Details on posterior predictive checks in spAbundance

Jeffrey W. Doser

2023

Contents

1 Introduction

This vignette provides complete details on the calculation of posterior predictive checks in `spAbundance` with the `ppcAbund()` function. Here we discuss only the underlying statistical details and calculations used in the posterior predictive checks and do not discuss how to implement the posterior predictive checks using `ppcAbund()`. Examples of how to use `ppcAbund()` are provided in the additional model fitting vignettes on the package website. Note that here we present this in the context of single-species models. Posterior predictive checks are identical for multi-species models, with all variables now indexed by species.

2 Hierarchical distance sampling models

Let $l = 1, \ldots L$ denote the L MCMC samples obtained from the model fit (after discarding any samples for burn-in and thinning). The first step in performing a posterior predictive check is to generate a set of replicate data values from the posterior predictive distribution of the data. Let $\boldsymbol{y}_j$ denote the vector of count data at site j for each of the K distance bands. Next, let $\boldsymbol{y}_{\text{rep},j}^{(l)}$ denote the set of model generated/replicated counts at site j for each all K distance bins for MCMC sample l. For hierarchical distance sampling models, we calculate replicate data values according to

$$
\begin{aligned}
N^{(l)}_{\text{rep},j} &\sim \text{Poisson}(\mu^{(l)}_j) \\
\boldsymbol{y}^{*,(l)}_{\text{rep},j} &\sim \text{Multinomial}(N^{(l)}_{\text{rep},j}, \boldsymbol{\pi}^{*,(l)}),
\end{aligned} \tag{1}
$$

where $\boldsymbol{y}^{(l)}_{\text{rep},j}$ is then the $K \times 1$ vector of the first K values of $\boldsymbol{y}^{*,(l)}_{\text{rep},j}$.

`spAbundance` provides four different types of posterior predictive checks for HDS models, which differ based on two components (the `fit.stat` and `group` arguments in `ppcAbund()`). First, we can use either a Freeman-Tukey test statistic (`fit.stat = 'freeman-tukey'`) or a Chi-squared test statistic (`fit.stat = 'chi-squared'`). Second, we can calculate the fit statistic using either the raw counts for each site and distance bin (`group = 0`), or can first sum all the values at a given site across the K distance bins to generate a single value at each site, and then calculate the test statistic using that value (`group = 1`).

2.1 `fit.stat = 'freeman-tukey'` and `group = 0`

Let $T^{(l)}_{j,k}$ denote the test statistic calculated for the true data and $T^{(l)}_{\text{rep},j,k}$ the test statistic calculated for the replicate data at each MCMC sample l at site j and distance bin k. Here we have

$$
\begin{aligned}
T^{(l)}_{j,k} &= \left(\sqrt{y_{j,k}} - \sqrt{\pi^{(l)}_{j,k} \cdot \mu^{(l)}_j}\right)^2 \\
T^{(l)}_{\text{rep},j,k} &= \left(\sqrt{y^{(l)}_{\text{rep},j,k}} - \sqrt{\pi^{(l)}_{j,k} \cdot \mu^{(l)}_j}\right)^2.
\end{aligned} \tag{2}
$$

The test statistics above can provide information on what locations and/or distance bins are showing inadequate model fit. Posterior quantiles for these values are available in the `fit.y.group.quants` and `fit.y.rep.group.quants` components of the resulting list that comes from `ppcAbund()`. For an overall GoF statistic across all sites and distance bands, we calculate

$$
\begin{aligned}
T^{(l)}_{\cdot} &= \sum_{j=1}^{J}\sum_{k=1}^{K} T^{(l)}_{j,k} \\
T^{(l)}_{\text{rep},\cdot} &= \sum_{j=1}^{J}\sum_{k=1}^{K} T^{(l)}_{\text{rep},j,k}
\end{aligned} \tag{3}
$$

