Poisson model with adjusted confidence intervals for overdispersed plankton counts

November 20, 2016

1 choice of model

Goal: we'd like a way of constructing a seasonal climatology of plankton concentration, but without worrying about how interannual variation may distort or obscure the underlying seasonal pattern (i.e. cell abundance for a given season is really high one year, but zero the next year, such that averaging really doesn't do a good job). This can be accomplished using a model that accounts for both a season effect and year affect. Normally, the Poisson distribution does a job of representing count data. However, plankton counts at MVCO tend to be overdispersed relative to the Poisson distribution.

Andy had mentioned that there are a few options to account for overdispersion (which may or may not simultaneously account for the over-occurrence of zeros). One option is a negative binomial model, of which the Poission distribution is a special case. This model would incorporate a seasonal and yearly effect, but also an overdispersion estimate for each season and year. While straightforward to design, it results in a large number of parameters that need to be estimated simultaneously, for example 26 seasons + 10year effects + $10 \cdot 26$ overdispersion effects. There could be work arounds to fitting ~ 300 parameters, but these would involve assumptions of some type of pattern to the overdispersion (i.e., constant, seasonal or scales with plankton abundance), which we could not find much evidence for in the data.

Another concern would be the robustness of the estimate for each overdispersion parameter. Some season/year combinations have very few observations, making one skeptical of any fit that solely relies on a few observations. In lieu of this approach, Andy recommended that we simply fit a Poisson model that incorporates a seasonal and year affect, and account for the overdispersion with corrected confidence intervals. As the Poisson model can account for observations of zero, this approach indirectly handles these observations. Below describes this model and subsequent confidence interval inflation as well as the matlab scripts for the fitting procedure.

2 the count model

We first organize our data by year (denoted by i) and by two-week 'seasons' (denoted by j). For a given plankton, let y_{ijk} be replicate count k in year i and season j. Likewise, we have corresponding volumes for each replicate count, denoted by v_{ijk}. Let Y_{ij} be the sum of all counts: $Y_{ij} = \sum_k y_{ijk}$ and likewise, $V_{ij} = \sum_k v_{ijk}$ be the sum of all volumes over the replicates in a season and year. Let's then assume that the counts vary according to a Poisson model:

$$Y_{ij} \sim \text{Poisson}(\mu_{ij} \cdot V_{ij}),$$

where μ_{ij} is the density of cells per volume for season j and year i. We can then parameterize this density to have both a seasonal component and a yearly component:

$$\mu_{ij} = \exp(\beta_i + \gamma_j).$$

The exponential is there to ensure that the mean is positive, as required for the Poisson distribution. For the 2007-2016 data, this will give us 10 yearly effects (β's) and 26 seasonal effects (γ's). To fit this model with the data, we use a maximum likelihood approach. The Poisson probability mass function (pmf) is:

$$P(\text{m events in an interval}) = \frac{\lambda^m e^{-\lambda}}{m!}$$

where λ is the mean number of events/counts in an interval and m is the number of events/counts for that interval. In our case, for a given season/year, the pmf becomes:

$$P(Y_{ij}) = \frac{(\mu_{ij} \cdot V_{ij})^{Y_{ij}} e^{-(\mu_{ij} \cdot V_{ij})}}{Y_{ij}!} = \frac{(\exp(\beta_i + \gamma_j) \cdot V_{ij})^{Y_{ij}} \exp(-(\exp(\beta_i + \gamma_j) \cdot V_{ij}))}{Y_{ij}!}$$

We then calculate the log likelihood function as:

$$L(\beta, \gamma) = \prod_{i=1}^{10} \prod_{j=1}^{26} P(Y_{ij})$$

$$\log L(\beta, \gamma) = \sum_{i=1}^{10} \sum_{j=1}^{26} \log P(Y_{ij})$$

$$\log L(\beta, \gamma) = \sum_{i=1}^{10} \sum_{j=1}^{26} [Y_{ij}(\beta_i + \gamma_j + \log V_{ij}) - \exp(\beta_i + \gamma_j)V_{ij}]$$

We search over 36 parameters to find the values that maximize this function, given the data. However, there is an additional constraint that must be imposed to be able to find a distinct solution. Because the mean density is an additive combination between

two sets of parameters, it is the sum or difference between them that will determine the likelihood value. Because of this, you could have several different sets of parameters that would equally generate the same "difference". For example, $\beta_1 = 2$ and $\gamma_1 = -1$ gives $\mu_{1,1} = \exp(1)$, but so does $(\beta_1 = 4, \gamma_1 = -3)$ or $(\beta_1 = -4, \gamma_1 = 5)$. To avoid this problem, we set an 'identifiability' constraint. This can take a few forms, such as $\beta_1 = -\sum_{i=2}^{10} \beta_i$, or $\beta_1 = 0$. The maximum likelihood estimates of parameters are therefore interpreted relative to this constriant (i.e. parameters are relative to a mean year effect or relative to the first year). We will set $\beta_1 = -\sum_{i=2}^{10} \beta_i$ as this is a bit easier to interpret (more on that later).

