \geq i} \delta_i q_{jk}.$$

with the fact that the nest is already active when it is first found in the study area. The denominator above could be rewritten as:

$$\sum_{j \geq i} \delta_i \; q_{jk} = \sum_{i=1}^{J} \sum_{j=i}^{J+1} \delta_i \; q_{jk}$$

We will also define

$$\boldsymbol{\delta} = (\delta_1 , \delta_2 , \delta_3 , \dots \dots , \delta_J)'$$

and

$$\mathbf{q} = (q_{11} , \dots \dots q_{(J+1)1} , \dots \dots q_{1n} , \dots \dots q_{(J+1)n})'$$

Therefore the Cao(2009) concluded that the likelihood function of $\boldsymbol{\delta}$ and $\mathbf{q}$ given observed data and variables is:

$$\boldsymbol{L}(\boldsymbol{\delta} , \mathbf{q} \text{ ; data, variables}) = \prod_{k=1}^{n} \frac{\delta_{U_k} q_{T_k k}}{\sum_{j \geq i} \delta_i q_{jk}}$$

Cao(2009) also introduced two variables to help people understand the double interval censoring: Let Z_{1k} be the number of days from the encounter of the *k*th nest to its second-to-last visit and let Z_{2k} be the number of days for the kth nest from the second-to-last visit to its last visit. If the nest is observed to be a failure, then:

$$1 \leq U_k \leq J - Z_{1k} \text{ and } U_k + Z_{1k} \leq T_k \leq \min(U_k + Z_{1k} + Z_{2k} , J).$$

If the nest is observed to be a success, then:

$$J - Z_{1k} - Z_{2k} \leq U_k \leq J - Z_{1k} \quad \text{and} \quad T_k = J + 1.$$

Cao(2009) have given us a simple example to help us understand the setting above. It supposed that $Z_{1k} = J - 4$ and $Z_{2k} = 2$, then

$$J - Z_{1k} = J - J + 4 = 4,$$
$$U_k + Z_{1k} = U_k + J - 4,$$
$$U_k + Z_{1k} + Z_{2k} = U_k + J - 2$$

and

$$J - Z_{1k} - Z_{2k} = 2,$$
$$J - Z_{1k} = J - J + 4 = 4,$$

so if we put the results above back to the inequality for U_k and T_k , we have:

$$1 \leq U_k \leq 4 \text{ , } U_k + J - 4 \leq T_k \leq \min(U_k + J - 2, J)$$

for the nest is observed to fail and :

$$2 \leq U_k \leq 4 \text{ , } T_k = J + 1$$

for the nest is observed to succeed.

From here Cao(2009) used a set V_k which is supported from (U_k , T_k) that are defined from the inequalities above. It is said that V_k then is a set of encounter and termination ages that could not be cancelled out by the observed data of the *k*th nest. With (u_0, t_0) is a possible realization in the V_k to determine the probability mass function:

$$P(U_k = u_0, T_k = t_0) = \frac{\delta_{u_0} q_{t_0 k}}{\sum_{i,j \in V_k} \delta_i \, q_j k} \text{ ,}$$

And Cao (2009) defined a multinomial logit transformation, which is:

$$\log \frac{\delta_i}{\delta_1} = E_i \text{ , for i = 2, 3, ... , J,}$$

$$\log\frac{q_{jk}}{q_{(J+1)k}} = A_j + x_k' \beta \text{ , for j = 1, 2, ... , J .}$$

where E_i is the age effect on the encounter probabilities and A_j is the age effect on the failure probabilities, x_k is the vector of covariates the β is the vector of regression parameters. Also it is easy to see that this transformation has given the right hand side's parameters the range from minus infinity to plus infinity (-∞. + ∞).

With some calculations we have the followings:

$$\delta_i = \frac{e^{E_i}}{1+\sum_{i=2}^{J} e^{E_i}} \text{ , for i = 2, 3, ..., J ,}$$

$$\delta_1 = \frac{1}{1+\sum_{i=2}^{J} e^{E_i}} ,$$

$$q_{jk} = \frac{e^{Aj + x_k' \beta}}{1+ \sum_{j=1}^{J} e^{Aj + x_k' \beta}} \text{ , for j = 1, 2, ..., J ,}$$

$$q_{(J+1)k} = \frac{1}{1+ \sum_{j=1}^{J} e^{Aj + x_k' \beta}} .$$

2. Information in Cao et al. (2009)

2.1 Introduction on the Hierarchical Bayesian Method:

According to Bayes'theorm, we have the following conditional probability:

$$P(H|E) = P(E|H)P(H) / P(E)$$

where

- H represents a specific hypothesis, which may or may not be some null hypothesis.
- P(H) is called the prior probability of H that was inferred before new evidence, E, became available.
- P(E | H) is called the conditional probability of seeing the evidence E if the hypothesis H happens to be true. It is also called a likelihood function when it is considered as a function of H for fixed E.
- P(E) is called the marginal probability of E: the a priori probability of witnessing the new evidence E under all possible hypotheses.
- P(H | E) is called the posterior probability of H given E.

And the hierarchical Bayes method is a useful and powerful tool for expressing the rich statistical models that could more fully show many given problems than a simper model could. In other words:

For given data **x** and parameter **β**, the simple Bayesian analysis will start with a prior probability p(**β**) and likelihood p(**x**|**β**) to calculate a posterior probability p(**β**|**x**) by using its relation to p(**x**|**β**) p(**β**).

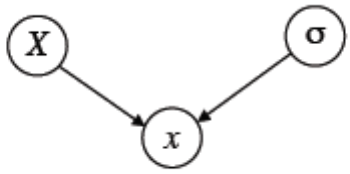
Usually the prior on **β** depends on another parameter **y** that are not being noticed in the likelihood. Then we must replace a prior p(**β**) with a prior p(**β**|**y**), and then the a posterior probability could be rewritten as :

p(**β**,**y**|**x**) related to p(**x**|**β**) p(**β**|**y**)p(**y**)

This is the simplest example for hierarchical Bayesian model. Therefore we know the basic idea in a hierarchical model is that when you look at the likelihood function, and decide on the right

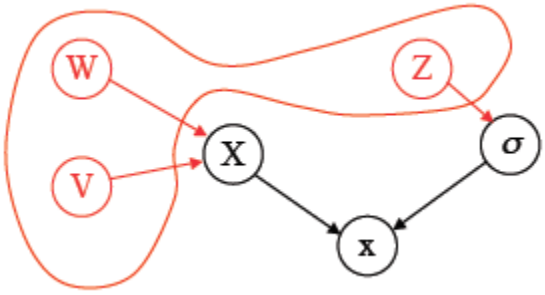
priors, it may be appropriate to use priors that themselves depend on other parameters not mentioned in the likelihood. These parameters themselves will require priors, which themselves may (or may not) depend on new parameters. Eventually the process terminates when we no
longer introduce new parameters.

Two sample illusions to show the simple Bayesian model and the hierarchical Bayesian model by using DAG(Directed Acyclic Graph):
Sample 1



x is stochastically dependent on X and σ in this model above.

Sample 2



The new red part of the above diagram indicates the new hierarchical structure, and we can clearly find out that W and V are not going to be part of the likelihood.

2.2 Priors

2.2.1 Something about Inverse gamma prior

The probability density function for inverse gamma distribution is :

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\left(\frac{1}{x}\right)^{\alpha+1}\exp\left(\frac{-\beta}{x}\right)$$

Where we have two parameters α and β, α is the shape parameter and β is the scale parameter. It is called inverse gamma because if X ∼ gamma(α, β), then 1 / X ∼ inv-gamma(α, 1/β). Let Y = 1 / X, with application from the transformation theorem, we will get:

$$f_Y(y) = f_X(1/y) \cdot \left|\frac{d}{dy}y^{-1}\right|$$
$$= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \cdot y^{-\alpha+1} \cdot \exp(-\beta y)\, y^{-2}$$
$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot y^{-\alpha+1} \cdot \exp(-\beta/y)$$

And for moments of inverse gamma, we could calculate for X ∼ inv-gamma(α, β) and if α > n:

$$E(X^n) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \int_0^{\infty} x^n x^{-\alpha-1}\exp(-\beta/x)\, dx$$
$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \int_0^{\infty} x^{n-\alpha-1}\exp(-\beta/x)\, dx$$
$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \int_0^{\infty} x^{n-\alpha-1}\exp(-\beta/x)\, dx \cdot \frac{\beta^{(\alpha-n)}}{\beta^{(\alpha-n)}} \cdot \frac{\Gamma(\alpha-n)}{\Gamma(\alpha-n)}$$
$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha-n)}{\beta^{\alpha-n}} \cdot \int_0^{\infty} \frac{1}{x^{(\alpha+n)+1}} \frac{\beta^{(\alpha-n)}}{\Gamma(\alpha-n)} \exp(-\beta/x)\, dx$$
$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha-n)}{\beta^{\alpha-n}} \cdot 1$$
$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha-n)}{\beta^{\alpha-n}}$$
$$= \frac{\beta^{n} \cdot \Gamma(\alpha-n)}{(\alpha-1)\cdots(\alpha-n)\Gamma(\alpha-n)}$$
$$= \frac{\beta^{n}}{(\alpha-1)\cdots(\alpha-n)}$$

It is easy for us now to get the expectation and variance from here:

$$E(X) = \frac{\beta}{\alpha-1},\ E(X^2) = \frac{\beta^2}{(\alpha-1)(\alpha-2)}, \text{ and}$$
$$Var(X) = E(X^2) - (E(X))^2 = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}.$$

Inverse gamma distributions are often used to be a conjugate priori in Bayesian studies when likelihood is related with exponential families. For example, if we have an observation with $X|\mu \sim$ exponential(μ), and μ is an inverse gamma

distributed, we will get the posterior distribution on μ given X = x is proportional to:

$$\frac{1}{\mu}\exp(-x/\mu)\frac{1}{\mu^{\alpha+1}}\exp(-\beta/\mu) = \frac{1}{\mu^{\alpha+2}}\exp(-(\beta+x)/\mu)$$

2.2.2 Intrinsic autoregressive priors or IAR(2) prior

Before we continue to discuss about the selecting of the priors for the hierarchical Bayesian model, we have to make some explanations on the term of IAR(2) prior, or so called intrinsic autoregressive priors.

