PBD

2022-07-11

Base exploration

Before we start, first a quick sanity check to verify that the PBD model can simulate birth-death trees. To do so, we switch off the protracted part and fit the birth-death model.

```
num_repl <- total_num_repl
lambda   <- 1
found <- rep(NA, num_repl)

for (r in 1:num_repl) {
  focal_tree <- ape::rphylo(n = 100,
                            birth = lambda, death = 0)
  brts <- treestats::branching_times(focal_tree)
  testthat::expect_output(
    bd_estimate <- DDD::bd_ML(brts = brts)
  )
  found[r] <- bd_estimate$lambda0
}
mean(found)
```

```
## [1] 1.078506
```

```
median(found)
```

```
## [1] 1.048436
```

```
quantile(found, probs = c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.8448379 1.4891504
```

The mean, median and quantiles seem to include 1.0 and be close to that, so I think we can be satisfied that the PBD model simulates accurate in the limit of the birth-death model.

Exploring the impact of tau

The protracted birth-death model was created to provide a possible solution of the 'pull of the present': because close to the present, there are many incipient species that are not yet developed into 'good' species, this may cause an apparent increase in speciation. Thus, one of the expectations we can have, is that as the time until species completion increases, this effect becomes stronger. Or in other words, as species completion

becomes more instantaneous, this effect becomes weaker. Thus, if we vary the speciation completion rate, we expect this to affect tau, and to affect branching-time related summary statistics.

To do so, we vary lambda (the speciation completion rate) in 10[1], and we choose b1 and b2 a bit higher, as this tends to lead to more pronounced effects. Simultaneously, we choose the crown age shorter, to avoid having huge trees. We condition the tree size to be in [100, 200], to avoid tree size effects to a large degree.

We track the following summary statistics that relate to branching times: Gamma, Mean Branch Length, Pigot's Rho and the base nLTT statistic.

Gamma statistic expectation

The Gamma statistic indicates a deviation from birth-death accumulation of lineages, where negative values indicate a deceleration of diversification (e.g. this indicates that the majority of branching events is closer to the root than expected under the bd model), and a positive value indicates an acceleration of diversification, e.g. the majority of branching events is closer to the tips than expected under the bd model. We expect that if the speciation completion rate is very high (and tau low), the Gamma statistic is close to zero, as in this case, the model approaches the birth-death model. As lambda decreases (and tau increases), we expect the branching events to move more towards the root, as many incipient species are pruned upon completion of the tree. Thus, lower lambda should yield negative Gamma values.

Mean Branch Length expectation

We expect that as lambda decreases (and tau increases), the mean branch length becomes longer, as it will take longer for speciation to complete, which will cause longer terminal branches.

Pigot's Rho expectation

Pigot's Rho calculates the change in diversification rate between the first half and the second half of the extant phylogeny. Negative values indicate a slow down (analogous to the Gamma statistic), and positive values indicate a speed up. With decreasing lambda (and increasing tau), we expect Pigot's Rho to become more negative. Furthermore, for high lambda (and tau ~ 0), we expect Pigot's Rho to be close to zero.