Posterior quantiles (2.5, 25, 50, 75, and 97.5) of $T^{(l)}_{j,k}$ (`fit.y.group.quants`) and $T^{(l)}_{\text{rep},j,k}$ (`fit.y.rep.group.quants`) are included in the resulting model object when calling `ppcAbund()`, and thus can be visualized to understand where the model is fitting well and where it is not. The full sets of MCMC samples is returned for $T^{(l)}_{\cdot}$ (`fit.y`) and $T^{(l)}_{\text{rep},\cdot}$ (`fit.y.rep`) are included in the resulting object from `ppcAbund()`, which are then used to calculate a Bayesian p-value according to

$$
\text{Bayesian p-value} = \frac{\sum_{l=1}^{L} I(T^{(l)}_{\text{rep},\cdot} > T^{(l)}_{\cdot})}{L}, \tag{4}
$$

where $I(\cdot)$ is the indicator function. Thus, the Bayesian p-value is the proportion of the total L samples in which the overall fit statistic calculated using the replicate data is greater than the value calculated using the observed data. Values around 0.5 indicate adequate model fit, while values close to the extremes of 1 and 0 indicate inadequate model fit.

2.2 `fit.stat = 'chi-squared'` and `group = 0`

Values are defined analogously to the previous section with the only difference being the form of the test statistic

$$\begin{aligned} T_{j,k}^{(l)} &= \frac{\left(y_{j,k} - \pi_{j,k}^{(l)} \cdot \mu_j^{(l)}\right)^2}{\pi_{j,k}^{(l)} \cdot \mu_j^{(l)} + c} \\ T_{\text{rep},j,k}^{(l)} &= \frac{\left(y_{\text{rep},j,k}^{(l)} - \pi_{j,k}^{(l)} \cdot \mu_j^{(l)}\right)^2}{\pi_{j,k}^{(l)} \cdot \mu_j^{(l)} + c} \end{aligned} \tag{5}$$

where $c = 0.0001$ is a very small constant to avoid dividing by 0 when expected counts are very close to 0. As before, we then have

$$\begin{aligned} T_{\cdot}^{(l)} &= \sum_{j=1}^{J} \sum_{k=1}^{K} T_{j,k}^{(l)} \\ T_{\text{rep},\cdot}^{(l)} &= \sum_{j=1}^{J} \sum_{k=1}^{K} T_{\text{rep},j,k}^{(l)} \end{aligned} \tag{6}$$

and

$$\text{Bayesian p-value} = \frac{\sum_{l=1}^{L} I(T_{\text{rep},\cdot}^{(l)} > T_{\cdot}^{(l)})}{L}, \tag{7}$$

2.3 `fit.stat = 'freeman-tukey'` and `group = 1`

When setting `group = 1`, a posterior predictive check is calculated using the total number of observations across all K distance bands at a given site, instead of using the individual count value in distance band k at each site j. Define $y_{j,\cdot}$ and $y_{\text{rep},j,\cdot}^{(l)}$ as the total number of individuals at site j for the observed data and the replicate data during iteration l, respectively. More specifically,

$$\begin{aligned} y_{j,\cdot} &= \sum_{k=1}^{K} y_{j,k} \\ y_{\text{rep},j,\cdot}^{(l)} &= \sum_{k=1}^{K} y_{\text{rep},j,k}^{(l)}. \end{aligned} \tag{8}$$

Our posterior predictive check then proceeds according to

$$
\begin{aligned}
T_j^{(l)} &= \left(\sqrt{y_{j,\cdot}} - \sqrt{\sum_{k=1}^{K}(\pi_{j,k}^{(l)} \cdot \mu_j^{(l)})}\right)^2 \\
T_{\text{rep},j}^{(l)} &= \left(\sqrt{y_{\text{rep},j,\cdot}^{(l)}} - \sqrt{\sum_{k=1}^{K}(\pi_{j,k}^{(l)} \cdot \mu_j^{(l)})}\right)^2. \\
T_{\cdot}^{(l)} &= \sum_{j=1}^{J} T_j^{(l)} \\
T_{\text{rep},\cdot}^{(l)} &= \sum_{j=1}^{J} T_{\text{rep},j}^{(l)} \\
\text{Bayesian p-value} &= \frac{\sum_{l=1}^{L} I(T_{\text{rep},\cdot}^{(l)} > T_{\cdot}^{(l)})}{L}.
\end{aligned}
\tag{9}
$$

2.4 `fit.stat = 'chi-squared'` and `group = 1`

Our final posterior predictive check is analogous to the previous section, except we now use a chi-squared test statistic.