Clayton(1994) defined IAR(2) prior as: an autoregressive prior specification for the baseline rates, in which the expected value for each $\lambda_0^{(t)}$ is predicted by a log-linear extrapolation from its two immediate predecessors, $\lambda_0^{(t-1)}$ and $\lambda_0^{(t-2)}$, plus a random perturbation $\varepsilon^{(t)}$. In the mathematical form, we could write this like:

$$\log \lambda_0^{(t)} = 2 \log \lambda_0^{(t-1)} - \log \lambda_0^{(t-2)} + \varepsilon^{(t)},\ t = 1, 2, \ldots \ldots, T \text{ and } t > 2.$$

And the side condition is that $\varepsilon^{(t)}$ is not too large, with:

$$\varepsilon^{(t)} \sim N(0, \sigma^2)$$

The hyperparameter σ^2 means the smoothness; the small values allow baseline rate to be smoother, while the large values allow rough variation. If the value of σ is 0, which tells that a log-linear relationship between baseline rates and time. Gelman(2006) says that if σ^2 has an inverse-gamma prior distribution, then the conditional posterior distribution is also inverse-gamma. This is a very good choice for the hierarchical Bayesian model, and indeed from Cao(2009), $\sigma^2 = \tau$ is assmumed to be inverse-gamma prior.

2.2.3 Specification of Prior Hyper-parameters

Cao(2009) assumed many things, and for the hyper-parameter specification, it set (a_i, b_i) from the inverse gamma prior for τ_i to be (2.0, 1.0), to give inverse gamma prior an infinite variance. And Gelman(2006) recommended that uniform priors on $\tau^{\frac{1}{2}}$ and τ itself could be useful for hierarchical variance parameters, it stated that 'in fitting hierarchical models, we recommend starting a non-informative uniform prior density on standard deviation parameter'. And during the simulation and data analysis in Cao(2009), the two choices of uniform priors are resulting almost the same outcome.

Variance for β is also a hyper-parameter, we use s_β to notify. Cao(2009) set s_β = 10 to serve as large variance. Gelman(2006) said that for the inverse-gamma(a,b) family of non-informative prior distribution, if the variance parameter is too small(near zero), the result will be very sensitive. And Cao (2009) also examined different values of $s_\beta's$, and little inference sensitivity were found after a reasonable large s_β. Therefore, a normal prior with zero as mean and 10 as variance is chosen as a non-informative prior for the regression parameters

2.3 All about the full conditional posterior distributions

There are different types of priors in Bayesian method; usually we have informative priors and uninformative priors. An informative priors could be explained from its name, this kind of priors have definite information about variables. A simple example is that a prior distribution for the people died in traffic accident next year, the reasonable way to estimate is that we could make the prior to be a normal distribution with expected value equal to this years' death from traffic accident and the variance equal to a fixed value (an average value we choose from year-to-year traffic accident death variance). An uninformative prior expresses vague or general information about a variable, it could express the variable's information such as the variable is less than average or the variable is positive.

Three priors have been pointed out in Cao (2009), they are:

$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots\ldots, \beta_p)'$, $\mathbf{A} = (A_1, A_2, \ldots\ldots, A_J)'$, and $\mathbf{E} = (E_2, E_3, \ldots\ldots, E_J)'$

And Cao assumed that β's are independent and one stage normal prior, which is written as:

$$\beta_i \sim N(0, s_\beta), \quad i = 1, 2, \ldots\ldots, p,$$

where β is the vector of regression parameters and we set s_β to be a fixed value.

A and **E** are the so called second-order difference IAR(2) priors on the age effects of the nests, and the nest survival curse is mainly estimated by the nest age. **A** and **E** are basically the second-order random walk smoothness priors, written as:

$$A_j = 2A_{j-1} - A_{j-2} + \varepsilon_j, \; j = 3, \ldots\ldots, J$$

A is the age effect on the failure probabilities and **A** prior assumes that there is an unknown smooth function fits the nest survival curve.

$$E_j = 2E_{j-1} - E_{j-2} + \theta_j, \; j = 4, \ldots\ldots, J$$

E is the age effect on the encounter probabilities and we also assume smooth nest encounter probabilities.

With i.i.d. Gaussian errors:

$$\varepsilon_j \sim N(0, \tau_1), \text{ and } \theta_j \sim N(0, \tau_2),$$

And the diffuse priors (Diffuse prior definition: In Bayesian inference, a prior probability density function that reflects little or no information regarding the value of an unknown parameter):

$$p(A_1) \propto 1, \quad p(A_2) \propto 1,$$

The IAR(2) priors means that A_j and E_j are depend on its two immediate neighbors, this also tells us that the estimation could borrow strength and the

estimated survival curve is going to be smooth. The IAR(2) priors have the following density function written in the vector format:

For **A** prior, $[\mathbf{A} \mid \tau_1] \propto 1/(\tau_1)^{(J-2)/2} \cdot \exp(-1/2\tau_1 \cdot \mathbf{A}' \mathbf{V_A} \mathbf{A})$,

And for **E** prior, $[\mathbf{E} \mid \tau_2] \propto 1/(\tau_2)^{(J-3)/2} \cdot \exp(-1/2\tau_2 \cdot \mathbf{E}' \mathbf{V_E} \mathbf{E})$,

Where $\mathbf{V_A}$ and $\mathbf{V_E}$ are the matrixes which could be written as:

$$\mathbf{V_A} = \mathbf{C^T C}\ ,\ \mathbf{V_E} = \mathbf{D^T D}\ ,$$

C and **D** are tridiagonal matrixes with constant diagonal elements (In linear algebra, a **tridiagonal matrix** is a matrix that is "almost" a diagonal matrix. To be exact: a tridiagonal matrix has nonzero elements only in the main diagonal, the first diagonal below this, and the first diagonal above the main diagonal).

In Cao(2009), it also states that $(\mathbf{V_A} / \tau_1)$ and $(\mathbf{V_E} / \tau_2)$ are two IAR(2) precision matrixes, with $\mathbf{V_A}$ has rank J – 2 and $\mathbf{V_E}$ has rank J – 3. Additionally, Cao(2009) defined that the IAR(2) priors are improper.

The variance parameters τ_1 and τ_2 are controlling the degree of smoothness of the survival and encounter curves, and they are assumed to be inverse gamma priors, which are written as:

$$\tau_i \sim IG(a_i, b_i)\ ,\ i = 1, 2,$$

Where a_i and b_i (i = 1, 2) are fixed values.

Example of calculations about priors for j = 1, 2, 3 (J=3):

$$A_j = 2A_{j-1} - A_{j-2} + \varepsilon_j$$

$$\varepsilon_j = A_j - 2A_{j-1} - A_{j-2}$$

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} = \begin{pmatrix} A_1 \\ A_2 - 2A_1 \\ A_3 - 2A_2 + A_1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix}$$

the matrix C^T is $\begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & -2 & 1 \end{pmatrix}$, C then is $\begin{pmatrix} 1 & -2 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix}$

$$\mathbf{V_A} = \mathbf{C^T C} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & -2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & -2 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -2 & 1 \\ -2 & 5 & -4 \\ 1 & -4 & 6 \end{pmatrix}$$

because we have $[\ \mathbf{A} \mid \tau_1\ ] \propto 1 / (\tau_1)^{(J-2)/2} \cdot \exp(-1/2\tau_1 \cdot \mathbf{A}' \mathbf{V_A}\, \mathbf{A})$, so set in all the numbers and we will get :

$$[\ \mathbf{A} \mid \tau_1\ ] \propto 1 / (\tau_1)^{1/2} \cdot \exp(-1/2\tau_1 \cdot \mathbf{A}' \mathbf{V_A}\, \mathbf{A})$$

Where $\mathbf{A}' \mathbf{V_A}\, \mathbf{A}$ will be :

$$(A_1 \quad A_2 \quad A_3) \cdot \begin{pmatrix} 1 & -2 & 1 \\ -2 & 5 & -4 \\ 1 & -4 & 6 \end{pmatrix} \cdot \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix}$$

$$= 6A_3^2 + 5A_2^2 + A_1^2 - 8A_3A_2 - 4A_2A_1 + 2A_1A_3$$

Take a new example for J = 4 and we have:

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix} = \begin{pmatrix} A_1 \\ A_2 - 2A_1 \\ A_3 - 2A_2 + A_1 \\ A_4 - 2A_3 + A_2 \end{pmatrix}$$

The matrix C^T is:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \end{pmatrix}$$

And the matrix C is:

$$\begin{pmatrix} 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$\mathbf{V_A} = \mathbf{C^T C}$ will be:

$$\begin{pmatrix} 1 & -2 & 1 & 0 \\ -2 & 5 & 4 & 1 \\ 1 & -4 & 6 & -4 \\ 0 & 1 & -4 & 6 \end{pmatrix}$$

Then for $[\ \mathbf{A} \mid \tau_1\ ] \propto 1 / (\tau_1) \cdot \exp(-1/2\tau_1 \cdot \mathbf{A}' \mathbf{V_A}\, \mathbf{A})$

Where $\mathbf{A}' \mathbf{V_A}\, \mathbf{A}$ will be: $6A_3^2 + 5A_2^2 + A_1^2 - 8A_3A_2 - 4A_2A_1 + 2A_1A_3 + 6A_4^2 - 8A_3A_4 + 2A_2A_4$

2.3.1 The full conditional posterior distribution of the encounter age effect E

The full conditional posterior distribution is proportional to the product of the likelihood function and the parameter's prior. As we know from before, the joint prior for $\mathbf{E} = (E_2, E_3, \ldots\ldots, E_J)'$ is an IAR(2) prior, then the prior distribution of E_l ($l = 2, 3, \ldots\ldots, J$) is a normal distribution with:

$$E_{E_l} = -\frac{\sum_{j \neq l} v_{lj} E_j}{v_{ll}}, \quad Var_{E_l} = \frac{1}{v_{ll}}$$

where v_{ll} is the element of the precision matrix $\frac{\mathbf{V_E}}{\tau_2}$.

