Investigate MLE bias for the CRBD model

The Constant Rate Birth Death (or CRBD) model has two constant parameters:

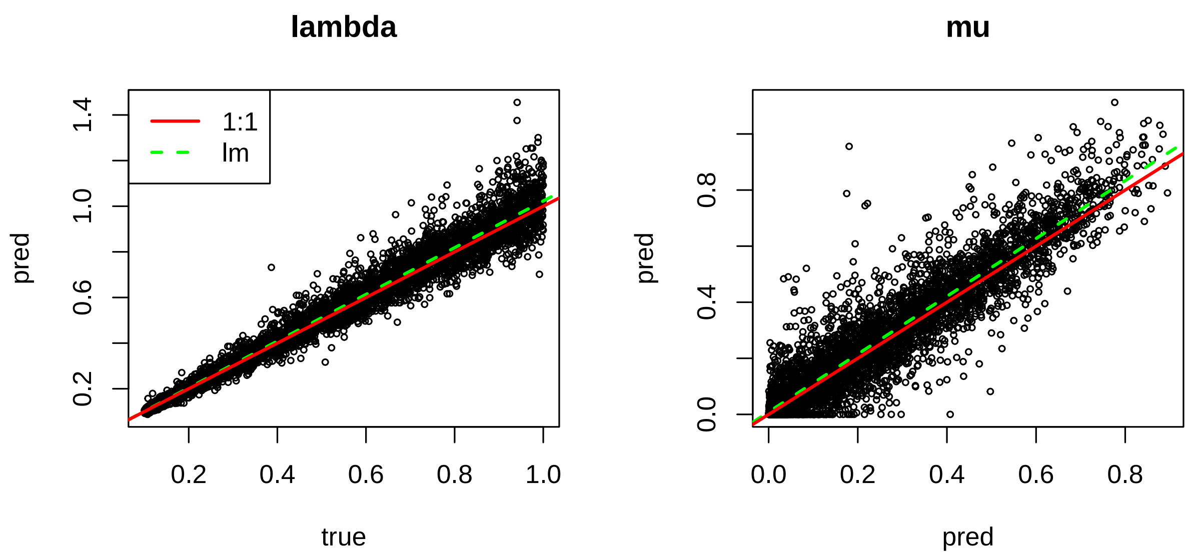
- $\lambda \in [0, 1]$ the speciation rate
- $\mu \in [0, 0.9]$ the extinction rate

Load data

```
set.seed(113)

params.crbd <- readRDS("params-testset-crbd.rds") # read data file
params.true <- params.crbd$true                   # true parameter values
params.mle  <- params.crbd$pred$mle               # predicted values by MLE
```

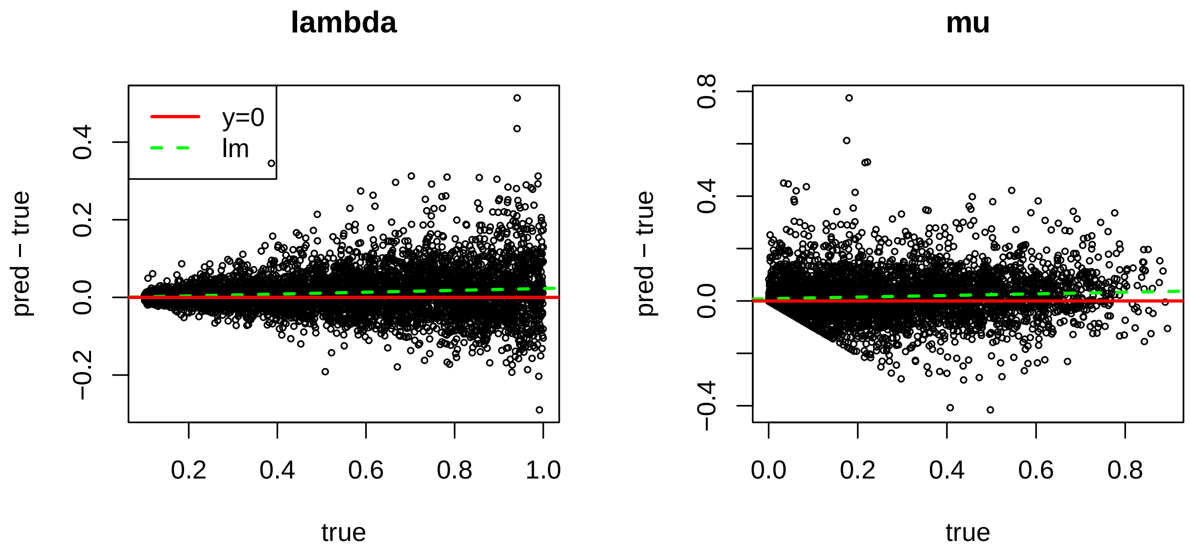
Plot true vs. predicted values by Maximum Likelihood Estimations (MLE)

```
par(mfrow = c(1,2)) # 1 row, 2 columns

# Scatter plot - Lambda
plot(params.true$lambda, params.mle$lambda, xlab = "true", ylab = "pred",
     main = "lambda", cex = .5)                          # scatter plot
lm.lambda <- lm(params.mle$lambda ~ params.true$lambda)  # fit lm
abline(lm.lambda, col = "green", lty = 2, lw = 2)        # plot lm
abline(0, 1     , col = "red"  , lty = 1, lw = 2)        # plot 1:1 line
legend("topleft", legend = c("1:1", "lm"), lty = c(1,2), # add legend
       col = c("red", "green"), lw = 2)

# Scatter plot - Mu
plot(params.true$mu    , params.mle$mu,     xlab = "pred", ylab = "pred",
     main = "mu", cex = .5)
lm.mu <- lm(params.mle$mu ~ params.true$mu)
abline(lm.mu, col = "green", lty = 2, lw = 2)
abline(0, 1 , col = "red"  , lty = 1, lw = 2)
```



