Accounting for imperfect detection in functional ecology using hierarchical models in R

Tobias Roth

2017-06-15

Introduction

Currently, most studies in functional ecology calculate measures of functional diversity directly from observations, which are usually the counts or observed presences of several species at a number of sites (a site x species matrix which we call the observed meta-community). However, if measurement errors occur and especially if the measurement errors depend on species' traits, functional diversity measures are likely to be affected (Cardoso et al. 2014; Mihaljevic, Joseph & Johnson 2015; van der Plas et al. 2017).

In their excellent book Kéry & Royle (2016) advocated hierarchical models as an unifying concept in ecology to infer biological processes independent of measurement errors. They also introduce the package `unmarked` that makes it easy to apply these methods in real-world applications. In this vignette we provide a working example how these methods could be applied in studies of functional diversity.

As a requirement to run the presented code the latest version of the `detecionfilter` package should be downloaded using the following code.

```
library(devtools)
install_github("TobiasRoth/detectionfilter")
library(detectionfilter)
```

The `detectionfilter` package provides the function `simcom()` to simulate meta-community data where the change of species' occurrence along a gradient depends on its trait expression (i.e. environmental filtering) and where species' detection during field surveys also depends on trait expression (i.e detection filtering). The function `simcom()` is adopted from the simulation of a meta-community as described in chapter 11.2 in Kéry and Royle (2016). It is convenient to use simulated data in a working example as it allows to compare the estimates with the truth (i.e. the values used for the parameters to run the simulations). See chapter 11.2 in Kéry and Royle (2016) and the vignette *Simulation of meta-community data subject to environmental and detection filtering* of the package `detectionfilter` that describes the simulation parameters and the concept of the simulation in more detail.

We start with simulating data for a study with 200 sites, each site was visited twice and the regional species pool contains 100 species.

```
set.seed(1234)
dat <- simcom(mu.FTfilter.lpsi = -0.5, mean.psi = 0.8,
              mu.FTfilter.lp = 1, mean.p = 0.8,
              nsite = 200, nspec = 100, nrep = 2)
```

In this simulation we assumed that an average species occurred in 80% of the sites (`mean.psi = 0.8`) and during one survey in average species were observed in 80% of the sites they actually occurred (`mean.p = 0.8`). These are rather high values for average species' occurrence and detection. Consequently, all species of the regional species pool were observed and except for one species that was observed only at four sites all other species were observed at more than 20 sites.

```
# Calculate observed meta-community (i.e. merge the observations from the two visits)
commat_obs <- apply(dat$y, c(1,2), max)
```

```r
# Show species with few observations
sort(apply(commat_obs, 2, sum))[1:10]
```

```
## sp88 sp82 sp80 sp65  sp3 sp35 sp46 sp36 sp15 sp16
##    4   24   29   44   48   64   68   71   75   78
```

Thus, in this example we do not face the problem of rarely observed species for which true occurrence might be difficult to estimate. Excluding many rarely observed species (either because they are indeed rare or because they are hard to detect) could bias the results on functional diversity because these species might be special in terms of their functional traits. In the last chapter of this vignette we therefore tested our approach under a simulation with many species that were observed at only few sites.

Estimating detection-corrected meta-communites

Estimators of functional diversity may be constructed from model-based estimators of occurrence of individual species that incorporate imperfect detection of species (Dorazio and Royle 2005). In the hierarchical models as defined by Kéry and Royle (2016) the model-based estimator of species' occurrence (or abundance) is denoted as the vector 'z' with length equal to the number of surveyed sites. Thus, `z` describes the true occurrence of a species and can be related to site variables using logistic regression. However, whether a species is observed at site `i` during visit `j` (`y[i,j]`) depends on variables affecting detection as well as on its occurrence (that is `z[i]`). Indeed, the hierarchical structure that the observations conditionally depend on the outcome of the biological process (`f(y|z)`) is the overarching principle in many different ecological model suche as capture-recapture models, distance sampling models, occupancy models or N-mixture models for abundance (Kéry and Royle 2016).

In short, we need to pick one of the many hierarchical models that best fits the structure of our data. Applying this model to all species separately reveals the true occurrence (or abundance) for all species and sites. Taking the true occurrences of all species (the 'z' for all species) results in the site x species matrix that we call the detection-corrected meta-community matrix. From this detection-corrected meta-community matrix we can then calculate functional diversity. We here use the term functional diversity to refer to one of the many indices of functional diversity that use the presence/absence or abundance of species at different sites (site x species matrix) and functional traits of the species to calculate some indices that describes the communities in terms of their functional space. Such measures could range from the simple average of the trait expression of species in the community to multi-dimensional measures of functional diversity (Mason et al. 2005). Independent of the index used to calculate functional diversity, if calculated from detection-corrected meta-community matrix the measure should be independent from detection filtering (Dorazio and Royle 2005). Therefore, comparison between the functional diversity calculated from the observations and functional diversity calculated from the detection-corrected meta-community matrix would reveal how detection filtering affects functional diversity.