We now rewrite the prior for E_l, as we noted before, we have:

$$[\ \mathbf{E} \mid \tau_2\ ] \propto 1 / (\tau_2)^{(J-3)/2} \cdot \exp\left(-1/2\tau_2 \cdot \mathbf{E}' \mathbf{V_E}\, \mathbf{E} \right)$$

When we think about the posterior, we only take consideration on the part of the prior that proportional to the posterior, which here is the exponential part:

$$\exp\left(-1/2\tau_2 \cdot \mathbf{E}' \mathbf{V_E}\, \mathbf{E} \right)$$

We take a close look, and we find out that:

$$-\frac{1}{2\tau_2}\ \mathbf{E}'\mathbf{V_E}\mathbf{E} = -\frac{V_E}{2\tau_2}\ \mathbf{E}'\mathbf{E}$$

$$= -2v_{ll}\ (\mathbf{E_l} - E_{E_l})'\ (\mathbf{E_l} - E_{E_l})$$

$$= -\frac{1}{2Var_{E_l}}\ (\mathbf{E_l} - E_{E_l})'\ (\mathbf{E_l} - E_{E_l})$$

$$= -\frac{1}{2Var_{E_l}}\ (\mathbf{E_l} - E_{E_l})^2$$

$$= -\frac{(\mathbf{E_l} - E_{E_l})^2}{2Var_{E_l}}$$

$$l = 2, 3, \ldots\ldots, J,$$

And we have the likelihood:

$$\boldsymbol{L}\ (\delta, q; \text{data}, \text{variables}) = \prod_{k=1}^{n} \frac{\delta_{U_k} q_{T_k k}}{\sum_{j \geq i} \delta_i q_{jk}}$$

so the conditional posterior distribution of encounter effect E_l given parameters is:

$$[E_l \mid .] \propto \prod_{k=1}^{n} \frac{\delta_{U_k} q_{T_k k}}{\sum_{j \geq i} \delta_i q_{jk}} \cdot \exp\{ -\frac{(\mathbf{E_l} - E_{E_l})^2}{2Var_{E_l}} \}$$

$$\propto \prod_{k=1}^{n} \frac{(I(U_k = l) e^{E_l})}{\sum_{j \geq i} e^{E_i} q_{jk}} \cdot \exp\{ -\frac{(\mathbf{E_l} - E_{E_l})^2}{2Var_{E_l}} \}$$

$$\propto \frac{e^{f_l E_l}}{\prod_{k=1}^{n} (c_{lk} e^{E_l} + d_{(-l)k})} \cdot \exp\{ -\frac{(\mathbf{E_l} - E_{E_l})^2}{2Var_{E_l}} \}$$

$$l = 2, 3, \ldots\ldots, J,$$

where:

$$f_l = \sum_{k=1}^{n}(I(U_k = l),$$

$$c_{lk} = \sum_{j=l}^{J+1} q_{jk},$$

$$d_{(-l)k} = 1.0 + \sum_{j=2}^{J} c_{jk}e^{E_j} - c_{lk}e^{E_l}$$

and $I(\cdot)$ is the indicator function.

2.3.2 The full conditional distribution of the survival age effect A

Like we did for encounter age effect E, we will do the exactly same steps to find the full conditional distribution of the survival age effect A. We know from before too that joint prior for $\mathbf{A} = (A_1, A_2, \ldots\ldots, A_i)'$ is an IAR(2) prior, then the prior distribution of A_j ($j = 2, 3, \ldots\ldots, J$) is also a normal distribution with:

$$E_{A_j} = -\frac{\sum_{i \neq j} v_{ij}A_i}{v_{jj}}, \quad Var_{A_j} = \frac{1}{v_{jj}}$$

where v_{jj} is the element of the precision matrix $\frac{\mathbf{V_A}}{\tau_1}$.

We rewrite prior of A:

$$[\ \mathbf{A} \mid \tau_1\ ] \propto 1 / (\tau_1)^{(J-2)/2} \cdot \exp(\ -1/2\tau_1 \cdot \mathbf{A}' \mathbf{V_A}\, \mathbf{A}\ ),$$

Then look at the exponential part:

$$-\frac{1}{2\tau_1}\mathbf{A'V_A A} = -\frac{V_A}{2\tau_1}\mathbf{A'A}$$
$$= -2v_{jj}\ (\mathbf{A_j} - E_{A_j})'\ (\mathbf{A_j} - E_{A_j})$$

$$= -\frac{1}{2\mathrm{Var}_{A_j}}(\mathbf{A_j} - \mathrm{E}_{A_j})' (\mathbf{A_j} - \mathrm{E}_{A_j})$$

$$= -\frac{1}{2\mathrm{Var}_{A_j}}(\mathbf{A_j} - \mathrm{E}_{A_j})^2$$

$$= -\frac{(\mathbf{A_j} - \mathrm{E}_{A_j})^2}{2\mathrm{Var}_{A_j}}$$

$$j = 1, 2, 3, \ldots\ldots, J,$$

And we have the likelihood:

$$\boldsymbol{L}\,(\delta, q; \text{data}, \text{variables}) = \prod_{k=1}^{n} \frac{\delta_{U_k} q T_k k}{\sum_{j \geq i} \delta_i q_{jk}}$$

So the full conditional posterior distribution of survival age effect A_j is:

$$[A_j \mid .] \propto \prod_{k=1}^{n} \frac{\delta_{U_k} q T_k k}{\sum_{j \geq i} \delta_i q_{jk}} \cdot \exp\{ -\frac{(\mathbf{A_j} - \mathrm{E}_{A_j})^2}{2\mathrm{Var}_{A_j}} \}$$

$$\propto \prod_{k=1}^{n} \frac{(I(T_k = j) e^{A_j})}{\sum_{j \geq i} e^{A_j} q_{jk}} \cdot \exp\{ -\frac{(\mathbf{A_j} - \mathrm{E}_{A_j})^2}{2\mathrm{Var}_{A_j}} \}$$

$$\propto \frac{e^{g_j A_j}}{\prod_{k=1}^{n} (h_{jk} e^{A_j} + l_{jk})} \cdot \exp\{ -\frac{(\mathbf{A_j} - \mathrm{E}_{A_j})^2}{2\mathrm{Var}_{A_j}} \}$$

$$j = 1, 2, \ldots\ldots, J,$$

where:

$$g_j = \sum_{k=1}^{n} (I(T_k = j),$$

$$h_{jk} = e^{x_k' \beta} \sum_{i=l}^{j} \delta_i,$$

$$l_{jk} = 1.0 + \sum_{i=1}^{J} h_{ik} e^{A_i} - h_{jk} e^{A_j}$$

2.3.3 The full conditional posterior distribution of β

From the priors we know that β's are set as independent and they have one-stage normal prior with expectation zero and a fixed variance value s_β.

Because the normal distribution are conjugate distribution, the posterior distribution of β is also normal distribution. Then we have:

$$[\beta_i \mid .] \propto \frac{e^{o_i \beta_i}}{\prod_{k=1}^{n}(r_{ik} e^{x_{ki}\beta_i + 1.0)}} \times \exp(-\frac{\beta^2}{2s_\beta}), \quad i = 1,2,\cdots\cdots,p,$$

Where :

$$o_i = \sum_{k=1}^{n} I(T_k \leq J) x_{ki}$$

And

$$r_{ik} = (\sum_{j \neq i} x_{kj}\beta_j) \sum_{j \geq m} \delta_m e^{A_j}$$

x_{ki} is the ith element of $\mathbf{x}_k$.

2.3.4 The full conditional posterior distribution of the τ_i

We recall that τ_i are the variance components that control the smoothness of the survival and encounter curves and they are assumed to be:

$$\tau_i \sim IG(a_i, b_i), \quad i = 1, 2,$$

From hyperparameter specification we know that a_i and b_i are already assumed in order to give τ_i an infinite variance. Combine these information with the two IAR(2) priors we have, the followings are shown:

$$(\tau_1 \mid .) \sim IG(\frac{J-3}{2} + a_1, \frac{1}{2}\sum_{j=3}^{J-1}(E_j - 2E_{j-1} + E_{j-2})^2 + b_1)$$

$$(\tau_2 \mid .) \sim IG(\frac{J-2}{2} + a_2, \frac{1}{2}\sum_{j=3}^{J}(A_j - 2A_{j-1} + A_{j-2})^2 + b_2)$$

2.3.5 Existence of posterior distribution

Because very little has been done by way of verifying the existence of posterior distributions resulting from improper priors, therefore it is also hard for us to find the necessary and sufficient conditions that could prove the existence of posterior distribution. However, Cao (2009) did some proof under certain conditions, which is not very relevant to this thesis' main topic. I will explain it in Appendix at last for the readers who have the interests.