$$
\begin{aligned}
T_j^{(l)} &= \frac{\left(y_{j,\cdot} - \sum_{k=1}^{K}(\pi_{j,k}^{(l)} \cdot \mu_j^{(l)})\right)^2}{\sum_{k=1}^{K}(\pi_{j,k}^{(l)} \cdot \mu_j^{(l)}) + c} \\
T_{\text{rep},j}^{(l)} &= \frac{\left(y_{\text{rep},j,\cdot}^{(l)} - \sum_{k=1}^{K}(\pi_{j,k}^{(l)} \cdot \mu_j^{(l)})\right)^2}{\sum_{k=1}^{K}(\pi_{j,k}^{(l)} \cdot \mu_j^{(l)}) + c} \\
T_{\cdot}^{(l)} &= \sum_{j=1}^{J} T_j^{(l)} \\
T_{\text{rep},\cdot}^{(l)} &= \sum_{j=1}^{J} T_{\text{rep},j}^{(l)} \\
\text{Bayesian p-value} &= \frac{\sum_{l=1}^{L} I(T_{\text{rep},\cdot}^{(l)} > T_{\cdot}^{(l)})}{L}.
\end{aligned}
\tag{10}
$$

3 N-mixture models

Posterior predictive checks proceed analogously for N-mixture models. Let $y_{j,k}$ denote the observed number of individuals at site j during replicate visit k. We will use the same notation as detailed previously. For N-mixture models, the argument `type` controls how the replicate data are calculated. The argument `type` takes two values. For what we call conditional replicate values (`type = 'conditional'`), we have

$$
y_{\text{rep},j,k}^{(l)} \sim \text{Binomial}(N_j^{(l)}, p_{j,k}^{(l)}), \tag{11}
$$

where $N_j^{(l)}$ denotes the posterior distribution of latent abundance values that is estimated directly when fitting the model. The second approach calculates "marginal" replicate values (`type = 'marginal'`), where we first generate replicate data by predicting a value of latent abundance at site j using the expected abundance at site j at MCMC sample l (i.e., $\mu_j^{(l)}$) and then subsequently generating a replicate data point for each replicate k at site j. More specifically, we have

$$
\begin{aligned}
N^{(l)}_{\text{rep},j} &\sim \text{Poisson}(\mu^{(l)}_j), \\
y^{(l)}_{\text{rep},j,k} &\sim \text{Binomial}(N^{(l)}_{\text{rep},j}, p^{(l)}_{j,k}).
\end{aligned} \tag{12}
$$

See the section on posterior predictive distributions in the N-mixture modeling vignette for more details.

Once the replicate data values are generated, the posterior predictive checks and calculation of Bayesian p-values is essentially identical to HDS. Below we give the equations used to perform each type of posterior predictive check that is available for N-mixture models. See the previous section on posterior predictive checks in HDS models for full definitions of all variables.

3.1 `fit.stat = 'freeman-tukey'` and `group = 0`

$$
\begin{aligned}
T^{(l)}_{j,k} &= \left(\sqrt{y_{j,k}} - \sqrt{p^{(l)}_{j,k} \cdot \mu^{(l)}_j}\right)^2 \\
T^{(l)}_{\text{rep},j,k} &= \left(\sqrt{y^{(l)}_{\text{rep},j,k}} - \sqrt{p^{(l)}_{j,k} \cdot \mu^{(l)}_j}\right)^2. \\
T^{(l)}_{\cdot} &= \sum_{j=1}^{J} \sum_{k=1}^{K_j} T^{(l)}_{j,k} \\
T^{(l)}_{\text{rep},\cdot} &= \sum_{j=1}^{J} \sum_{k=1}^{K_j} T^{(l)}_{\text{rep},j,k} \\
\text{Bayesian p-value} &= \frac{\sum_{l=1}^{L} I(T^{(l)}_{\text{rep},\cdot} > T^{(l)}_{\cdot})}{L},
\end{aligned} \tag{13}
$$

