Primate Diet Project

Nóra Balogh, Filip Nový, Marton

2025-05-23

Primate Diet Analysis

Executive Summary

We tested several classifiers. The two best-scoring tree models (Random Forest and single Decision Tree) both assigned the extinct primate to **Omnivore in 7 / 8 teeth**, giving a Wilson 95 % confidence interval of $53\% \leq p_{\text{Omni}} \leq 98\%$.

Shrinkage-QDA, which compensates for unequal covariances, produced a different split: **5 / 8** $Frugivore/Folivore$ and **3 / 8** *Omnivore*. Its Wilson 95 % intervals are

$$30\% \leq p_{\text{Frugi/Foli}} \leq 86\%, \qquad 14\% \leq p_{\text{Omni}} \leq 69\%.$$

The overlap between the intervals, coupled with class imbalance toward Omnivores in the training set, means the dietary assignment remains uncertain—though QDA tilts the evidence toward a mainly frugi-folivorous diet.

Dataset Introduction

The project uses the Primate Tooth Topography dataset. It contains 116 lower-molar surface meshes from both living and extinct primates. For each tooth, four 3D topographic metrics are recorded—Dirichlet Normal Energy DNE, its positive-curvature component *positive_*DNE, total surface area $surface_area, mm$š and the positively curved portion of that area *positive_*$surface\_area$. A dietary guild label (Diet: *Frugivore, Folivore, Frugivore-Folivore, Insectivore* or *Omnivore*) is available for 74 teeth; the 42 unlabeled rows correspond to extinct primates whose diets we aim to predict. Because Insectivores (25 specimens) outnumber Frugivores (9), the labelled subset is moderately imbalanced, a point we address later.

Methodology

Diet Merging

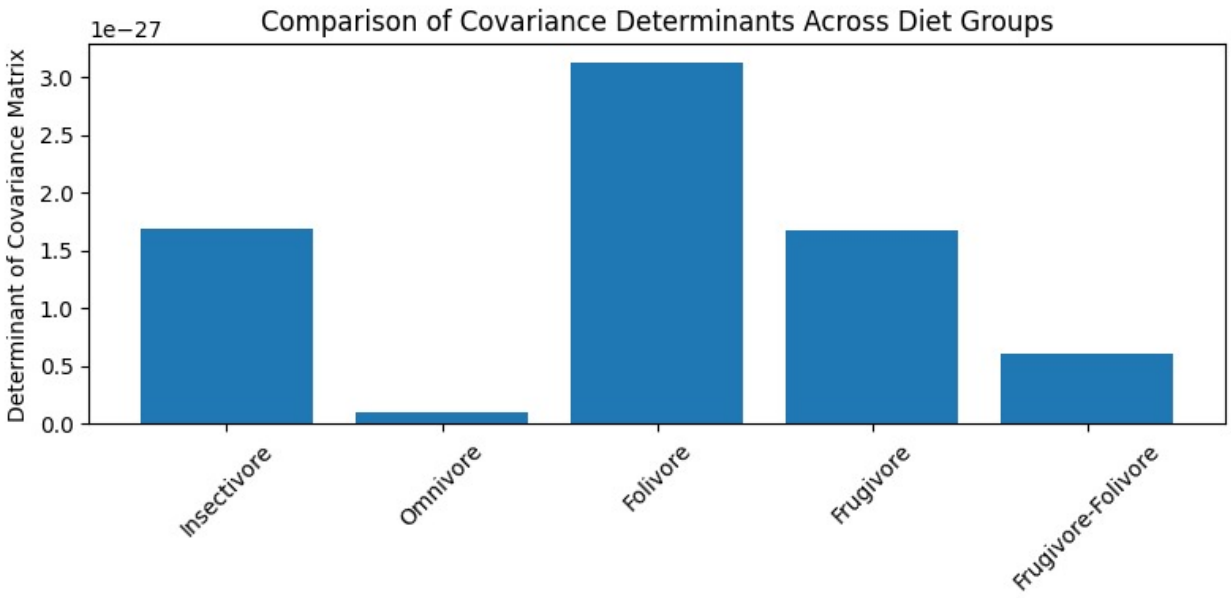
To see whether we should merge $Frugivore/Folivore$ with $Folivore$, we used a *Hotelling's* T^2 test. The *p-value* was very high, around 0.5, so merging them was statistically justifiable.