2.4 Simulation study results and analyses from Cao(2009)

The simulation study consist a sample size equal to 300 of each type of bird, and Cao(2009) generated 100 samples from the pool. Cao(2009) assumed there were two independent covariates and 300 pairs of the covariates observation were extracted from uniform distribution $U(-0.5, 0.5)$. And the regression parameters β_1, β_2 were assumed to be 0.9 and -1.1 respectively.

Beside the generated values, Cao(2009) used the true values of the survival age effect **A** and the encounter age effect **E** based on the real data analysis results. And the true encounter and failure probabilities were calculated from:

$$\delta_i = \frac{e^{E_i}}{1+\sum_{i=2}^{J} e^{E_i}} \text{ , for i = 2, 3, ..., J ,}$$

$$\delta_1 = \frac{1}{1+\sum_{i=2}^{J} e^{E_i}} ,$$

$$q_{jk} = \frac{e^{Aj + x_k' \beta}}{1+ \sum_{j=1}^{J} e^{Aj + x_k' \beta}} \text{ , for j = 1, 2, ..., J ,}$$

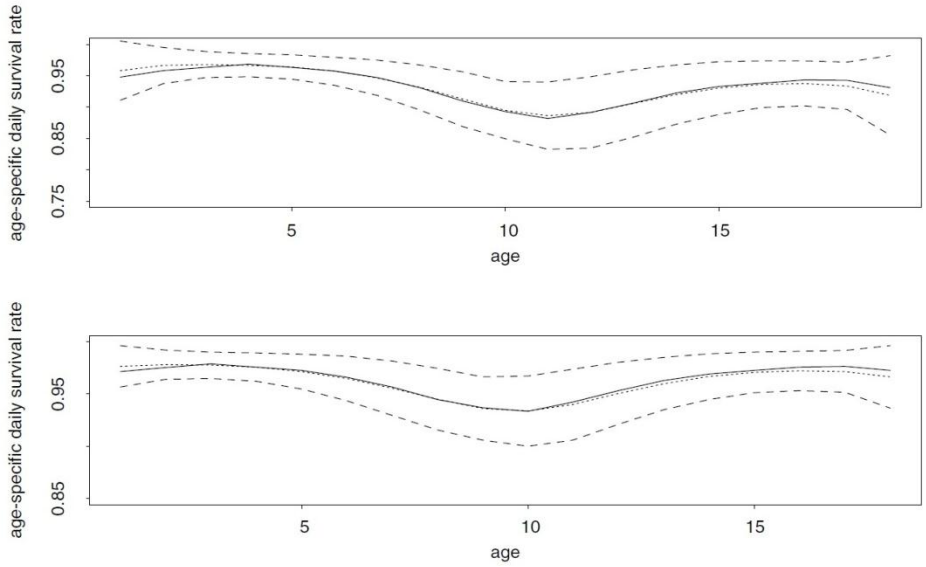
$$q_{(J+1)k} = \frac{1}{1+ \sum_{j=1}^{J} e^{Aj + x_k' \beta}} .$$

And there was another condition that in each sample, the data were generated for 300 nests under a schedule of visit-every-three-days.

Gibbs sampling with 51000 cycles were selected as the computation tool, with the burn-in was 1000. The result was:

β	True value	Mean	$\sqrt{\text{mse}}$
β_1	0.9	0.8960	0.3524
β_2	-1.1	-1.0848	0.3599

The mean and $\sqrt{\text{mse}}$ values from above table were the Bayesian estimates of β_1 and β_2 over the 100 samples. It is straight to see that the estimates are unbiased and significant.



As we see from the figure about two estimated survival probabilities of two different nests, we could easily get the difference, the first nest had lower survival probabilities than the second one. And another good thing showed from this figure is that the estimated value followed the true survival curve quite precisely.

There were the simulation results for the nest success probabilities:

Nest NO.	True q_{J+1}	Mean est. of q_{J+1}	$\sqrt{\text{MSE}}$

1	0.2673	0.2710	0.0462
2	0.4899	0.4848	0.0553
3	0.3765	0.3738	0.0332
4	0.2182	0.2199	0.0292
5	0.2574	0.2577	0.0319

There were only the first five nests' result on the table, but the true value and the estimated value were all within 1% difference, these were very well estimated.

Cao(2009) also did another simulation with lesser sample. The lesser sample conducted with 100 as the sample size and the results for the estimations of $\beta's$ were still unbiased, but the $\sqrt{\text{mse}}$
Increased from 0.35 to 0.52, however, the estimated survival curves matched the true survival curve closely again. This showed that this was a good model (Cao(2009) told that all 300 nests had a very satisfactory estimated survival curves).

The data collected by Wells(2005) was the core to this section and Cao(2009) used Bayesian hierarchical model with nest-specific covariates to analyze it. In this data set, there were 217 observations valid in total and there were six nest specific covariate measurements recorded for the surrounding vegetation of the nest immediately when an outcome occurred.
These six were:
X_1: percentage of grass cover,
X_2: percentage of litter cover,
X_3: percentage of forbs(a type of herb) cover,
X_4: percentage of woody cover,
X_5: height (cm) of the tallest plant,
X_6: distance to the nearest woody plant within one meter of the nest,

In Cao (2009), the deviance information criterion (DIC) is chosen to select the proper subset of the covariates. The DIC provides a Bayesian measure of model fit and complexity and the smaller value of DIC means the better models. To

understand the DIC, we first need to find deviance D(β), where β is the unknown parameter:

$$D(\beta) = -2\log(f(\mathbf{y}|\boldsymbol{\beta})) + C$$

$f(\mathbf{y}|\boldsymbol{\beta})$ is the likelihood function and C is a constant but will be cancelled out during the calculation.

Then we need to know the expectation of deviance which is:

$$Đ = \mathbf{E}^{\beta}[D(\beta)]$$

And the effective number of parameters of the model is :

$$p_D = Đ - D(\beta)$$

Finally the DIC is :

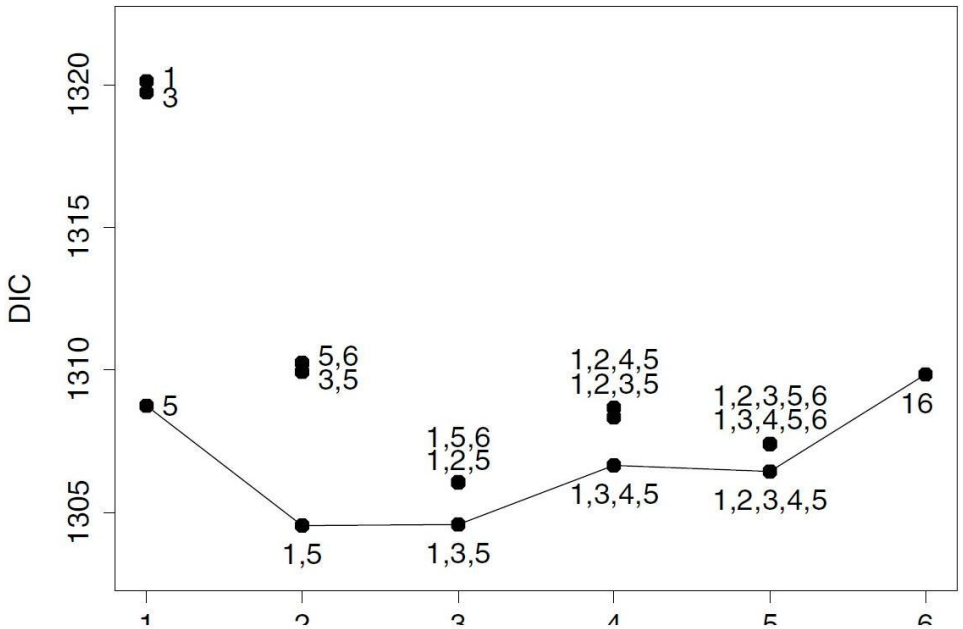
$$DIC = p_D + Đ = Đ - D(\beta) + Đ = 2\,Đ - D(\beta)$$

This is only a brief definition about DIC, we will explain it with more details later in this thesis. Now we look back to the six different covariates we have here, and we examine thought all the different combinations of the covariates in the linear model:

$$\log\frac{q_{jk}}{q_{(J+1)k}} = A_j + x_k'\ \beta \text{ , for j = 1, 2, ... , J .}$$

The figure below shows the different DIC scores with different numbers of covariates. As we have shortly noticed above: lower DIC score suggest better model. From the figure, it appears that subsets (X_1, X_5) and (X_1, X_3, X_5) are the better fit. Cao(2009) also examined models with interaction term and quadratic terms, the result turned out that those models with different terms were not good enough, this was the proof that a linear model was adequate. Then Cao(2009) chose the model with covariate X_1 and X_5 to be the final one not only because it

had the lowest DIC score, but also it was simpler compare to another model candidate.(model with covariates X_1, X_3 and X_5)



When the model is chosen, we have to consider the regression parameter β, and the estimate of it. To help readers understand the meaning of regression parameter, the following expression was given in Cao(2009):

$$\frac{q_{jk}/q_{(J+1)k}}{q_{jl}/q_{(J+1)l}} = \frac{\exp(A_j + x_k'\beta)}{\exp(A_j + x_l'\beta)} = \frac{\exp(x_k'\beta)}{\exp(x_l'\beta)}$$

This expression showed that the value e^{β_i} was an odds ratio, this means that if we assume that the other covariates remain the same, the value of e^{β_i} would be the ratio of the odds that a nest failing at a certain age against the nest failing at the same age but with one less covariate x_i.