Visualize bias: plot pred - true vs. true

```
par(mfrow = c(1,2)) # 1 row, 2 columns

# Scatter plot - Lambda
plot(params.true$lambda, params.mle$lambda - params.true$lambda,
     xlab = "true", ylab = "pred - true",
     main = "lambda", cex = .5)                           # scatter plot
lm.lambda <- lm(params.mle$lambda - params.true$lambda
                ~ params.true$lambda)                     # fit lm
abline(lm.lambda, col = "green", lty = 2, lw = 2)         # plot lm
abline(0, 0      , col = "red"   , lty = 1, lw = 2)       # plot y=0 line
legend("topleft", legend = c("y=0", "lm"), lty = c(1,2), # add legend
       col = c("red", "green"), lw = 2)

# Scatter plot - Mu
plot(params.true$mu, params.mle$mu - params.true$mu,
     xlab = "true", ylab = "pred - true",
     main = "mu", cex = .5)                       # scatter plot
lm.mu <- lm(params.mle$mu - params.true$mu
                ~ params.true$mu)                 # fit lm
abline(lm.mu, col = "green", lty = 2, lw = 2)     # plot lm
abline(0, 0 , col = "red"  , lty = 1, lw = 2)     # plot y=0 line
```




```
summary(lm.lambda)
```

```
##
## Call:
## lm(formula = params.mle$lambda - params.true$lambda ~ params.true$lambda)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31239 -0.02743 -0.00253  0.02318  0.49251
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -0.0008967  0.0018694  -0.480    0.632
## params.true$lambda    0.0236743  0.0030446   7.776 9.04e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05631 on 4998 degrees of freedom
## Multiple R-squared:  0.01195,    Adjusted R-squared:  0.01176
## F-statistic: 60.46 on 1 and 4998 DF,  p-value: 9.044e-15
```

```
summary(lm.mu)
```

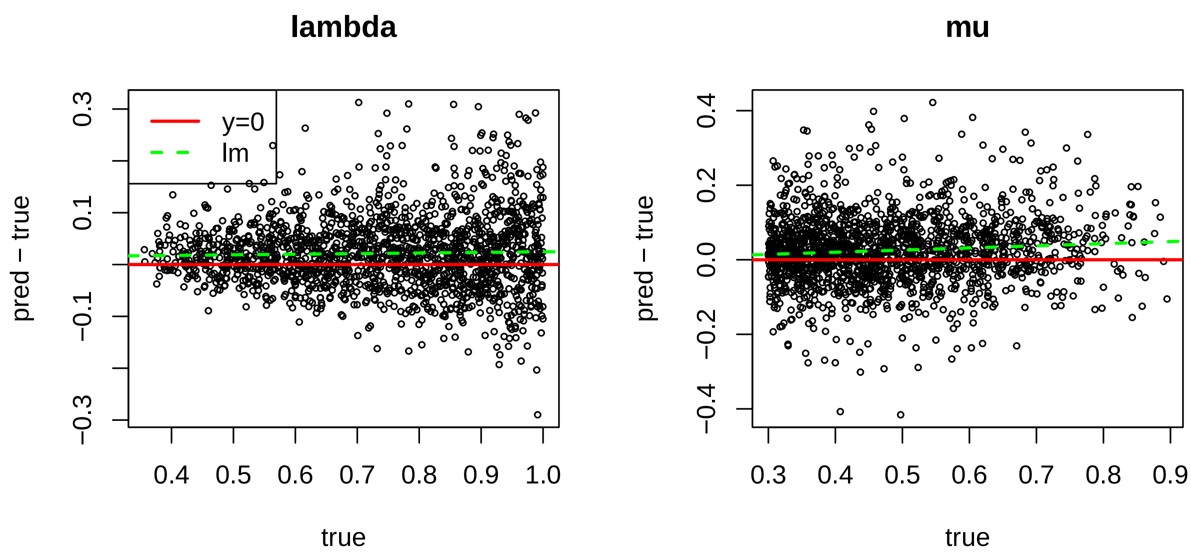
```
##
## Call:
## lm(formula = params.mle$mu - params.true$mu ~ params.true$mu)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43977 -0.03867 -0.00764  0.03356  0.76094
##
## Coefficients:
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.008576   0.001824   4.702 2.65e-06 ***
## params.true$mu 0.030898   0.005747   5.377 7.93e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08028 on 4998 degrees of freedom
## Multiple R-squared:  0.005751,   Adjusted R-squared:  0.005552
## F-statistic: 28.91 on 1 and 4998 DF,  p-value: 7.933e-08
```