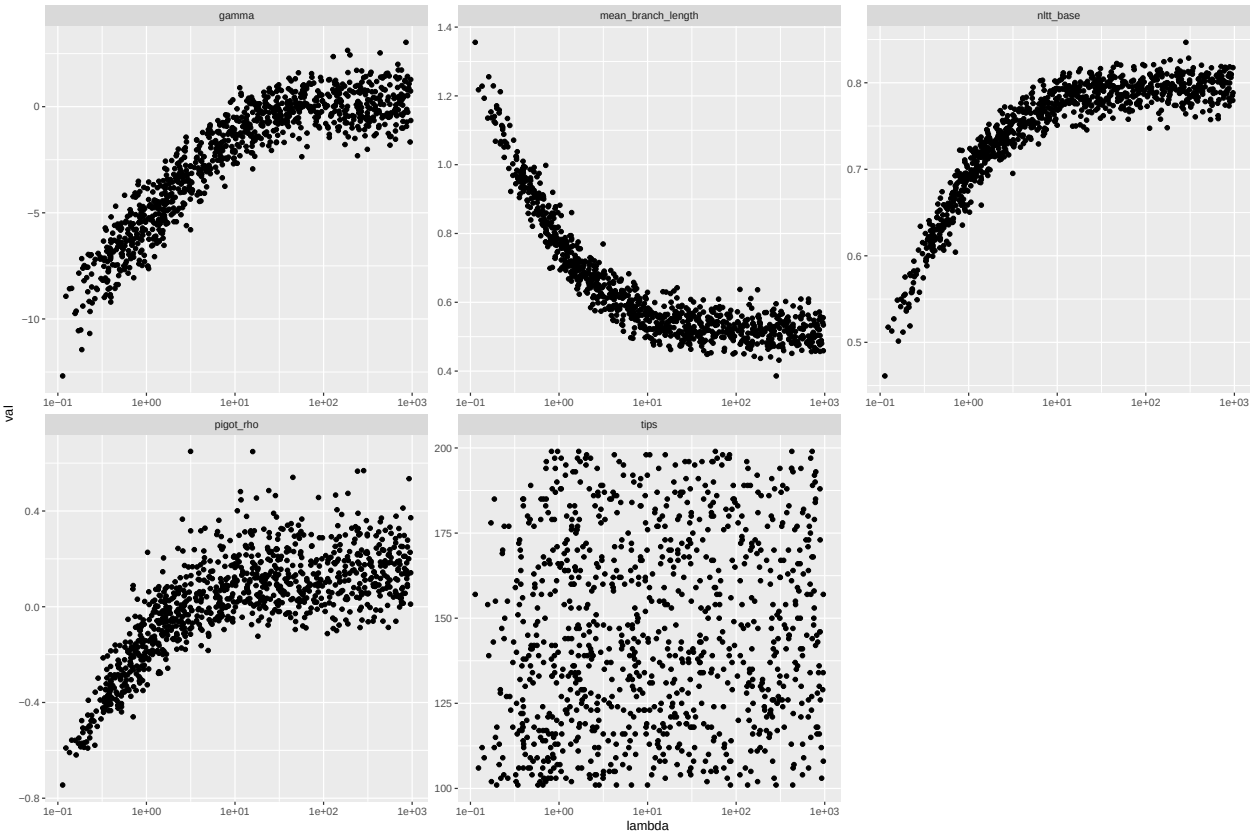
nLTT base expectation

The nLTT base statistic compare the normalized Lineage Through Time plots between the focal tree and an 'empty' tree consisting of only two crown lineages that don't diversify. As such, this statistic measures the surface under the nLTT curve. As diversification shifts towards the root, we expect this surface to increase, thus, decreasing lambda should lead to larger values of the nLTT base statistic.