Since most of the methods to account for measurement errors in hierarchical models described in Kéry and Royle (2016) are implemented in the R-package `unmarked`, to calculate the detection corrected community matrix `z` only needs few lines of R-code. The function `unmarkedFrameOccu()` organizes detection, non-detection data along with the covariates. The function `occu()` fits the single season occupancy model of MacKenzie et al. (2002). A single season occupancy model is a hierarchical model in the form of `f(y|z)` were y are observed species presence and z is the true species presence. The function `ranef()` estimates posterior distributions of the `z` using empirical Bayes methods. Finally, we use the function `bup()` to extract the mode of the posterior probability. All these functions are from the package `unmarked`.

```r
# Package to fit hierarchical models of occurrence and abundance data
library(unmarked)

# Prepare detection-corrected meta-community matrix
z <- array(NA, dim = c(dim(dat$y)[1], dim(dat$y)[2]))
```

```
# Apply hierarchical model to all species seperate
for(k in 1:dim(dat$y)[2]) {
  d <- unmarkedFrameOccu(y = dat$y[,k,], obsCovs = list(date = dat$date),
                         siteCovs = data.frame(gradient = dat$gradient))
  res <- occu(~ date ~ gradient, data = d, se = TRUE)
  z[,k] <- bup(ranef(res), stat = "mode")
}
```

Note, that the functions `ranef()` and `bup()` can be applied to all hierarchical models implemented in the package `unmarked`. Thus, if abundance data instead of presence/absence should be analysed one only needs to choose the respective function to organize the data and one of the functions that implement hierarchical model for abundance data.

But let's go on with the presence/absence data example. Say, we are interested in how the mean trait expression of the species in a community (i.e. community mean CM) changes along the gradient. We calculate CMs from observed meta-community, from the detection-corrected meta-community and from the true meta-community. Note that the later we only know because we simulated the data.

```
# Function to calculate mean trait expression of species in a community (CM)
CM <- function(x) mean(dat$traitmat[x])

# Calculate CMs from observed meta-community matrix, from detection-corrected
# meta-community matrix and from true mata-community matrix
CM_obs <- apply(commat_obs==1, 1, CM)
CM_cor <- apply(z>0.5, 1, CM)
CM_true <- apply(dat$z_true==1, 1, CM)

# Plot CM along gradient
plot(dat$gradient, CM_obs, pch = 16, cex = 0.7, ylim = c(-1, 1),
     xlab = "Gradient", ylab = "Community mean")
points(dat$gradient, CM_true, cex = 0.7, col = "orange")
points(dat$gradient, CM_cor, pch = "+", cex = 0.7, col = "red")
abline(lm(CM_obs ~ dat$gradient))
abline(lm(CM_true ~ dat$gradient), col = "orange")
abline(lm(CM_cor ~ dat$gradient), col = "red")
```

