Bayesian Estimation of Cooccurrence *Affinity* with Dyadic Regression

Introduction

The analysis of patterns of cooccurrence between taxa is an important and active area of ecological research [8, 1, 19, 10, 20]. Mainali et al. [12] have recently shown that of the numerous ways of measuring cooccurrence relationships between species pairs (reviewed in [12]) the correct and unbiased method is to make use of Fisher's noncentral hypergeometric distribution, as suggested by Veech [17] and Griffith et al. [9]. Mainali et al. [12] derived a cooccurrence metric called *affinity* (or $\hat{\alpha}$) based on this distribution. This is a significant step forward in the analysis of cooccurrence relationships. There are still several challenges remaining however. For instance, when species pairs of very low or very high prevalence are analysed with this method they will often be assigned very high or low affinity scores, but with low confidence (high p-value and wide confidence interval). For downstream analysis then the researcher may (i) treat all data points as equal, (ii) remove data above some high p-value threshold or (iii) devise a scheme to weight data appropriately. (i) wastes information can yield misleading results. (ii) again wastes information and could severely bias results depending on the reasons for the differences in prevalence/p-values. If an appropriate, unbiased scheme for (iii) should be devised, then this would be a welcome development. However, adopting a Bayesian approach that builds upon the work of Mainali et al. [12] not only yields more accurate estimates of the *affinity* between species, but also naturally propagates uncertainty through the analysis, i.e., it accounts for the different levels of confidence we have out the cooccurrence relationships between different species pairs. In the present work I illustrate this point and provide the model description and code to use this Bayesian method in practice.

A significant advantage of this framework is that it allows cooccurrence data to analysed as a response data in a Bayesian general linear model (GLM). By That is, by supplying species occurrence data as the prevalences of individual species, the number of times a given species pair cooccur and the total number of sites considered, it is now no more complicated to construct a regression model (with a Fisher's noncentral hypergeometric likelihood function) than it would be to perform binomial regression with count data. Similarly, this method makes the best use of all available information and weighs data points appropriately. Thus, just as it is not appropriate to convert count data to proportions and conduct linear regression, it is no longer best practice to summarize cooccurrence data as point estimates for linear regression.

Results

Mainali et al. [12] show that the log of the odds ratio term in Fisher's noncentral hypergeometric distribution (a quantity they term α) can be used to appropriately describe the extent

to which two species will tend to cooccur more or less than would be expected based just on the prevalences of the species. They go on to propose that the maximum likelihood estimate of this parameter $\hat{\alpha}$ or affinity should be used as a pairwise coocurrence metric. However, Maximum likelihood estimates can yield values of positive or negative infinity. This causes difficulties for downstream analyses (one cannot sensibly do something as simple as calculating the mean of a set containing infinite values). Furthermore, we know *a priori* that and infinitely large or small affinity is not sensible for most cases of interest to ecologists. Bayesian analysis uses prior knowledge to avoid the estimation of physically or biologically implausible values. Mainali et al. [12] reassign these infinite estimates an absolute value of $\log(2N^2)$ (from an argument made based on the Jeffreys' prior for the beta distribution). While this figure comes from sound argument, there are at least two problems with this approach. Firstly, not all data are treated the same way, i.e., no regularisation is applied to finite affinity estimates, only to these extreme values. Secondly, the value $\log(2N^2)$ is only a function of N, not the species prevalences. Thus, it is not influenced by our actual state of knowledge about the species in question.

To make these ideas more concrete and show the practical implications, I simulated species pairs using the affinity model. For each pair there are $N = 30$ sites they can inhabit, species A has a prevalence mA and species B has prevalence mB. The number of sites at which they cooccurr k was drawn from a Fishers noncentral hypergeometric distribution