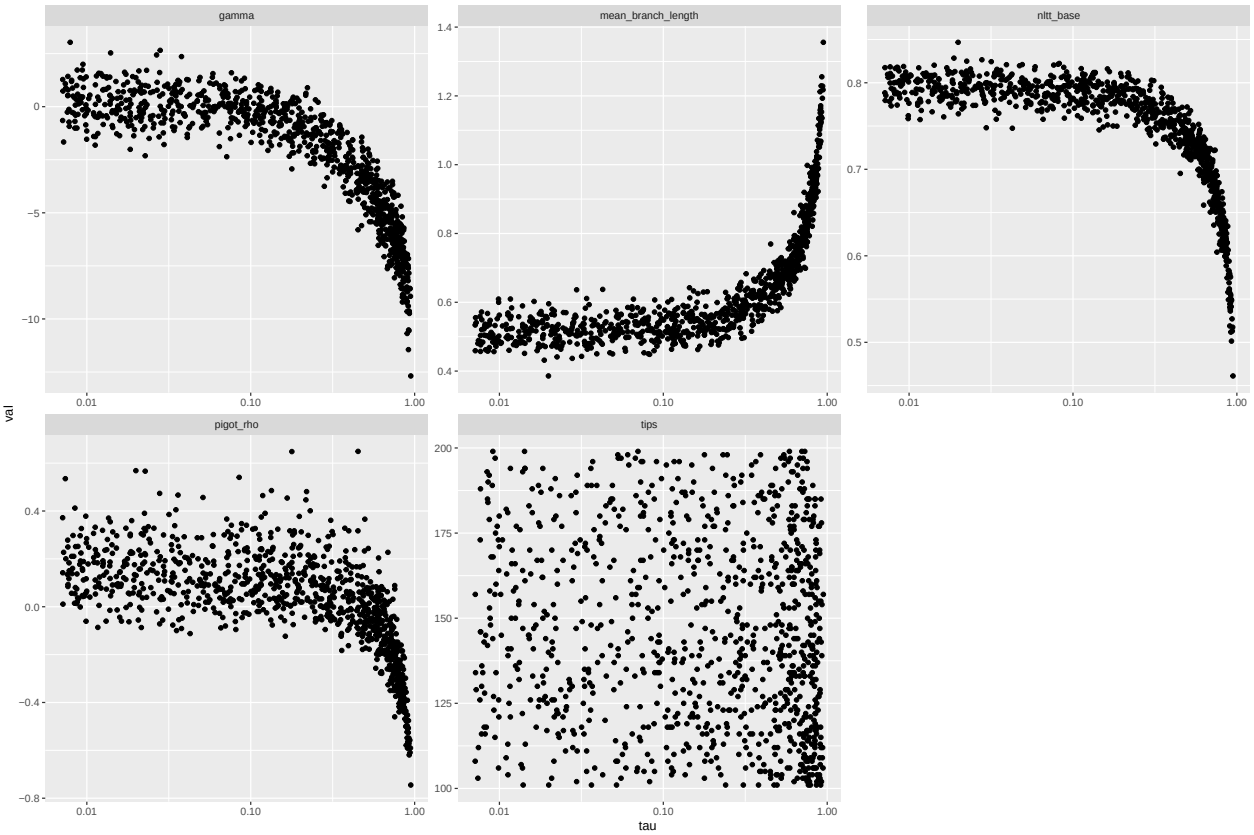
```
found <- c()
num_repl <- total_num_repl
while (TRUE) {
  mu1 <- mu2 <- 0.0
  b1 <- 1 # speciation rate good species
  b2 <- 1 # speciation rate incipient species
  lambda <- 10^runif(1, -1, 3)

  m2 <- mu2
  la2 <- b2
```

[1] -1, 3

```
  la3 <- lambda

  local_d <- sqrt((la2 + la3) ^ 2 + 2 * (la2 - la3) * m2 + (m2) ^ 2)
  local_frac <- (la2 - la3 + m2) / local_d
  tau <- (2 / (local_d - la2 + la3 - m2)) * log(2 / (1 + local_frac))

  focal_tree <- pbd_sim(pars = c(b1, lambda, b2, mu1, mu2), age = 5)

  n_lin <- treestats::number_of_lineages(focal_tree$stree_random)

  if (n_lin > 100 && n_lin < 200) {

    stats <- treestats::calc_brts_stats(focal_tree$stree_random)
    to_add <- c(lambda, tau, n_lin, unlist(stats), "random")
    found <- rbind(found, to_add)
    if (length(found[, 1]) >= num_repl) {
      break
    }
  }
}


colnames(found) <- c("lambda", "tau", "tips", names(unlist(stats)), "type")
found <- tibble::as_tibble(found)
found <- found %>% mutate_at(1:7, as.numeric)

found %>%
  gather(key = "statistic", value = "val", -c(tau, lambda, type)) %>%
  ggplot(aes(x = lambda, y = val)) +
  geom_point() +
  scale_x_log10() +
  facet_wrap(~statistic, scales = "free")
```




```
found %>%
  gather(key = "statistic", value = "val", -c(tau, lambda, type)) %>%
  ggplot(aes(x = tau, y = val)) +
  geom_point() +
  scale_x_log10() +
  facet_wrap(~statistic, scales = "free")
```



We see that the Gamma statistic indeed decreases with increasing tau, and is ~0 for tau ~0. Furthermore, the mean branch length indeed increases as tau increases. Pigot's Rho is also ~0 for tau ~0, and decreasing with increasing tau, indicating that the majority of diversification is moving towards the root. Lastly, the nLTT statistic seems to do the opposite of our expectation: with increasing tau, the statistic becomes smaller. I do not know why this is the case.