With maximum likelihood estimators in hand, we can then construct confidence intervals for each parameter. We do this with use of the standard likelihood ratio test with profile log likelihoods for each parameter:

$$2\left[\log L(\hat{\beta}, \hat{\gamma}) - \log_{\mathrm{prof}}(\beta_i \text{ or } \gamma_j)\right] \sim \chi_1^2,$$

where $(\hat{\beta}, \hat{\gamma})$ are the maximum likelihood estimates of the parameters, $\log_{\mathrm{prof}}$ refers to the profile log likelihood of a given parameter (β_i or γ_j; constructed by holding the parameter at a certain value and resolving for the rest of the parameters that maximize the likelihood function for this value), and χ_1^2 is the chi-squared distribution with 1 degree of freedom. We then search over the range of a given parameter to find the values that satisfy this inequality at significance level $\alpha = 0.05$ (see Figure 1). Note that profile likelihoods are also constructed with the identifiability constraint of $\beta_1 = 0$.
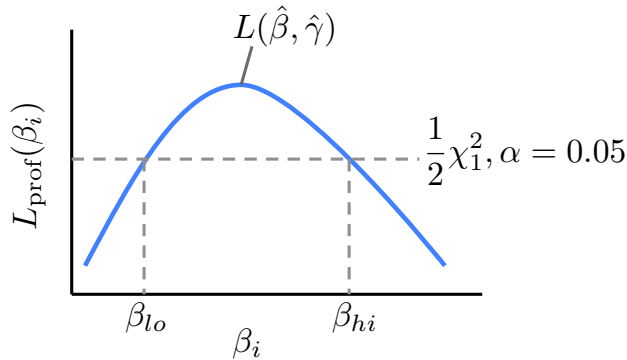

Figure 1: Schematic of profile likelihood and relationship to confidence intervals.

If the counts are over-dispersed, we inflate the confidence intervals with a calculated metric for overdispersion. To do this, we first form the Pearson statistic (also denoted by χ^2):

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(y_{ij} - \hat{y}_{ij})^2}{\hat{y}_{ij}},$$

where $\hat{y}_{ij} = \exp(\hat{\beta}_i + \hat{\gamma}_j)V_{ij}$, and I is the total number of years and J is the total number of seasons for which there is at least one non-zero count. We leave out seasons or years that have all zero counts as these will not contribute to calculating an overdispersion metric (and we cannot actually evaluate the Pearson statistic in these cases (the denominator will be 0)). When then define a quantity c as a multiplier for the χ_1^2 $\alpha = 0.05$ significance value:

$$c = \max \left[1, \frac{\chi^2}{(I \cdot J - (I + J))} \right]$$

The corrected 0.95 confidence interval for say β_1, is then given by the values of β_1 that satisfy:

$$2 \left[\log L(\hat{\beta}, \hat{\gamma}) - \log L_{\mathrm{prof}}(\beta_1)) \right] < \chi_1^2 \cdot c,$$

where the $\log L_{\mathrm{prof}}$ is the profile log likelihood based on the Poisson assumption. Note that this approach gives the option of not having over dispersed data if the maximum value between 1 or $\frac{\chi^2}{(I \cdot J - (I+J))}$ is indeed 1. For the data that has all zero counts for a given season or year, only an upper bound of the confidence interval can be found, where the lower bound will try to go to $-\infty$ (the seasonal or yearly parameter will numerically try to get the quantity $\exp(\beta_i + \gamma_j)$ as close to zero as possible to yield a mean density of 0, but as only $\lim_{x \to -\infty} \exp(x) = 0$, a true lower bound does not exist). We also are not able to construct a confidence interval for β_1 as we have used this for our reference for other parameters.

3 matlab scripts

The matlab scripts to fit this model and construct confidence intervals are listed below along with short descriptions. These are called from the main wrapper script, **opt_max_logL.m**, which in addition to the functions below, also sets up the data in expected format and generates some figures to compare data to fitted model. Nested lists indicate script dependencies.