3.2 `fit.stat = 'chi-squared'` and `group = 0`

$$
\begin{aligned}
T^{(l)}_{j,k} &= \frac{\left(y_{j,k} - p^{(l)}_{j,k} \cdot \mu^{(l)}_j\right)^2}{p^{(l)}_{j,k} \cdot \mu^{(l)}_j + c} \\
T^{(l)}_{\text{rep},j,k} &= \frac{\left(y^{(l)}_{\text{rep},j,k} - p^{(l)}_{j,k} \cdot \mu^{(l)}_j\right)^2}{p^{(l)}_{j,k} \cdot \mu^{(l)}_j + c} \\
T^{(l)}_{\cdot} &= \sum_{j=1}^{J} \sum_{k=1}^{K} T^{(l)}_{j,k} \\
T^{(l)}_{\text{rep},\cdot} &= \sum_{j=1}^{J} \sum_{k=1}^{K} T^{(l)}_{\text{rep},j,k} \\
\text{Bayesian p-value} &= \frac{\sum_{l=1}^{L} I(T^{(l)}_{\text{rep},\cdot} > T^{(l)}_{\cdot})}{L},
\end{aligned} \tag{14}
$$

3.3 `fit.stat = 'freeman-tukey'` and `group = 1`

$$
\begin{aligned}
y_{j,\cdot} &= \sum_{k=1}^{K_j} y_{j,k} \\
y^{(l)}_{\text{rep},j,\cdot} &= \sum_{k=1}^{K_j} y^{(l)}_{\text{rep},j,k}. \\
T^{(l)}_j &= \left(\sqrt{y_{j,\cdot}} - \sqrt{\sum_{k=1}^{K_j}(p^{(l)}_{j,k} \cdot \mu^{(l)}_j)}\right)^2 \\
T^{(l)}_{\text{rep},j} &= \left(\sqrt{y^{(l)}_{\text{rep},j,\cdot}} - \sqrt{\sum_{k=1}^{K_j}(p^{(l)}_{j,k} \cdot \mu^{(l)}_j)}\right)^2. \\
T^{(l)}_{\cdot} &= \sum_{j=1}^{J} T^{(l)}_j \\
T^{(l)}_{\text{rep},\cdot} &= \sum_{j=1}^{J} T^{(l)}_{\text{rep},j} \\
\text{Bayesian p-value} &= \frac{\sum_{l=1}^{L} I(T^{(l)}_{\text{rep},\cdot} > T^{(l)}_{\cdot})}{L}.
\end{aligned}
\tag{15}
$$

3.4 `fit.stat = 'chi-squared'` and `group = 1`

$$
\begin{aligned}
T^{(l)}_j &= \frac{\left(y_{j,\cdot} - \sum_{k=1}^{K_j}(p^{(l)}_{j,k} \cdot \mu^{(l)}_j)\right)^2}{\sum_{k=1}^{K_j}(p^{(l)}_{j,k} \cdot \mu^{(l)}_j) + c} \\
T^{(l)}_{\text{rep},j} &= \frac{\left(y^{(l)}_{\text{rep},j,\cdot} - \sum_{k=1}^{K_j}(p^{(l)}_{j,k} \cdot \mu^{(l)}_j)\right)^2}{\sum_{k=1}^{K_j}(p^{(l)}_{j,k} \cdot \mu^{(l)}_j) + c} \\
T^{(l)}_{\cdot} &= \sum_{j=1}^{J} T^{(l)}_j \\
T^{(l)}_{\text{rep},\cdot} &= \sum_{j=1}^{J} T^{(l)}_{\text{rep},j} \\
\text{Bayesian p-value} &= \frac{\sum_{l=1}^{L} I(T^{(l)}_{\text{rep},\cdot} > T^{(l)}_{\cdot})}{L}.
\end{aligned}
\tag{16}
$$

3.5 `fit.stat = 'chi-squared'` and `group = 2`

For N-mixture models, `spAbundance` also has functionality for grouping the data by replicate prior to performing the posterior predictive check. When setting `group = 2`, a posterior predictive check is calculated using the total number of observations across all J sites for each replicate k, instead of using the individual count value in replicate k at each site j. Define $y_{\cdot,k}$ and $y^{(l)}_{\text{rep},\cdot,k}$ as the total number of individuals detected during visit k across all J sites for the observed data and the replicate data during iteration l, respectively. More specifically,

$$
\begin{aligned}
y_{\cdot,k} &= \sum_{j=1}^{J} y_{j,k} \\
y^{(l)}_{\text{rep},\cdot,k} &= \sum_{j=1}^{J} y^{(l)}_{\text{rep},j,k}.
\end{aligned}
\tag{17}
$$

Our posterior predictive check and calculation of the Bayesian p-values proceeds as follows, where $K_{\max}$ is the maximum number of replicates at any given site.