Chasing balance statistics

The goal we are chasing, is to find whether the PBD model can, in a reliable way, cause imbalance in a phylogenetic tree. Conversely, imbalance in empirical trees could then be indicative of the PBD process.

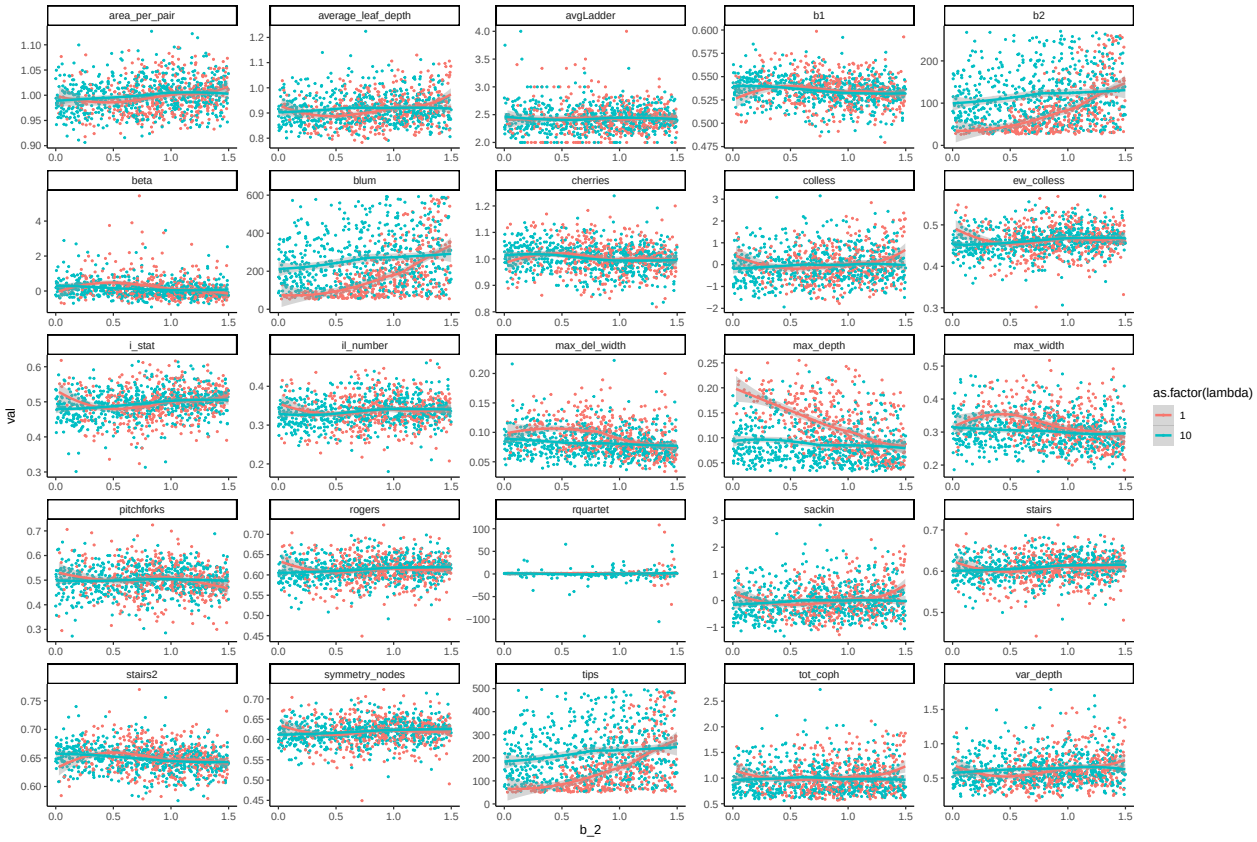
Incipient speciation rate

So far, we have assumed the incipient speciation rate (b2) to be identical to the 'good' speciation rate. E.g. incipient species can undergo speciation at the same rate as complete species. Changing this rate may cause an imbalance, as parts of the tree with many incipient species may explode into more and more incipient species. We again simulate, this time conditional on the number of tips in [50, 500], and using size adjusted balance statistics. We choose b1 = 1, and vary b2 in U[0, 1.5]. Results are shown for two speciation completion rates: 1 and 10.

```
found <- c()
num_repl <- total_num_repl
while (TRUE) {
  mu1 <- mu2 <- 0.0
  b1 <- 1 # speciation rate good species
```

```
  b2 <- runif(1, 0.0, 1.5) # speciation rate incipient species
  lambda <- sample(c(1, 10), 1)

  m2 <- mu2
  la2 <- b2