- maximum likelihood estimate of parameters:

  poisson_logL.m function called by optimization routine; calculates a negative log likelihood value given a set of parameters and count data

- confidence interval construction:

  poisson_est_CIs.m script that sets up root solver to find parameter values that satisfy likelihood ratio test inequality with inflated value based on overdispersion metric

  eval_param_for_CIs.m function called by root solver; with specified parameter value calculates a profile log likelihood by resolving for parameter values that minimize negative log likelihood

  poisson_logL_prof.m function called by optimization to find parameter values that minimize negative profile log likelihood

4   interpretation of seasonal and year effects

One of the main goals of this approach is to be able to produce a seasonal climatology without the influence of interannual variation. If we fit with the model with the identifiability constraint of $\beta_1 = -\sum_{i=2}^{10} \beta_i$ (or another year of choice), a climatology is retrieved from the exponential of the fitted seasonal parameters ($\exp(\gamma_j)$). Because the fitted year parameters are fit such that they average zero, the value of the fitted seasonal parameters are relative to a 'zero' year affect, exactly what we would like for the climatology. For example, with simulated dat, we see that the estimated year effects are equally distributed around zero (as required by the identifiability constraint) and shifted by some amount from the true year parameters used to simulated the data. The seasonal parameters are also shifted by this same amount, only in the opposite direction.
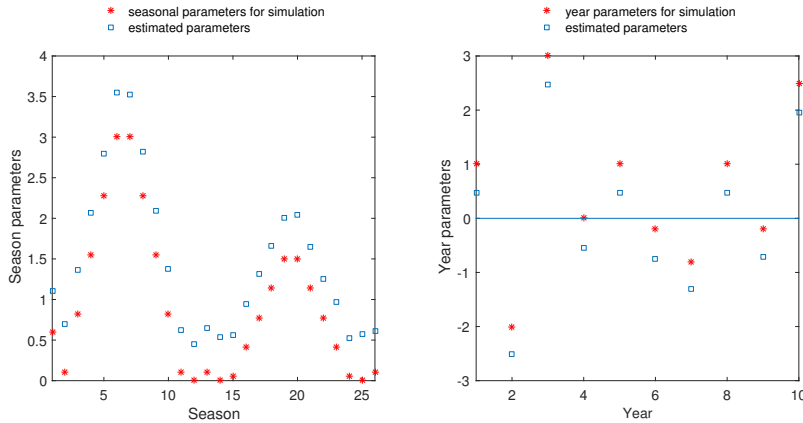

Figure 2: True and estimated parameters of simulated data. A. Seasonal parameters. B. Year parameters.

5

We can write this as:

$$\beta_i = \hat{\beta}_i - \bar{\beta} \tag{1}$$

$$\gamma_j = \hat{\gamma}_j + \bar{\beta} \tag{2}$$

where β_i and γ_j are the true parameters, $\hat{\beta}_i$ and $\hat{\gamma}_j$ are the parameter estimates fitted with the identifiability constraint, and $\bar{\beta}$ is the mean of the true year parameters. These relationships also show that it is only the absolute distance between the year and seasonal parameters that matter:

$$\beta_i + \gamma_j = \hat{\beta}_i - \bar{\beta} + \hat{\gamma}_j + \bar{\beta} \tag{3}$$

$$\beta_i + \gamma_j = \hat{\beta}_i + \hat{\gamma}_j \tag{4}$$

4.1   units

The parameters are linear in the log space, such that both β_i and γ_j should be in units of $\log(\text{cells mL}^{-1})$. This ends up being a rather tricky interpretation if the exponential is re-arragned:

$$\mu_{ij} = \exp(\beta_i + \gamma_j) = \exp(\log(\frac{\text{cells}}{\text{mL}}) + \log(\frac{\text{cells}}{\text{mL}})) \tag{5}$$

$$\mu_{ij} = \exp(\beta_i)\exp(\gamma_j) = \exp(\log(\frac{\text{cells}}{\text{mL}}))\exp(\log(\frac{\text{cells}}{\text{mL}})) \tag{6}$$

$$\mu_{ij} = \left(\frac{\text{cells}}{\text{mL}}\right)^2 \;\; ??? \tag{7}$$

Alternatively, one could just simply do the addition before taking the exponential. However, the relationship in Eqn. 4 can help with an alternative unit interpretation, where the units are attached to the fitted seasonal parameters and the year parameters are interpreted as 'scale' factors. For simply an estimated seasonal climatology we have:

$$\mu_j = \exp(\hat{\gamma}_j) \tag{8}$$

$$\mu_{ij} = \exp\left(\gamma_j - \bar{\beta}\right) \tag{9}$$

$$\mu_{ij} = \exp(\gamma_j + \hat{\beta}_i - \beta_i) \tag{10}$$

$$\mu_{ij} = \frac{\exp(\hat{\beta}_i)}{\exp(\bar{\beta})}\exp(\hat{\gamma}_j + \bar{\beta}) = \frac{\text{cells/mL}}{\text{cells/mL}}\left(\frac{\text{cells}}{\text{mL}}\right) \tag{11}$$