One example was given in Cao(2009): Assume that we have two nests A and B, A has X_1 = 20% and X_5 = 67cm, where B has the same X_1, but X_5 = 66cm. The failing probability at age one for A is 0.04348 and the nest success probability is 0.26109. For nest B, the failing probability at age one is 0.04374 and the nest success probability is 0.25661. We take those numbers back into the expression we have, it becomes:

$$\frac{q_{jk}/q_{(J+1)k}}{q_{jl}/q_{(J+1)l}} = \frac{0.04348/0.26109}{0.04374/0.25661} = 0.976,$$

And:

$$0.976 = e^{\beta_5} = e^{-0.024},$$

Then we have β_5 = -0.024. The estimated regression parameter is negative means two things: One is that a negative β is favorable for the nest survival; in this example, X_5 is the height of the tallest plant in the observing area, and it is correct for Dickcissels to have a better survival rate with taller plant in the neighborhoods, because higher plants would minimize the chances for predators to catch their nests. Second thing to notice is that with a negative estimated regression parameter, the larger value of the covariate, the higher survival probability.

Cao(2009) also estimated β_1 , which is the regression parameter for X_1, the grass cover percentage of the observing area. β_1 was positive and have the value 0.012, it may look strange to our common knowledge that for a grassland bird that the grass covering percentage have negative effect on its nest surviving rate. The truth is the main threat to the bird nests in that area are the predators like snakes and small mammals, which usually are observed on moving outside or on the edge of the grass, but well hidden in the grass cover. This may explain why grass cover is negatively related to the nest survival rate. And from these two estimated regression parameters, Cao(2009) pointed out that the highlight of their study was the need of higher plant or vegetation with minimal percentage of grass cover were the key to the nest survival for Dickcissels.

3. Further Studies

3.1 Deviance information criterion

There are many models today that are used to estimate the real world complexities of data, but not all of them are good enough. We wish that there could be a way which would compare the different models for us and eventually could identify the most fit model appear to describe the data information adequately. Generally within the classical modeling framework, this kind comparison normally takes place by defining a measure of fit, most of the time it is a deviance statistic; and by estimating the number of free parameters in the model, so called complexity of the model.

When we briefly introduced the DIC in Cao(2009), the complexity measure ρ_D for the effective number of parameters in a model was mentioned. This quantity is the difference between the posterior mean of deviance and the deviance for the posterior estimates of the parameters of interest, Spiegelhalter(2002) also stated that ρ_D could be trivially determined by using MCMC.

The posterior mean deviance $\overline{D}$ is meant to be the Bayesian measure of fit, it shows the adequacy or how adequate the model could be.

The complexity of Bayesian model or even hierarchical Bayesian model could be very different if we choose the different prior distribution. The example is simple, if we choose a prior and parameterize it with unknown hyper-parameters φ, the hierarchical Bayesian model we created will be:

$$p(y, \theta, \varphi) = p(y, \theta)p(\theta|\varphi)\ p(\varphi)$$

For this model, if we choose prior $p(\theta)$ and likelihood function $p(y \mid \theta)$:

$$p(\theta) = \int_{\varphi}\ p(\theta|\varphi)p(\varphi)\ d\varphi$$

Or we can choose prior $p(\varphi)$ and likelihood function $p(y \mid \varphi)$:

$$p(y \mid \varphi) = \int_{\theta} p(y|\theta)p(\theta|\varphi)\, d\theta$$

Whatever we choose, it will lead us to the same marginal distribution:

$$p(y) = \int_{\theta} p(y|\theta)p(\theta)\, d\theta$$

But they two choices have different complexity because they do not have the same number of parameters. As a consequence for hierarchical Bayesian model, Gelfand and Trevisani(2002) stated that we cannot find a likelihood without defining the level of hierarchy of the model. This means that we would rather choose the existing parameter as prior and likelihood than the hyper-parameter, this is a way to reduce all models to non-hierarchical structure and reduce model's complexity.

Now here comes a new question: 'How could we choose the better model to get the most accurate results?'. We know that it is very useful for us to have measures of fit and complexity, and try to combine them into overall criteria which would have better theoretical justification. However, we also feel that there won't be a formula for model 'selection' because there are too many things we have to take into consideration before we could even use it. Spiegehalter(2002) have discussed this in section 7(A model comparison criterion). In his theory, both classical and Bayesian approaches will start with a concept of an independent replicate data set, this is not the observed data but by using the same data-generating system which gave the observed data. If we suppose the loss of a data set Y with a probability $p(Y|\tilde{\theta})$ is $L(Y, \tilde{\theta})$, it is nature for us to select the model for $p(Y|\tilde{\theta})$ with the least $L(Y, \tilde{\theta})$. Then a criterion is based on the estimate of

$$E_{Y_{replicated}|\theta^t}\left[L\{Y_{replicated}, \tilde{\theta}(y)\}\right].$$

With an optimistic estimated loss $L\left(y, \tilde{\theta}(y)\right)$ that is suffered on re-predicting the observed y which gave rise to $\tilde{\theta}(y)$. Efron(1986) defined the 'optimism' with an estimator C_θ, then we have:

$$E_{Y_{replicated}|\theta^t}\left[L\{Y_{replicated}, \tilde{\theta}(y)\}\right] = L\left(y, \tilde{\theta}(y)\right) + C_\theta\left(y, \theta^t, \tilde{\theta}(y)\right)$$

Spieghalter(2002) explained from here that both classical and Bayesian approaches to estimate the C_θ would now be examined as a logarithmic loss function $L(Y, \tilde{\theta}) = -2log\{p(Y|\tilde{\theta})\}$. And the main difference for the classical and the Bayesian approaches were: The classical approach will attempt to estimate the sampling expectation of C_θ, whereas the Bayesian approach will concentrate on a direct calculation of the posterior expectation of C_θ.

Although the Bayesian approach is the main point we should focused on, the classical approach has some foundations we have to take a look. From the Speeghalter(2002), we have an approximate forms for the expected optimism:

$$\pi(\theta^t) = E_{Y|\theta^t}\left[C_\theta\left(Y, \theta^t, \tilde{\theta}(Y)\right)\right]$$

Put this back to the expectation of the replicated data loss, we will have:

$$E_{Y_{replicated}|\theta^t}\left[L\{Y_{replicated}, \tilde{\theta}(y)\}\right] = L\left(y, \tilde{\theta}(y)\right) + \tilde{\pi}(\theta^t)$$

Efron(1986) again defined the expression for $\pi(\theta^t)$ both for exponetial families and for general loss functions, and paticularly for the logarithmic loss function which is very useful here:

$$\pi_E(\theta^t) = 2\sum_i cov^t\left(\hat{Y}_i, Y_i\right) \approx 2p$$

We could rewrite it as we could generalize Akaike(1973) under broad conditions:

$$\pi(\theta^t) = 2p$$

These classical criteria for general model comparison are thus all based on the equation of the expectation of the replicated loss functions, and more importantly could be considered as corresponding to a plugged estimate of fit, plus twice the effective number of parameters in the model. This is the basic structure we should adapt in the Bayesian context.

As we have discussed before, a deviance information criterion (DIC) could be defined as a classical estimate of fit, plus twice the effective number of parameters, which has the simplest form below:

$$\text{DIC} = \text{D}(\bar{\theta}) + 2p_D$$

If we look at this definition, it is very similar to the AIC and has the same structure of the classical criteria. However, if we rewrite it as:

$$\text{DIC} = \bar{D} + p_D$$

This is how we define DIC with a Bayesian measure of fit, added by an extra complexity term p_D.

The following content will try to prove the DIC definition equations; it might be hard to understand for the readers.

We have defined the equation of expectation of the replicated loss function, which is:

$$E_{Y_{replicated}|\theta^t}\left[L\{Y_{replicated}, \tilde{\theta}(y)\}\right] = L\left(y, \tilde{\theta}(y)\right) + C_\theta\left(y, \theta^t, \tilde{\theta}(y)\right)$$

By mimicking the Ripley(1996) and Burnham and Anderson(1998), and using the logarithmic loss function on the equation above, we get:

$$C_\theta\left(y, \theta^t, \tilde{\theta}(y)\right) = E_{Y_{replicated}|\theta^t}\left[D_{replicated}(\tilde{\theta})\right] - D(\tilde{\theta})$$

Where we have that:

$$-2log\left[p\{Y_{replicated}|\tilde{\theta}(y)\}\right] = D_{replicated}(\tilde{\theta})$$

Because we now are taking a Bayesian perspective, we could replace the true θ^t with a random θ. And with the condition that D is an unstandardized deviance ($f(\cdot) = 1$), we could now expand C_θ into three terms:

$$C_\theta = L_1 + L_2 + \{D(\theta^t) - D(\tilde{\theta})\},$$

$$L_1 = L_1(\theta, \tilde{\theta}) \approx$$
$$E_{Y_{replicated}|\theta}\left\{-2(\tilde{\theta}-\theta)^T L'_{replicated,\theta} - (\tilde{\theta}-\theta)^T L''_{replicated,\theta}(\tilde{\theta}-\theta)\right\},$$

$$L_2 = L_2(y,\theta) = E_{Y_{replicated}|\theta}\left[-2log\{p(Y_{replicated}|\theta)\}\right] + 2log\{p(y|\theta)\}$$

We could rewrite L_1 with the knowledge that $L_{replicated,\theta} = log\{p(Y_{replicated}|\theta)\}$ and $E_{Y_{replicated}|\theta}(L'_{replicated,\theta}) = 0$, it then become:

$$L_1 = L_1(\theta, \tilde{\theta}) \approx tr\left\{I_\theta(\tilde{\theta}-\theta)(\tilde{\theta}-\theta)^T\right\},$$
$$I_\theta = E_{Y_{replicated}|\theta}(L''_{replicated,\theta})$$

The I_θ is supposed to be the Fisher information in $Y_{replicated}$, and hence also in y. L_1 then again could be approximately rewrite as:

$$L_1 = L_1(\theta, \tilde{\theta}) \approx tr\left\{-L''_{\tilde{\theta}}(\tilde{\theta}-\theta)(\tilde{\theta}-\theta)^T\right\}$$

