Stickle!

Matthew Stuart
Department of Mathematics and Statistics
Loyola University Chicago
Chicago, IL 60660
mstuart1@luc.edu

Akhil Ghosh
Department of Mathematics and Statistics
Loyola University Chicago
Chicago, IL 60660
aghosh@luc.edu

Yoel E. Stuart
Department of Biology
Loyola University Chicago
Chicago, IL 60660
ystuart@luc.edu

Gregory J. Matthews
Center for Data Science and Consulting; Department of Mathematics and Statistics
Loyola University Chicago
Chicago, IL 60660
gmatthews1@luc.edu

Abstract

Evewryone loves the stickle

Keywords: Stickle

1 Introduction

Sexual Dimorphism is interesting because. . . . . .

Sexual Dimoprhism lit review: Saitta et al. (2020)

Studying sexual dimorphism of phenotypes in modern species is a straightforward endeavor: measure a phenotype of interest and statistically test for differences between the sexes. However, examining sexual dimorphism in fossils is complicated substantially by the issue that sex may not be able to be directly observed from fossil specimens. For this study, we have two sources of data: 1) Modern stickleback specimens with observed sex and 2) fossil specimens with sex unobserved. We view the sex of the fossils as missing data and use multiple imputation (Little and Rubin (2002)) to impute the sex of the fossils using the modern stickleback fish with observed sex to model the relationship between sex and observed phenotypes. Once sex is imputed, an Ornstein–Uhlenbeck (OU) model is fit using a Bayesian framework to look for sexual dimorphism across a variety of stickleback phenotypes (e.g. length, vertebrae count?, etc).

The remainder of this manuscript contains a description of the data in section 2 and a description of our models in 3. Section 4 presents a summary of our results, and we end with our conclusion and future work in section 5.

2 Data

The data used here consists of a total of 367 extant specimens with known sex all collected in the last 30 years. Of these, there are 202 and 165 female and male specimens, respectively.

In addition there are 814 fossil specimens from approximately 10.3 million years ago with unknown sex over 18 time periods spaced about 1000 years apart. Table 1 shows the sample size at each of the 18 time periods. There are at least 22 specimens at each time period with a high of 67 specimens in period 7.

What covariates do we have in the data: length, what else,

3 Models

3.1 Imputation Model

Let $\boldsymbol{W}$ be an $(n_{extant} + n_{fossil}) \times 1$ vector of the covariate gender of the stickleback fish, $\boldsymbol{X}$ be an $(n_{extant} + n_{fossil}) \times K$ matrix of the K continuous phenotypes of interest, and $\boldsymbol{Y}$ be

time	count
1	43
2	41
3	51
4	41
5	46
6	48
7	67
8	55
9	42
10	33
11	37
12	22
13	41
14	43
15	46
16	47
17	56
18	55

Table 1: The number of stickleback fossils at each time point. Sample size at each time point ranges from a low of 22 to a high of 67.

an $(n_{extant} + n_{fossil}) \times L$ matrix of the L discrete phenotypes of interest. Because the gender of the fossilized stickleback fish is unobservable, we further define $\boldsymbol{W} = (\boldsymbol{W}_{extant}^T, \boldsymbol{W}_{fossil}^T)^T$ where $\boldsymbol{W}_{extant}$ and $\boldsymbol{W}_{fossil}$ are the $n_{extant} \times 1$ and $n_{fossil} \times 1$ vectors of the observed extant gender and missing fossil gender, respectively.

We impute the missing gender for the fossil data by sampling from the posterior predictive distribution $P(\boldsymbol{W}_{fossil}|\boldsymbol{W}_{extant}, \boldsymbol{X}, \boldsymbol{Y})$ using the multiple imputation with chained equations (MICE) algorithm (Buuren and Groothuis-Oudshoorn (2011)) with predictive mean matching. The imputation algorithm is run to obtain a total of $M = 100$ completed datasets.

3.2 Completed Data Model

For a given imputed dataset, let W_{ti} be the imputed gender, $\boldsymbol{X}_{ti}$ be the $K \times 1$ vector of continuous phenotypes, and $\boldsymbol{Y}_{ti}$ be the $L \times 1$ vector of discrete phenotypes for $t = 1, \ldots, T$ and $i = 1 \ldots, n_t$).

Define the last element of $\boldsymbol{X}_{ti}$ ($X_{K,ti}$) as the length of the stickleback. For $k = 1, \cdots, K-1$,

we will assume

$$X_{k,ti} \overset{iid}{\sim} \begin{cases} \mathcal{N}(\mu_{k,ft} \quad +\beta_k(X_{K,ti} - \mu_{K,ft}), \quad \sigma_k^2), & W_{ti} = \text{Female} \\ \mathcal{N}(\mu_{k,mt} \quad +\beta_k(X_{K,ti} - \mu_{K,mt}), \quad \sigma_k^2), & W_{ti} = \text{Male} \end{cases} \tag{1}$$

$$X_{K,ti} \overset{iid}{\sim} \begin{cases} \mathcal{N}(\mu_{K,ft}, \quad \sigma_K^2), & W_{ti} = \text{Female} \\ \mathcal{N}(\mu_{K,mt}, \quad \sigma_K^2), & W_{ti} = \text{Male} \end{cases} \tag{2}$$