We note a significant bias for both parameters.

Moreover observe an asymmetry in the scatter plot of μ because MLE doesn't predict negative values. Let's check if the bias comes from this asymmetry by considering only the points such that $\mu \geq 0.3$.

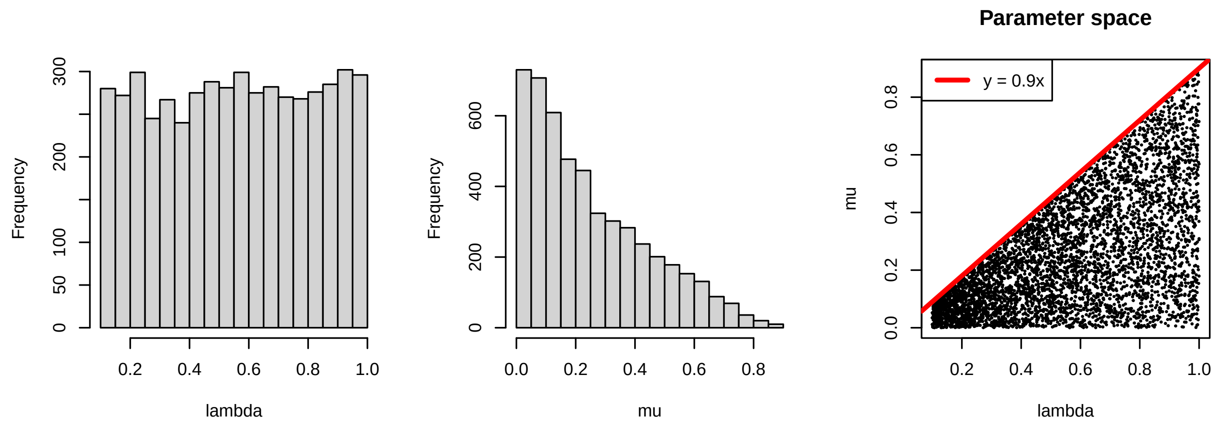
```r
par(mfrow = c(1,2)) # 1 row, 2 columns

ind <- params.true$mu > .3

# Scatter plot - Lambda
plot(params.true$lambda[ind], params.mle$lambda[ind] - params.true$lambda[ind],
     xlab = "true", ylab = "pred - true",
     main = "lambda", cex = .5)                              # scatter plot
lm.lambda <- lm(params.mle$lambda[ind] - params.true$lambda[ind]
                  ~ params.true$lambda[ind])                      # fit lm
abline(lm.lambda, col = "green", lty = 2, lw = 2)         # plot lm
abline(0, 0     , col = "red"  , lty = 1, lw = 2)         # plot y=0 line
legend("topleft", legend = c("y=0", "lm"), lty = c(1,2), # add legend
       col = c("red", "green"), lw = 2)

# Scatter plot - Mu
plot(params.true$mu[ind], params.mle$mu[ind] - params.true$mu[ind],
     xlab = "true", ylab = "pred - true",
     main = "mu", cex = .5)                              # scatter plot
lm.mu <- lm(params.mle$mu[ind] - params.true$mu[ind]
                  ~ params.true$mu[ind])                      # fit lm
abline(lm.mu, col = "green", lty = 2, lw = 2)         # plot lm
abline(0, 0 , col = "red"  , lty = 1, lw = 2)         # plot y=0 line
```



The bias doesn't come from the asymmetry for $\mu \simeq 0^+$.