Now under a particular model assumption we could calculate a posterior distribution $p(\theta|y)$, and then our posterior expected optimism under this model and the estimator $\tilde{\theta}$ is:

$$E_{\theta|y}(C_\theta) \approx tr\left[-L''_{\tilde{\theta}} E_{\theta|y}\left\{(\tilde{\theta}-\theta)(\tilde{\theta}-\theta)^T\right\}\right] + E_{\theta|y}\{L_2(y,\theta)\} + E_{\theta|y}\{D(\theta) - D(\tilde{\theta})\}$$

By using the posterior mean $\bar{\theta}$ as our estimator could change the expected optimism as below:

$$E_{\theta|y}(C_\theta) \approx tr(-L''_{\bar{\theta}} V) + E_{\theta|y}\{L_2(y,\theta)\} + p_D,$$

Where V is defined as the posterior covariance of θ, and as we mentioned before

$$p_D = \bar{D} - D(\bar{\theta}),$$

$$E_Y\left[E_{\theta|Y}\{L_2(y,\theta)\}\right] = E_\theta\left[E_{Y|\theta}\{L_2(y,\theta)\}\right] = 0,$$

$$p_D \approx tr\left(-L''_{\bar{\theta}}V\right),$$

With all those conditions above, the expected posterior loss when adopting a particular model then would be:

$$D(\bar{\theta}) + E_{\theta|y}(C_\theta) \approx D(\bar{\theta}) + 2p_D = DIC$$

This proof above shows that the main idea behind the DIC from Speighalter(2002): Common standardization across models will leave unchanged the property that difference in DIC are estimates of differences in expected loss in prediction.

The conclusion for DIC is, it could be treated as a Bayesian analogue of AIC because it has similar justification, however, the DIC has a much wider applicability than the AIC and therefore the DIC could be applicable to almost any class of model which involves negligible additional analytic work or contains Monte Carlo sampling. The DIC today is still a new thing but it deserves further investigation and promotion to be a tool for model selection and comparison.

3.2 Mayfield method

Mayfield method is another way of estimating the nest survive rate. Although it is not wildly being used by either biologist or statistician, it is the method that among those focused on the truncated data. Mayfield (1960) stated that the data that gathered for estimation were only those data we could observe, and there were data we could not able to observe and thus the predicted results were often over-estimated; he introduced a new observing units: nest days d and a simple

method. If the mortality rate is being calculated as

$$\frac{\text{number of nests are lost}}{\text{total number of nests are observed}} = r,$$

then the probability of survival is $(1 - r)^d$. He also declared that there were five different parts of surviving calculation during the nest fledging, however the simple thing was by using his method to determine the survival probabilities for each five parts: P1, P2, P3, P4 and P5, and multiplied them together.

The mathematics of this method seems quite friendly to us compare to the Bayesian method, but how accurate could Mayfield method be? Hensler and Nichols (1981) present an experimental situation for Mayfield; they used the maximum likelihood estimators of this experimental model and also used the Monte Carlo simulation to test them in order to compare them with the traditional method before Mayfield. They assumed that every nest they observe could be considered as a vector: $X_k = (Y_k, T_k)$, where Y_k is a random variable with value to be either 1 if the *k*th nest was successful or 0 if it failed during the observation; T_k is a random variable that indicates the number of days that the *k*th nest needed to be either successful or failed. Under this conditions, the joint distribution of X_k is :

$$f(y, t|p) = \left[\theta_{J-t+1} p^t\right]^y \left[p^{t-1}(1-p)\sum_{j=1}^{J-t+1}\theta_j\right]^{1-y},$$

Where J is the total number of days that all nests need for their nesting process based on the Mayfield concept of 'nest days'; $p(0<p<1)$ is a constant probability for a nest to survive from day j to day j+1, this probability is unknown. Mayfield also called the value of p^t to be the nest succeeding probability; and θ_j is also an unknown nest encounter probability for the Mayfield method.

When we find out the log-likelihood function of this joint distribution, which is:

$$\log \prod_{k=1}^{K} \left[\theta_{J-T_k+1} p^t\right]^{Y_k}$$

$$+ \left[\sum_{k=1}^{K} T_k Y_k\right] \log p$$

$$+ \sum_{k=1}^{K} (T_k - 1)(1 - Y_k) \log p + \left(K - \sum_{k=1}^{K} Y_k\right) \log(1 - p)$$

$$+ \log \prod_{k=1}^{K} \left[\sum_{j=1}^{J-T_k+1} \theta_j\right]^{1-Y_k}$$

If we differentiate this with respect to p and we solve it, just like we do when we try to determine the maximum likelihood estimator, we could find $\hat{p} = \frac{\sum_{k=1}^{K} T_k + \sum_{k=1}^{K} Y_k - K}{\sum_{k=1}^{K} T_k}$.

Compare this to the Mayfield method; we easily notice that the Mayfield estimator is in fact the same as the maximum likelihood estimator. (Mayfield's total number of nest days observed is T_k and the total number of failures is $K - \sum_{k=1}^{K} Y_k$)

The results from the test example on Hensler and Nichols (1981) indicated that the accuracy of the Mayfield method is limited. If the model we assumed at the beginning was not far from the real situation and the overall probabilities of survival were not low, the Mayfield method would be a better estimator compare to the traditional method. However, a basic assumption of the Mayfield method is that the daily failure hazard rate is constant. This assumption is obviously very unrealistic from the world we are lived in. If we want to get more accurate result from the data and thus find a better estimator, we have to take every small detail into consideration.

3.3 Comparisons between Bayesian method and Mayfield method with some examples

Like we have mentioned before, an unknown parameter is often estimated by its posterior mean in the Bayesian analysis because the posterior mean is the most accurate estimator of a parameter under a squared error loss. But this kind of

analysis often needs high dimensional integrations which are not so friendly for most of people. With the help of computer, He(2003) presented some examples to show the difference between the Bayesian method and the Mayfield method.

Bayesian and Mayfield estimates for bobwhite example with a sample size of 36

	Encounter probability	Survival rate	
Age	Bayes est. (std)	Mayfield est. (std)	Bayes est. (std)
1	.49252 (.12725)	.95472 (.00009)	.97455 (.03652)
2	.09323 (.07103)	.95472 (.00009)	.95685 (.04941)
3	.04023 (.05300)	.95472 (.00009)	.96006 (.05035)
4	.13457 (.08794)	.95472 (.00009)	.96189 (.04757)
5	.09129 (.06917)	.95472 (.00009)	.91532 (.07581)
6	.02293 (.03269)	.95472 (.00009)	.93578 (.07125)
7	.02098 (.03105)	.95472 (.00009)	.95340 (.05670)
8	.01616 (.02484)	.95472 (.00009)	.92589 (.07276)
9	.01484 (.02368)	.95472 (.00009)	.93291 (.07564)
10	.01160 (.01856)	.95472 (.00009)	.89046 (.09181)
11	.01797 (.02560)	.95472 (.00009)	.93582 (.07376)
12	.01481 (.02520)	.95472 (.00009)	.95751 (.05321)
13	.00960 (.01703)	.95472 (.00009)	.88931 (.09491)
14	.00483 (.00993)	.95472 (.00009)	.94860 (.06349)
15	.00364 (.00784)	.95472 (.00009)	.95881 (.05276)
16	.00264 (.00569)	.95472 (.00009)	.93226 (.07746)
17	.00200 (.00458)	.95472 (.00009)	.93497 (.07467)
18	.00156 (.00418)	.95472 (.00009)	.94082 (.06788)
19	.00343 (.00650)	.95472 (.00009)	.96106 (.05079)
20	.00060 (.00201)	.95472 (.00009)	.96303 (.04865)
21	.00028 (.00104)	.95472 (.00009)	.92838 (.07392)
22	.00014 (.00070)	.95472 (.00009)	.95950 (.05349)
23	.00008 (.00053)	.95472 (.00009)	.96599 (.04584)
24	.00007 (.00040)	.95472 (.00009)	.81294 (.09949)
Total	1.0000	.32891 (.00070)	.20176 (.05528)

This is the first example with J = 24 and n = 30. The Mayfield estimate of the survival rate is much higher than the Bayesian one, and the Mayfield survival rate value indeed has smaller difference when compare to the sample proportion, which has the tendency to overestimate the survival rate. And here we could clearly look the Mayfield method assumed the daily hazard rate to be the same (0.0009); while the Bayesian method estimated different daily hazard rate. When we look at the day 23 and day 24, we notice that the significant survival rate drop for the Bayesian method, which is indeed indicates that the birds' behaviors: when the first chick comes out from its egg, it starts to peck egg shells which would eventually kill other chicks by destroying the others' eggs or get killed by predators who attracted by its pecking sound.

This example was from an observed data set, He(2003) also introduced another example of two simulated data sets.

