Extensions of the Underhill-Zucchini Model for Avian Molt Data Allowing Random Duration

Guangyu Wang

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2020

Reading Committee:

Youyi Fong, Chair

James Dai

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Abstract

Extensions of the Underhill-Zucchini Model for Avian Molt Data Allowing Random Duration

Guangyu Wang

Chair of the Supervisory Committee:
Associate Professor Youyi Fong
Department of Biostatistics

Avian molting process is worth studying for its role in reflecting climate change or human activities. Two quantities, initiation molt time and duration time, are of great significance when describing avian molt. Existing models for the molting process use either a linear-based model, which treats both parameters as constants for all birds, or the Underhill-Zucchini model, which treats the initiation molt time as a random variable. In this thesis, we study extended models that allow both initiation time and duration to be random. The identifiability of the new model and its relationship with other models are then assessed. Using simulation studies, we compare the results from various models and evaluate conditions when some models are preferable. Finally, we show that we can get more accurate estimates from study designs with multiple observations per individual.

TABLE OF CONTENTS

Page

LIST OF FIGURES

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to his academic advisor, Youyi Fong, second reader, James Dai, and all supervisory committee members for their consistent help and advice in the completion of this thesis.

The author also shows his gratitude to all faculties, staff, and students in the department of biostatistics, University of Washington, who provided supports throughout his entire master's program.

Last but not least, thanks to the author of TEX, Donald Knuth, who makes it possible for the author to work with the TEX formatting system.

DEDICATION

Dedicated to my family and friends who always give encouragement and trust to me.

Chapter 1

INTRODUCTION

Feather molt plays an essential role during the whole life of a bird. The feather provides vital biological functions for birds, such as controlling the body temperature and influencing flight. The avian annual molt progress is regulated by environmental factors including climate, distribution of resources, and growth history, all of which are subject to change in recent years due to human activities. Therefore, analyzing avian molt can help people study the effects of human activities on birds through these environmental factors. Comparing molt timing between species also give hints about the evolutionary history of different bird species. One of the most direct ways of describing the molting process is to construct a statistical model to reveal how avian molt progresses over time in a year.

1.1 Construct the proportion of feather mass grown (PFMG)

As an annual event, field studies using mist-net or specimens in museums serve as the source of avian molt data. A quantity indicating molt progression needs to be constructed from the original samples for further analysis. With the knowledge that the molting process starts from the innermost primary on birds' wings and moves outwards (Figure 1.1), the molt progression can be represented by a sequence of non-increasing integers (s_i), each of which indicates the degree of completion for primary at that location, with 0 denoting molt not started yet and 5 denoting the completeness of molt (Ginn and Melville, 1983; Underhill and Zucchini, 1988). The sum of these integers is a primary measure of avian molt, referred to as the molt score.

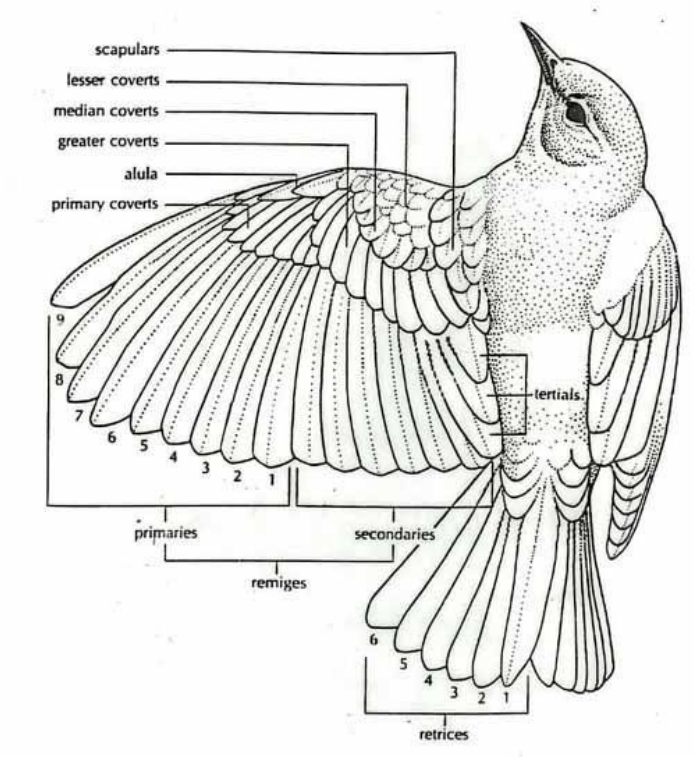


Figure 1.1: Bird wings diagram with 1 indicating the innermost primary or rectrix.

Most birds have longer outer primaries than inner primaries. The molt score is approximately linear on time for species whose outer primaries molt faster than inner primaries. However, Summers (2003) pointed out this does not hold for all species, especially birds with extremely large outer primaries that are several times longer than the innermost ones. It is also more reasonable to assume a constant molting rate (Summers, 2003; Summers et al., 1983; Underhill, 1985). To guarantee the linearity, an adjustment is made on the molt score and then brings another measure of molting progression, the proportion of feather mass grown (PFMG)

$$y = PFMG = \sum_{i=1}^{n} m_i p(s_i),$$

where m_i is the proportional weight of the ith primary relative to the total weight of all primaries, and $p(s_i)$ is the proportional weight of a feather with score s_i relative to its weight when the feather is fully grown, also known as molt index. When estimating the m_i and

$p(s_i)$, $p(s_i)$ is assumed to be the same for all feathers, and s_i are allocated in a way such that the change in score (s_i) is proportional to the change in feather growth (Evans, 1966). Defined in such a way, PFMG is controlled between 0 and 1, which is concise and intuitive as a measure describing the molting process. A bird that has not started molt should have PFMG=0; a bird with fully new feathers should have PFMG=1; birds undergoing molt should have PFMG in between.

1.2 Linear-based model

After having a metric representing molt progress, we can figure out the property of molting process by proposing the primary question that how the PFMG changes over time throughout a year for a species. Since we made the assumption regarding PFMG that molt score, s_i, are distributed in a way such that the change in score is proportional to the change in feather growth, and the molting rate is constant, it is logical to impose a linear relationship between PFMG and time in a model. Therefore, the most straightforward approach to model the molting process, i.e., how PFMG changes over dates (starting from July 1^{st} and ending in June 30^{th} in the following year), is applying a linear regression only upon birds in molt when captured. It is plausible to construct a threshold model to imitate the avian molting process: the molting process starts at the initiation molt time and ends at the completion time; the PFMG is 0 for birds before molt starts and 1 for birds with completely new feathers. In accordance with how we defined PFMG, the linear model is forced to be 0 and 1 for PFMG less than 0 or greater than 1 on the regression line, respectively. Therefore, a linear-based model gives a model with two parameters and two platforms. Each observation takes the form (t_i, y_i), where t_i and y_i represent date and PFMG, respectively. Let $Y(t)$ represent PFMG at time t; the linear-based model takes the form

$$Y(t) = f(t;\mu,\tau) = \begin{cases} 0 & \text{if} \quad t \le \mu \\ \frac{t-\mu}{\tau} + \epsilon & \text{if} \quad \mu < t < \mu + \tau \\ 1 & \text{if} \quad t \ge \mu + \tau \end{cases}, \tag{1.1}$$

$$\epsilon \sim N(0, \sigma_\epsilon^2)$$

where τ is the molting duration, and μ is the initiation time. τ and μ are assumed to be constant for all birds. The parameters to be estimated are μ, τ, and σ_ϵ.

In a linear-based model, the variance in error is assumed constant at any time. The initial molt time and duration are treated as unknown parameters, and the PFMG is assumed to increase linearly with time. The estimates are computed under the criteria that the mean squared error (MSE) is minimized:

$$(\hat{\mu}, \hat{\tau}) = \arg\min_{\mu,\tau} \sum_{i=1}^{n} (y_i - f(t_i; \mu, \tau))^2, 0 < y_i < 1.$$

Summers et al. (1983) implemented the linear model in practice by solving μ and τ from normal equations or optimization algorithms. A grid search among all possible initiation date and duration pairs serves as an alternative approach to getting the estimates of μ and τ.

An issue with this model is that instead of following up with birds to get their PFMGs, we can only get snapshot data once for each bird. Pimm et al. (1976) pointed out that linear regression usually gave incorrect results due to data's heteroscedasticity. If creating a scatterplot for all available observations, it is clear that the data cloud has a parallelogram shape and that the variability in molt index for birds near the commencement and conclusion of the molting process is way less than the variability for birds in the middle of their molt process (Figure 1.2). Therefore, the regression line tends to provide the starting and ending molt time for birds in the whole population instead of the average starting and ending time. Though some posterior adjustments were made by Summers et al. (1983) to improve the linear-based model's performance, the mean-variance relation violates the basis of linear regression and should not be applied from a statistical perspective.