where $\mu_{k,ft}$ and $\mu_{k,mt}$ represent the time-t specific mean of phenotype k for female and male stickleback fish, respectively, and β_k is an additional parameter to account for the biological phenomenon that other continuous phenotypes for a particular animal are positively correlated with their length **CITATION NEEDED**. We further set

$$\mu_{k,gt} = \theta_{k,g} + u_{k,gt}, \tag{3}$$

for $g \in \{f, m\}$ where $\theta_{k,g}$ is the overall mean of phenotype X_k for each gender, and $u_{k,gt}$ measures the difference between $\mu_{k,gt}$ and $\theta_{k,g}$. We then fit $\boldsymbol{u}_{k,g} = \{u_{k,g1}, \cdots, u_{k,gT}\}$ to an Ornstein-Uhlenback (OU) process ((**OUprocess?**)) where we assume $u_{k,gT}$ have a marginal mean of 0. Because each time period is ~1000 years, we will discretize the OU process without loss of generality. For $t = 2, \ldots, T$, we assume

$$u_{k,gt} \overset{iid}{\sim} \mathcal{N}(\kappa_k u_{k,g(t-1)}, \tau_k^2). \tag{4}$$

$\kappa_{k,g}$ represents the correlation between $\mu_{k,gt}$ and $\mu_{k,g(t+1)}$, and it also follows that $\text{cor}(u_{g,t}, u_{g,t+h}) = \kappa^h$ for $h \in \mathbb{N}$. In addition, we assume

$$u_{k,g1} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\tau_k^2}{1 - \kappa_k^2}\right). \tag{5}$$

This model choice is to preserve stationarity; i.e. $p(u_{k,gs}) = p(u_{k,gt})$ for $s \neq t$.

For the discrete phenotypes, we will assume

$$Y_{l,ti} \sim \begin{cases} Poisson(\lambda_{l,ft}), & W_{ti} = \text{Female} \\ Poisson(\lambda_{l,mt}), & W_{ti} = \text{Male} \end{cases}, \tag{6}$$

for $l = 1, \cdots, L$. Similar to the continuous phenotypes, we set

$$\log(\lambda_{l,gt}) = \gamma_{l,g} + v_{l,gt}. \tag{7}$$

We use the log function to allow $v_{l,gt}$ to take all values on the real number line so we can

properly fit an OU process to these effects. We further assume

$$v_{l,gt} \overset{iid}{\sim} \mathcal{N}(\phi_l v_{l,g(t-1)}, \omega_l^2), \tag{8}$$

and

$$v_{l,g1} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\omega_l^2}{1-\phi_l^2}\right)\,. \tag{9}$$

Because we are fitting a dataset with a stochastic structure on the means of the phenotypes, we analyze the data via a Bayesian analysis. Bayesian data analysis is also more naturally used when we have to impute data (**CITATION**).

Priors: For $k = 1, \cdots, K$ and $l = 1, \cdots, L$,

$$\begin{aligned}
\sigma_k &\overset{iid}{\sim} \mathcal{N}(0, 100) I_{\{\sigma > 0\}} \\
\tau_k &\overset{iid}{\sim} \mathcal{N}(0, 100) I_{\{\tau > 0\}} \\
\kappa_k &\overset{iid}{\sim} \mathcal{N}(0, 1) I_{\{-1 < \kappa_g < 1\}} \\
\beta_k &\overset{iid}{\sim} \mathcal{N}(0, 5) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (10) \\
\theta_{k,g} &\overset{iid}{\sim} \mathcal{N}(0, 10000)\omega_l \qquad\qquad \overset{iid}{\sim} \mathcal{N}(0, 100) I_{\{\tau > 0\}} \\
\phi_l &\overset{iid}{\sim} \mathcal{N}(0, 1) I_{\{-1 < \kappa_g < 1\}} \\
\gamma_{l,g} &\overset{iid}{\sim} \mathcal{N}(0, 10000) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (11)
\end{aligned}$$

All models were built using R Core Team (2022)

Cornuault (2022) Bayesian OU model.

Bayesian Analysis after multiple imputation Zhou and Reiter (2010): They recommend using a large number of imputations. 5 or 10 is too small. We are using M = 100.

4 Results

5 Future work and conclusions

Acknowledgements

Stickle!

Supplementary Material

All code for reproducing the analyses in this paper is publicly available at https://github.com/Akhil-Ghosh/SticklebackProject

References

Buuren, Stef van, and Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in r." *Journal of Statistical Software* 45 (3): 1–67. https://doi.org/10.18637/jss.v045.i03.

Cornuault, Josselin. 2022. "Bayesian Analyses of Comparative Data with the Ornstein–Uhlenbeck Model: Potential Pitfalls." *Systematic Biology* 71 (6): 1524–40. https://doi.org/10.1093/sysbio/syac036.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Mathematical Statistics. Probability and Mathematical Statistics. Wiley. http://books.google.com/books?id=aYPwAAAAMAAJ.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Saitta, Evan T, Maximilian T Stockdale, Nicholas R Longrich, Vincent Bonhomme, Michael J Benton, Innes C Cuthill, and Peter J Makovicky. 2020. "An effect size statistical framework for investigating sexual dimorphism in non-avian dinosaurs and other extinct taxa." *Biological Journal of the Linnean Society* 131 (2): 231–73. https://doi.org/10.1093/biolinnean/blaa105.

Zhou, X., and J. Reiter. 2010. "A Note on Bayesian Inference After Multiple Imputation." *The American Statistician* 64 (2): 159–63.