  la3 <- lambda

  local_d <- sqrt((la2 + la3) ^ 2 + 2 * (la2 - la3) * m2 + (m2) ^ 2)
  local_frac <- (la2 - la3 + m2) / local_d
  tau <- (2 / (local_d - la2 + la3 - m2)) * log(2 / (1 + local_frac))

  focal_tree <- pbd_sim(pars = c(b1, lambda, b2, mu1, mu2), age = 5)

  n_lin <- treestats::number_of_lineages(focal_tree$stree_random)

  if (n_lin > 50 && n_lin < 500) {

    stats <- treestats::calc_balance_stats(focal_tree$stree_random,
                                           normalize = TRUE)
    to_add <- c(b2, tau, lambda, n_lin, unlist(stats))
    found <- rbind(found, to_add)
    if (length(found[, 1]) >= num_repl) {
      break
    }
  }
}

colnames(found) <- c("b_2", "tau", "lambda", "tips", names(unlist(stats)))
found <- tibble::as_tibble(found)

found %>%
  gather(key = "statistic", value = "val", -c(b_2, tau, lambda)) %>%
  ggplot(aes(x = b_2, y = val, col = as.factor(lambda))) +
  geom_point(size = 0.5) +
  stat_smooth() +
  facet_wrap(~statistic, scales = "free") +
  theme_classic()
```

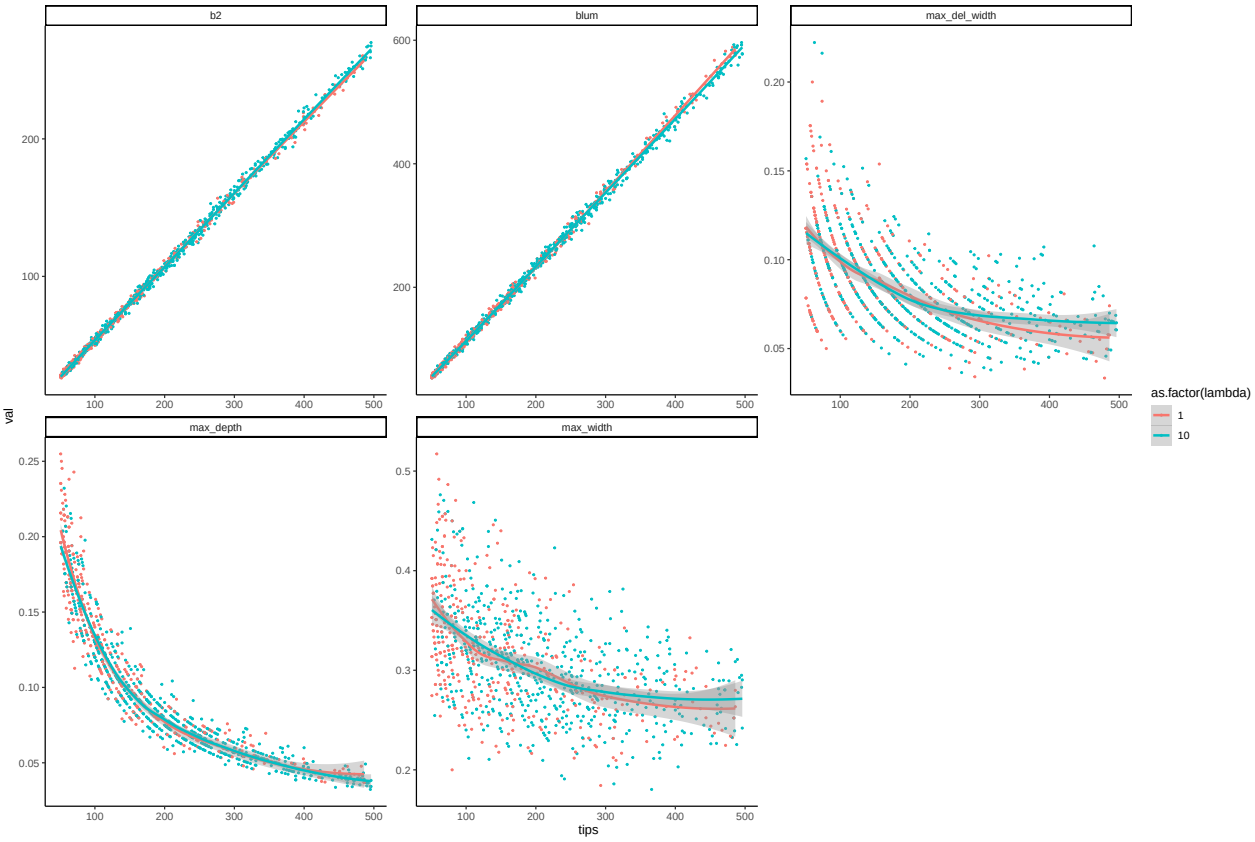
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