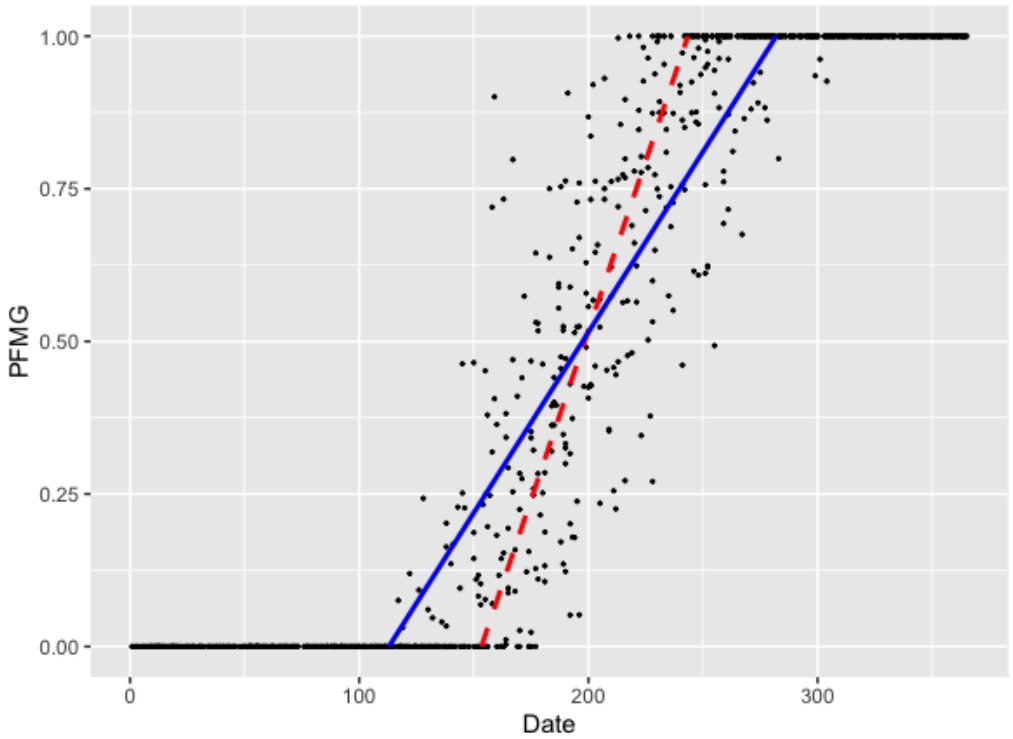


Figure 1.2: Linear-based regression line for simulated data. The dashed red line reflects the actual average molt process from which the data were generated, but the model based on linear regression will give the solid blue line, which has a large bias for birds near the commencement and completion of the molting process.

1.3 Reversed linear model

We have noticed that the heteroscedasticity of the data makes the general linear-based approach invalid. Some statisticians proposed an idea that switches the role of PFMG and time in the linear model to overcome this issue. Pimm et al. (1976) proposed a reversed version of linear regression for all observations in molt. The model takes the form

$$t_j = \mu + \tau y_j + e_j, \ j = 1, 2, ..., J,$$

where e_i is the error term with mean zero, and assumed to be independent of each other with identical distribution. In this expression, time is treated as the outcome and PFMG acts as an independent variable.

This assumption is correct when PFMG acts as the explanatory covariates. In practice, however, researchers are only able to determine the time of collecting samples (t_j) but not PFMG (y_j). The uncertainty contributes to the PFMG computation instead of selecting the sampling dates. As y_j takes value $0 < y_j < 1$, the error should have the restriction $t_j - \mu - \tau < e_j < t_j - \mu$ (Underhill and Zucchini, 1988). This implies e_j has different boundaries for different choices of t_j, hence violating the *i.i.d.* assumption. Underhill and Zucchini (1988) showed the starting date μ will be underestimated, and the duration τ will be overestimated on average. The bias tends to be minimized when observations are in $(\mu + 2\sigma, \mu + \tau - 2\sigma)$, which generally contains no sample data.

Therefore, ϵ in linear-based model (1.1) should have variable variance over time, i.e. $\epsilon(t) \sim N(0, g(t) \cdot \sigma_\epsilon^2)$ where $g(t)$ serves as a scalar. $g(t)$ should be positive and have smaller values for observed data near boundaries but larger values in between. One possible choice is $g(t) = sin(\pi \cdot f(t))$. Since the error term depends on PFMG, the optimization process is convoluted. On the contrary, if the variability can be attributed to the variance in starting time or duration, we can treat starting time or duration as random variables. Such a conversion gives us the regular and extended Underhill-Zucchini model.

1.4 Regular Underhill-Zucchini Model

To solve the drawbacks of the general and reversed linear model, Underhill and Zucchini (1988) constructed a new model that treats initiation time and duration differently. It differs from the linear-based model in the way that the initiation time, X, is treated as a random variable, which contains the model's variability previously presented in ϵ. Duration is still assumed to be constant for all individual birds. Let $Y(t)$ denotes the PFMG at time t. The model takes the form

$$Y(t) = f(t; X, \tau) = \begin{cases} 0 & \text{if} \quad t \leq X \\ \frac{t-X}{\tau} & \text{if} \quad X < t < X + \tau \\ 1 & \text{if} \quad t \geq X + \tau \end{cases}, \tag{1.2}$$
$$X \sim N(\mu_X, \sigma_X^2)$$

The parameters to be estimated are μ_X,σ_X, and τ. This model is referred to as the

regular Underhill-Zucchini model (Underhill and Zucchini, 1988; Underhill et al., 1990). The onset molt date follows a parametric distribution (typically a normal distribution with mean μ_X and variance σ_X^2 assumed). The length of the molt duration is still treated as a parameter. These parameters are estimable using the maximum likelihood method.

In the linear-based model, both initiation time and duration are constant, thus we have to use an error term to capture the model's variability. The non-constant variance in that error term leads to a problem when using linear regression to fit the data. In the Underhill-Zucchini model, however, the variability can be captured in the random variable X. There is no error term in the model at all. We attributed the variability from error term to initiation time and did not use a linear regression approach.

By comparing the regular Underhill-Zucchini model with the linear-based model using both simulated and empirical data, we noticed that the estimates' accuracy was improved in the regular Underhill-Zucchini model, mostly when insufficient observations were available. The regular Underhill-Zucchini model takes benefits by incorporating birds before, in, and after molt while estimating parameters, whereas the linear-based model only uses partial data, i.e., all birds in molt. In terms of how the regular Underhill-Zucchini model incorporates all observation, we will include more details in our extended model of which the regular Underhill-Zucchini model is a reduced case.

Chapter 2

THE EXTENDED UNDERHILL-ZUCCHINI MODEL

2.1 The extended Underhill-Zucchini model

In existing methods, molt duration is always treated as a parameter, which is the same for all birds but may not reflect the truth in the real world. The starting molt date of birds may also affect the length of molt by influencing temperature or sunlight. Sometimes instead of the mean of the variables, we are much more interested in their deviation that cannot be accurately estimated using the regular Underhill-Zucchini model. Therefore, the model needs to be extended to fit in more conditions.

The extended Underhill-Zucchini model differs from the regular Underhill-Zucchini model in that both starting date and duration, X and W, are treated as random quantities. As in the previous section, the error term in linear regression is implied in the random variables, and the PFMG is thereby not modeled as per the linear-based approach. X and W can be jointly distributed as any reasonable parametric distribution. In the following discussion, we assumed they follow bivariate normal distribution to simplify the calculation. The model then takes the following form, in which $\mu_X, \sigma_X, \mu_W, \sigma_W$, and ρ will be estimated via the maximum likelihood approach.

$$Y(t) = f(t; X, W) = \begin{cases} 0 & \text{if} \quad t \leq X \\ \frac{t-X}{W} & \text{if} \quad X < t < X + W \\ 1 & \text{if} \quad t \geq X + W \end{cases}, \tag{2.1}$$

$$\begin{pmatrix} X \\ W \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_X \\ \mu_W \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_W \\ \rho\sigma_X\sigma_W & \sigma_W^2 \end{pmatrix} \right].$$

To construct a likelihood function, the mass function or density function for PFMG at a given date is necessary. For birds in molt, the PFMG can be expressed as

$$Y(t) = \frac{t - X}{W}.$$

The density function of such ratio needs to be calculated. Geary (1930) proved that the ratio of two jointly normally distributed variables (normal ratio) could be approximated to a new random variable, provided that the coefficient of variation for denominator tends to zero. Hinkley (1969) and Marsaglia et al. (2006) then demonstrated the relationship between the cumulative density functions of the normal ratio and the standard normal variable. In our setting, the molt duration should have a positive mean. If $\sigma_W/\mu_W \to 0$, then we get

$$F_{Y(t)}(y) = P(Y(t) \leq y) = P(\frac{t-X}{W} \leq y) \to P(t - X - yW \leq 0).$$

Though we listed the assumption that $\sigma_W/\mu_W \to 0$, the derivation is still valid when the coefficient of variation is slightly larger. The only reason we made this assumption is to ensure W is positive. According to normal distribution's property, the probability of $W < 0$ is low as long as the coefficient of variation is not too large. Since X and W are jointly normally distributed, the cumulative function of $Y(t)$ can be approximated as follows:

$$F_{Y(t)}(y) \to \Phi(\frac{y\mu_W - t + \mu_X}{\sqrt{\sigma_X^2 + 2y\rho\sigma_X\sigma_W + y^2\sigma_W^2}}).$$

The density function of $Y(t)$ is available by taking the derivative of the cumulative function,

$$\begin{aligned} f_{Y(t)}(y) = \phi(&\frac{y\mu_W - t + \mu_X}{\sqrt{\sigma_X^2 + 2y\rho\sigma_X\sigma_W + y^2\sigma_W^2}}) \\ &\times \frac{\mu_W\sigma_X^2 + \rho\sigma_X\sigma_W(t-\mu_X) + y[(t-\mu_X)\sigma_W^2 + \rho\sigma_X\sigma_W\mu_W]}{(\sigma_X^2 + 2y\rho\sigma_X\sigma_W + y^2\sigma_W^2)^{3/2}}. \end{aligned}$$