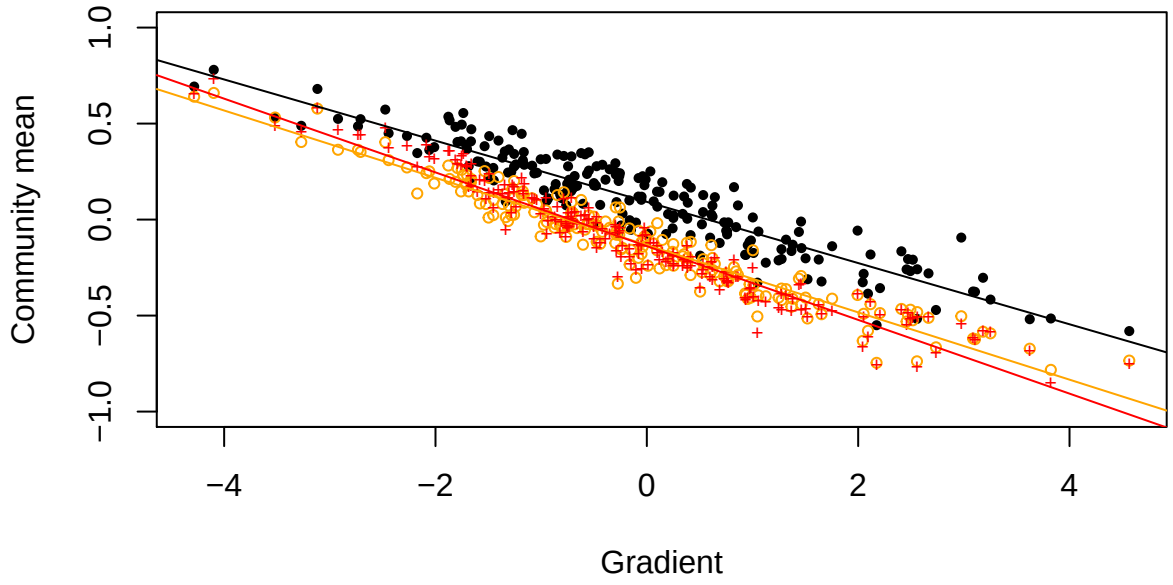

Fig. 1: Mean trait expression (CM) of true communities (open orange dots, only known because data are simulated), CMs of observed communities (black dots) and CMs from detection-corrected communities (red +) along elevational gradient.

In this example, community means calculated from observed meta-communities are biased high because species with low trait values are more likely to remain undetected than species with larger trait values. Clearly, calculating CMs from the detection-corrected meta-community reveals results that are less biased than CMs calculated from the observed meta-community. Still, however, how steep CMs are decreasing along the gradient is similar for observed and detection-corrected meta-communities and both are similar to the decrease of CMs along the gradient of the true meta-community. In this example detection filtering is thus unlikely to distort the main conclusion that species with higher trait expressions are more likely to be filtered from the communities at the higher end of the gradient (and species with low trait value at the lower end of the gradient).

Effect of unobserved or rarely observed species

We now simulate data where some of the species from the species pool remain undetected and a larger proportion of the species are only observed at few sites.

```
set.seed(1234)
dat <- simcom(mu.FTfilter.lpsi = -0.5, mean.psi = 0.5,
              mu.FTfilter.lp = 2, mean.p = 0.5,
              nsite = 200, nspec = 100, nrep = 2)

# Observed community matrix
commat_obs <- apply(dat$y, c(1,2), max)

# Number of observed species
ncol(commat_obs)
```

```
## [1] 98
```

From the 100 species that are in the regional species pool two species were not observed at all. Evidently it is not possible to apply the single-species hierarchical models that we use in our approach to species that were never observed. Note however, that hierarchical models exists that are applied to the entire meta-community and not to each species separately. Some of these multi-species hierarchical models can account for species of the regional species pool that were not observed. Such multi-species models, however, are difficult to be implemented in a frequentist framework and are thus not (yet) included in the package `unmarked`.

Unobserved species may arise because they were rare and did (by chance) not occur on the studied sites. In this case they are also not part of the true communities and will thus not affect the estimates of functional diversity. However, unobserved species may also arise because they were overseen at all the sites they occurred. In that case these specie are likely to bias estimates of functional diversity and we will not be able to account for this bias using our approach.

```
sum(apply(dat$z_true, 2, sum)>0)
```

```
## [1] 100
```

Since we use simulated data we know the true meta-community. Indeed all 100 species were present in at least on of the sites. Thus the two unobserved species were missed in all the sites they occured.

Additionally to the problem that two species were not observed at all, we also face the problem that some species were observed only at few sites.

```
# Proportion of  species observed at least once but in less than 10% of sites
mean(apply(commat_obs, 2, sum) < 20)
```

```
## [1] 0.2653061
```

Indeed more than a quarter of the species were observed in less than 10% of the sites. For species with few observations it might be difficult to accurately estimate the true occurrence and it is thus not clear whether our approach is able to accurately account for detection filtering in case when a larger proportion of the species are rare.

As above we estimate the detection corrected meta-community using the functions form the package `unmarked`, calculate CMs from detection-corrected meta-communities and compare it with the CMs from observations and the true meta-community known from the simulation.

```
# Estimate detection-corrected meta-community matrix
z <- array(NA, dim = c(dim(dat$y)[1], dim(dat$y)[2]))
for(k in 1:dim(dat$y)[2]) {
  d <- unmarkedFrameOccu(y = dat$y[,k,], obsCovs = list(date = dat$date),
                         siteCovs = data.frame(gradient = dat$gradient))
  res <- occu(~ date ~ gradient, data = d, se = TRUE)
  z[,k] <- bup(ranef(res), stat = "mode")
}

# Calculate CMs
CM_obs <- apply(commat_obs==1, 1, CM)
CM_cor <- apply(z>0.5, 1, CM)
CM_true <- apply(dat$z_true==1, 1, CM)

# Plot CMs along gradient
plot(dat$gradient, CM_obs, pch = 16, cex = 0.7, ylim = c(-1, 1),
     xlab = "Gradient", ylab = "Community mean")
points(dat$gradient, CM_true, cex = 0.7, col = "orange")
points(dat$gradient, CM_cor, pch = "+", cex = 0.7, col = "red")
abline(lm(CM_obs ~ dat$gradient))
abline(lm(CM_true ~ dat$gradient), col = "orange")
abline(lm(CM_cor ~ dat$gradient), col = "red")
```

