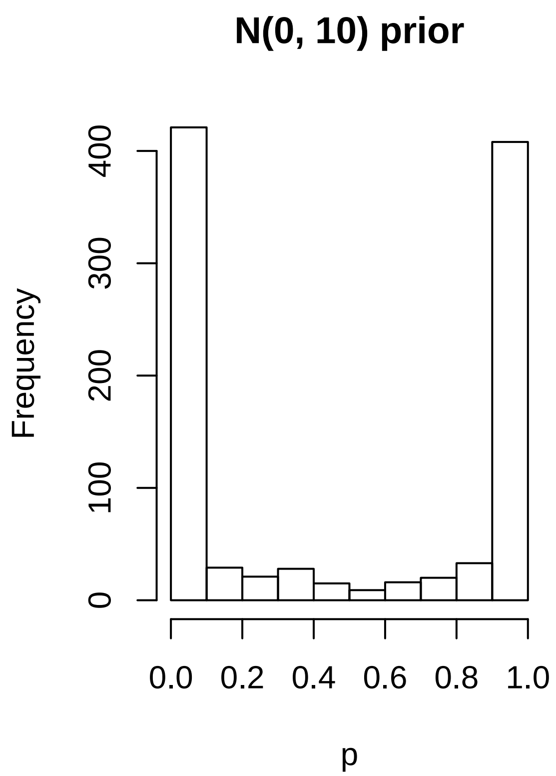
Review: Combining a pollen and macrofossil synthesis and climate simulations for spatial reconstructions of European climate using Bayesian filtering

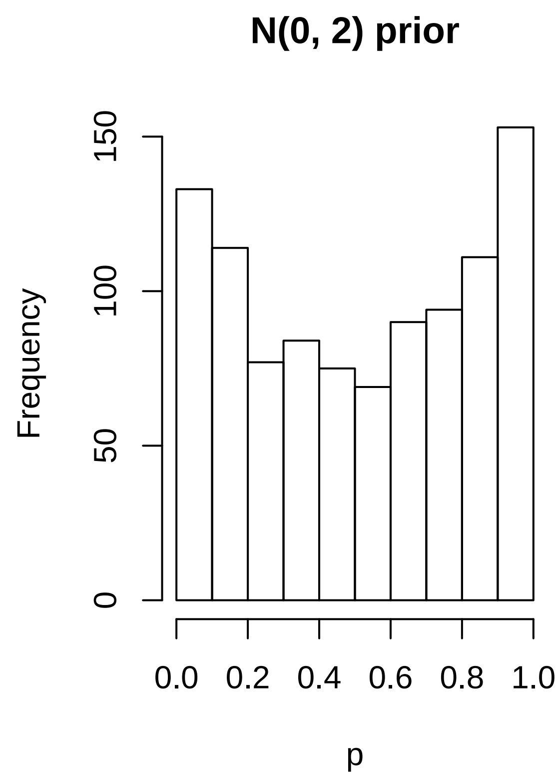
I appreciate the effort of the authors to address the comments of their manuscript in the revision. I think this revision is thoughtful and useful. I still have reservations about the interpretation of the statistics with specific comments below. I think the issues can be easily addressed with either small changes in the modeling or modifications in the writing.

I **very strongly** thank the authors for the inclusion of their code on Bitbucket and applaud them on their time putting this together. This is a major, and important, undertaking and is greatly appreciated.

- I find the fact that the regularization of the covariance matrix influences prediction more than model choice to be concerning (page 19, lines 30-32). I understand that this is common in the literature; however, I do think the difference in prediction based on regularization method suggest that the covariance matrix is not necessarily representing what it is supposed to. This is especially true given the effects of regularization suggest that $\mathbf{\Sigma}$ isn't really representing ensemble variability - just prior assumptions about the covariance structure. More of a discussion about how the impact of the regularization implies the interpretation of what $\mathbf{\Sigma}$ isn't as clear as inter-model variability even though in the ideal world this is what $\mathbf{\Sigma}$ would represent.
- page 20, lines 29-30 - significant if the **posterior** probability to exceed the…
- Page 8, line 21 - The prior for θ of $N(0, 10)$ is somewhat too vague for a logistic regression (in fact, it is actually somewhat informative). For example, it puts most of the prior mass on 0 or 1 and little in between. A $N(0, 2)$ is often a better choice for logistic models.

```r
library(LaplacesDemon)
layout(matrix(1:2, 1, 2))
## Normal (0, 10)
theta <- rnorm(1000, 0, 10)
p <- invlogit(theta)
hist(p, main = "N(0, 10) prior")
## Normal (0, 2)
theta <- rnorm(1000, 0, 2)
p <- invlogit(theta)
hist(p, main = "N(0, 2) prior")
```