There are no obvious correlations. It seems there are some differences between lambda = 1 and lambda = 10, and for some statistics (b2, blum, max_del_width, max_depth, max_width), it seems as if there is some kind of relationship between b2 and the statistic, for lambda = 1. However, it appears that this is an artefact of tree size:

```
found %>%
  gather(key = "statistic", value = "val", -c(b_2, tau, lambda, tips)) %>%
  filter(statistic %in% c("b2", "blum", "max_del_width", "max_depth", "max_width")) %>%
  ggplot(aes(x = tips, y = val, col = as.factor(lambda))) +
  geom_point(size = 0.5) +
  stat_smooth() +
  facet_wrap(~statistic, scales = "free") +
  theme_classic()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



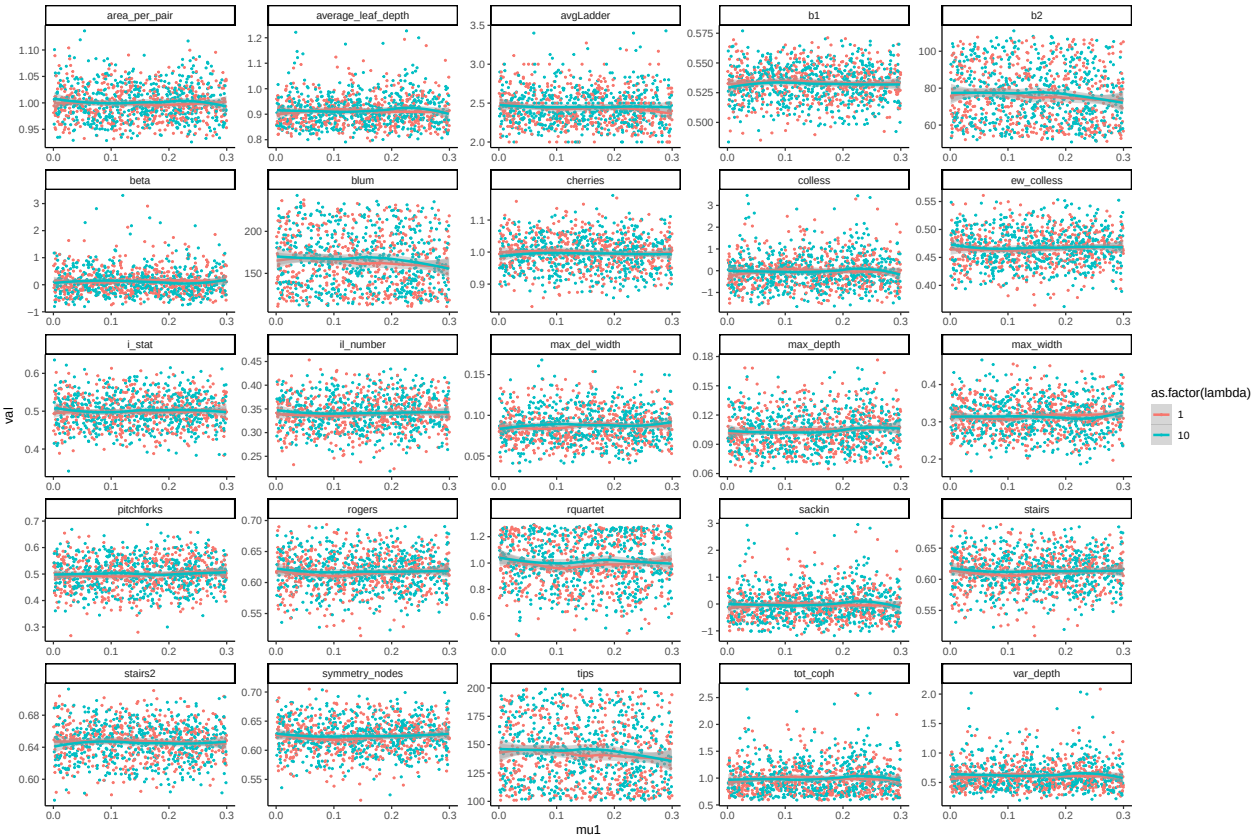
Extinction rate

So far, we have only considered Yule trees, e.g. trees without extinction. Differential extinction between good and incipient species might alternatively drive (im)balance. First, we explore the effect of good species having a higher extinction rate than incipient species. We set b1 = b2 = 1, lambda = [1, 10], mu2 = 0 and mu1 = U[0, 0.3]. Trees are conditioned on [100, 200] tips.