An other possibility is that the bias comes from the non-uniform distribution of μ:

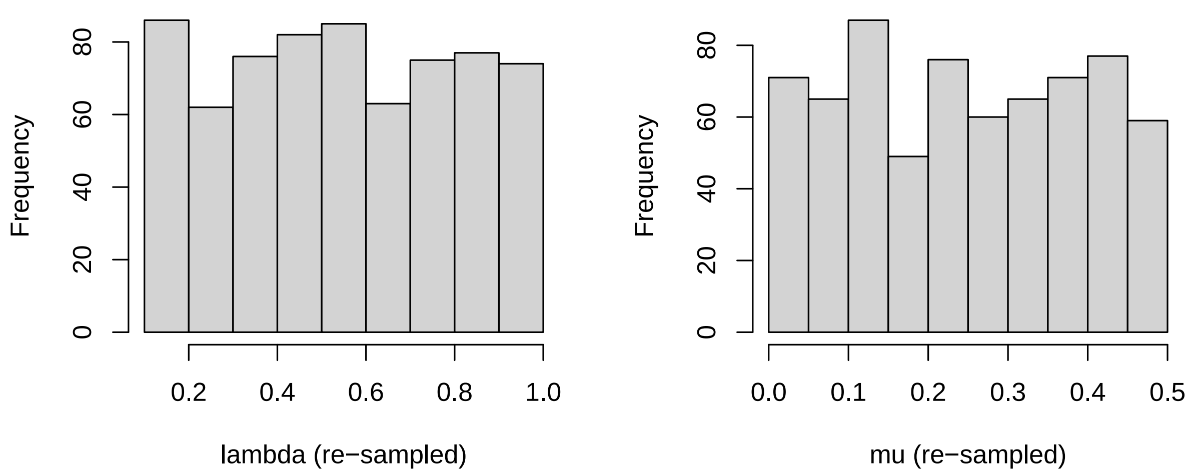
```
par(mfrow = c(1,3))
hist(params.true$lambda, main = "", xlab = "lambda")
hist(params.true$mu     , main = "", xlab = "mu")
plot(params.true$lambda, params.true$mu, xlab = "lambda", ylab = "mu", cex = .2,
     main = "Parameter space")
abline(0, .9, lty = 1, lw = 3, col = "red")
legend("topleft", legend = "y = 0.9x", lty = 1, col = "red", lw = 3)
```



Let's check that by re-sampling uniformly μ

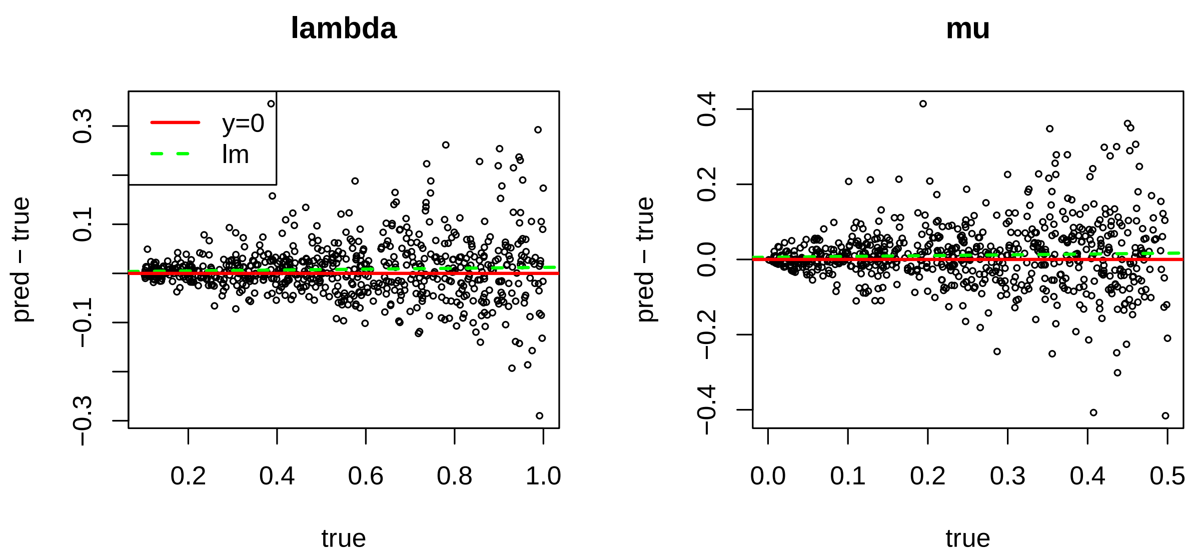
```
# Re-sampling parameter space s.t. lambda and mu have a uniform distribution
resamplingMu <- function(true.mu, true.lambda, nbin,
                         mu.min     = 0., mu.max     = .5,
                         lambda.min = .1, lambda.max = 1.){
```

```
  bin.border.mu     <- seq(mu.min    , mu.max    , length.out = nbin + 1)
  bin.border.lambda <- seq(lambda.min, lambda.max, length.out = nbin + 1)
  n.per.bin  <- length(true.mu)
  ind.all    <- vector(mode = "list", length = nbin)

  for (i in 1:nbin){
    ind.bin.mu     <- true.mu > bin.border.mu[i] & true.mu < bin.border.mu[i + 1]
    ind.bin.lambda <- true.lambda > bin.border.lambda[i] &
                      true.lambda < bin.border.lambda[i + 1]
    ind.bin        <- ind.bin.mu & ind.bin.lambda
    ind.bin        <- which(ind.bin == TRUE)
    ind.all[[i]]   <- ind.bin
    n.per.bin      <- min(c(n.per.bin, length(ind.bin)))
  }

  ind.samp <- c()
  for (i in 1:nbin){ind.samp <- c(ind.samp, sample(ind.all[[i]], n.per.bin))}

  return(ind.samp)
}


# Plotting histograms of the re-sampled parameter space
# to check if lambda and mu are now uniformly distributed

par(mfrow = c(1,2))

ind.resample <- resamplingMu(params.true$mu, params.true$lambda , 5)
hist(params.true$lambda[ind.resample], xlab = "lambda (re-sampled)", main = "")
hist(params.true$mu[ind.resample]    , xlab = "mu (re-sampled)"    , main = "")
```