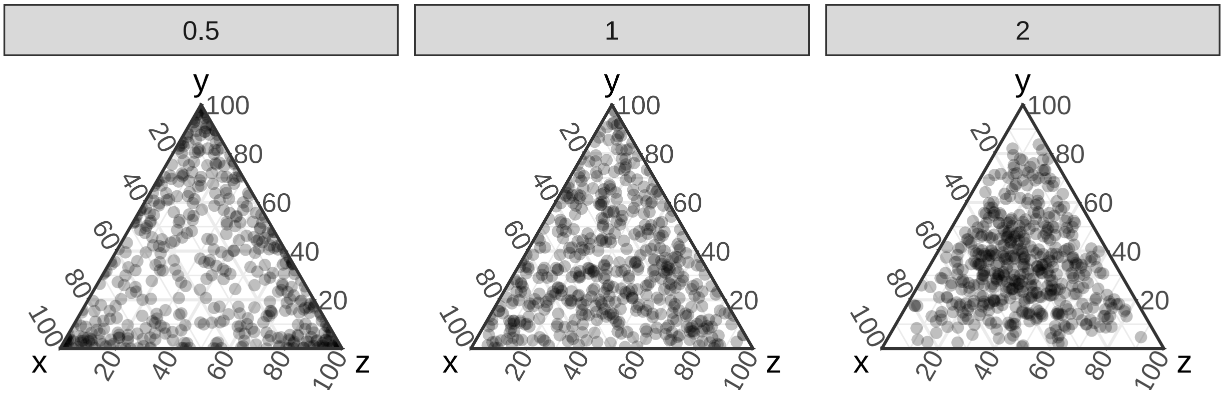
- Equation 8 - Why the (1/2 , … 1/2)? A (1, …, 1) prior produces an *a priori* uniform distribution over mixtures. I believe a (1/2, …, 1/2) prior pushes the weights more towards the extremes of the composition which could influence the collapsing of the weights to a small number of ensemble members. A better choice would be to use an α (1, …, 1) prior and assign α some prior (e.g., a gamma(1, 1)). Notice how in the first plot, there are more samples at the verices of the plot suggesting a collapsing of weights. A more relaxed prior might also improve MCMC mixing.

```r
library(ggtern)
```

```
## Loading required package: ggplot2

## --
## Remember to cite, run citation(package = 'ggtern') for further info.
## --

##
## Attaching package: 'ggtern'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, aes, annotate, calc_element, ggplot, ggplot_build,
##     ggplot_gtable, ggplotGrob, ggsave, layer_data, theme,
##     theme_bw, theme_classic, theme_dark, theme_gray, theme_light,
##     theme_linedraw, theme_minimal, theme_void
```

```r
N <- 500
draws <- rbind(LaplacesDemon::rdirichlet(N, 0.5 * c(1,1,1)),
               LaplacesDemon::rdirichlet(N, c(1,1,1)),
               LaplacesDemon::rdirichlet(N, 2 * c(1,1,1)))
ggtern(data = data.frame(x = draws[, 1],
                         y = draws[, 2],
                         z = draws[, 3],
```

```
                          alpha = rep(c(0.5, 1, 2), each = N)),
     aes(x, y, z)) +
  facet_wrap( ~ alpha) +
  geom_point(alpha = 0.25) +
  theme(legend.position = "NULL") +
  theme_bw()
```



- Equation 21 - Skill scores in general are not proper scoring rules. I know these are common in the literature, but I find the widespread use to be concerning, especially where there are many alternatives that convey the same information but don't lose propriety (e.g., just present the scores, present the difference in scores (not scaled by BS(prior)), etc.).