```
found <- c()
num_repl <- total_num_repl
while (TRUE) {
  mu1 <- runif(n = 1, min = 0, max = 0.3)
  mu2 <- 0.0
  b1 <- 1 # speciation rate good species
  b2 <- 1 # speciation rate incipient species
  lambda <- sample(c(1, 10), 1)

  m2 <- mu2
  la2 <- b2
  la3 <- lambda

  local_d <- sqrt((la2+la3)^2+2*(la2 - la3)*m2+(m2)^2)
  local_frac <- (la2 - la3 + m2) / local_d
  tau <- (2 / (local_d - la2 + la3 - m2)) * log(2 / (1 + local_frac))

  focal_tree <- pbd_sim(pars = c(b1, lambda, b2, mu1, mu2), age = 5)
```

```
  n_lin <- treestats::number_of_lineages(focal_tree$stree_random)

  if (n_lin > 100 && n_lin < 200) {

    stats <- treestats::calc_balance_stats(focal_tree$stree_random,
                                           normalize = TRUE)
    to_add <- c(mu1, tau, lambda, n_lin, unlist(stats))
    found <- rbind(found, to_add)
    if (length(found[, 1]) >= num_repl) {
      break
    }
  }
}

colnames(found) <- c("mu1", "tau", "lambda", "tips", names(unlist(stats)))
found <- tibble::as_tibble(found)

found %>%
  gather(key = "statistic", value = "val", -c(mu1, tau, lambda)) %>%
  ggplot(aes(x = mu1, y = val, col = as.factor(lambda))) +
  geom_point(size = 0.5) +
  stat_smooth() +
  facet_wrap(~statistic, scales = "free") +
  theme_classic()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

