

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 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 Dimorphism 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 \mathbf{W} be an $(n_{\text{extant}} + n_{\text{fossil}}) \times 1$ vector of the covariate gender of the stickleback fish and \mathbf{X} be an $(n_{\text{extant}} + n_{\text{fossil}}) \times k$ matrix of the k phenotypes of interest of the fish. Because \mathbf{W}

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.

contains missing data, we further define $\mathbf{W} = (\mathbf{W}_{obs}^T, \mathbf{W}_{mis}^T)^T$ where \mathbf{W}_{obs} and \mathbf{W}_{mis} are the observed and missing data of the covariate gender, respectively. In this setting, \mathbf{W}_{obs} is a $n_{extant} \times 1$ vector and \mathbf{W}_{mis} is a $n_{fossil} \times 1$ vector, perfectly corresponding to the observable gender of the modern extants and the unobservable gender of the fossil data, respectively.

We impute the missing gender for the fossil data by sampling from the posterior predictive distribution $P(\mathbf{W}_{mis} | \mathbf{W}_{obs}, \mathbf{X})$ 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 OU model

After imputing sex, we fit an OU model for different phenotypes of interest. WLOG we choose one of the columns of \mathbf{X} to be the target phenotype and our analysis is repeated for all phenotypes of interest.

Let W_{ti} and X_{ti} be the gender and phenotype of interest for the i -th observation in year t for $t = 1, \dots, 18$ and $i = 1 \dots, n_t$. We assume that

$$X_{ti} \stackrel{iid}{\sim} \begin{cases} \mathcal{N}(\theta_f + u_{ft}, \sigma^2), & W_{ti} = \text{Female} \\ \mathcal{N}(\theta_m + u_{mt}, \sigma^2), & W_{ti} = \text{Male} \end{cases}$$

In this setting, for $g \in \{f, m\}$, θ_g is defined as the overall mean of the phenotype of interest for each gender, and u_{gt} are the gender specific random effects, measuring the difference between the gender and time-specific mean and the overall gender-specific mean, which we fit to an autoregressive process of lag 1, which is the discretization of an OU process. For $t = 2, \dots, T$ and $g \in \{f, m\}$, the AR1 model for is defined as

$$u_{g,t} \stackrel{iid}{\sim} \mathcal{N}(\kappa_g u_{g,t-1}, \tau^2).$$

κ_g represents the gender-specific correlation between the previous and current time periods' random effects. In addition, $\text{cor}(u_{g,t}, u_{g,t+h}) = (\kappa_g)^h$ for $h \in \mathbb{N}$.

We also set

$$u_{g,1} \stackrel{iid}{\sim} \mathcal{N}\left(0, \frac{\tau^2}{1 - \kappa_g^2}\right).$$

This model choice is to make the distribution of the random effects stationary, i.e. $p(\mu_{g,s}) = p(\mu_{g,t})$ for $s, t \in \{1, \dots, T\}$.

Priors:

$$\begin{aligned} \sigma &\sim \mathcal{N}(0, 100)I_{\{\sigma > 0\}} \\ \tau &\sim \mathcal{N}(0, 100)I_{\{\tau > 0\}} \\ \kappa &\stackrel{iid}{\sim} \mathcal{N}(0.5, 1)I_{\{-1 < \kappa_g < 1\}} \\ \theta_g &\stackrel{iid}{\sim} \mathcal{N}(54, 400)I_{\{\theta_g > 0\}} \end{aligned}$$

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.