$$
\begin{aligned}
T^{(l)}_{k} &= \left(\sqrt{y_{\cdot,k}} - \sqrt{\sum_{j=1}^{J} (p^{(l)}_{j,k} \cdot \mu^{(l)}_{j})} \right)^2 \\
T^{(l)}_{\text{rep},k} &= \left(\sqrt{y^{(l)}_{\text{rep},\cdot,k}} - \sqrt{\sum_{j=1}^{J} (p^{(l)}_{j,k} \cdot \mu^{(l)}_{j})} \right)^2. \\
T^{(l)}_{\cdot} &= \sum_{k=1}^{K_{\max}} T^{(l)}_{k} \\
T^{(l)}_{\text{rep},\cdot} &= \sum_{k=1}^{K_{\max}} T^{(l)}_{\text{rep},k} \\
\text{Bayesian p-value} &= \frac{\sum_{l=1}^{L} I(T^{(l)}_{\text{rep},\cdot} > T^{(l)}_{\cdot})}{L}.
\end{aligned}
\tag{18}
$$

3.6 `fit.stat = 'freeman-tukey'` and `group = 2`

$$
\begin{aligned}
T^{(l)}_{k} &= \frac{\left(y_{\cdot,k} - \sum_{j=1}^{J} (\pi^{(l)}_{j,k} \cdot \mu^{(l)}_{j}) \right)^2}{\sum_{j=1}^{J} (\pi^{(l)}_{j,k} \cdot \mu^{(l)}_{j}) + c} \\
T^{(l)}_{\text{rep},k} &= \frac{\left(y^{(l)}_{\text{rep},\cdot,k} - \sum_{j=1}^{J} (\pi^{(l)}_{j,k} \cdot \mu^{(l)}_{j}) \right)^2}{\sum_{j=1}^{J} (\pi^{(l)}_{j,k} \cdot \mu^{(l)}_{j}) + c} \\
T^{(l)}_{\cdot} &= \sum_{k=1}^{K_{\max}} T^{(l)}_{k} \\
T^{(l)}_{\text{rep},\cdot} &= \sum_{k=1}^{K_{\max}} T^{(l)}_{\text{rep},k} \\
\text{Bayesian p-value} &= \frac{\sum_{l=1}^{L} I(T^{(l)}_{\text{rep},\cdot} > T^{(l)}_{\cdot})}{L}.
\end{aligned}
\tag{19}
$$

4 Generalized linear mixed models

Posterior predictive checks proceed analogously for GLMMs. Let y_j denote the observed number of individuals at site j. `spAbundance` supports two types of posterior predictive checks for GLMMs.

4.1 `fit.stat = 'freeman-tukey'` and `group = 0`

$$
\begin{aligned}
T_j^{(l)} &= \left(\sqrt{y_j} - \sqrt{\mu_j^{(l)}}\right)^2 \\
T_{\text{rep},j}^{(l)} &= \left(\sqrt{y_{\text{rep},j}^{(l)}} - \sqrt{\mu_j^{(l)}}\right)^2. \\
T_\cdot^{(l)} &= \sum_{j=1}^{J} T_j^{(l)} \\
T_{\text{rep},\cdot}^{(l)} &= \sum_{j=1}^{J} T_{\text{rep},j}^{(l)} \\
\text{Bayesian p-value} &= \frac{\sum_{l=1}^{L} I(T_{\text{rep},\cdot}^{(l)} > T_\cdot^{(l)})}{L},
\end{aligned}
\tag{20}
$$

4.2 `fit.stat = 'chi-squared'` and `group = 0`

$$
\begin{aligned}
T_j^{(l)} &= \frac{\left(y_j - \mu_j^{(l)}\right)^2}{\mu_j^{(l)} + c} \\
T_{\text{rep},j}^{(l)} &= \frac{\left(y_{\text{rep},j}^{(l)} - \mu_j^{(l)}\right)^2}{\mu_j^{(l)} + c} \\
T_\cdot^{(l)} &= \sum_{j=1}^{J} T_j^{(l)} \\
T_{\text{rep},\cdot}^{(l)} &= \sum_{j=1}^{J} T_{\text{rep},j}^{(l)} \\
\text{Bayesian p-value} &= \frac{\sum_{l=1}^{L} I(T_{\text{rep},\cdot}^{(l)} > T_\cdot^{(l)})}{L},
\end{aligned}
\tag{21}
$$

References