$$k \sim \text{fnchypg}(mA, N - mA, mB, e^{\alpha}),$$

with 10 draws per combination of mA and mB for each of 41 different values of α. Then, given the values for N, k, mA and mB I estimated α using two methods. Firstly, I used the original maximum likelihood estimate (MLE) of Mainali et al. [12]. Next I obtained maximum *a posteriori* (MAP) estimates with a Gaussian prior $\text{N}(0, 3)$ for α. Note that these are not strongly regularising priors as when exponentiated in the likelihood function a standard deviation of $3 \approx 20$ and two standard deviations $6 \approx 403$ which is a very large odds ratio for most applications. Priors used for analysing real data should be chosen after simulation, to demonstrate that they do not bias against feasible parameter values for the specific research. In order to compare the two methods, for each combination of mA and mB I calculated the root mean squared error (RMSE) for each inference method. Figure 1 shows that only when both mA and mB were equal to 15 was the RMSE approximately equivalent between MLE and MAP methds. Whenever one or both species had a high or low prevalence, and particularly as the absolute value of α became larger, the MLE method produced very poor estimates, and the extreme estmates were always the same $\log(2N^2) = 7.496$. By contrast, for the MAP values, the prior provides regularisation which can be overcome by increasing confidence in the data, which is a function of mA and mB. Thus, the models *best guess* when $mA = 15$, $mB = 5$ and $k = 5$ is higher than the equivalent situation when $mB = 1$ and $k = 1$. Neither of these methods is perfect however. When asked for one, a model will give you it's buest guess point estimate, but we can make better use of the data we have collected if we can utilise not only the point estimates but also our unceratinty around them.

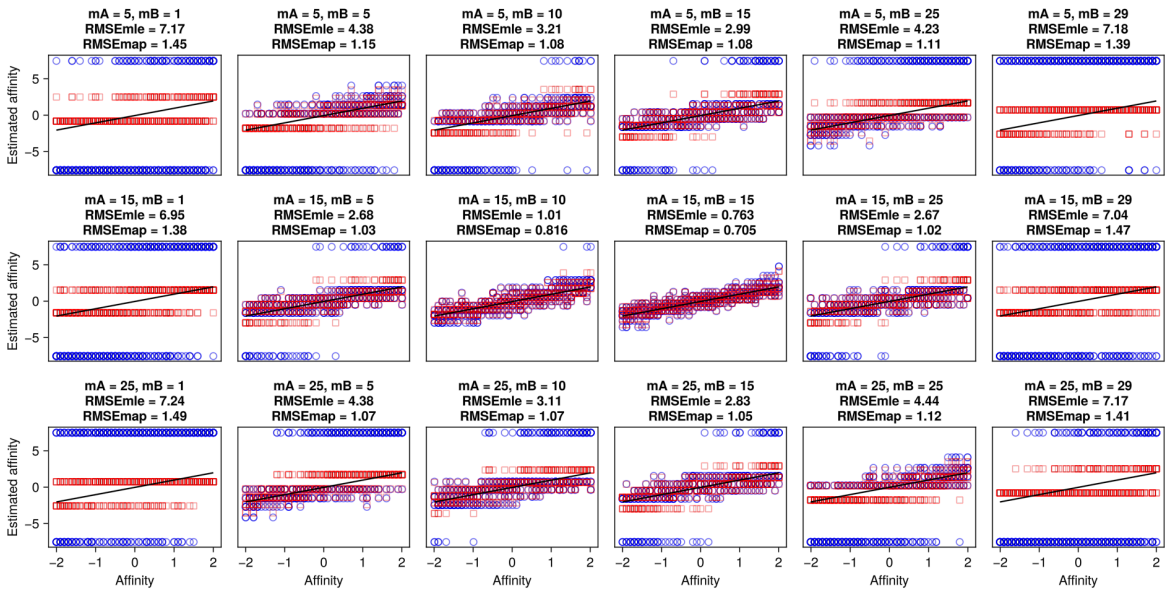

Figure 1: Actual and estimated affinity values for a range of species prevalences. Blue circles are estimates using the original maximum-likelihood method and red squares are maximum a posteriori estimates. Black lines indicate the actual affinity values used to generate the data. For each panel is shown the root mean squared error (RMSE) for both types of estimate.

Often we do not simply wish to report cooccurrence relationships, but to measure how they change with some other variables of interest. For many types of data there are well understood and regularly used probability distributions which can be used in Bayesian and frequentist GLMs. For cooccurrence data this has not been the case. Given the issues with deriving point estimates highlighted above (Figure 1) it seems unlikely that simply fitting a linear model to such point estimates of cooccurrence affinity will yield reliable results. Thus, rather than supplying a second inference model with uninformative point estimates from a previous model, we can provide our regression model with all the data on species prevalences and cooccurrences instead. To demonstrate the impact of this I simulated 41 sets of predictor data $\vec{x}$, each consisting of 30 draws from $\mathrm{N}(0, 1)$. Affinity values were obtained by multiplying the predictor data by a regression coefficient β. For each affinity value - species prevalences mA and mB were chosen randomly between 1 and 29 inclusive and a k value was drawn from the Fisher's noncentral hypergeometric distribution as above. For each generated data set pairwise affinity values were estimated by the MAP and MLE methods. Then linear regression analysis was conducted on these point estimates $\alpha \sim \beta\vec{x}$. Additionally I obtained a maximum likelihood estimate of a GLM of the form

$$\begin{aligned} \vec{k}_i &\sim \text{fnchypg}(\vec{mA}_i, N - \vec{mA}_i, \vec{mB}_i, e^{\alpha}) \\ \alpha_i &= \gamma + \beta \vec{x}_i, \end{aligned} \tag{1}$$

where $\vec{k}$, $\vec{mA}$ and $\vec{mB}$ are vectors containing the values of k, mA and mB respectively and γ is the intercept.