For birds having completed or not started molt, their mass functions can be achieved from distribution of $X + W$ and X. Let $S = X + W$, then

$$S \sim N(\mu_X + \mu_W, \sigma_X^2 + \sigma_W^2 + 2\rho\sigma_X\sigma_W).$$

To keep alignment with current literature, we use the same notification introduced by Underhill and Zucchini (1988). Suppose there are I, J, K birds in the sample from subgroup *not commenced*, *in molt*, and *molt completed*, respectively, for which the sampling time are $t_1, t_2, ..., t_I$; $u_1, u_2, ..., u_J$; $v_1, v_2, ..., v_K$. The molt indices for birds *in molt* are denoted as $y = (y_1, y_2, ..., y_J)$. Thus the likelihood function is

$$L(t, u, y, v; \boldsymbol{\theta}) = \prod_{i=1}^{I} P(t_i) \prod_{j=1}^{J} q(u_j, y_j) \prod_{k=1}^{K} R(v_k),$$

where $\boldsymbol{\theta}^T = (\mu_X, \sigma_X, \mu_W, \sigma_W, \rho)$, and

$$P(t) = 1 - F_X(t),$$
$$R(v) = F_S(v),$$
$$q(u, y) = f_{Y(u)}(y),$$

where F_X and F_S represent the corresponding cumulative density functions. Finally, the normal equations are

$$\frac{\partial logL}{\partial \theta_a} = \sum_{i=1}^{I} \frac{\partial P(t_i)}{\partial \theta_a} \frac{1}{P(t_i)} + \sum_{j=1}^{J} \frac{\partial q(u_j, y_j)}{\partial \theta_a} \frac{1}{q(u_j, y_j)} + \sum_{k=1}^{K} \frac{\partial R(v_k)}{\partial \theta_a} \frac{1}{R(v_k)} = 0$$
$$(a = 0, 1, 2, \ldots, p).$$

Rather than finding its analytic solution, we can maximize the logarithm of the likelihood function via numerical approaches. Our algorithm used Broyden–Fletcher–Goldfarb–Shanno (BFGS) for optimization. The extended Underhill-Zucchini model can be modified by allowing parameters to depend on other explanatory variables. Furthermore, if evidence shows that X and W are independent, the model can be easily adjusted by setting $\rho = 0$, hence constraining the parameter space to a lower dimension and giving estimates for restricted extended Underhill-Zucchini model. To distinguish different models in later sections, we may use **regular model** to refer the 3-parameter regular Underhill-Zucchini model, **restricted extended model** to refer the 4-parameter extended Underhill-Zucchini model assuming $\rho = 0$, and **unrestricted extended model** to refer the 5-parameter extended Underhill-Zucchini model without constraint.

2.2 Model's Identifiability

The extended Underhill-Zucchini model involves more random quantities than the regular model from two additional parameters, σ_W, and ρ. Although the estimates are available using a numerical optimization algorithm, it bears the risk of parameter redundancy, leading to unreliable estimates for some parameters. So before using this model in practice, its identifiability should be checked systematically. Catchpole and Morgan (1997) proposed

the definition of parameter redundancy. Exhaustive summary, a term first used by Walter and Lecourtier (1982), is also defined formally for further analysis.

Definition 1. *A model is parameter redundant if we can write the model* $M(\boldsymbol{\theta})$ *as a function just of* $\boldsymbol{\beta}$*, where* $\boldsymbol{\beta} = g(\boldsymbol{\theta})$ *and* $dim(\boldsymbol{\beta}) < dim(\boldsymbol{\theta})$*, where dim is the dimension or length of the vector.*

Definition 2. *A parameter vector* $\boldsymbol{\kappa}(\boldsymbol{\theta})$ *is an exhaustive summary if knowledge of* $\boldsymbol{\kappa}(\boldsymbol{\theta})$ *uniquely determines* $M(\boldsymbol{\theta})$*, so that* $\boldsymbol{\kappa}(\boldsymbol{\theta}_1) = \boldsymbol{\kappa}(\boldsymbol{\theta}_2) \iff M(\boldsymbol{\theta}_1) = M(\boldsymbol{\theta}_2), \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$.

Cole et al. (2010) proposed a gold standard to test parameters redundancy using the symbolic method, which computes the derivative matrix of the model's exhaustive summary with respect to each parameter. If the rank of the matrix is smaller than the number of parameters, the model is parameter redundant and non-identifiable. Otherwise, if the rank of the matrix is equal to the number of parameters, the model is full rank and at least locally identifiable for part of the parameter space. The method is stated formally in Theorem 1.

Theorem 1. *Let* $\mathbf{D}$ *be the derivative matrix by differentiating each of the exhaustive summary terms with respect to each of the parameters. If the rank of* $\mathbf{D}$ *is equal to* r*, the deficiency of a model with* q *parameters is* $d = q - r$*. If* $d > 0$*, then the model is parameter redundant and will be non-identifiable. If* $d = 0$*, the model is full rank and will be at least locally identifiable for part of the parameter space.*

Similar results can also be obtained from the Fisher Information matrix or Hessian matrix. One benefit of the symbolic method is by solving the partial differential equations, the estimable parameter combinations are available if the model is non-identifiable, which directs statisticians on how to reparameterize original parameters to get reliable estimates. If the model is merely conditionally full rank, which indicates the model can be non-identifiable in specific parameter subspaces, the symbolic method can also find those subspaces for us.

Suppose there are only two observations in our study, and both of them are in molt (with PFMG between 0 and 1). One possible way of constructing an exhaustive summary (Ran and Hu, 2014) is to include all components of the log-likelihood function, which in

this case is a one-to-one transformation of

$$\boldsymbol{\kappa} = \begin{bmatrix} \sigma_X^2 + 2y_1\rho\sigma_X\sigma_W + y_1^2\sigma_W^2 \\ \mu_W \\ y_1\mu_W - t_1 + \mu_X \\ \rho\sigma_X\sigma_W + y_1\sigma_W \\ \sigma_X^2 + 2y_2\rho\sigma_X\sigma_W + y_2^2\sigma_W^2 \\ \mu_W \\ y_2\mu_W - t_2 + \mu_X \\ \rho\sigma_X\sigma_W + y_2\sigma_W \end{bmatrix}.$$

Since σ_X, σ_W, and ρ are only defined on subsets of $\mathbb{R}$, we apply non-linear transformations to allow all parameters to take values in $(-\infty, \infty)$:

$$\begin{aligned} \mu_X &= \nu_x \\ \mu_W &= \nu_w \\ \sigma_X &= e^{\tau_x} \\ \sigma_W &= e^{\tau_w} \\ \rho &= \frac{2}{\pi}\arctan(\xi) \end{aligned}$$

where $\nu_x, \nu_w, \tau_x, \tau_w$ and ξ can take any value from $(-\infty, \infty)$. The transformation is one to one, hence the vector represented by new parameters is still an exhaustive summary

$$\boldsymbol{\kappa} = \begin{bmatrix} e^{2\tau_x} + \frac{4y_1}{\pi}\arctan(\xi)\cdot e^{\tau_x+\tau_w} + y_1^2 e^{2\tau_w} \\ \nu_w \\ y_1\nu_w + \nu_x - t_1 \\ \frac{2}{\pi}\arctan(\xi)\cdot e^{\tau_x+\tau_w} + y_1 e^{2\tau_w} \\ e^{2\tau_x} + \frac{4y_2}{\pi}\arctan(\xi)\cdot e^{\tau_x+\tau_w} + y_2^2 e^{2\tau_w} \\ \nu_w \\ y_2\nu_w + \nu_x - t_2 \\ \frac{2}{\pi}\arctan(\xi)\cdot e^{\tau_x+\tau_w} + y_2 e^{2\tau_w} \end{bmatrix},$$

and the corresponding derivative matrix is

$$\mathbf{D}^T = \begin{bmatrix} 0 & 0 & 2e^{2\tau_x} + \frac{4y_1}{\pi}\arctan(\xi)\cdot e^{\tau_x+\tau_w} & \frac{4y_1}{\pi}\arctan(\xi)\cdot e^{\tau_x+\tau_w} + 2y_1^2 e^{2\tau_w} & \frac{4y_1 e^{\tau_x+\tau_w}}{\pi(\xi^2+1)} \\ 0 & 1 & 0 & 0 & 0 \\ 1 & y_1 & 0 & 0 & 0 \\ 0 & 0 & \frac{2}{\pi}\arctan(\xi)\cdot e^{\tau_x+\tau_w} & \frac{2}{\pi}\arctan(\xi)\cdot e^{\tau_x+\tau_w} + 2y_1 e^{2\tau_w} & \frac{2e^{\tau_x+\tau_w}}{\pi(\xi^2+1)} \\ 0 & 0 & 2e^{2\tau_x} + \frac{4y_2}{\pi}\arctan(\xi)\cdot e^{\tau_x+\tau_w} & \frac{4y_2}{\pi}\arctan(\xi)\cdot e^{\tau_x+\tau_w} + 2y_2^2 e^{2\tau_w} & \frac{4y_2 e^{\tau_x+\tau_w}}{\pi(\xi^2+1)} \\ 0 & 1 & 0 & 0 & 0 \\ 1 & y_2 & 0 & 0 & 0 \\ 0 & 0 & \frac{2}{\pi}\arctan(\xi)\cdot e^{\tau_x+\tau_w} & \frac{2}{\pi}\arctan(\xi)\cdot e^{\tau_x+\tau_w} + 2y_2 e^{2\tau_w} & \frac{2e^{\tau_x+\tau_w}}{\pi(\xi^2+1)} \end{bmatrix}.$$

This matrix has rank $r = 5$. Therefore, if the dataset contains only two observations and both are in molt, the extended Underhill-Zucchini model is identifiable. If we want to generalize the conclusion from two observations to an arbitrary larger sample size, we need the extension theorem (Catchpole and Morgan, 1997).

Theorem 2. *The derivative matrix of the extended model is*

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1(\boldsymbol{\theta}_1) & \mathbf{D}_{2,2}(\boldsymbol{\theta}_1) \\ \mathbf{0} & \mathbf{D}_{2,2}(\boldsymbol{\theta}_2) \end{bmatrix} \text{ with } \mathbf{D}_{2,1} = \begin{bmatrix} \frac{\partial \boldsymbol{\kappa}_2}{\partial \boldsymbol{\theta}_1} \end{bmatrix} \text{ and } \mathbf{D}_{2,2} = \begin{bmatrix} \frac{\partial \boldsymbol{\kappa}_2}{\partial \boldsymbol{\theta}_2} \end{bmatrix}.$$

If $\mathbf{D}_1$ *is full rank and* $\mathbf{D}_{2,2}$ *is full rank, then the extended model will be full rank.*

Remark 1. *If zero or one new parameter is added to a full rank model, the model will still be full rank.*

Since the full rank model does not include new parameters when the sample size grows, the model is generally identifiable regardless of the number of observations (but should contain at least two observations in molt).