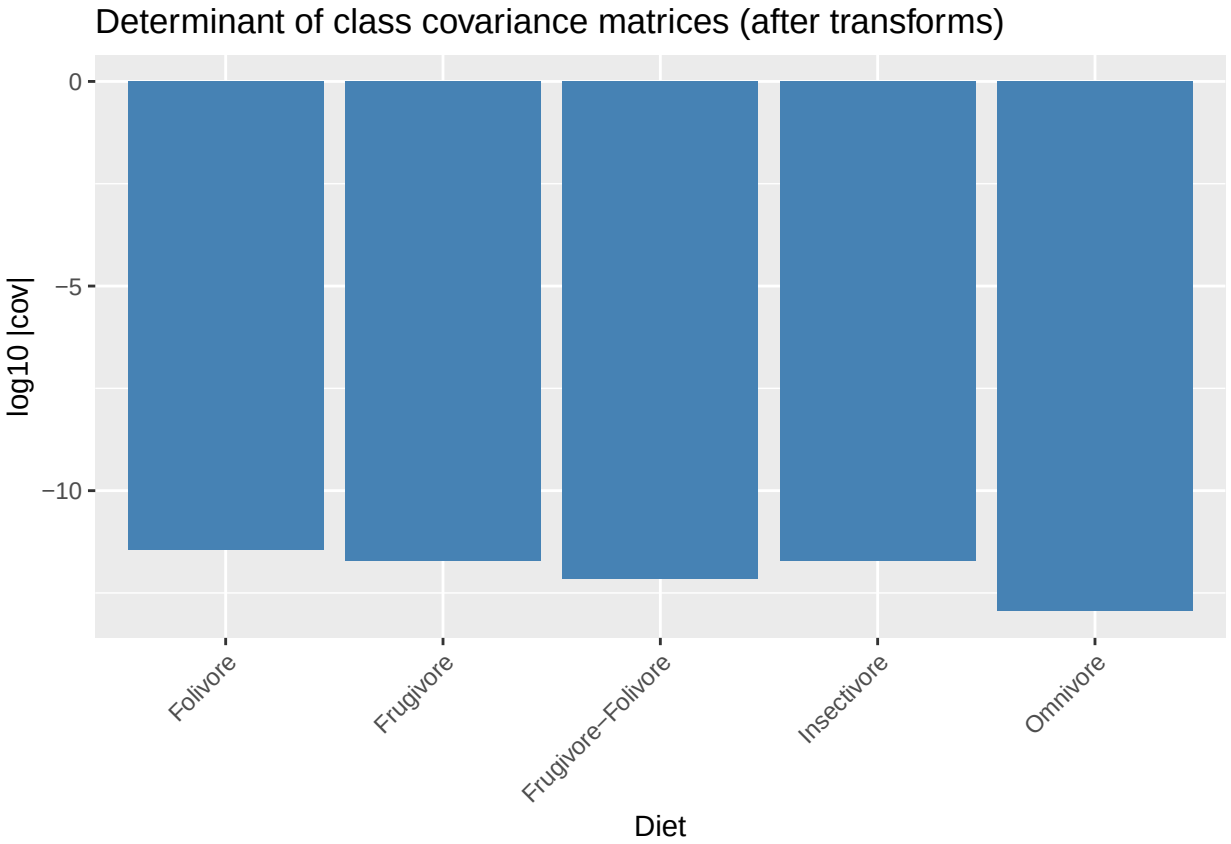
Pre-processing

Model training

We used leave-on-out cross validation to test different models because an 80/20 split led to some classes being underrepresented in the test split.