Bayesian and Mayfield estimates for simulated data set I with a sample size of 300

	Encounter probability		Survival rate		
Age	True	Bayes est. ($\sqrt{\text{MSE}}$)	True	Mayfield est. ($\sqrt{\text{MSE}}$)	Bayes est. ($\sqrt{\text{MSE}}$)
1	.40001	.32411 (.09035)	.95	.96982 (.01994)	.95757 (.02489)
2	.24001	.31039 (.08952)	.95	.96982 (.01994)	.95761 (.02043)
3	.14400	.14869 (.05033)	.99	.96982 (.02029)	.97330 (.02103)
4	.08640	.08384 (.03767)	.99	.96982 (.02029)	.98555 (.00851)
5	.05184	.06095 (.02943)	.99	.96982 (.02029)	.98877 (.00542)
6	.03110	.03095 (.01658)	.99	.96982 (.02029)	.98706 (.00707)
7	.01866	.01489 (.00995)	.99	.96982 (.02029)	.98760 (.00667)
8	.01120	.01002 (.00741)	.99	.96982 (.02029)	.98998 (.00500)
9	.00672	.00770 (.00637)	.99	.96982 (.02029)	.98979 (.00518)
10	.00403	.00367 (.00331)	.99	.96982 (.02029)	.98744 (.00704)
11	.00242	.00218 (.00230)	.99	.96982 (.02029)	.98784 (.00595)
12	.00145	.00122 (.00141)	.99	.96982 (.02029)	.98846 (.00537)
13	.00087	.00066 (.00094)	.99	.96982 (.02029)	.98622 (.00796)
14	.00052	.00032 (.00055)	.99	.96982 (.02029)	.98587 (.00824)
15	.00031	.00019 (.00037)	.99	.96982 (.02029)	.98873 (.00544)
16	.00019	.00010 (.00023)	.99	.96982 (.02029)	.98807 (.00557)
17	.00011	.00005 (.00013)	.99	.96982 (.02029)	.97909 (.01504)
18	.00007	.00003 (.00008)	.99	.96982 (.02029)	.96312 (.03224)
19	.00004	.00002 (.00007)	.85	.96982 (.11984)	.91971 (.08050)
20	.00002	.00001 (.00002)	.85	.96982 (.11984)	.82990 (.06530)
21	.00001	.00001 (.00002)	.85	.96982 (.11984)	.84138 (.05557)
Total	1.0000	1.0000	.47192	.52601 (.05924)	.45394 (.03464)

This is for the first data set, with J=21 and n=300. Because the Mayfield method assumes that the encounter probabilities are unknown, therefore only the Bayesian estimates are presented. From the figure, both methods have no significant difference on the estimated survival rate from day 1 to day 18, but from day 19, the Mayfield method starts to overestimate the survival rate which has over 10% difference compare to the true values; the Bayesian estimates have the same tendency as the true values.

Another simulated data set with J=21 and n=300:

Bayesian and Mayfield estimates for simulated data set II with a sample size of 300

	Encounter probability		Survival rate		
Age	True	Bayes est. ($\sqrt{\text{MSE}}$)	True	Mayfield est. ($\sqrt{\text{MSE}}$)	Bayes est. ($\sqrt{\text{MSE}}$)
1	.40001	.32102 (.10798)	.95	.95235 (.00307)	.95469 (.02493)
2	.24001	.32175 (.11035)	.95	.95235 (.00307)	.96058 (.02061)
3	.14400	.15883 (.06724)	.99	.95235 (.03770)	.97335 (.02116)
4	.08640	.08852 (.04643)	.99	.95235 (.03770)	.98568 (.00803)
5	.05184	.05136 (.02805)	.99	.95235 (.03770)	.98897 (.00493)
6	.03110	.02545 (.01768)	.99	.95235 (.03770)	.98708 (.00747)
7	.01866	.01243 (.01116)	.99	.95235 (.03770)	.98755 (.00683)
8	.01120	.00851 (.00784)	.99	.95235 (.03770)	.98990 (.00504)
9	.00672	.00583 (.00549)	.99	.95235 (.03770)	.98985 (.00496)
10	.00403	.00280 (.00311)	.99	.95235 (.03770)	.98758 (.00655)
11	.00242	.00158 (.00197)	.99	.95235 (.03770)	.98715 (.00670)
12	.00145	.00093 (.00148)	.99	.95235 (.03770)	.98742 (.00628)
13	.00087	.00047 (.00083)	.99	.95235 (.03770)	.98453 (.00903)
14	.00052	.00024 (.00049)	.99	.95235 (.03770)	.98354 (.01015)
15	.00031	.00014 (.00032)	.99	.95235 (.03770)	.98613 (.00723)
16	.00019	.00007 (.00021)	.99	.95235 (.03770)	.98416 (.00896)
17	.00011	.00004 (.00012)	.99	.95235 (.03770)	.96973 (.02606)
18	.00007	.00002 (.00008)	.99	.95235 (.03770)	.93322 (.06797)
19	.00004	.00001 (.00005)	.65	.95235 (.30236)	.81438 (.18620)
20	.00002	.00001 (.00003)	.65	.95235 (.30236)	.60715 (.11286)
21	.00001	.00001 (.00003)	.65	.95235 (.30236)	.64662 (.09705)
Total	1.0000	1.0000	.21103	.35901 (.14881)	.20326 (.02430)

Maybe the estimates will change if the sample sizes become smaller? Let's cut the sample size to half:

Bayesian and Mayfield estimates for simulated data set II with a sample size of 150

	Encounter probability		Survival rate		
Age	True	Bayes est. ($\sqrt{\text{MSE}}$)	True	Mayfield est. ($\sqrt{\text{MSE}}$)	Bayes est. ($\sqrt{\text{MSE}}$)
1	.40001	.33415 (.11905)	.95	.95231 (.00359)	.94348 (.03451)
2	.24001	.32083 (.13068)	.95	.95231 (.00359)	.96063 (.02142)
3	.14400	.16018 (.08185)	.99	.95231 (.03779)	.97249 (.02186)
4	.08640	.08502 (.05239)	.99	.95231 (.03779)	.98260 (.01053)
5	.05184	.04815 (.03268)	.99	.95231 (.03779)	.98631 (.00701)
6	.03110	.02297 (.01866)	.99	.95231 (.03779)	.98607 (.00753)
7	.01866	.01157 (.01214)	.99	.95231 (.03779)	.98578 (.00886)
8	.01120	.00761 (.00869)	.99	.95231 (.03779)	.98758 (.00642)
9	.00672	.00455 (.00566)	.99	.95231 (.03779)	.98747 (.00664)
10	.00403	.00234 (.00330)	.99	.95231 (.03779)	.98592 (.00793)
11	.00242	.00124 (.00204)	.99	.95231 (.03779)	.98536 (.00826)
12	.00145	.00063 (.00120)	.99	.95231 (.03779)	.98481 (.00866)
13	.00087	.00035 (.00081)	.99	.95231 (.03779)	.98197 (.01142)
14	.00052	.00019 (.00053)	.99	.95231 (.03779)	.98050 (.01308)
15	.00031	.00010 (.00034)	.99	.95231 (.03779)	.98096 (.01277)
16	.00019	.00005 (.00020)	.99	.95231 (.03779)	.97773 (.01608)
17	.00011	.00003 (.00011)	.99	.95231 (.03779)	.96139 (.03510)
18	.00007	.00001 (.00007)	.99	.95231 (.03779)	.92639 (.07448)
19	.00004	.00001 (.00006)	.65	.95231 (.30232)	.81363 (.18668)
20	.00002	.00000 (.00003)	.65	.95231 (.30232)	.63576 (.12016)
21	.00001	.00000 (.00002)	.65	.95231 (.30232)	.64616 (.12175)
Total	1.0000	1.0000	.21103	.35899 (.14955)	.19443 (.03479)

Still not too much difference compare to the one with sample size 300. Let's make it even smaller,

Bayesian and Mayfield estimates for simulated data set II with a sample size of 50

	Encounter probability		Survival rate		
Age	True	Bayes est. ($\sqrt{\text{MSE}}$)	True	Mayfield est. ($\sqrt{\text{MSE}}$)	Bayes est. ($\sqrt{\text{MSE}}$)
1	.40001	.36885 (.16042)	.95	.95175 (.00524)	.91192 (.07368)
2	.24001	.31926 (.16649)	.95	.95175 (.00524)	.95331 (.02561)
3	.14400	.15616 (.09856)	.99	.95175 (.03856)	.96599 (.02904)
4	.08640	.07628 (.05582)	.99	.95175 (.03856)	.97440 (.01917)
5	.05184	.04160 (.03812)	.99	.95175 (.03856)	.97769 (.01551)
6	.03110	.01819 (.02171)	.99	.95175 (.03856)	.97863 (.01425)
7	.01866	.00864 (.01367)	.99	.95175 (.03856)	.97928 (.01381)
8	.01120	.00511 (.00883)	.99	.95175 (.03856)	.98013 (.01299)
9	.00672	.00281 (.00559)	.99	.95175 (.03856)	.97963 (.01378)
10	.00403	.00149 (.00345)	.99	.95175 (.03856)	.97807 (.01566)
11	.00242	.00081 (.00221)	.99	.95175 (.03856)	.97692 (.01631)
12	.00145	.00041 (.00133)	.99	.95175 (.03856)	.97613 (.01676)
13	.00087	.00019 (.00077)	.99	.95175 (.03856)	.97252 (.02133)
14	.00052	.00010 (.00049)	.99	.95175 (.03856)	.96992 (.02474)
15	.00031	.00005 (.00030)	.99	.95175 (.03856)	.96887 (.02529)
16	.00019	.00003 (.00022)	.99	.95175 (.03856)	.96218 (.03251)
17	.00011	.00002 (.00012)	.99	.95175 (.03856)	.93892 (.06040)
18	.00007	.00001 (.00007)	.99	.95175 (.03856)	.89943 (.10470)
19	.00004	.00000 (.00004)	.65	.95175 (.30179)	.79682 (.18023)
20	.00002	.00000 (.00002)	.65	.95175 (.30179)	.68016 (.12842)
21	.00001	.00000 (.00001)	.65	.95175 (.30179)	.64555 (.14896)
Total	1.0000	1.0000	.21103	.35601 (.15008)	.15871 (.06923)

From the data above, the tendency suggest that the estimates under the Bayesian method are better compare to the Mayfield, especially when the nest age are at their lastest stages.

4. Conclusion

In this thesis, I study about the hierarchal Bayesian model and try to get a clear picture of it. The model itself is a better model compare to other existing models today; it clearly states that the more parameters we take into consideration, the more accurate result we will get; and furthermore because we could choose the priors and therefore we could also choose the difficult level of the model.