One property of the extended Underhill-Zucchini model is that it also works when constraints exist for parameters, which may lead to a reduction of parameter space's dimension. Therefore, the next step is to prove the model's identifiability in its nested (restricted) models.

Theorem 3. *If we write the derivative matrix as* $\mathbf{D} = \mathbf{PLUR}$*, where* $\mathbf{P}$ *is a permutation matrix,* $\mathbf{L}$ *is a lower triangular matrix with ones on the diagonal,* $\mathbf{U}$ *is an upper triangular*

matrix and $\mathbf{R}$ *is a matrix in reduced echelon form, then the model is parameter redundant at* $\boldsymbol{\theta}$ *if and only if* $Det(\mathbf{U}) = 0$ *at a point* $\boldsymbol{\theta} \in \Omega$*, as long as* $\mathbf{L}$*,* $\mathbf{U}$ *and* $\mathbf{R}$ *are defined at* $\boldsymbol{\theta}$ *(Corless and Jeffrey, 1997).*

Theorem 4. *The model with constraint* $\boldsymbol{\theta} = g(\boldsymbol{\theta})$*, nested with a full rank model with derivative matrix* $\mathbf{D} = \mathbf{PLUR}$*, will be parameter redundant if* $Det(\mathbf{U}|_{\boldsymbol{\theta}=g(\boldsymbol{\theta})}) = 0$*. The deficiency of the nested model is equal to the deficiency of* $\mathbf{U}|_{\boldsymbol{\theta}=g(\boldsymbol{\theta})}$ *(Cole et al., 2010).*

To check the identifiability of the nested (restricted) model with constraints in parameters space, the derivative matrix is decomposed via **PLUR** approach. The determinant of $\mathbf{U}$ is

$$Det(\mathbf{U}) = -\frac{8(y_1 - y_2)^2}{\pi(\xi^2 + 1)} \cdot e^{3(\tau_x + \tau_w)}.$$

If $y_1 \neq y_2$, the determinant will not be zero. In practice, τ_x and τ_w are very unlikely to take negative values (in which case σ_x and σ_w are smaller than 1); hence the determinant will even not be close to zero. The parameters in our model are then proved theoretically identifiable, with or without constraints in parameter space.

In some situations, the exact derivative matrix of exhaustive summaries, or the Hessian matrix, is difficult to obtain. In complex models, the Hessian matrix is almost always found numerically, which may not be singular even in a non-identifiable model. Therefore, the existence of extremely small standardized eigenvalues (divided by the largest eigenvalue) for the approximate Hessian matrix often indicates the non-identifiability (Chis et al., 2016). By checking the associated eigenvectors for small eigenvalues, Zhou et al. (2019) and Cole (2019) pointed that any entries close to zero for all these eigenvalues correspond to parameters that are at least locally identifiable. Simulated data are generated to test model-based identifiability (or intrinsic identifiability) caused by the parameterization. Since the non-identifiable parameters do not have standard error (Fisher Information is singular) but we get them using numerical analysis, the coefficient of variance may be large for non-identifiable parameters. Though the model is proved to be identifiable in theory, the estimates may not perform well in practice, which gives rise to the issue of practical non-identifiability, near-redundancy, or sloppiness. **Definition 3-5** are how these terms defined by Cole et al. (2010). We will visualize the estimates with these issues by observing their likelihood-based

confidence regions on their likelihood profiles in chapter 3 (Figure 3.1). Raue et al. (2009) provided an approach to check the practical non-identifiability using likelihood-based confidence region for parameter estimates. For a model with q parameters, a $100(1-\alpha)\%$ likelihood-based confidence region for parameter vector, $\boldsymbol{\theta}$, is a region in parameter space that follows

$$\{\boldsymbol{\theta} | l(\hat{\boldsymbol{\theta}}) - l(\boldsymbol{\theta}) < \frac{1}{2}\chi^2_{\alpha,q}\},$$

where $\chi^2_{\alpha,q}$ is the upper α point in a chi-squared distribution with q degrees of freedom.

Definition 3. *A parameter is practically non-identifiable if the log-likelihood has a unique maximum, but the parameter's likelihood-based confidence region tends to infinity in either or both directions.*

Definition 4. *A near-redundant model is one that is formally full rank, but might behave as a parameter redundant model because it is very similar to a parameter redundant model for particular data sets (Catchpole et al., 2001; Cole et al., 2010).*

Definition 5. *A model is sloppy if the condition number of the Hessian or Fisher information matrix is large, where the condition number is the largest eigenvalue of a matrix divided by the smallest eigenvalue (Chis et al., 2016; Dufresne et al., 2016).*

2.3 Relationship between different models

In general, if we assume that PFMG grows linearly with time, the variability of the molting process comes from three parts: initiation time, duration, and measurement error. To ensure the discussion's integrity, we introduce a full random model treating all these three quantities as random variables. The model can be easily expressed by appending a measurement error

term. The most flexible full random model takes the form

$$
\begin{aligned}
Y(t) &= f\left(t; X, W\right) + \delta \\
f\left(t; X, W\right) &= \begin{cases} 0 & \text{if} \quad t \leq X \\ \frac{t-X}{W} & \text{if} \quad X < t < X + W \\ 1 & \text{if} \quad t \geq X + W \end{cases}, \\
\begin{pmatrix} X \\ W \end{pmatrix} &\sim N\left[\begin{pmatrix} \mu_X \\ \mu_W \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_W \\ \rho\sigma_X\sigma_W & \sigma_W^2 \end{pmatrix} \right] \\
\delta \mid X, W &\sim N(0, h(f) \cdot \sigma_\delta^2).
\end{aligned}
\tag{2.2}
$$

The parameters to be estimated are $\mu_X, \sigma_X, \mu_W, \sigma_W$, ρ, and σ_δ. The measurement error is related to the PFMG in a way that its variance (controlled by h function) is smaller for PFMG near 0 or 1 and larger for PFMG somewhere in between. It makes sense because the measurement tends to be stable when molting progression is near its boundaries.

The full random model (extended Underhill-Zucchini model with measurement error) regards all three variables as random quantities. Such a model is reasonable because it is closest to the actual molt process and the data collection process. However, the model is too complicated to be estimated using the available data because we have only one record for each individual (see an extension in chapter 4 when multiple records are available). To ensure the model is estimable, measurement error δ in model (2.2) is ignored because its influence should be negligible compared to other random quantities. In addition, we can always improve measurement accuracy by conceiving reliable measuring techniques or transformation formula from raw observation to PFMG. The model without measurement error is the (unrestricted) extended Underhill-Zucchini model introduced in section 2.1. The molt duration can be further fixed as a parameter to simplify the model. Likewise, without considering the measurement error, such a model is equivalent to the regular Underhill-Zucchini model. If we continue reducing the model's flexibility by treating initiation time as a parameter and only use birds in molt to construct the model, the variability in data can no longer be implied in any random variables. Therefore, we have to add the error term back (as in model 1.1) to take the variability of the model and construct the linear model. As discussed in section 1.3, such an error term should have a variable standard

deviation to validate the analysis. All of these models assume linearity as well as platforms at $PFMG = 0$ and $PFMG = 1$. See Figure 2.1 for a brief overview of the relationship between different models.

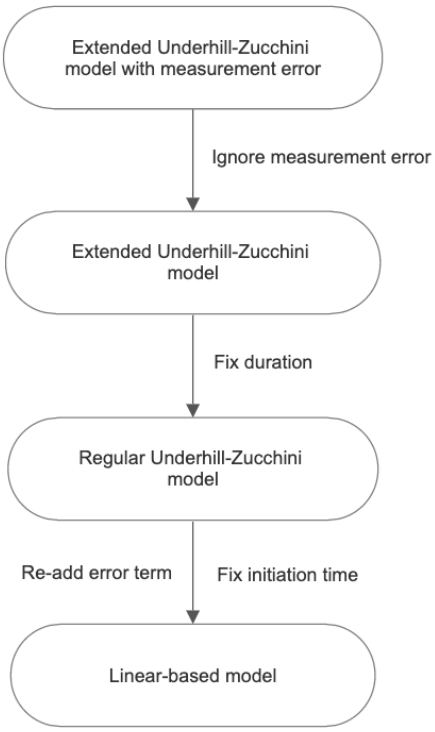


Figure 2.1: Relationship between models with different flexibility.

In practice, since the measurement precision can be improved via precise measuring tools and has less impact on PFMG than the uncertainty of initiation time, the extended Underhill-Zucchini model reflects the truth well even though it ignores the measurement error.

Chapter 3

NUMERICAL STUDIES

3.1 Data Generating Mechanism for Simulation

To investigate the extended Underhill-Zucchini models' performance, especially the maximum likelihood estimates' consistency and efficiency, we undertook a series of simulation studies. We first sampled time t_i from a uniform distribution

$$T \sim Uniform[1, 365]$$

Then initiation time x_i and duration w_i are generated from the bivariate normal distribution

$$\begin{pmatrix} X \\ W \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_X \\ \mu_W \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_W \\ \rho\sigma_X\sigma_W & \sigma_W^2 \end{pmatrix} \right].$$

Due to linearity, the PFMG can be calculated from x_i, w_i, and t_i. It is set to be 0 or 1 for all observed dates before the selected initiation date (x_i) or after the conclusion date ($x_i + w_i$). For a time in between, PFMG is computed through a linear function that connects the two platforms at 0 and 1. The measurement error term can thenceforth be added to the computed molt index. Since PFMG ranges from 0 to 1, any results beyond the boundaries after adding error terms should be forced as 0 or 1. However, in our studies, we ignored measurement error while simulating data because the extended Underhill-Zucchini model does not incorporate measurement error in its model. All simulated data (before, after, or in molt) were used to optimize the parameters.