The results in Figure 2 show how poorly fitting a linear model to $\hat{alpha}$ point estimates does, typically overestimating the absolute value of β by a large margin. Using MAP estimates of α does better here, exhibiting the opposite behaviour of slightly underestimating the absolute value of β. However, by cutting out the step of generating point estimates for each pair the GLM retains all pertinent information and accurately recaptures the parameters of the data generating model. It is of course expected that the GLM should be able to discover the correct parameter values, since they were generated by an identical model. What is important is the way the other two models fail by comparison, and of course the fact that we now have the correct likelihood function for such a cooccurrence GLM.

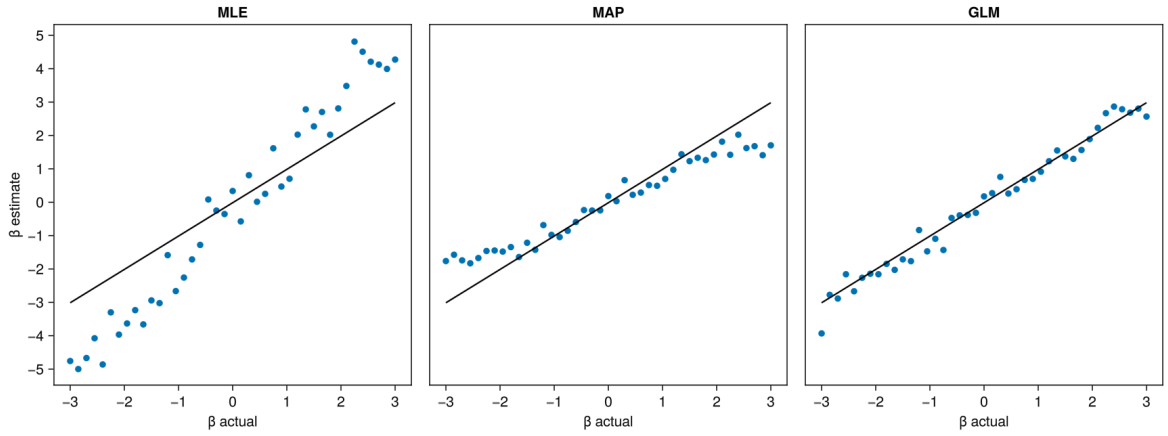

Figure 2: Estimated regression coefficients β according to three different methods: fit linear model to αMLE, fit linear model to αMAP, fit GLM to raw data. Black lines indicate the true β values.

In many if not most cases when working with cooccurrence data it will be in the form of a square cooccurrence matrix similar to the distance and dissimilarity matrices used to record e.g. phylogenetic distances between species or community dissimilarities between sample sites. As with these other types of matrices, if we wish to perform regression analysis treating each entry in the matrix as data point, we must account for non-independence of data coming from the same row or column, e.g. same site, species etc.. To deal with this we must simply include a random effect λ for each species [3, 7]. Thus, whereas the GLM above contains the term

$$\alpha_i = \gamma + \beta\vec{x}_i$$

we would now have

$$\alpha_i = \lambda_i + \lambda_j + \beta X_{ij}.$$

Where X is a (possibly dissimilarity or distance) matrix in which element X_{ij} is a quantity of interest relating species i to species j. Given a similarly arranged matrix K which holds the k values for all species pairs and a vector $\vec{m}$ containing species prevalences, our dyadic GLM becomes

$$\begin{aligned} K_{ij} &\sim \text{fnchypg}(\vec{m}_i, N - \vec{m}_i, \vec{m}_j, e^{\alpha}) \\ \alpha_{ij} &= \lambda_i + \lambda_j + \beta X_{ij} \end{aligned} \tag{2}$$