Hence the s

The expected seasonal mean density effect for *Laboea strobila* is in Figure 1. The actual densities for each timepoint appear in grey and the expected seasonal effect $(\exp(\gamma_j))$

appears as a black line. To get expected densities for each year, this would be multiplied by $\exp(\beta_i)$. Notice that averaging aggregate seasonal densities across the years gives a different pattern to the one achieved with this method.
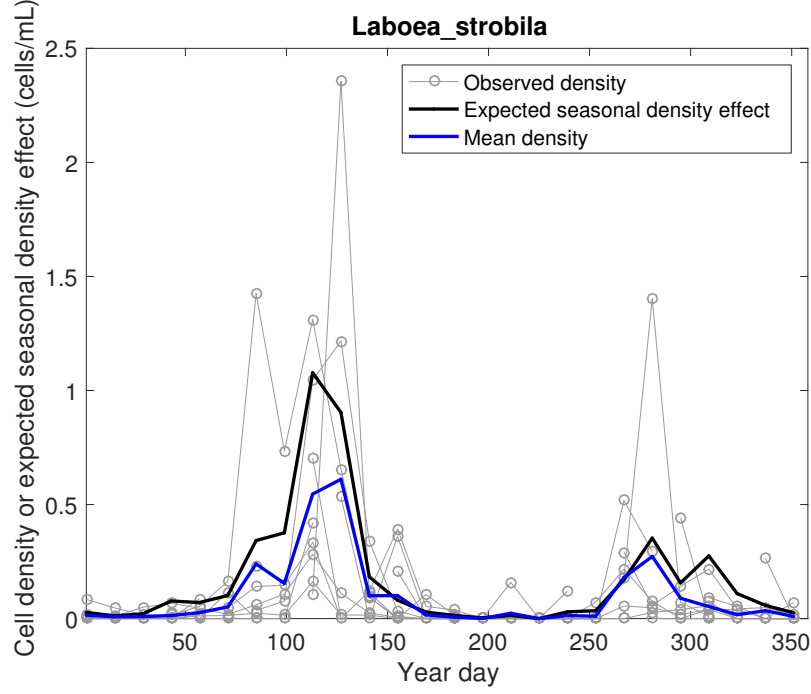

Figure 3: Cell densities of observed *Laboea strobila* from 2007-2016 (gray lines), expected seasonal effect on cell density from maximum likelihood estimates of seasonal parameters assuming Poisson model (black line), and mean cell density of season across all years (blue line).

One could also investigate the year affects and if these parameters show any patterns among the years or any relationship to environmental variables.

5   test for overdispersion

How did we find out that our counts were overdispersed in the first place? Andy provided two tests to investigate this: one a full significance test and the other metric based.

Significance test: According to Andy, under the Poisson assumption, the conditional distribution of the replicate counts within a season/year is multinomial, with the number of trials being the replicates k, and success probability p, where:

$$p_k = \frac{v_k}{\sum_{k=1}^{K} v_k},$$

depends on volume sampled for each replicate (K is total number of replicates for one season/year). The Poisson assumption can be assessed by testing the null hypothesis that these probabilities are as given, or agree with the observed data. To do this, we use a standard Chi-squared goodness of fit test between the observed and expected. The test statistic is:

$$\chi^2 = \sum_{k=1}^{K} \frac{(y_k - np_k)^2}{np_k},$$

where n is the total counts of plankton across all replicates, $n = \sum_{k=1}^{K} y_k$. Under the null hypothesis, this metric has an approximate χ^2 distribtuion with $K - 1$ degrees of freedom. This approximation won't work well if a lot of the quantities np_k are small. One could assess significance with a parametric bootstrap in that case.

Less formal approach: A less formal way to make an assessment of the Poisson assumption is based on the quantity:

$$\frac{\chi^2}{(K - 1)}$$

If this quantity is systematically or much greater than 1, then the data is likely overdispersed. This is especially useful for getting a feel for how overdispersion varies over the timeseries, although Andy cautioned not to get too excited by a few peaks. The matlab script that performs these calculations is **test_poisson_overdisp.m**