Monte-Carlo replicates were used to test the consistency and efficiency of the estimates. Given a sample size, we repeated the simulating procedure 10000 times, calculated the mean and standard error for each parameter, and then examined each parameter's bias and convergence rate.

3.2 Consistency of estimates

We used the Monte-Carlo method to check the maximum likelihood estimates' consistency for all parameters using different models. First of all, data were generated from the regular Underhill-Zucchini model. The true parameters used to generate data are $\mu_{X,0} = 150$, $\mu_{W,0} = 100$, and $\sigma_{X,0} = 20$. We separately applied the regular Underhill-Zucchini model, restricted extended Underhill-Zucchini model, and unrestricted extended Underhill-Zucchini model to the simulated data. Different sample sizes were chosen from 50 to 6400. The estimates and corresponding standard errors are shown in Table 3.1.

n	3 - parameter UZ model						4 - parameter UZ model								5 - parameter UZ model									
	μ_W		μ_X		σ_X		μ_W		μ_X		σ_W		σ_X		μ_W		μ_X		σ_W		σ_X		ρ	
	est	se	est	se	est	se	est	se	est	se	est	se	est	se	est	se	est	se	est	se	est	se	est	se
50	101	12.6	150	8.1	18.6	3.59	101	12.9	149	8.1	7.9	10.53	16.6	5.14	100	13.4	150	8.7	17.9	15.40	19.4	8.44	-0.15	0.91
100	100	8.6	150	5.4	19.3	2.51	100	8.6	150	5.4	6.0	7.90	18.3	3.06	99	8.8	150	5.5	14.4	12.57	20.4	5.08	-0.22	0.82
200	100	6.0	150	3.7	19.7	1.76	100	6.0	150	3.7	5.6	6.69	19.0	2.10	99	6.1	150	3.8	11.7	10.35	20.3	3.30	-0.20	0.76
400	100	4.1	150	2.6	19.9	1.24	100	4.1	150	2.6	4.8	5.72	19.4	1.47	100	4.2	150	2.7	9.7	8.71	20.4	2.31	-0.17	0.74
1600	100	2.1	150	1.3	20.0	0.60	100	2.1	150	1.3	3.4	3.87	19.7	0.70	100	2.1	150	1.3	6.8	6.45	20.2	1.12	-0.11	0.64
6400	100	1.0	150	0.7	20.0	0.31	100	1.0	150	0.7	2.4	2.72	19.9	0.36	100	1.0	150	0.7	4.6	4.50	20.1	0.58	-0.07	0.52

Table 3.1: MC means (est) and standard errors (se) of parameter estimates using regular model, restricted extended model, and unrestricted extended model, for simulated data generated from regular Underhill-Zucchini model. True parameter values: $\mu_{W,0} = 100$, $\mu_{X,0} = 150$, and $\sigma_{X,0} = 20$.

The results show that when data ware simulated from the regular model, the estimates and standard error for mean initiation (μ_X) time and mean duration (μ_W) are similar for all three models. This makes sense because none of these models impose a false structure on the data. For standard deviation of initiation time (σ_X), the unrestricted model has the lowest bias but largest variance when the sample size is small due to its flexibility, but the regular model outperforms others as the sample size grows. The estimated correlation coefficient (ρ) from the unrestricted extended model has a small bias but again a large variance. Neither

restricted nor unrestricted extended model provides a reasonable estimate of the standard deviation of duration (σ_W). These conclusions indicate that when the underlying truth is the regular model, the regular model performs well enough since we should not expect unreasonable estimates for σ_W and ρ. However, the actual model is very unlikely to be a regular model in the real world.

More datasets were generated from the extended Underhill-Zucchini model. The true parameters used to generate data are $\mu_{X,0} = 150$, $\mu_{W,0} = 100$, $\sigma_{X,0} = 20$, and $\sigma_{W,0} = 10$. The correlation coefficient (ρ_0) takes four different values (-0.5, -0.3, 0, and 0.3), each of which was then fitted by regular model, restricted extended model with constraint $\rho = 0$, and unrestricted extended model. The estimates and corresponding standard errors are given in Table 3.2.

n	3 - parameter model						4 - parameter model								5 - parameter UZ model									
	μ_W		μ_X		σ_X		μ_W		μ_X		σ_W		σ_X		μ_W		μ_X		σ_W		σ_X		ρ	
	est	se	est	se	est	se	est	se	est	se	est	se	est	se	est	se	est	se	est	se	est	se	est	se
$\rho_0 = -0.5$																								
50	100	11.5	149	7.4	16.9	3.35	100	11.9	149	7.7	6.0	8.89	15.4	4.65	99	12.6	150	8.2	17.8	15.62	18.7	8.11	-0.24	0.89
100	100	8.2	150	5.2	17.6	2.36	100	8.2	150	5.2	3.8	6.15	17.0	2.60	99	8.3	150	5.4	13.8	11.75	19.6	4.71	-0.37	0.78
200	100	5.6	150	3.8	18.1	1.71	100	5.6	150	3.8	2.5	4.66	17.8	1.84	100	5.7	150	3.8	12.7	9.84	20.1	3.20	-0.49	0.65
400	100	4.0	150	2.5	18.2	1.16	100	4.0	150	2.5	1.6	3.38	18.1	1.22	100	4.0	150	2.5	11.3	8.76	20.2	2.35	-0.52	0.58
1600	100	2.0	150	1.3	18.3	0.58	100	2.0	150	1.3	0.3	1.21	18.3	0.59	100	2.1	150	1.3	9.6	6.21	20.1	1.17	-0.63	0.30
6400	100	1.0	150	0.7	18.3	0.29	100	1.0	150	0.7	0.0	0.08	18.4	0.29	100	1.0	150	0.7	9.3	4.23	20.0	0.61	-0.59	0.17
$\rho_0 = -0.3$																								
50	100	12.2	150	7.5	17.9	3.53	100	12.4	149	7.7	8.1	10.23	15.8	4.99	99	13.3	150	8.4	18.1	16.35	18.4	8.48	-0.13	0.91
100	100	8.4	150	5.3	18.6	2.48	100	8.5	150	5.3	5.8	7.60	17.7	2.88	99	8.6	150	5.4	14.3	12.23	19.6	4.80	-0.25	0.80
200	100	5.8	150	3.8	19.1	1.80	100	5.8	150	3.8	5.0	6.43	18.5	2.06	100	5.9	150	3.9	12.8	10.22	20.1	3.29	-0.26	0.73
400	100	4.1	150	2.5	19.3	1.20	100	4.1	150	2.5	4.1	5.16	18.9	1.36	100	4.2	150	2.6	11.2	9.28	20.2	2.34	-0.27	0.67
1600	100	2.1	150	1.3	19.4	0.61	100	2.1	150	1.3	2.8	3.58	19.2	0.68	100	2.1	150	1.3	9.3	6.90	20.1	1.19	-0.30	0.52
6400	100	1.1	150	0.7	19.4	0.30	100	1.1	150	0.7	1.6	2.22	19.3	0.33	100	1.1	150	0.7	8.8	4.90	20.0	0.62	-0.30	0.30
$\rho_0 = 0$																								
50	100	12.5	150	7.6	19.3	3.79	100	12.7	150	7.8	10.9	11.53	16.5	5.40	99	13.1	150	8.1	19.3	16.47	18.4	8.15	-0.06	0.90
100	100	8.7	150	5.3	20.1	2.66	99	8.7	150	5.3	9.2	9.07	18.5	3.16	99	8.9	150	5.5	15.3	12.85	19.7	4.90	-0.05	0.82
200	100	6.1	150	3.8	20.6	1.92	100	6.1	150	3.9	9.1	7.90	19.3	2.32	99	6.1	150	3.9	13.6	10.90	20.1	3.39	0.01	0.75
400	100	4.3	150	2.5	20.8	1.29	100	4.3	150	2.5	8.8	6.49	19.7	1.59	100	4.3	150	2.6	11.9	9.67	20.2	2.33	0.05	0.70
1600	100	2.2	150	1.3	20.8	0.65	100	2.2	150	1.3	9.1	4.29	19.9	0.87	100	2.2	150	1.3	9.7	6.87	20.1	1.19	0.19	0.55
6400	100	1.1	150	0.7	20.9	0.33	100	1.1	150	0.7	9.7	1.91	20.0	0.45	100	1.1	150	0.7	9.0	4.76	20.0	0.62	0.18	0.41
$\rho_0 = 0.3$																								
50	100	12.7	150	7.8	20.7	4.07	100	12.8	150	7.7	13.5	12.54	17.0	5.88	99	13.4	150	8.2	20.2	16.85	18.3	8.37	0.10	0.89
100	100	9.0	150	5.5	21.5	2.84	99	9.0	150	5.5	12.7	10.00	19.0	3.42	99	9.1	150	5.6	16.4	13.21	19.7	4.93	0.11	0.80
200	100	6.4	150	4.0	22.0	2.08	100	6.4	150	4.0	13.4	8.29	19.8	2.51	99	6.5	150	4.0	15.0	10.94	20.0	3.47	0.21	0.72
400	100	4.5	150	2.6	22.1	1.39	100	4.5	150	2.6	13.1	6.56	20.2	1.76	99	4.6	150	2.6	12.8	9.54	20.2	2.35	0.29	0.65
1600	100	2.3	150	1.3	22.2	0.69	100	2.3	150	1.3	14.1	3.29	20.4	0.93	100	2.3	150	1.3	10.8	6.26	20.1	1.16	0.42	0.49
6400	100	1.2	150	0.7	22.3	0.36	100	1.2	150	0.7	14.5	1.46	20.5	0.47	100	1.2	150	0.7	9.8	4.30	20.0	0.63	0.44	0.40

Table 3.2: MC means (est) and standard errors (se) of parameter estimates using regular model, restricted extended model, and unrestricted extended model, for simulated data generated from extended Underhill-Zucchini model. With various values for ρ, true values for other parameters are $\mu_{W,0} = 100$, $\mu_{X,0} = 150$, $\sigma_{W,0} = 10$, and $\sigma_{X,0} = 20$.