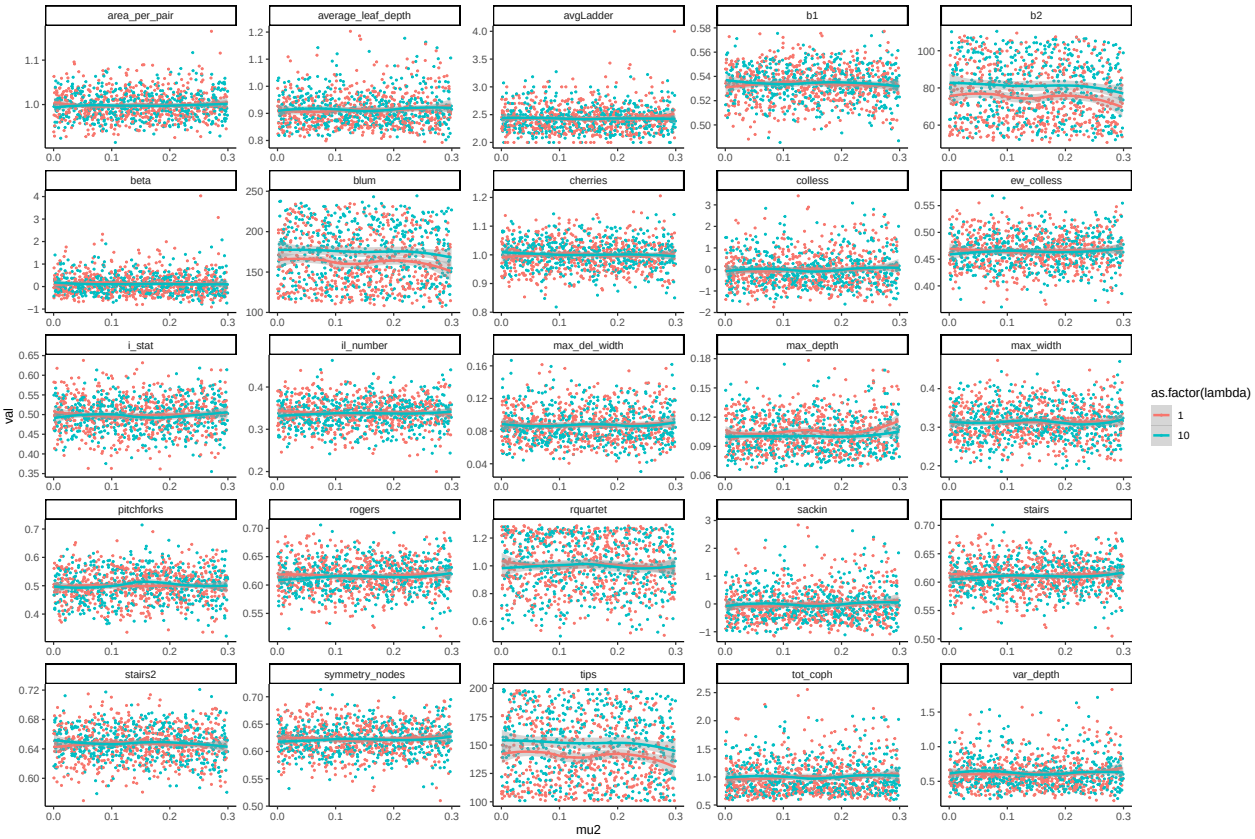


This, again, does not lead to any generated imbalance.

Now, we can do the same thing for the extinction rate of incipient species:

```
found <- c()
num_repl <- total_num_repl
while (TRUE) {
  mu1 <- 0.0
  mu2 <- runif(n = 1, min = 0, max = 0.3)
  b1 <- 1 # speciation rate good species
  b2 <- 1 # speciation rate incipient species
  lambda <- sample(c(1, 10), 1)

  m2 <- mu2
  la2 <- b2
  la3 <- lambda

  local_d <- sqrt((la2 + la3) ^ 2 + 2*(la2 - la3) * m2 + (m2) ^ 2)
  local_frac <- (la2 - la3 + m2) / local_d
  tau <- (2 / (local_d - la2 + la3 - m2)) * log(2 / (1 + local_frac))

  focal_tree <- pbd_sim(pars = c(b1, lambda, b2, mu1, mu2), age = 5)

  n_lin <- treestats::number_of_lineages(focal_tree$stree_random)

  if (n_lin > 100 && n_lin < 200) {

    stats <- treestats::calc_balance_stats(focal_tree$stree_random,
                                           normalize = TRUE)
    to_add <- c(mu2, tau, lambda, n_lin, unlist(stats))
    found <- rbind(found, to_add)
    if (length(found[, 1]) >= num_repl) {
      break
    }
  }
}

colnames(found) <- c("mu2", "tau", "lambda", "tips", names(unlist(stats)))
found <- tibble::as_tibble(found)

found %>%
  gather(key = "statistic", value = "val", -c(mu2, tau, lambda)) %>%
  ggplot(aes(x = mu2, y = val, col = as.factor(lambda))) +
  geom_point(size = 0.5) +
  stat_smooth() +
  facet_wrap(~statistic, scales = "free") +
  theme_classic()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Again, this does not lead to any discernable effect on balance.

With that, I would like to conclude that in practice, the PBD model does not seem to be able to generate imbalance.