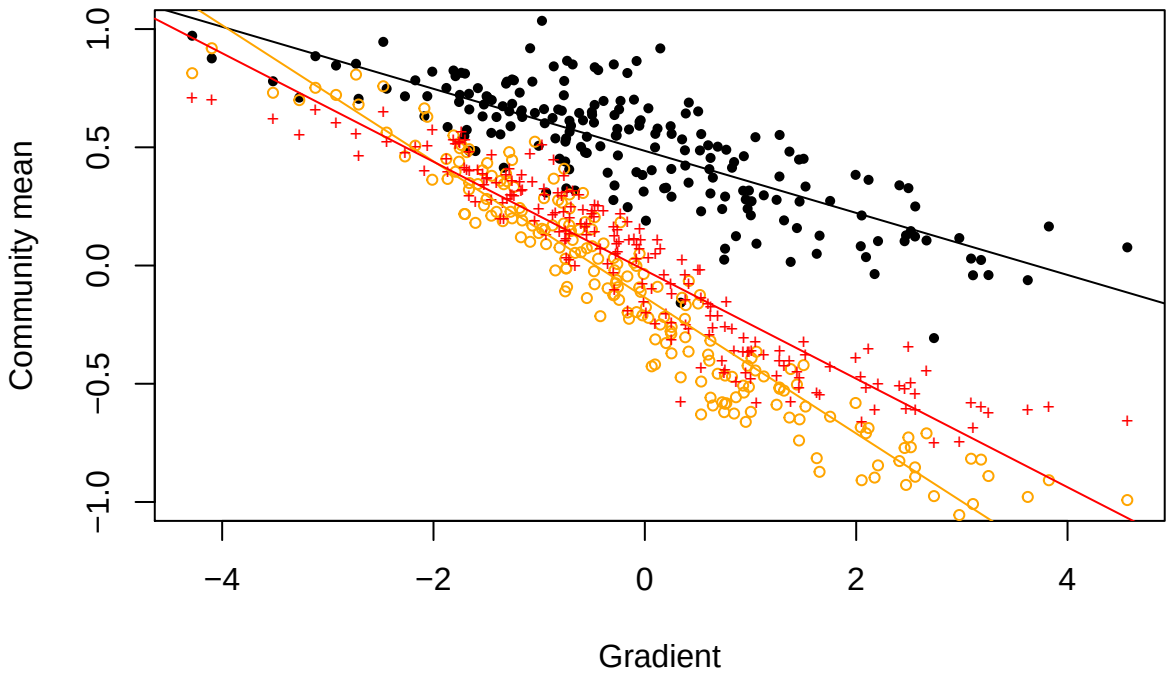

Fig. 2: Mean trait expression (CM) of true communities (open orange dots, only known because data are simulated), CMs of observed communities (black dots) and CMs from detection-corrected communities (red +) along elevational gradient.

In this example, the change in CMs along the gradient is steeper in the true communities than in the observed communities. Our approach that accounts for detection filtering is doing quit well in recovering the true pattern of CMs along the gradient. Still, however, the slope of CMs along the gradient estimated from the detection-corrected meta-communities is slightly less steep than the slope of true CMs. This is likely because two species were never observed and our approach is not able to account for the true occurrences of theses species.

References

Cardoso, P.; Rigal, F.; Borges, P. A. V.; Carvalho, J. C.; Faith, D., 2014. A new frontier in biodiversity inventory: a proposal for estimators of phylogenetic and functional diversity. - Methods in Ecology and Evolution 5: 452-461.

Dorazio, R. M.; Royle, J. A., 2005. Estimating size and composition of biological communities by modeling the occurrence of species. - Journal of the American Statistical Association 100: 389-398.

Kéry, M.; Royle, J. A., 2016. Applied Hierarchical Modeling in Ecology. Academic Press.

MacKenzie, D. I.; Nichols, J. D.; Lachman, G. B.; Droege, S.; Royle, J. A.; Langtimm, C. A., 2002. Estimating site occupancy rates when detection probabilities are less than one. - Ecology 83: 2248-2255.

Mason, N. W. H.; Mouillot, D.; Lee, W. G.; Wilson, J. B., 2005. Functional richness, functional evenness and functional divergence: the primary components of functional diversity. - Oikos 111: 112-118.

Mihaljevic, J. R.; Joseph, M. B.; Johnson, P. T. J., 2015. Using multispecies occupancy models to improve the characterization and understanding of metacommunity structure. - Ecology 96: 1783-1792.

van der Plas, F.; van Klink, R.; Manning, P.; Olff, H.; Fischer, M., 2017. Sensitivity of functional diversity metrics to sampling intensity. - Methods in Ecology and Evolution.