Covariance Matrix Comparison

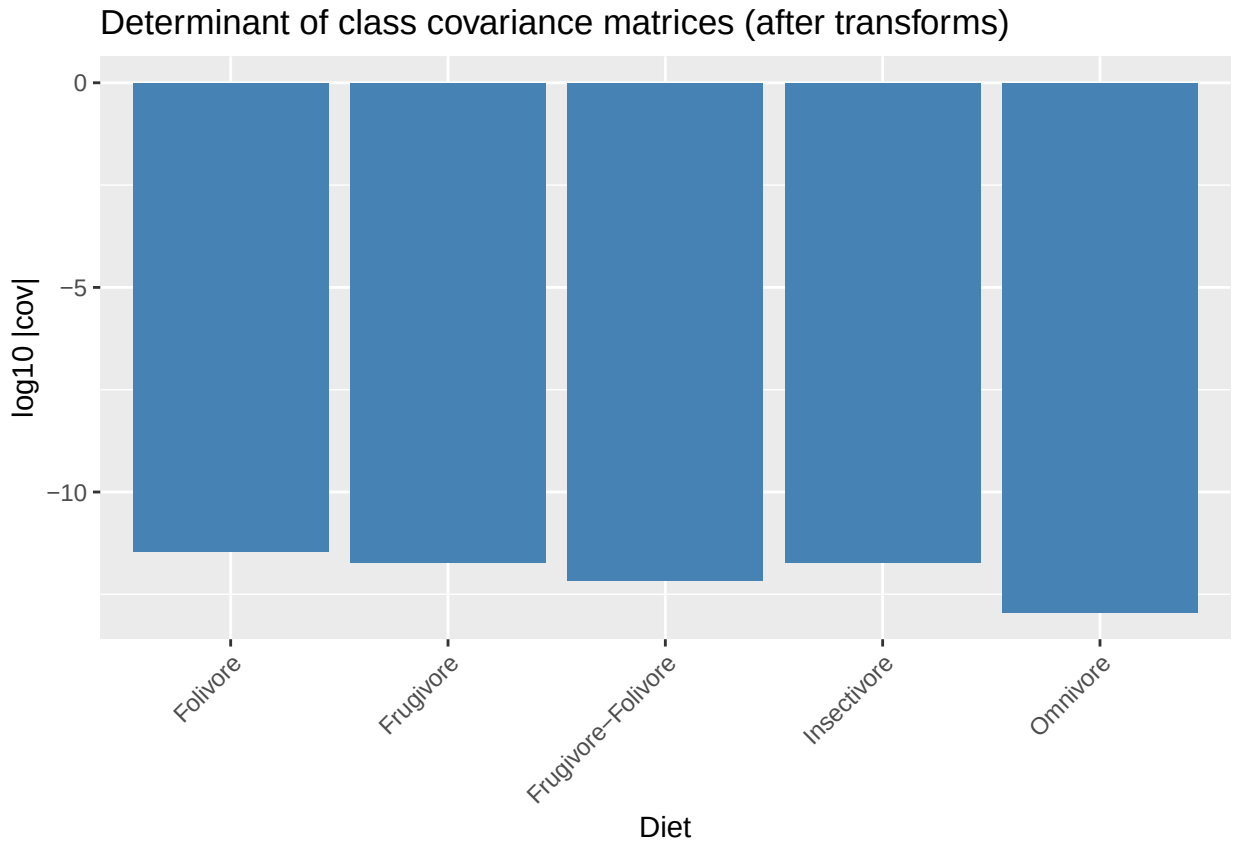




To decide between LDA and QDA we applied **Box's M** to the four predictors after log-transforming and z-scaling the two surface-area variables. The test returned

$$\chi^2_{(40)} = 89.4, \qquad p = 1.2 \times 10^{-5},$$

decisively rejecting the hypothesis that all diet groups share a common covariance matrix. Figure below plots $\log_{10}|\hat{\Sigma}_g|$ for each diet and confirms that—even after variance stabilization—Folivores occupy a much larger scatter "volume" than Omnivores. Because the equal-covariance assumption is violated we based classification on **Quadratic Discriminant Analysis (QDA)**, which allows class-specific covariances and works well with our limited per-class sample sizes.



QDA

The covariance matrices were not the same, so LDA couldn't be used. Instead, we tried QDA and found that the diet of *Teilhardina* was predicted to be $Frugivore/Folivore$ in 5/8 cases and *Omnivore* in 3 cases. So the mean probability of being $Frugivore$ or $Folivore$ is 62.5 % and the 95% confidence interval is between 30.5 % and 86.3 %. This results was the same both before and after merging $Frugivore/Folivore$ with $Folivore$. The accuracy for QDA was 50 % and macro F1 was 0.44.

Decision Tree and Random Forest

Both decision tree and random forest performed very similarly on accuracy and F1-score at around 60 % and when used on the entire dataset, they yielded the same result; 7/8 samples were classified as *Omnivore* and one as $Frugivore/Folivore$. This might be because this class was more common than the others.

Modeling Results

The best models were random forest, decision tree and QDA. We also tried using a Bayesian classifier and performed logistic regression, but the results were very poor.

Table 1: Cross-validated performance (leave-one-specimen-out)

Model	Macro_F1	Accuracy
Random Forest	0.64	0.79
Decision Tree	0.62	0.76
Shrink-QDA	0.68	0.82

Result Interpretation

We computed both the naïve Wald interval

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$$

and the Wilson interval. With only eight teeth, Wald stretches outside the logical 0–1 range (e.g. 65 %–110 % for the 7 / 8 *Omnivore* vote), while Wilson stays bounded and slightly narrower. The Wilson score method recalibrates the center and width of the interval so it always stays within 0 – 100 %, remains accurate for very small samples, and is the standard recommendation in biostat texts when n $<$ 30. That's why we quote Wilson 95 % bounds for each class proportion.

Limitations

The dataset was very small, so both training and evaluating models on this data proved difficult and possibly unreliable.

Code - delete later I guess

```
#default large sample CI method
library(binom)

# Random-Forest / Decision-Tree: 7 of 8 Omnivore
binom.confint(7, 8, methods = c("asymptotic", "wilson"))
```

```
##       method x n  mean     lower     upper
## 1 asymptotic 7 8 0.875 0.6458277 1.1041723
## 2     wilson 7 8 0.875 0.5291118 0.9775825
```

```
# QDA: 5 of 8 Frugivore/Folivore
binom.confint(5, 8, methods = c("asymptotic", "wilson"))
```

```
##       method x n  mean     lower     upper
## 1 asymptotic 5 8 0.625 0.2895261 0.9604739
## 2     wilson 5 8 0.625 0.3057424 0.8631557
```

```
# QDA: 3 of 8 Omnivore
binom.confint(3, 8, methods = c("asymptotic", "wilson"))
```

```
##       method x n  mean     lower     upper
## 1 asymptotic 3 8 0.375 0.0395261 0.7104739
## 2     wilson 3 8 0.375 0.1368443 0.6942576
```

```
## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## ---
## biotools version 4.3

##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  labelled[, vars]
## Chi-Sq (approx.) = 89.404, df = 40, p-value = 1.221e-05

## [1] 1.220951e-05
```