Bayesian or hierarchal Bayesian methods are not easy for people without mathematical knowledges such as the higher dimensions of integration, this is the main reason that why it does not become a popular model. If we could develop a friendly IT software program that could make this model relatively easy for most people, this model could evolve into a new valuable tool for statistical analyses.

But the suggestion at the moment is that for most nest data analysis today, the survival rates are having minor different from day to day until the last few days; it is reasonable to use easier model to estimate the first part of days' survival probabilities, like the Mayfield estimates. And for the last few days survival probabilities, we could either use the Hierarchal Bayesian model if we require very accurate result or just lower the Mayfield estimates with respect to historical datas' basic hazard rate which were estimated by the Cox model.

DIC is a useful tool to analyse what's the most important parameters for data. As we said before: DIC could be recognized as an Bayesian update of AIC, with similar justification but much wider applicability. Despite to its difficulty for most people to understand, we could use DIC and the hierarchical Baysesian model togather to find the real parameters that affect the data.

There might be some kind of link between nest data and medical data, like the extensive usage from brid survival analysis to human survival analysis in the future when the method becomes more mature. Due to the lack of information on the medical data, the realistic usefulness of this model for medical statitistical analysis remains unknown. However, with more research and more effort, I personally believe that the Bayesian model would serve us better in the future.

Appendix

Appendix A: Simple proof of the existence of posterior distribution

Before we start the proof, we need to make two conditions:
Condition 1 is that there are at least three nests discovered at age one and there are at least three successful nests at the end;
Condition 2 is that we define two matrix $\mathbf{Q}_E$ and $\mathbf{Q}_A$, both to be full rank matrix with some special requirement(check Wells(2007) for details);

We know that to prove the existence of the posterior distribution is the same to prove that:

$$\int H_1(E, A, \beta, \tau_1, \tau_2) dE dA < C$$

Where $H_1(E, A, \beta, \tau_1, \tau_2) = (data, U, T|E, A, \beta)(E|\tau_1)(A|\tau_2)$ and C is a constant independent of $\beta, \tau_1 \; and \; \tau_2$. We learn that the likelihood is actually from a

multinomial distribution with the given parameters: $\delta_{U_k} q_{T_k k} \Big/ \sum_{j \geq i} \delta_i q_{jk} \leq 1$, and

let $C_x = \min_k \left(e^{x_k' \beta} \right) > 0$,

$n_1 = \sum_k I(U_{k=1})$ is the NO. of nests discovered at age one;

$n_2 = \sum_k I(T_k = J + 1)$ is the NO. of successful nests at the end.

For n_1, we then determine that:

$$H_1(E, A, \beta, \tau_1, \tau_2)$$

$$\propto \prod_{k=1}^{n} \frac{e^{E_{U_k}} e^{A_{T_k}}}{1 + \sum_{i=2}^{J} e^{E_i} + e^{x_k'\beta} \left(\sum_{j=1}^{J} e^{A_j} + \sum_{j \geq i \geq 2}^{J} e^{E_i} e^{A_j} \right)}$$

$$\cdot \tau_1^{-(J-3)/2} exp\left\{ -\frac{1}{2\tau_1} \sum_{i=3}^{J-1} \xi_i D_i^2 \right\} \tau_2^{-(J-2)/2} exp\left\{ -\frac{1}{2\tau_2} \sum_{i=3}^{J} \eta_j F_j^2 \right\}$$

$$\leq \frac{1}{C_x^{n_1} (e^{E_l} + 1)^{\frac{n_1}{2}} (e^{E_m} + 1)^{\frac{n_1}{2}}} \tau_1^{-(J-3)/2}$$

$$\cdot exp\left\{ -\frac{1}{2\tau_1} \sum_{i=3}^{J-1} \xi_i D_i^2 \right\} \tau_2^{-(J-2)/2} exp\left\{ -\frac{1}{2\tau_2} \sum_{i=3}^{J} \eta_j F_j^2 \right\}$$

And for n_2, we have:

$$H_1(E, A, \beta, \tau_1, \tau_2)$$

$$\propto \prod_{k=1}^{n} \frac{e^{E_{U_k}} e^{A_{T_k}}}{1 + \sum_{i=2}^{J} e^{E_i} + e^{x_k'\beta} \left(\sum_{j=1}^{J} e^{A_j} + \sum_{j \geq i \geq 2}^{J} e^{E_i} e^{A_j} \right)}$$

$$\cdot \tau_1^{-(J-3)/2} exp\left\{ -\frac{1}{2\tau_1} \sum_{i=3}^{J-1} \xi_i D_i^2 \right\} \tau_2^{-(J-2)/2} exp\left\{ -\frac{1}{2\tau_2} \sum_{i=3}^{J} \eta_j F_j^2 \right\}$$

$$\leq \frac{1}{(C_x e^{A_s} + 1)^{\frac{n_2}{2}} (C_x e^{A_t} + 1)^{\frac{n_2}{2}}} \tau_1^{-(J-3)/2}$$

$$\cdot exp\left\{ -\frac{1}{2\tau_1} \sum_{i=3}^{J-1} \xi_i D_i^2 \right\} \tau_2^{-(J-2)/2} exp\left\{ -\frac{1}{2\tau_2} \sum_{i=3}^{J} \eta_j F_j^2 \right\}$$

After this, we rewrite it as:

$$\int H_1(E, A, \beta, \tau_1, \tau_2) dE dA < C_1 \gamma_1 \gamma_3 \gamma_4 \text{ , if } n_1 \geq 3$$

and

$$\int H_1(E, A, \beta, \tau_1, \tau_2) dE dA < C_2 \gamma_2 \gamma_3 \gamma_4 \text{ , if } n_2 \geq 3$$

Where C_1 and C_2 are normalizing constants.
It is easy from here to show that $\gamma_1, \gamma_2, \gamma_3$ and γ_4 are all less than infinty, which proves the existence of the posterior distribution.

This is just the shorter version of the proof that Wells(2007) had showed, if readers wish to acquire more details about the existence of the posterior distribution, I recommand to read the full proof from it.

Appendix B: Additonal facts of the DIC

P_D is the effective number of parameters in DIC, as we have learnt that to choose the different prior would have change the P_D and thus change the value of the DIC. But is the chioce of parameterzation would to have a strong effect for different type of priors?

Section 5 and section 8 in Spiegehalter(2002) have tested this by choosing binomial, Poisson and Bernoulli priors and by observing the corresponding results of P_D, the conclusion was a mixture. It seemed that for binomial and Possion priors, the parameterization didnt show any strong effect, however, as for the Bernoulli model, the result was different and the P_D did affect much. The reason for this special behaviour of P_D may have many explainations; P_D may be only approximately invariant to the chosen parameterization, because the different fitted deviance $\mathbf{D}(\bar{\theta})$ could arise from replacing posterior means of an alternative θ. This is like in the section 8 of Spiegehalter(2002) and could be important for Bernoulli data. If we use the posterior median as an estimator and use it to find P_D, it may have little effect as there are two possible disadvantages if we do so: we do not know that P_D will be postive and additionally there are some computational difficulties theoretically because the approximate properties which are based on the Taylor expansions may not hold. So DIC today are recommanded that the calculations are based on several different estimators with a preference for posterior means that its parameterization obeying approximate likelihood normality.

References

Cao, J., He, C.Z., Wells, K.M.S., Millspaugh, J.J. and Ryan, M.R. (2009). Modeling age and nest-specific survival using a hierarchical Bayesian approach. *Biometrics*, **Vol.65**, pp. 1052-1062.

Clayton, David and Berzuini, Carlo. (1994). Bayesian analysis of survival on multiple time scales. *Statistics in medicine*, **Vol.13**, pp. 823-838.

Gelman, Andrew. (2006). Prior distributions for variance parameters in hierarchical models(Comment on article by Browne and Draper). *Bayesian analysis*, **1**, NO.3, pp. 515-534.

He, Chong Z. (2003). Bayesian modeling of age-specific survival in bird nesting studies under irregular visits. *Biometrics*, **Vol. 59**, pp. 962-973.

He, Chong Z., Sun, Dongchu and Tra, Yolande. (2001). Bayesian modeling of age-specific survival in nesting studies under dirichlet priors. *Biometrics*, **Vol. 57**, pp. 1059-1066.

Heisey, Dennis M. and Nordheim, Erik V. (1995). Modeling age-specific survival in nesting studies, using a general approach for doubly censored and truncated data. *Biometrics*, **Vol.51**, NO.1, pp. 51-60.

Hensler, Gary L. and Nichols, James D. (1981). The Mayfield method of estimating nest success: a model, estimators and simulation results. *The Wilson bulletin*, **Vol. 93(1)**, pp. 42-53.

Mayfield, Harold. (1961). Nesting success, *The Wilson bulletin*, **Vol. 73**, NO.3.

Mayfield, Harold. (1975). Calculating nest success, *The Wilson bulletin*, **Vol. 87**, NO.4.

Pollock, K.H. and Cornelius, W.L. (1988). A distribution-free nest survival model. *Biometrics*, **Vol.44**, NO.2, pp. 397-404.

Speckman, Paul L. and Sun, Dongchu. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors, *Biometrics*, **Vol. 90**, pp. 289-302.

Spiegelhalter, David J., Best, Nicola G., Carlin, Bradley P., and Van Der Linde, Angelika. (2002). Bayesian measures of model complexity and fit. *J.R.Statist.Soc.B*, **64**, Part 4, pp. 583-639.

Wells, K.M.S., Ryan, M.R., Millspaugh, J.J., Thompson III, F.R. and Hubbard, M.W. (2007). Survival of postfledging grassland birds in Missouri. *The Condor*, **Vol.109**, pp. 781-794.