In general, Table 3.2 shows the estimates of mean duration ($\hat{\mu}_W$), mean and standard deviation of initiation time ($\hat{\mu}_X$ and $\hat{\sigma}_X$) obtained from the unrestricted extended model are consistent. Though the model that assumes independence ($\rho = 0$) perform well when the underlying assumption is correct, it loses consistency when the correlation exists. An interesting finding is that the estimates of mean initiation time ($\hat{\mu}_X$), mean duration ($\hat{\mu}_W$), and standard deviation of initial time ($\hat{\sigma}_X$) are accurate even the sample size is not sufficient. On the contrary, the standard deviation of duration ($\hat{\sigma}_W$) and the correlation coefficient ($\hat{\rho}$) need larger datasets to get stable estimates and often have larger biases. Another attempt was then applied to interpret this phenomenon by starting the optimization algorithm at true parameter values but gave similar results. Though it may because the optimization algorithm gets stuck in some local maxima, it also implies the extended model is difficult to identify practically, given the augmented standard errors for those two parameters.

This situation can be better explained by plotting the likelihood profile for each of the five parameters using the unrestricted extended model. The log-likelihood profile in Figure 3.1 demonstrates how likelihood changes as specific parameter changes while holding all other parameters at their true values. The log-likelihood has been set on the same scale. As pointed out in Definition 3, the likelihood-based confidence region tends to infinity in both directions for σ_W and ρ. The model, therefore, is not practically identifiable. This is a consequence of a flat likelihood function, which means multiple combinations of σ_W and ρ will give similar likelihood, rendering a plane or ridge on the likelihood surface. One reason explaining the occurrence of the flat likelihood is only one observation is available for each individual, leading to the model's failure of distinguishing the influence of σ_W and ρ on PFMG (See Figure 3.2b and 3.3). To obtain accurate estimates for the standard deviation of duration or correlation coefficient, multiple observations (at least two) are required for some of the individuals. More details are included in chapter 4.

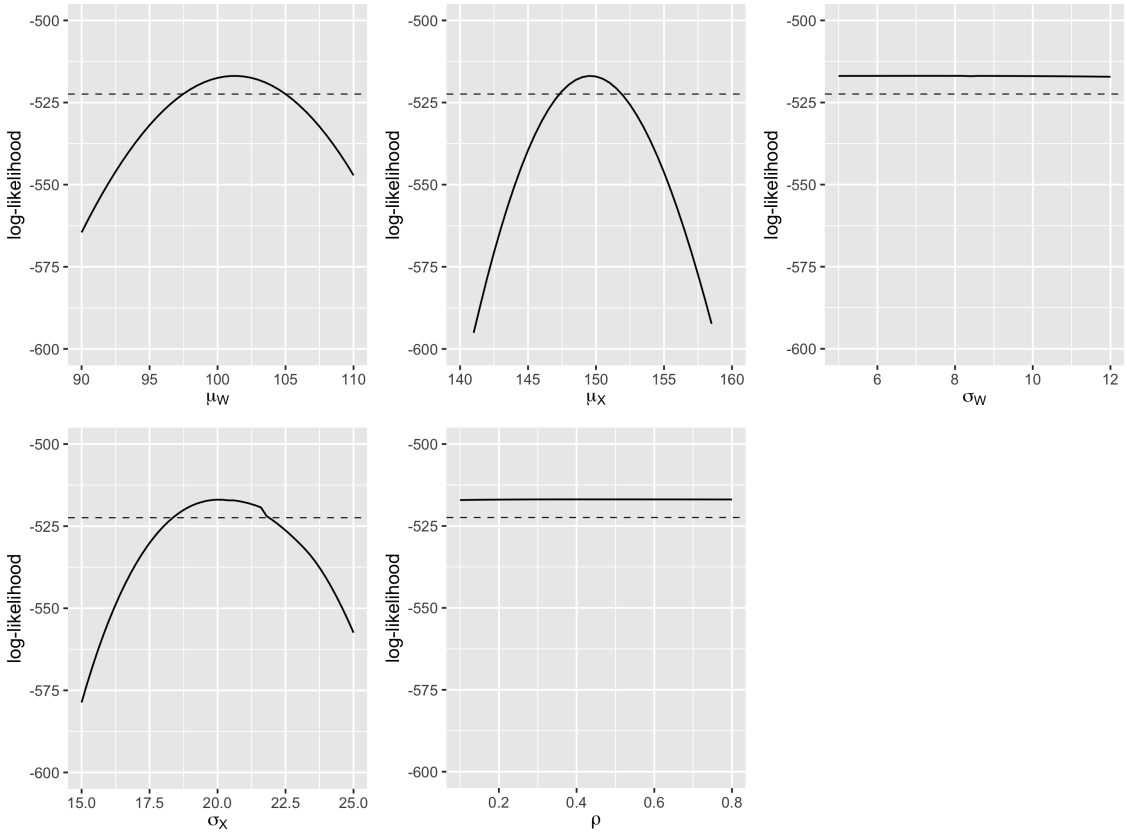
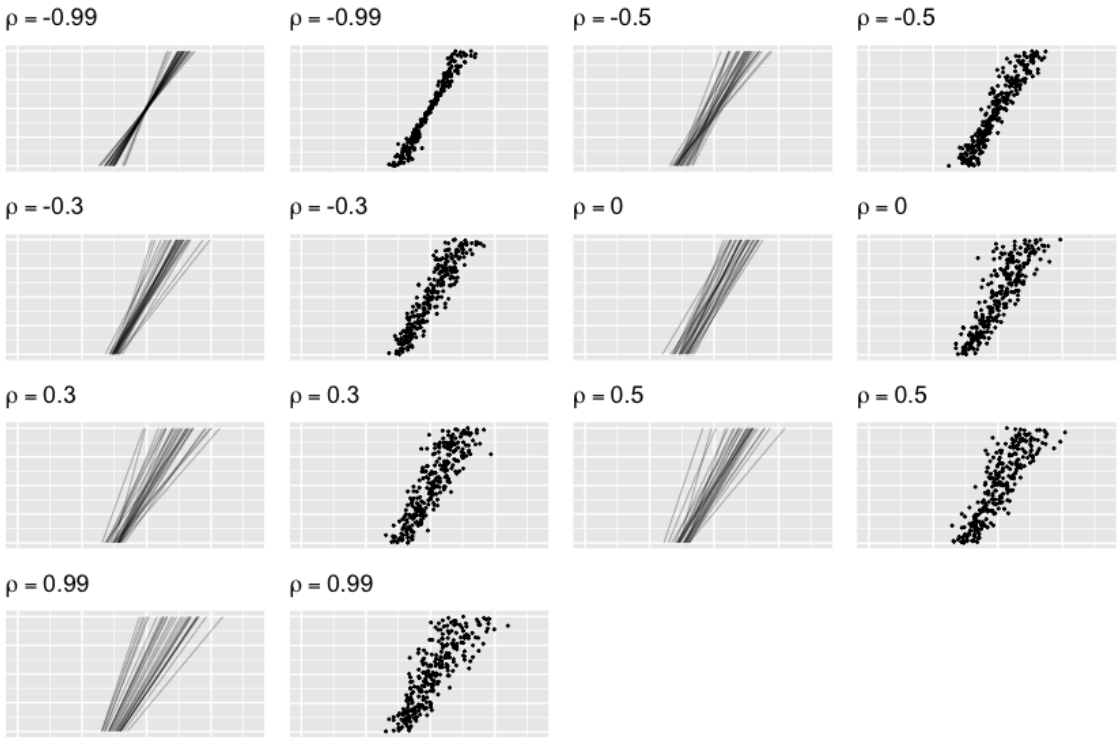
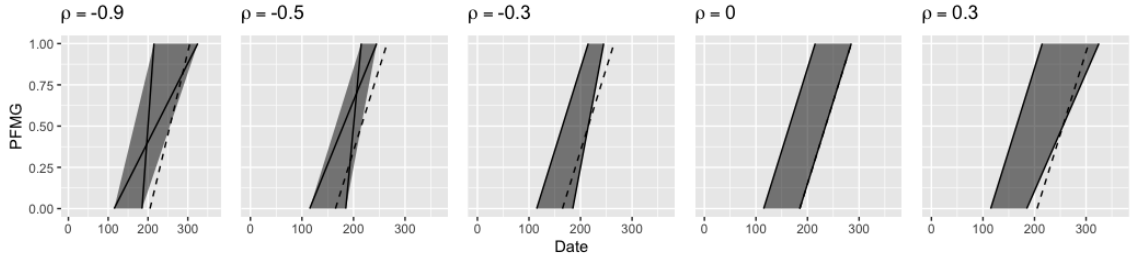


Figure 3.1: Likelihood profile for each of the five parameters using simulated data. The dashed lines show distance of $0.5\chi^2_{\alpha,q}$ from maximum log-likelihood. Simulation parameters are $\mu_{W,0} = 100$, $\mu_{X,0} = 150$, $\sigma_{W,0} = 10$, $\sigma_{X,0} = 20$, $\rho_0 = 0.3$. Sample size used for creating this plot is $n = 6400$. There are no likelihood confidence limits for σ_W and ρ.