for each index ij in either the lower or upper triangle of K. While Maximum likelihood estimates of the λ and β parameters can be obtained, we still need a way to properly account for our uncertainty in our estimates. Bayesian inference provides an intuitive framework for this, grounded in probability theory. Thus, we can construct a Bayesian GLM by assigning priors to the unknown parameters. Assuming Gaussian priors for all parameters we have

$$\begin{aligned} K_{ij} &\sim \text{fnchypg}(\vec{m}_i, N - \vec{m}_i, \vec{m}_j, e^{\alpha}) \\ \alpha_{ij} &= \lambda_i + \lambda_j + \beta X_{ij} \\ \beta &\sim \text{N}(0, \beta_\sigma) \\ \lambda_i &\ \text{N}(0, \lambda_\sigma) \end{aligned} \tag{3}$$

where β_σ and λ_σ are to be chosen according to the specific of the system being analysed.

Discussion

Here, I have shown how to construct a Bayesian dyadic GLM for the analysis of cooccurrence data. This builds on the work of Mainali et al. [12] as well as Veech [17] and Griffith et al. [9]. The identification of Fisher's noncentral hypergeometric distribution (or mathematically equivalent formulations) as the correct distribution for modelling cooccurrence led first to null model approaches to cooccurrence analysis [9, 17], then to a useful cooccurrence metric [12] and now to general model capable of analysing raw cooccurrence data as a response variable even when data points are not independent (which will generally be the case).

Part of the motivation for this work was the failure to recapture known regression coefficients when fitting linear models to pairwise affinity estimates (Figure 2). There may be other ways

of combating this failure. For example, the removal of data points for which we have low confidence may be an option. For instance if $N = 30$, $mA = 29$ and $mB = 1$, then a k of 1 tells us very little, since our null expectation is that k will very likely = 1. This will lead to a high affinity estimate but with a wide confidence interval and a high p-value. Using a cut-off threshold e.g. only using data for which $p < 0.05$ may lead to better results. However, the potential pitfalls involved in removing data are numerous, nuanced and beyond the scope of the present work. Suffice to say it is dangerous and unnecessary to risk the possible bias associated with systematically removing data points when the Bayesian analysis framework naturally accounts for differing levels of confidence between data.

The method proposed here provides ecologist with an important new tool for the analysis of cooccurrence, and in particular discovering the relationships between cooccurrence and other variables e.g. phylogenetic distance, which is an active area of research [6] and has been the subject of much research effort over the past few decades [18] but only now has an analysis framework based on the simple application of probability theory [13] with correct modelling of cooccurrence probabilities [9, 17, 12] while accounting for non-independence of data in matrices of pairwise species measurements [3, 7].

Methods

Maximum likelihood estimates of cooccurrence affinity were obtained using the R [16] package `CooccurrenceAffinity` [12]. All other analyses and visualisations were carried out in Julia [2]. All models were constructed in the probablistic programming language `Turing` [5] with MLE estimates of regression coefficients fit to data using the Nelder-Mead method [15] and MAP estimates of α fit using L-BFGS [11] implemented in `Optim` [14] and results visualised in `Makie` [4].

Code availability

All code to produce the figures and the manuscript can be found at

Supplement

Model implementations in Turing

Bayesian GLM

$$
\begin{aligned}
\vec{k}_i &\sim \text{fnchypg}(\vec{mA}_i, N - \vec{mA}_i, \vec{mB}_i, e^{\alpha}) \\
\alpha_i &= \gamma + \beta\vec{x}_i \\
\beta &\sim \text{N}(0, \beta_\sigma) \\
\gamma &\sim \text{N}(0, \beta_\gamma)
\end{aligned}
$$

```
@model function reg(x, N, mA, mB, k, priors)
    γ ~ Normal(0,prior[1])
    β ~ Normal(0,prior[2])
    for i in eachindex(x)
        α = γ + β*x[i]
        k[i] ~ FisherNoncentralHypergeometric(mA[i],N-mA[i],mB[i],exp(α))
    end
end
```

Dyadic Bayesian GLM

$$
\begin{aligned}
K_{ij} &\sim \text{fnchypg}(\vec{m}_i, N - \vec{m}_i, mj, e^{\alpha}) \\
\alpha_{ij} &= \lambda_i + \lambda_j + \beta X_{ij} \\
\beta &\sim \text{N}(0, \beta_\sigma) \\
\lambda_i &\ \text{N}(0, \lambda_\sigma)
\end{aligned}
$$

```
@model function dyadic_glm(X, N, m, k, priors)

    n = size(k,1)
    β ~ Normal(0,prior[1])
    λ ~ filldist(Normal(0,prior[2]),n)

    for j in 1:n-1
        mB = m[j]
        for i in j+1:n
            mA = m[i]
            α =  β*X[i,j] +λ[i] +λ[j]
            k[i] ~ FisherNoncentralHypergeometric(mA,N-mA,mB,exp(α))
```

```
        end
    end
end
```

References

[1] Albert Barberán et al. "Using network analysis to explore co-occurrence patterns in soil microbial communities". In: *The ISME journal* 6.2 (2012), pp. 343–351.