Now that we have uniformly distributed values for both parameters, we can check if the bias is still here:

```r
par(mfrow = c(1,2)) # 1 row, 2 columns

ind <- ind.resample

# Scatter plot - Lambda
plot(params.true$lambda[ind], params.mle$lambda[ind] - params.true$lambda[ind],
     xlab = "true", ylab = "pred - true",
     main = "lambda", cex = .5)                             # scatter plot
lm.lambda <- lm(params.mle$lambda[ind] - params.true$lambda[ind]
                ~ params.true$lambda[ind])                        # fit lm
abline(lm.lambda, col = "green", lty = 2, lw = 2)         # plot lm
abline(0, 0      , col = "red"  , lty = 1, lw = 2)         # plot y=0 line
legend("topleft", legend = c("y=0", "lm"), lty = c(1,2), # add legend
       col = c("red", "green"), lw = 2)

# Scatter plot - Mu
plot(params.true$mu[ind], params.mle$mu[ind] - params.true$mu[ind],
     xlab = "true", ylab = "pred - true",
     main = "mu", cex = .5)                         # scatter plot
lm.mu <- lm(params.mle$mu[ind] - params.true$mu[ind]
                ~ params.true$mu[ind])                        # fit lm
abline(lm.mu, col = "green", lty = 2, lw = 2)         # plot lm
abline(0, 0 , col = "red"  , lty = 1, lw = 2)         # plot y=0 line
```



The bias seems to disappear. Thus that latter could result from the non-uniform distribution of μ. But further investigations need to be done to confirm this.

```r
summary(lm.lambda)
```

```
##
## Call:
## lm(formula = params.mle$lambda[ind] - params.true$lambda[ind] ~
##     params.true$lambda[ind])
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30192 -0.02801 -0.00369  0.02115  0.33873
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.003194   0.005289   0.604    0.546
## params.true$lambda[ind] 0.008979   0.008733   1.028    0.304
##
## Residual standard error: 0.05976 on 678 degrees of freedom
## Multiple R-squared:  0.001557,   Adjusted R-squared:  8.422e-05
## F-statistic: 1.057 on 1 and 678 DF,  p-value: 0.3042
```

```
summary(lm.mu)
```

```
##
## Call:
## lm(formula = params.mle$mu[ind] - params.true$mu[ind] ~ params.true$mu[ind])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43232 -0.03805 -0.00303  0.03499  0.40433
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.005816   0.006520   0.892    0.373
## params.true$mu[ind] 0.021461   0.022759   0.943    0.346
##
## Residual standard error: 0.08541 on 678 degrees of freedom
## Multiple R-squared:  0.00131,    Adjusted R-squared:  -0.0001632
## F-statistic: 0.8892 on 1 and 678 DF,  p-value: 0.346
```