Figure 3.3 amplifies and illustrates the difficulty of extracting the influence of a single parameter on PFMG. Even though the initial time and duration are highly negatively correlated ($\rho = -0.9$), the data cloud has a similar shape as if $\rho = 0$ but with a smaller σ_W. Consequently, the likelihood surface is flat in the parametric space, and σ_W tends to be slightly underestimated. Furthermore, if we force $\rho = 0$ by selecting the restricted model, σ_W will tend to be largely overestimated when $\rho_0 > 0$ and underestimated when $\rho_0 < 0$. All of these findings conform to the results shown in Table 3.2. When using a restricted model, i.e., forcing $\rho = 0$, the dashed lines in Figure 3.2b were treated as the boundary of the data



(a) Simulated data points and the model from which they were generated.



(b) Conceptual shape of data clouds (grey areas) for various ρ ($\rho_0 = -0.9, -0.5, -0.3, 0, 0.3$ from left to right). Solid lines show the true underlying model; dashed lines show how the data was treated when forcing $\rho = 0$.

Figure 3.2: In plot(a), we notice some different models may generate data points with a similar shape. From the first and last graph in the plot(b), we understand it is unlikely to figure out the relation between σ_W and ρ solely from the observed data.

points. The plots help us analyze how the estimated parameters perform compared to their actual values.

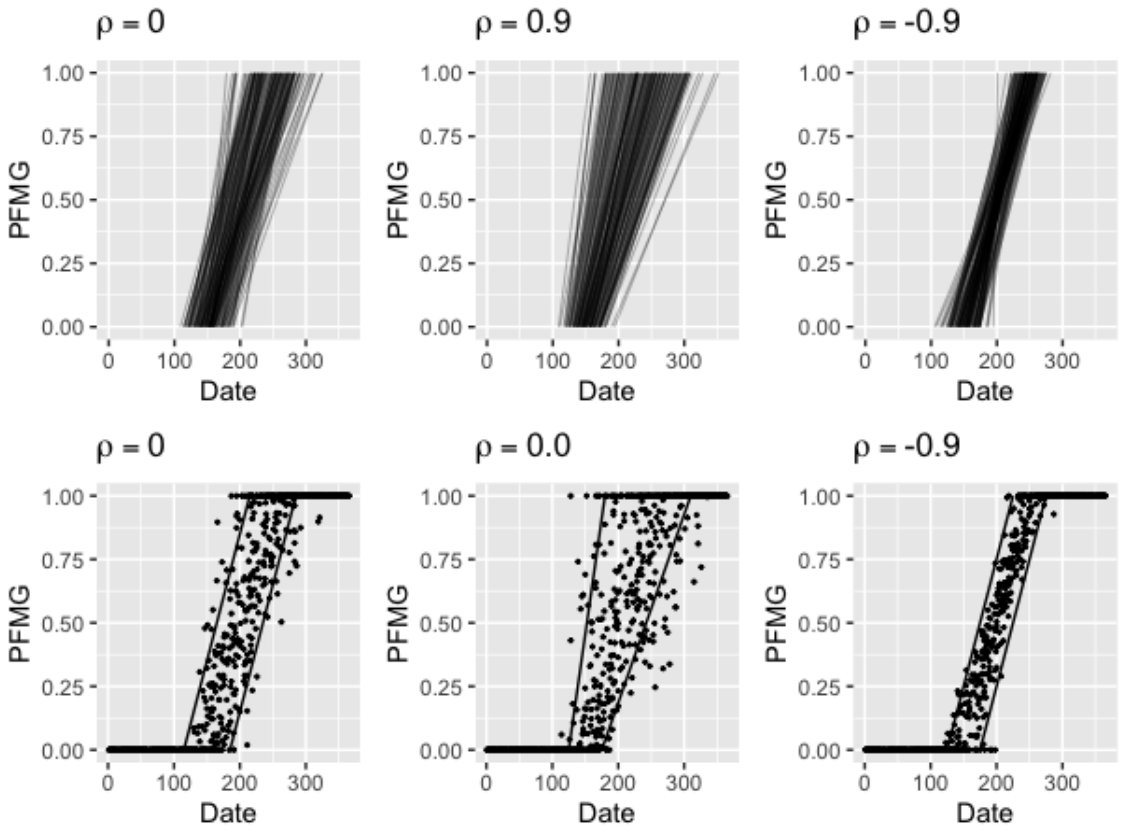


Figure 3.3: Molt lines and data points generated with different values of ρ.

In Tables 3.2, we see that if the correlation between X and W does exist, the unrestricted extended model gives less biased estimates for σ_X than the regular model and restricted extended model, and less biased estimates for σ_W than the restricted model. It also shows that if the correlation between σ_W and ρ is indeed equal to zero, parameters obtained using the restricted extended model are the most accurate. The regular model gives reliable estimates of μ_X, μ_W for medium-sized sample regardless of the correctness of model specification, but produces an estimate for σ_X with small but significant bias. These consequences suggest that: 1) regular model has a good performance estimating μ_X and μ_W even if the actual

duration is a random variable; 2) extended models also provide accurate estimates for μ_X and μ_W with a slightly large sample; 3) unrestricted extended model gives an unbiased estimate for σ_X when the duration is random; 4) it is possible to estimate the variability of duration and the correlation coefficient between duration and starting time from such data using extended models.

Due to its flexibility, the unrestricted extended model instead of the restricted model is encouraged to use if it is not clear whether the molt pace and the starting time should be independent. However, if there is not much doubt that they are independent, using the restricted model allows more efficient inference. For the sake of getting reasonable estimates, hypothesis testing can be performed to provide a hint of model selection.

Chapter 4

MULTIPLE RECORDS FOR EACH INDIVIDUAL

This section discusses how to adjust our study design to get more reliable estimates for the standard deviation of duration (σ_W) and correlation coefficient (ρ). We proved in chapter 2 that the extended Underhill-Zucchini model is identifiable. However, the likelihood profile for σ_W and ρ were nearly flat, and their likelihood-based confidence regions tend to infinity in both directions, which gave rise to an issue of practically non-identifiability. Since the primary cause of non-identifiability is the snapshot data (one observation for each individual), which confuses the model in how to distinguish the effects of σ_W and ρ on the PFMG, the problem should be resolved by including multiple records for each bird in sample data. If more than one observation is available for a bird, the model can predict the duration more stably by combining their information. An intuition explaining this is that the slope is arbitrary for a line through one point but fixed for a line through two points. This helps the model separate the influence of σ_W out and therefore simultaneously improves the estimation for ρ.

In practice, one way of getting multiple records is adopting a mark-recapture study design. When birds are captured for the first time, we put marks on them and then release them. While collecting sample data on the other day, it is likely to recapture some pre-observed birds and record the new data. The likelihood is calculated for each individual, and different weights are assigned to birds with different numbers of observations. For birds with two records, six scenarios may occur as per the locations of those observations on the molting procedure, which are shown in Figure 4.1. For example, if a bird was captured twice and the bird was in molt in both cases, it fell into scenario six. In some circumstances, the likelihood functions for birds with repeated measures are apparently easy to compute and optimize. In later studies, we will see only several birds with two observations in molt will significantly reduce the estimates' standard errors (for instance, $\hat{se}(\hat{\sigma_W})$).

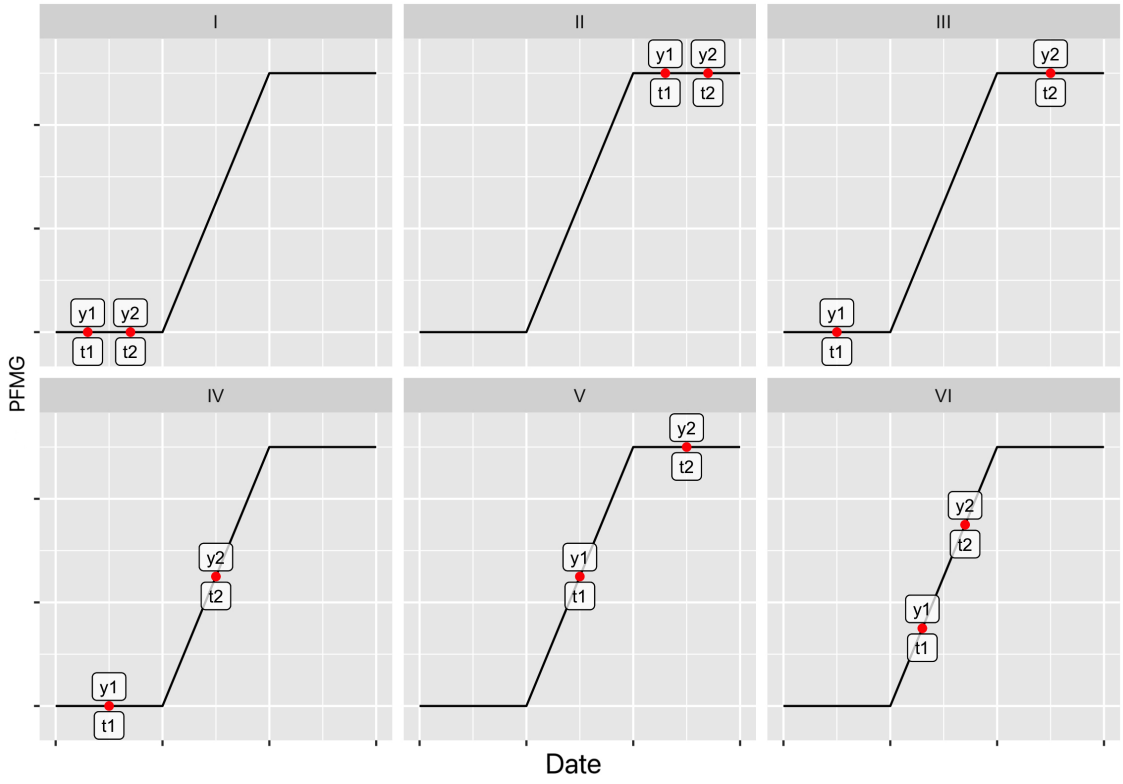


Figure 4.1: Six scenarios for individuals with two records. Locations of red dots depend on the molt progression when birds were captured and recorded.