[2] Jeff Bezanson et al. "Julia: A fresh approach to numerical computing". In: *SIAM review* 59.1 (2017), pp. 65–98. URL: https://doi.org/10.1137/141000671.

[3] Ralph T. Clarke, Peter Rothery, and Alan F. Raybould. "Confidence Limits for Regression Relationships between Distance Matrices: Estimating Gene Flow with Distance". In: *Journal of Agricultural, Biological, and Environmental Statistics* 7.3 (2002), pp. 361–372. ISSN: 10857117. URL: http://www.jstor.org/stable/1400681 (visited on 03/03/2023).

[4] Simon Danisch and Julius Krumbiegel. "Makie.jl: Flexible high-performance data visualization for Julia". In: *Journal of Open Source Software* 6.65 (2021), p. 3349. DOI: 10.21105/joss.03349. URL: https://doi.org/10.21105/joss.03349.

[5] Hong Ge, Kai Xu, and Zoubin Ghahramani. "Turing: a language for flexible probabilistic inference". In: *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*. 2018, pp. 1682–1690. URL: http://proceedings.mlr.press/v84/ge18b.html.

[6] Marta Goberna et al. "Incorporating phylogenetic metrics to microbial co-occurrence networks based on amplicon sequences to discern community assembly processes". In: *Molecular ecology resources* 19.6 (2019), pp. 1552–1564.

[7] Zachariah Gompert et al. "Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants". In: *Molecular Ecology* 23.18 (2014), pp. 4555–4573. DOI: https://doi.org/10.1111/mec.12811. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.12811. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.12811.

[8] Nicholas J Gotelli and Declan J McCabe. "Species co-occurrence: a meta-analysis of JM Diamond's assembly rules model". In: *Ecology* 83.8 (2002), pp. 2091–2096.

[9] Daniel M Griffith, Joseph A Veech, and Charles J Marsh. "Cooccur: probabilistic species co-occurrence analysis in R". In: *Journal of Statistical Software* 69 (2016), pp. 1–17.

[10] Casper Kraan, Simon F Thrush, and Carsten F Dormann. "Co-occurrence patterns and the large-scale spatial structure of benthic communities in seagrass meadows and bare sand". In: *BMC ecology* 20.1 (2020), pp. 1–8.

[11] Dong C Liu and Jorge Nocedal. "On the limited memory BFGS method for large scale optimization". In: *Mathematical programming* 45.1-3 (1989), pp. 503–528.

[12] Kumar P. Mainali et al. “A better index for analysis of co-occurrence and similarity”. In: *Science Advances* 8.4 (2022), eabj9204. DOI: 10.1126/sciadv.abj9204. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.abj9204. URL: https://www.science.org/doi/abs/10.1126/sciadv.abj9204.

[13] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2018.

[14] Patrick Kofod Mogensen and Asbjørn Nilsen Riseth. “Optim: A mathematical optimization package for Julia”. In: *Journal of Open Source Software* 3.24 (2018), p. 615. DOI: 10.21105/joss.00615.

[15] John A Nelder and Roger Mead. “A simplex method for function minimization”. In: *The computer journal* 7.4 (1965), pp. 308–313.

[16] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[17] Joseph A Veech. “A probabilistic model for analysing species co-occurrence”. In: *Global Ecology and Biogeography* 22.2 (2013), pp. 252–260.

[18] Campbell O Webb et al. “Phylogenies and community ecology”. In: *Annual review of ecology and systematics* 33.1 (2002), pp. 475–505.

[19] Ryan J Williams, Adina Howe, and Kirsten S Hofmockel. “Demonstrating microbial co-occurrence pattern analyses within and between ecosystems”. In: *Frontiers in microbiology* 5 (2014), p. 358.

[20] Wentao Zhu et al. “Different responses of bacteria and microeukaryote to assembly processes and co-occurrence pattern in the coastal upwelling”. In: *Microbial Ecology* 86.1 (2023), pp. 174–186.