Since $t_2 > t_1$, the likelihood functions in scenario I and II are computed as

$$f_1 = P(X > t_1, X > t_2) = P(X > t_2) = 1 - F_X(t_2),$$

$$f_2 = P(X + W \leq t_1, X + W \leq t_2) = P(X + W \leq t_1) = F_{X+W}(t_1).$$

In scenario IV and V, the likelihood functions involve the conditional density functions, conditioning on subspaces of the random variables, which have intricate densities (in which case the joint density of X and W is not bivariate normal anymore). Computing the likelihood in these cases involves the Bayes formula for mass/density functions in the mixed case. Nevertheless, we will see that estimates' properties can be improved by only consid-

ering birds in scenario VI.

$$\begin{aligned} f_4 &= P(X > t_1) \cdot f(y_2|X > t_1) \\ &= (1 - F_X(t_1)) \cdot \frac{f_Y(y_2) \cdot P(X > t_1|Y = y_2)}{1 - F_X(t_1)} \\ &= (1 - F_X(t_1)) \cdot \frac{f_Y(y_2) \cdot P(W < \frac{t_2 - t_1}{y_2})}{1 - F_X(t_1)}; \end{aligned}$$

$$\begin{aligned} f_5 &= P(X + W \leq t_2) \cdot f(y_1|X + W \leq t_2) \\ &= F_{X+W}(t_2) \cdot \frac{f_Y(y_1) \cdot P(X + W \leq t_2|Y = y_1)}{F_{X+W}(t_2)} \\ &= F_{X+W}(t_2) \cdot \frac{f_Y(y_1) \cdot P(X > \frac{t_1 - y_1 t_2}{1 - y_1})}{F_{X+W}(t_2)}. \end{aligned}$$

Finally in scenario VI, the relation between Y_1 and Y_2 is

$$Y_2 = Y_1 + \frac{t_2 - t_1}{W},$$

then the conditional probability density function $f_{Y_2|Y_1}(y_2|y_1)$ can be derived as

$$\begin{aligned} f_{Y_2|Y_1}(y_2|y_1) &= \frac{\partial}{\partial y_2} F_{Y_2|Y_1}(y_2|y_1) \\ &= \frac{\partial}{\partial y_2} P(W \geq \frac{t_2 - t_1}{y_2 - y_1}) \\ &= \frac{\partial}{\partial y_2} (1 - \Phi(\frac{t_2 - t_1}{\sigma_W (y_2 - y_1)} - \frac{\mu_W}{\sigma_W})). \end{aligned}$$

The likelihood function of a individual bird with two observations in molt then can be computed as

$$\begin{aligned} f_6 &= f_{Y_1, Y_2}(y_1, y_2) \\ &= f_Y(y_1) \cdot f(y_2|y_1) \\ &= f_Y(y_1) \cdot \frac{t_2 - t_1}{\sigma_W (y_2 - y_1)^2} \cdot \phi(\frac{t_2 - t_1}{\sigma_W (y_2 - y_1)} - \frac{\mu_W}{\sigma_W}). \end{aligned}$$

Fortunately, the estimated results of the 5-parameter model can be promoted when only incorporating data in scenario VI into our sample. The above equation gives the likelihood for individuals with two observations in molt. For all individuals with single record, their

likelihood can be calculated through formula instructed in chapter 2. Comparison of the extended model with multiple records and single record is shown in Table 4.1. In the table, n_1 and n_2 denotes the number of generated individuals with one record and two records, respectively. The first three estimates were obtained by using model adjusted for repeated records. We then generated dataset with same sample size but one record for all individuals, using the same underlying true parameters. The extended model (unadjusted for repeated records) was then applied to obtain the estimates of the same three parameters. The results demonstrated that with same sample size, only 10% of birds with two records in the dataset can reduce the bias of the estimated σ_W and ρ, as well as their standard errors.

n_1	n_2	Repeated records						Single record					
		σ_W		σ_X		ρ_{XW}		σ_W		σ_X		ρ_{XW}	
		est	se	est	se	est	se	est	se	est	se	est	se
167	15	9.5	2.23	20.0	2.80	0.33	0.56	13.2	11.07	19.8	3.51	0.33	0.72
334	29	9.7	1.33	20.0	2.07	0.35	0.42	11.1	9.36	19.9	2.44	0.45	0.63
670	58	9.8	0.92	19.9	1.45	0.35	0.32	10.7	7.92	20.1	1.61	0.47	0.56
1340	118	9.9	0.64	19.9	1.08	0.35	0.23	9.6	6.32	20.0	1.19	0.54	0.49
2679	235	9.9	0.45	20.0	0.74	0.34	0.17	8.8	5.31	20.0	0.81	0.57	0.44

Table 4.1: MC means (est) and standard errors (se) of some parameter estimates using unrestricted extended model while some individuals have two records as scenario 6. True parameters used to generate data are $\mu_{W,0} = 100$, $\mu_{X,0} = 150$, $\sigma_{W,0} = 10$, $\sigma_{X,0} = 20$, and $\rho_0 = 0.3$. n_1, n_2 denote the numbers of individuals with one or two records, respectively.

The simulation study showed that data containing only observations in scenario VI would significantly improve the algorithm's performance. In practice, depending on the collected data, one can incorporate observations in scenario VI, V, or any combination of them to get more reliable estimates.

Chapter 5

DISCUSSION

The regular Underhill-Zucchini model shows good performance for the avian molt data. However, the regular Underhill-Zucchini only treats the initial molt time as a random quantity and treats the duration as a parameter. This thesis extended the regular Underhill-Zucchini model by adding two extra parameters and regarding the duration as a random quantity to increase the flexibility and reflect the actual molting process better.

The extended model is proved to be identifiable in theory, but the likelihood surface is almost flat with respect to some parameters. Using simulated snapshot data, we show that for the mean duration and mean initiation time parameters, the extended model can provide reasonable estimates with a small to medium sample size, though sometimes not as good as the regular model even when the data is simulated from an extended model. For the standard deviation of the initiation time, the extended model has a remarkable improvement (less bias) compared to the regular model when the duration and initiation time are correlated. For the standard deviation of the duration and the correlation coefficient, the extended model gives estimates that are more biased and less precise than the other parameters. The bias and precision of the parameter estimates do improve as the sample size grows, but we need extremely large datasets to obtain satisfactory estimates.

The limitation of the extended Underhill-Zucchini models can be attributed to the fact that the data is limited to one observation for each bird in our study. If we have more than one observation for a bird, the model can much more easily distinguish the effects of σ_W and ρ on PFMG, and therefore give less-biased estimates with smaller standard errors. For example, our simulation study shows that 10% of individuals with two observations in scenario VI would noticeably increase the estimates' accuracy.

BIBLIOGRAPHY

Catchpole, E.A. and Morgan, B.J. (1997), "Detecting parameter redundancy," *Biometrika*, 84, 187–196.

Catchpole, E.A., Kgosi, P. and Morgan, B.J. (2001), "On the near-singularity of models for animal recovery data," *Biometrics*, 57, 720–726.

Chis, O.T., Villaverde, A.F., Banga, J.R. and Balsa-Canto, E. (2016), "On the relationship between sloppiness and identifiability," *Mathematical biosciences*, 282, 147–161.

Cole, D.J. (2019), "Parameter redundancy and identifiability in hidden Markov models," *Metron*, 77, 105–118.

Cole, D.J., Morgan, B.J. and Titterington, D. (2010), "Determining the parametric structure of models," *Mathematical biosciences*, 228, 16–30.

Corless, R.M. and Jeffrey, D.J. (1997), "The Turing factorization of a rectangular matrix," *ACM SIGSAM Bulletin*, 31, 20–30.

Dufresne, E., Harrington, H.A. and Raman, D.V. (2016), "The geometry of sloppiness," *arXiv preprint arXiv:1608.05679*.

Evans, P. (1966), "Autumn movements, moult and measurements of the Lesser Redpoll Carduelis flammea cabaret," *Ibis*, 108, 183–216.

Geary, R.C. (1930), "The frequency distribution of the quotient of two normal variates," *Journal of the Royal Statistical Society*, 93, 442–446.

Ginn, H. and Melville, D. (1983), "Moult in birds. BTO Guide 19," *British Trust for Ornithology, Tring*, 67.

Hinkley, D.V. (1969), "On the ratio of two correlated normal random variables," *Biometrika*, 56, 635–639.

Marsaglia, G. et al (2006), "Ratios of normal variables," *Journal of Statistical Software*, 16, 1–10.

Pimm, S. et al (1976), "Estimation of the duration of bird molt." *Condor*, 78, 550.

Ran, Z.Y. and Hu, B.G. (2014), "Determining structural identifiability of parameter learning machines," *Neurocomputing*, 127, 88–97.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U. et al (2009), "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood," *Bioinformatics*, 25, 1923–1929.

Summers, R. (2003), "On the rate of change of moult scores in waders," *BULLETIN-WADER STUDY GROUP*, 69, 79–79.

Summers, R., Swann, R. and Nicoll, M. (1983), "The effects of methods on estimates of primary moult duration in the Redshank Tringa totanus," *Bird Study*, 30, 149–156.

Underhill, L. (1985), "Estimating the parameters for primary moult-A new statistical model," *Wader Study Group Bulletin*, 44, 27–29.

Underhill, L., Zucchini, W. and Summers, R. (1990), "A model for avian primary moult-data types based on migration strategies and an example using the Redshank Tringa totanus," *Ibis*, 132, 118–123.

Underhill, L.G. and Zucchini, W. (1988), "A model for avian primary moult," *Ibis*, 130, 358–372.

Walter, E. and Lecourtier, Y. (1982), "Global approaches to identifiability testing for linear and nonlinear state space models," *Mathematics and Computers in Simulation*, 24, 472–482.

Zhou, M., McCrea, R.S., Matechou, E., Cole, D.J. and Griffiths, R.A. (2019), "Removal models accounting for temporary emigration," *Biometrics*, 75, 24–35.